Two Sample Problem: Permutation Test

Paweł Polak

March 29 & 31, 2016

STAT W4413: Nonparametric Statistics - Lecture 14

Permutation tests: training example

Let's again discuss the "new training example".

Table : The grades of the employees after receiving different trainings.

New method	37, 49, 55, 57
Traditional method	23, 31,46

- X_1, X_2, X_3, X_4 represent the grades of the people that took the second training and
- Y_1, Y_2, Y_3 represent the grades of the employees that took the old training.
- In the first step we assume that

$$X_1, \ldots, X_4 \sim F(x - \Delta) \quad \text{and} \quad Y_1, Y_2, Y_3 \sim F(x).$$

- Now we cast the company's problem as testing

$$H_0 : \Delta = 0 \quad \text{versus} \quad H_1 : \Delta > 0.$$

According to H_0 the two instructions are essentially the same, while the alternate hypothesis says that the new instruction improves the average grade of the employees (hence it is better than the previous one). How can we do this test?

Permutation tests: training example under the null

Suppose that the null hypothesis is true.

- Then since $\Delta = 0$ the distribution of $X_1, \ldots, X_4$ and $Y_1, \ldots, Y_3$ are exactly the same.
- Therefore, all these samples are drawn from the same distribution.
- Therefore, if we permute the data points and obtain new X and Y samples, e.g.,

$$X_1^{new} = X_1, X_2^{new} = X_2, X_3^{new} = Y_1, X_4^{new} = Y_2,$$

$$\text{and } Y_1^{new} = X_3, Y_2^{new} = X_4, Y_3^{new} = Y_4,$$

- then the new samples are coming from the same distribution.
- Therefore, we expect that the mean difference of the observed samples $\bar{X} - \bar{Y}$ and the new sample (permutation sample) $\bar{X}^{new} - \bar{Y}^{new}$ to be close.

Permutation tests: training example under the alternative

Now, suppose that the alternative is true.

- Then we expect $\bar{X} - \bar{Y}$ to be greater than $\bar{X}^{new} - \bar{Y}^{new}$ new (to see this more clearly consider an extreme case in which $\min_i X_i > max_i Y_i$).

This simple idea is the main principle of what is known as the permutation test.

In the *permutation test*, instead of picking only one permutation of the data we consider all the possible permutations and we expect $\bar{X} - \bar{Y}$ to be greater than most of the $\bar{X}^{new} - \bar{Y}^{new}$ new (that are drawn from permutations) if H_1 is true.

Permutation tests

To understand the process look at Table from Higgins' book displayed in the next slide. This table has listed all the possible permutations for the data of Table 1. How many different permutations do we have? $\binom{7}{4}$. For each of these samples that I would like to call X^{new}, Y^{new}, you can see the difference between the means $\bar{X}^{new} - \bar{Y}^{new}$ new below. The histogram of these differences is also shown in the next slide.

Permutation tests: permutations

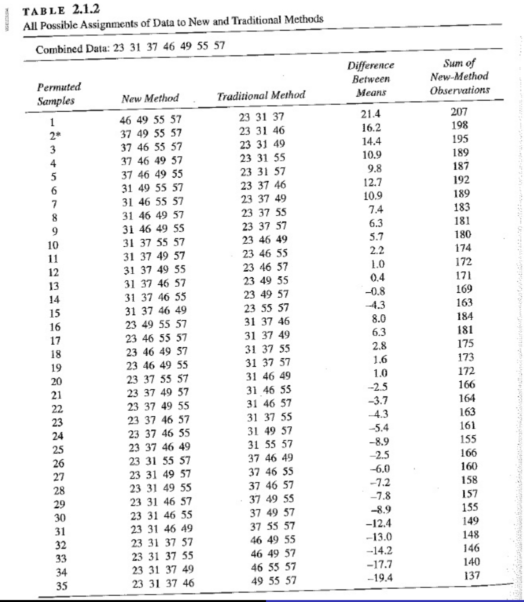


TABLE 2.1.2
All Possible Assignments of Data to New and Traditional Methods

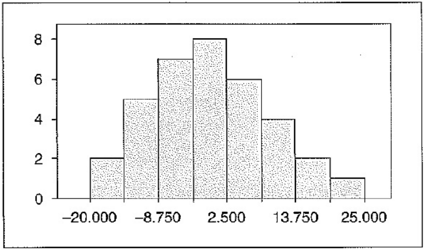
Combined Data: 23 31 37 46 49 55 57

Permuted Samples	*New Method*	*Traditional Method*	*Difference Between Means*	*Sum of New-Method Observations*
1	46 49 55 57	23 31 37	21.4	207
2*	37 49 55 57	23 31 46	16.2	198
3	37 46 55 57	23 31 49	14.4	195
4	37 46 49 57	23 31 55	10.9	189
5	37 46 49 55	23 31 57	9.8	187
6	31 49 55 57	23 37 46	12.7	192
7	31 46 55 57	23 37 49	10.9	189
8	31 46 49 57	23 37 55	7.4	183
9	31 46 49 55	23 37 57	6.3	181
10	31 37 55 57	23 46 49	5.7	180
11	31 37 49 57	23 46 55	2.2	174
12	31 37 49 55	23 46 57	1.0	172
13	31 37 46 57	23 49 55	0.4	171
14	31 37 46 55	23 49 57	–0.8	169
15	31 37 46 49	23 55 57	–4.3	163
16	23 49 55 57	31 37 46	8.0	184
17	23 46 55 57	31 37 49	6.3	181
18	23 46 49 57	31 37 55	2.8	175
19	23 46 49 55	31 37 57	1.6	173
20	23 37 55 57	31 46 49	1.0	172
21	23 37 49 57	31 46 55	–2.5	166
22	23 37 49 55	31 46 57	–3.7	164
23	23 37 46 57	31 37 55	–4.3	163
24	23 37 46 55	31 49 57	–5.4	161
25	23 37 46 49	31 55 57	–8.9	155
26	23 31 55 57	37 46 49	–2.5	166
27	23 31 49 57	37 46 55	–6.0	160
28	23 31 49 55	37 46 57	–7.2	158
29	23 31 46 57	37 49 55	–7.8	157
30	23 31 46 55	37 49 57	–8.9	155
31	23 31 46 49	37 55 57	–12.4	149
32	23 31 37 57	46 49 55	–13.0	148
33	23 31 37 55	46 49 57	–14.2	146
34	23 31 37 49	46 55 57	–17.7	140
35	23 31 37 46	49 55 57	–19.4	137

Permutation tests: histogram

FIGURE 2.1.2

Permutation Distribution of Difference Between Means of Data in Table 2.1.1



- For the set of grades that we actually observed $\bar{X} - \bar{Y}$ is 16.2.
- If we check this value in the histogram we see that only a few of the permutation samples have a higher "mean difference".
- Therefore, according to our intuitive discussion we should reject the null hypothesis.

But how do we calculate the *p*-value of this test?

Permutation tests: p-values

- Suppose that the null hypothesis H_0 is true.
- Then each of the permuted samples that we construct are coming from the same distribution.
- Suppose that D_i is the difference between the means of the i_{th} permuted samples.
- Then we can estimate the CDF of D, the difference of the means, from all the permutation samples that we collected.
- This CDF can be written as

$$\hat{F}_D(t) = \frac{1}{\binom{7}{4}} \sum_{i=1}^{\binom{7}{4}} \mathbb{I}(D_i \leq t).$$

Permutation tests: *p*-values

- Now based on this estimate of the CDF, we can provide an estimated *p*-value for our test. If $D_{obs} = \bar{X} - \bar{Y}$, then the estimated p-value that we represent as p_{perm}, for the fact that it is based on the permutation samples, is given by

$$p_{perm} = \frac{1}{\binom{7}{4}} \sum_{i=1}^{\binom{7}{4}} \mathbb{I}(D_i \geq D_{obs}).$$

Permutation tests: the steps

The steps of the permutation test:

- Suppose that we have observed

$$X_1, X_2, \ldots, X_n \sim F(x - \Delta) \quad \text{and} \quad Y_1, Y_2, \ldots, Y_m \sim F(x)$$

- and we would like to test

$$H_0 : \Delta = 0 \quad \text{versus} \quad H_1 : \Delta > 0.$$

- The permutation test follows these three simple steps:
  1. Permute $m + n$ observations. By doing this we obtain $\binom{m+n}{m}$ permuted samples. Call the ith permuted sample $X_1^{new_i}, X_2^{new_i}, \ldots, X_n^{new_i}$ and $Y_1^{new_i}, Y_2^{new_i}, \ldots, Y_m^{new_i}$.
  2. For each permuted sample calculate $D_i = \bar{X}^{new_i} - \bar{Y}^{new_i}$.
  3. Calculate the permutation p-value as

$$p_{perm} = \frac{1}{\binom{m+n}{n}} \sum \mathbb{I}(D_i \geq D_{obs}),$$

where D_{obs} is the difference of the means of the original samples.

Permutation tests

- We will show later that this p-value is even theoretically quite accurate.
- But before we move on to the theory section let's explore some of the advantages and disadvantages of the permutation test.
- We would also like to see other applications of this test.
- The first and the most important advantage of the permutation test is its flexibility.
- As you see in our discussion so far, we have never used the fact that D_i is the difference between the averages of the samples.
- In fact this test can perform as well with any other statistic that captures the shift of the distribution.
- In the rest of this lecture we explore some of these options and briefly describe the permutation test for each statistic.

Permutation test with other criteria

The most intriguing feature of the permutation test is its flexibility in the choice of the test statistic that is employed in the test:

- We used the difference of the mean to test for $H_0 : \Delta = 0$.
- Clearly, the median can also be used as a measure of the shift, meaning that if $\Delta = 0$, then $Med(X) = Med(Y)$, and
- under H_1, $Med(X) > Med(Y)$.
- Therefore, we can employ the statistic $Med(X) - Med(Y)$ in the permutation test to evaluate the validity of H_0.

Permutation test with other criteria: median

- Here is what the permutation test with median statistic looks like:
  1. Permute $m+n$ observations. By doing this we obtain $\binom{m+n}{m}$ permuted samples. Call the i^{th} permuted sample $X_1^{new_i}, X_2^{new_i}, \ldots, X_n^{new_i}$ and $Y_1^{new_i}, Y_2^{new_i}, \ldots, Y_m^{new_i}$.
  2. For each permuted sample calculate $D_i = Med(X^{new_i}) - Med(Y^{new_i})$.
  3. Calculate the permutation p-value as

$$p_{perm} = \frac{1}{\binom{m+n}{n}} \sum \mathbb{I}(D_i \geq D_{obs}),$$

where D_{obs} is the difference of the medians of the original samples.

As you can see the amount of effort we have to do for this new test is minimal here. Instead, computer is doing all the work for us.

Permutation test with other criteria: Wilcoxon rank-sum statistic

- Based on this calculation one would argue that if W is much larger than $\frac{mn}{2} + \frac{m(m+1)}{2}$ this is a good indication of $\Delta > 0$.
- Therefore, one would again use the permutation test with this statistic.
  1. Permute $m + n$ observations. By doing this we obtain $\binom{m+n}{m}$ permuted samples. Call the i^{th} permuted sample $X_1^{new_i}, X_2^{new_i}, \dots, X_n^{new_i}$ and $Y_1^{new_i}, Y_2^{new_i}, \dots, Y_m^{new_i}$.
  2. For each permuted sample calculate $D_i = \sum_{j=1}^{m} rank(Y_j^{new_i})$.
  3. Calculate the permutation p-value as

$$p_{perm} = \frac{1}{\binom{m+n}{n}} \sum \mathbb{I}(D_i \geq D_{obs}),$$

where D_{obs} is the Wilcoxon rank-sum statistic of the original samples.

Permutation test with other test statistics

Other test statistics:

- Clearly you can come up with many other test statistics that can be incorporated in the permutation test.
- You can think of one sided Kolmogorov-Smirnov, or some other test statistics.
- But for almost all the applications the statistics we have mentioned so far perform very well.
- There is usually no need to come up with new statistics, except for homeworks or exam problems.

Comparing different test statistics

So far we have mainly used three different test statistics, i.e.,

- the mean difference,
- median difference,
- Wilcoxon rank-sum or Mann-Whitney statistic.

Which one should we use in practice?

This is based on your judgement of the data as a statistician. However, here we provide a few guidelines that might help you in this direction:

Comparing different test statistics: guidelines

1. If the underlying distribution of the observations seem to be Gaussian (you can check it with one of the techniques you learned in goodness of fit lectures), then the mean difference is optimal (among all unbiased tests). Therefore, you should use that.

   However, note that this statistic is not robust to outliers, therefore if you suspect that your data might have outliers probably trimmed mean or median might be a better choice.

2. If the Gaussianity is violated, Wilcoxon rank-sum test is more reliable and it might be a better choice. In fact when mean difference outperforms the Wilcoxon test the difference is not much. But in some cases Wilcoxon can gain huge improvements over the mean-difference test.

Extensions of permutation test

So far we have studied the permutation test on the following two-sample problem:

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} F(x-\Delta) \quad \text{and} \quad Y_1, Y_2, \ldots, Y_m \overset{iid}{\sim} F(y).$$

We would like to test

$$H_0 : \Delta = 0 \quad \text{vs.} \quad H_1 : \Delta > 0.$$

However, the flexibility of the permutation test enables us to apply it a wide range of problems of interest. Below we briefly discuss some of these problems.

Two-sided location (shift) test

Now we would like to discuss a new problem and that is the problem of two-sided location parameter test. As before we observe two samples that satisfy

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} F(x - \Delta) \quad \text{and} \quad Y_1, Y_2, \ldots, Y_m \overset{iid}{\sim} F(y).$$

But this time we would like to test the hypothesis:

$$H_0 : \Delta = 0 \quad \text{vs.} \quad H_1 : \Delta \neq 0.$$

As you can imagine this problem is called two-sided since the alternate hypothesis includes both $\Delta > 0$ and $\Delta < 0$. We can again use permutation test with many different statistics. Here are a few statistics that we can use:

1. $D_i = |\bar{X}^{new_i} - \bar{Y}^{new_i}|$.
2. $D_i = |Med(X^{new_i}) - Med(Y^{new_i})|$.
3. $D_i = |\sum_{j=1}^{n} rank(Y_j^{new_i}) - \frac{m(m+1)}{2} - \frac{mn}{2}|$.

Two-sided location (shift) test

Once you decide on the statistic, the permutation test can be performed in exactly the same way that we performed it before, i.e.,

1. Permute $m+n$ observations. By doing this we obtain $\binom{m+n}{m}$ permuted samples. Call the i^{th} permuted sample $X_1^{new_i}, X_2^{new_i}, \ldots, X_n^{new_i}$ and $Y_1^{new_i}, Y_2^{new_i}, \ldots, Y_m^{new_i}$.
2. For each permuted sample calculate D_i.
3. Calculate the permutation p-value as

$$\boxed{p_{perm} = \frac{1}{\binom{m+n}{n}} \sum \mathbb{I}(D_i \geq D_{obs}),}$$

where D_{obs} is calculated for the original samples.

As you can see when we use the permutation test the problem of designing a test will reduce to the problem of designing a new statistic. Before we compare the statistics that have been proposed for the above two problems let us discuss two closely related but more general testing problems for which we can use the same statistics.

One-sided domination problem

Consider the two samples

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} F(x) \quad \text{and} \quad Y_1, Y_2, \ldots, Y_m \overset{iid}{\sim} G(y).$$

We still believe that the X_i's tend to be larger than Y_i (better grades in the example of the company with new instruction), but we do not have any reason to believe that the two samples have the same distribution except for a shift.

Therefore, we design our hypothesis in a more general way:

$$H_0 : F(x) = G(x)\ \forall x \ \text{ versus } \ H_1 : F(x) > G(x)\ \forall x. \tag{1}$$

One-sided domination problem

- As before, we intend to propose permutation test for this hypothesis and hence we only need to propose certain statistics that can distinguish H_0 and H_1.

- Let's start with the statistic involving the means. Can we use $\bar{X} - \bar{Y}$? To use this we require that $\bar{X} - \bar{Y}$ to be small when H_0 is true, and to be large when H_1 holds.

- When H_0 is true, $\bar{X} - \bar{Y}$ is small, since they are drawn from the same distribution.

- Therefore, the only other thing we require is that under the alternate hypothesis we expect $\bar{X} - \bar{Y} > 0$.

The following lemma provides this missing link:

One-sided domination problem

Lemma

Let the random variable X satisfy $\mathbb{E}(|X|) < \infty$ and has a pdf.[a] *Then*

$$\mathbb{E}(X) = \int_0^\infty (1 - F(x))dx - \int_{-\infty}^0 F(x)dx.$$

[a] Having pdf is not required. We make this assumption to avoid some technical details.

One-sided domination problem

Proof.

First, we emphasize that from the fact that $E(|X|)$ is bounded we can conclude that

$$\begin{aligned} \lim_{x\to\infty} x(1-F(x)) &= 0. \\ \lim_{x\to-\infty} xF(x) &= 0. \qquad (2) \end{aligned}$$

In case you are interested in probability theory try to prove these two statements. Otherwise, assume that our random variable is bounded above by some number M, i.e. $|X| \leq M$, and then prove the above two statements. Now we can characterize the mean in terms of the CDF. □

One-sided domination problem

Proof.

$$
\begin{aligned}
\mathbb{E}(X) &= \int_{-\infty}^{\infty} xf(x)dx = \int_{0}^{\infty} xf(x)dx + \int_{-\infty}^{0} xf(x)dx = -\int_{0}^{\infty} xd(1-F(x)) + \int_{-\infty}^{0} xdF(x) \\
&= \int_{0}^{\infty} (1-F(x))dx - \int_{-\infty}^{0} F(x)dx. \qquad (3)
\end{aligned}
$$

The last equality is due to integration by parts and (2).

□

Now that we have characterized the connection between CDF and the expected value we can easily prove the following proposition.

Proposition

Let $X \sim F$ and $Y \sim G$. If $F \geq G$, then $\mathbb{E}(X) \leq \mathbb{E}(Y)$.

Prove this result for yourself.
This Proposition justifies the application of $\bar{X} - \bar{Y}$ for testing problem of (1).

One-sided domination problem

We can also use other statistics that we discussed before such as the median difference and Wilcoxon rank-sum test.

You will show in the homework that these two statistics are appropriate with the testing problem of (1).

Now that we introduced the statistics that can be used in the permutation test, can you propose the full permutation test for this problem?

Two-sided domination problem

- As we mentioned before the one-sided domination problem is a generalization of the "one-sided shift problem".
- Is there any generalization for the two-sided shift problem?
- The answer is yes. Here is an example.
- Given two samples

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} F(x) \quad \text{and} \quad Y_1, Y_2, \ldots, Y_m \overset{iid}{\sim} G(y)$$

  we can test for

  $H_0 : F(x) = G(x) \quad \text{vs.} \quad H_1 : F(x) < G(x),\ \forall x \text{ or } F(x) > G(x)\ \forall x.$

  As before you can use any of the following statistics in the permutation test algorithm:

  1. $D_i = |\bar{X}^{new_i} - \bar{Y}^{new_i}|$.
  2. $D_i = |Med(X^{new_i}) - Med(Y^{new_i})|$.
  3. $D_i = |\sum_{j=1}^{n} rank(Y_j^{new_i}) - \frac{m(m+1)}{2} - \frac{mn}{2}|$.

Scale problem: Equal shift

To prove the flexibility of the permutation test here we study another hypothesis testing problem.

Suppose that

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} F\left(\frac{x-\mu}{\sigma_x}\right) \quad \text{and} \quad Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} F\left(\frac{x-\mu}{\sigma_y}\right).$$

We would like to test one of the following hypotheses:

1. One sided: $H_0 : \sigma_x = \sigma_y$ versus $H_1 : \sigma_x > \sigma_y$
2. One sided: $H_0 : \sigma_x = \sigma_y$ versus $H_1 : \sigma_x < \sigma_y$
3. Two sided: $H_0 : \sigma_x = \sigma_y$ versus $H_1 : \sigma_x \neq \sigma_y$.

- Note that here we assume that the shift parameter is the same for both samples.
- If we were in the parametric setting and F was Gaussian, we could use the F-test as described before.
- However, since F is not known, the F-test does not work. In fact, it is well-known that the F-test for comparing variances is very sensitive to the Gaussianity of the data.
- Now, the question is how can we distinguish the hypotheses described above under the nonparametric settings?
- Again, permutation test makes life very simple for us. Since under the null hypothesis X samples and Y samples are drawn from the same distribution, we can permute the data.
- Therefore we only need to come up with a statistic that distinguished between the null and alternate hypothesis.

Scale problem: Equal shift

- An example of such statistic is the statistic we used in F-test, i.e.,

$$D = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}{\frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \bar{Y})^2}$$

- Clearly if the null hypothesis holds we expect $D \approx 1$. If D is greater than 1 that is an indication of $\sigma_x > \sigma_y$ and if D is less than 1 that is an indication of $\sigma_x < \sigma_y$.
- Based on this statistic we can propose the permutation test in the following way:
  1. Permute $m+n$ observations. By doing this we obtain $\binom{m+n}{m}$ permuted samples. Call the i^{th} permuted sample $X_1^{new_i}, X_2^{new_i}, \dots, X_n^{new_i}$ and $Y_1^{new_i}, Y_2^{new_i}, \dots, Y_m^{new_i}$.
  2. For each permuted sample calculate D_i in the following way
     1. For $H_1 : \sigma_x > \sigma_y$, $D_i = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}{\frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \bar{Y})^2}$.
     2. For $H_1 : \sigma_x < \sigma_y$, $D_i = \frac{\frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \bar{Y})^2}{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$.
     3. For $H_1 : \sigma_x \neq \sigma_y$, $D_i = \max\left(\frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}{\frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \bar{Y})^2}, \frac{\frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \bar{Y})^2}{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)$.

Scale problem: Equal shift

- Calculate the permutation p-value as

$$p_{perm} = \frac{1}{\binom{m+n}{n}} \sum \mathbb{I}(D_i \geq D_{obs}),$$

where D_{obs} is calculated for the observed samples.

Scale problem: Equal shift

- The ratio of the variances that we used above is very sensitive to the outliers.
- Hence, one would like to use other measures that are more robust to outliers.
- As we mentioned before, the flexibility of permutations test in using different statistics makes this task very straightforward.
- For instance one would use the ratio of deviances as defined below.

$$RD = \frac{\frac{1}{n}\sum_{i=1}^{n}|X_i - Med(X)|}{\frac{1}{m}\sum_{i=1}^{m}|Y_i - Med(Y)|}$$

- Can you explain what the permutation test looks like with this new statistic?

Scale problem: Unequal shifts

- In the last slides we assumed that the the shift parameter μ is the same for both X and Y samples.
- Now, we would like to generalize the problem.

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} F\left(\frac{x-\mu_x}{\sigma_x}\right) \quad \text{and} \quad Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} F\left(\frac{x-\mu_y}{\sigma_y}\right).$$

- We would like to test one of the following hypotheses.
  1. One sided: $H_0 : \sigma_x = \sigma_y$ versus $H_1 : \sigma_x > \sigma_y$
  2. One sided: $H_0 : \sigma_x = \sigma_y$ versus $H_1 : \sigma_x < \sigma_y$
  3. Two sided: $H_0 : \sigma_x = \sigma_y$ versus $H_1 : \sigma_x \neq \sigma_y$.

Scale problem: Unequal shifts

- The first challenge that we face here is that under the null hypothesis the two datasets are not exchangeable, i.e., we cannot assume that X samples and Y samples are drawn from the same distribution.

- This is unfortunately one of the weaknesses of the permutation test.

- In fact permutation test is not suitable for such problems as we will discuss later.

- However, for this specific problem there exist some transformation of the data that makes it "approximately exchangeable" and hence we can still use permutation test.

- However, it is important to note that once we discuss the theory of permutation test that theory will not be applied to this problem, and it is only approximately correct for this problem.

Scale problem: Unequal shifts

The idea here is very simple:

- Based on the samples that we have we construct new samples $\tilde{X}_1, \ldots, \tilde{X}_n$ and $\tilde{Y}_1, \ldots, \tilde{Y}_n$ in the following way
$$\tilde{Y}_i = Y_i - Med(Y). \quad \tilde{X}_i = X_i - Med(X).$$

- You could also subtract the mean of X and Y instead of the median.

- By doing this we obtain samples that have the same mean. We pretend that $\tilde{X}_1, \ldots, \tilde{X}_n$ and $\tilde{Y}_1, \ldots, \tilde{Y}_m$ are independent samples and perform the permutation tests as we described above.

Limitations of the permutation test

Two important limitations of permutations test

- As we described, one of the limitations of the permutation test and that is the fact that we need the two samples to be exchangeable under the null hypothesis.
- In other words, we expect samples to be drawn from the same distribution under the null.
- In some cases even though this assumption is not true, by some transformation of the data we obtain a dataset for which exchangeability is approximately true.
- As the problem gets more complicated the situation gets worse.

Limitations of the permutation test

Example

Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} F\left(\frac{x-\mu_1}{\sigma_1}\right)$ and $Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} F\left(\frac{x-\mu_2}{\sigma_2}\right)$. We would like to test

$$H_0 : \mu_1 = \mu_2 \text{ versus } \quad H_1 : \mu_1 \neq \mu_2.$$

Clearly, under the null the datasets are not exchangeable. Why? Here you might need more complicated transformation to make the data exchangeable. Therefore, permutation test is not particularly useful for these problems. As we will see later in the course, Bootstrap can solve these problems. Bootstrap is a less accurate, but more flexible algorithm than the permutation test.

Limitations of the permutation test

Another limitation of the permutation test that we have ignored so far is its computational complexity.

Consider a small two-sample problem in which we have 10 treatment samples (X) and 10 control samples (Y).

The number of permutation samples that we have is $\binom{n+m}{m} \approx 185000$.

As we can see the number of permutation samples grows very rapidly as m and n increase.

Therefore, it seems that the computational complexity of permutation test becomes prohibitive for even small sample sizes such as $m = 20$, $n = 20$.

In the next slides we will show that Monte Carlo method addresses this issue of the permutation test.

Monte Carlo method for categorical random variables

While we have described the Monte Carlo method in the context of continuous random variables, the application of this method is not limited to such variables and it is extensively used for categorical random variables.

Consider categorical random vector $\{x_i\}_{i=1}^n \sim P(x_1, x_2, \dots, x_n)$, where $P(x_1, x_2, \dots, x_n)$ is a probability mass function. For notational simplicity assume that each $x_i \in \{0, 1\}$.

Suppose that we are interested in the numeric evaluation of

$$\mathbb{E}g(x_1, x_2, \dots, x_n).$$

According to the definition

$$\mathbb{E}g(x_1, x_2, \dots, x_n) = \sum_{x_n=0}^{x_n=1} \dots \sum_{x_2=0}^{x_2=1} \sum_{x_1=0}^{x_1=1} g(x_1, x_2, \dots, x_n) P(x_1, x_2, \dots, x_n)$$

Monte Carlo method for categorical random variables

Performing this calculation is equivalent to performing 2^n additions.

Even in low dimensional cases such as $n = 30$, this means one billion additions which is computationally demanding.

Therefore we should look for inexpensive alternatives. One of the most popular approaches is the Monte Carlo simulation.

Here is the summary of the Monte Carlo simulation.

1. Generate $\{x_i^1\}_{i=1}^n, \{x_i^2\}_{i=1}^n \dots \{x_i^I\}_{i=1}^n$ iid samples from $P(x_1, x_2, \dots, x_n)$.
2. Approximate $\mathbb{E}[g(x_1, x_2, \dots, x_n)]$ with

$$\mathbb{E}[g(x_1, x_2, \dots, x_n)] \approx \frac{1}{I} \sum_{i=1}^{I} g(x_1^i, x_2^i, \dots, x_n^i).$$

Application of MC method for the permutation test

Let's get back to our original problem: simplifying the computations in the permutation test. Recall the main steps of the permutation test:

1. Consider all $\binom{m+n}{n}$ permutation samples.
2. Calculate D_i for each permutated sample.
3. Estimate the p-value according to **Proportion**

$$p_{perm} = \frac{\#D_i \geq D_{obs}}{\binom{m+n}{n}}$$

Application of MC method for the permutation test

The final objective of the entire process described above is to calculate p_{perm}.

Can we calculate it using the Monte Carlo method and less permutation samples?

- Let D_i^* be a random variable that is generated from a random vector that has uniform distribution on all the permutations of the vector $(x_1, \dots, x_n, y_1, \dots, y_m)$.
- Since we have $\binom{m+n}{m}$ different permutations, the weight of each sample will be $\frac{1}{\binom{m+n}{m}}$.
- It is straightforward to confirm that

$$p_{perm} = \mathbb{E}[\mathbb{I}(D_i^* > D_{obs})].$$

- Now that p_{perm} is written in terms of an expectation we can easily use the Monte Carlo method to simplify it.

Application of MC method for the permutation test

Here is the Monte Carlo estimate of p_{perm}:

1. Draw B permutations of data at random from the total $\binom{m+n}{n}$ permutations.
2. For each sample, calculate D_i.
3. Estimate $\hat{p} = \frac{1}{B}\sum_{i=1}^{B} \mathbb{I}(D_i > D_{obs})$

As the last part of our discussion we evaluate how accurate our estimate of p_{perm} is.

According to CLT

$$\boxed{\sqrt{B}(\hat{p} - p_{perm}) \overset{d}{\approx} N(0, \underbrace{var(\mathbb{I}(D_i^* > D_{obs}))}_{p_{perm}(1-p_{perm})})}$$

Application of MC method for the permutation test

Therefore, the variance of the Monte Carlo estimate is $\frac{p_{perm}(1-p_{perm})}{B}$ and therefore the error is at the order of $\sqrt{\frac{p_{perm}(1-p_{perm})}{B}}$. In order to calculate the relative error in estimating p_{perm} we define the coefficient of variation as:

$$CV(\hat{p}) = \frac{1}{p_{perm}}\sqrt{\frac{p_{perm}(1-p_{perm})}{B}} = \sqrt{\frac{1-p_{perm}}{Bp_{perm}}}$$

This essentially tells us about the percentage of the error in estimating p_{perm}.

If p_{perm} is small, then you will need a higher B to estimate it with a given level of accuracy. On the other hand, if you have a bigger p_{perm}, then you don't require as many samples.

Practically speaking, in most cases 2000 samples are enough, unless the p-value is very low and you really like to know the p-value accurately.

Theoretical guarantees for the permutation test

We turn now to some of the nice theoretical properties of the permutation test.
Consider two sets of samples

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} F \quad \text{and} \quad Y_1, Y_2, \ldots, Y_m \overset{iid}{\sim} G.$$

We assume that the null hypothesis is in the form of

$$H_0: \; F(x) = G(x) \quad \forall x$$

(so the samples are exchangeable).

However, we will not make any assumption on the alternate hypothesis.

To simplify our exposition we also define the following two notations:

- let v represent the *combined ordered values* and
- g, the vector that indicates which group each ordered observation belongs to.

The following example clarifies these two notations.

Theoretical guarantees for the permutation test

Example

Let $X_1 = 1, X_2 = 2$ and $Y_1 = 1.5$, $Y_2 = 1.8$. Then we have

$$\mathbf{v} = (1, 1.5, 1.8, 2).$$

Also $\mathbf{g}$ for our observation looks like

$$\mathbf{g} = (X, Y, Y, X).$$

Lemma (Permutation lemma)

Under H_0, the vector $\mathbf{g}$ has probability $\frac{1}{\binom{m+n}{n}}$ of equating any one of its possible choices.

Theoretical guarantees for the permutation test

Proof.

Clearly, g can take on $\binom{m+n}{n}$ choices. Why? Suppose that $\mathbf{g}_0$ is one of those choices. Then,

$$\mathbb{P}_{H_0}(\mathbf{g} = \mathbf{g}_0) = \int_{\mathbf{v}} \mathbb{P}(\mathbf{g} = \mathbf{g}_0 | \mathbf{v}) f_{\mathbf{v}}(\mathbf{v}) d\mathbf{v}.$$

In the above equation the subscript H_0 emphasizes the fact that all the probabilities are calculated under the null hypothesis. $\mathbf{v}$ is the vector of combined ordered values and finally $f_{\mathbf{v}}(\mathbf{v})$ is the probability density function of vector $\mathbf{v}$. $d\mathbf{v}$ is the a short notation for $dv_1 dv_2 \dots dv_{m+n}$. We now claim that

$$\mathbb{P}(\mathbf{g} = \mathbf{g}_0 \mid \mathbf{v}) = \frac{1}{\binom{m+n}{m}}.$$

The proof of this claim is straightforward and is left for you. But to help you gain some intuition on why this is true and how you can prove it, we provide the proof in a simpler example. □

Theoretical guarantees for the permutation test

Example

Consider $\mathbf{v} = (1, 1.5, 1.8, 2)$. and $\mathbf{g}_0 = (X, Y, Y, X)$. Then

Theoretical guarantees for the permutation test

$$
\begin{aligned}
\mathbb{P}(\mathbf{g} = \mathbf{g}_0 | \mathbf{v}) = \lim_{\Delta \to 0} \mathbb{P}(\mathbf{g} = \mathbf{g}_0 \mid 1 - \Delta/2 \leq \mathbf{v}_1 \leq 1 + \Delta/2, \dots, 2 - \Delta/2 \leq \mathbf{v}_4 \leq 2 + \Delta/2) \\
= \quad & \lim_{\Delta \to 0} \frac{\mathbb{P}(\mathbf{g} = \mathbf{g}_0 \;\cap\; 1 - \Delta/2 \leq \mathbf{v}_1 \leq 1 + \Delta/2, \dots, 2 - \Delta/2 \leq \mathbf{v}_4 \leq 2 + \Delta/2)}{\mathbb{P}(1 - \Delta/2 \leq \mathbf{v}_1 \leq 1 + \Delta/2, \dots, 2 - \Delta/2 \leq \mathbf{v}_4 \leq 2 + \Delta/2)} \\
\overset{(a)}{=} \quad & \lim_{\Delta \to 0} \frac{\mathbb{P}(1 - \Delta/2 \leq X_1 \leq 1 + \Delta/2, 1.5 - \Delta/2 \leq Y_1 < 1.5 + \Delta/2, \dots, 2 - \Delta/2 \leq X_2 \leq 2 + \Delta/2)}{\mathbb{P}(1 - \Delta/2 \leq \mathbf{v}_1 \leq 1 + \Delta/2, \dots, 2 - \Delta/2 \leq \mathbf{v}_4 \leq 2 + \Delta/2)} \\
= \quad & \lim_{\Delta \to 0} \frac{f_x(X_1) f_y(Y_1) f_y(Y_2) f_x(X_2) \Delta^4}{\mathbb{P}(1 - \Delta/2 \leq \mathbf{v}_1 \leq 1 + \Delta/2, \dots, 2 - \Delta/2 \leq \mathbf{v}_4 \leq 2 + \Delta/2)} \\
\overset{(b)}{=} \quad & \lim_{\Delta \to 0} \frac{f_x(X_1) f_x(Y_1) f_x(Y_2) f_x(X_2) \Delta^4}{\mathbb{P}(1 - \Delta/2 \leq \mathbf{v}_1 \leq 1 + \Delta/2, \dots, 2 - \Delta/2 \leq \mathbf{v}_4 \leq 2 + \Delta/2)} \qquad (4)
\end{aligned}
$$

where (a) is due to the fact that by knowing $\mathbf{v}$ and $\mathbf{g}$ the samples are exactly specified. (b) is also due to the fact that the null hypothesis is true and there is no difference between the pdf of X and Y. As is clear from (4), $\mathbb{P}(\mathbf{g} = \mathbf{g}_0 | \ \mathbf{v})$ is independent of the choice of $\mathbf{g}_0$. Therefore, it is the same for all choices of $\mathbf{g}_0$.

Theoretical guarantees for the permutation test

Now that we have this lemma we can characterize the significance level of the permutation test.

- Suppose that we perform permutation test with statistic D.
- We perform the test in the following way.
- For each dataset we consider all $\binom{m+n}{n}$ different permutation and for each sample we calculate D_i. We construct our test in the following way:

$$\text{Reject } H_0 \text{ if and only if } \#(D_i \geq D) \leq k.$$

Theoretical guarantees for the permutation test

So far we have explained some heuristic way to set the p-values for such permutation test. But here we would like to characterize the probability of Type I error exactly and show that our heuristic approach is in fact very accurate.

Theorem

The significance level of the permutation test specified above is equal to $\frac{k}{\binom{m+n}{m}}$

Theoretical guarantees for the permutation test

Proof.

- As before let $\mathbf{v}$ denote the vector of combined ordered values. Also we use the notation $\mathbb{P}_{H_0}(\cdot)$ to emphasize that all the probabilities are calculated under the null hypothesis.

$$\mathbb{P}_{H_0}(\text{Type I Error}) = \mathbb{P}_{H_0}((\#D_i \geq D_{obs}) \leq k) = \int_{\mathbf{v}} \mathbb{P}((\#D_i \geq D_{obs}) \leq k|\mathbf{v}) f_{\mathbf{v}}(\mathbf{v}) d\mathbf{v}.$$

- In the rest of the proof, we would like to prove that

$$\mathbb{P}((\#D_i \geq D_{obs}) \leq k|\mathbf{v}) = \frac{k}{\binom{m+n}{m}}.$$

- Convince yourself that once we prove this claim we have

$$\mathbb{P}_{H_0}(\text{Type I Error}) = \frac{k}{\binom{m+n}{m}}.$$

□

Theoretical guarantees for the permutation test

Proof.

Why $\mathbb{P}((\#D_i \geq D_{obs}) \leq k|\mathbf{v}) = \frac{k}{\binom{m+n}{m}}$?

- Given $\mathbf{v}$, D_{obs} can only take $\binom{m+n}{m}$ different values. These different values correspond to different $\mathbf{g}$ vectors that we assign to the data.
- Let $\mathcal{D} = \{D_{obs}^1, \ldots, D_{obs}^{\binom{m+n}{m}}\}$ denote all the possible values of D_{obs}. For notational simplicity assume that $D_{obs}^1 > D_{obs}^2 \ldots, > D_{obs}^{\binom{m+n}{m}}$.

□

Theoretical guarantees for the permutation test

Proof.

According to the permutation lemma the probability that D_{obs} takes any of the values on $\mathcal{D}$ is $\frac{1}{\binom{m+n}{m}}$. Therefore,

$$\mathbb{P}_{H_0}((\#D_i \geq D_{obs}) \leq k|\mathbf{v}) \overset{(a)}{=} \sum_{D_{obs} \in \mathcal{D}} \frac{1}{\binom{m+n}{m}} \mathbb{I}((\#D_i \geq D_{obs}) \leq k) \overset{(b)}{=} \frac{k}{\binom{m+n}{m}}. \qquad (5)$$

- Note that equality (a) is due to the fact that once D_{obs} and $\mathbf{v}$ are known there is no randomness in the system and $\mathbb{P}_{H_0}(\#D_i \geq D_{obs} \leq k|\mathbf{v})$ is either zero or one depending on the number of D_is that are larger than D_{obs}.
- Equality (b) holds because:
  1. If $D = D^1_{obs}$ then $\mathbb{I}((\#D_i \geq D_{obs}) \leq k) = 1$ since there is only one D_i that is greater than or equal to D^1_{obs}.
  2. If we set D_{obs} to D^2_{obs} again $\mathbb{I}((\#D_i \geq D_{obs}) \leq k) = 1$ since there are only two of D_is that are larger than D_{obs}.
  3. This holds until $D_{obs} = D^k_{obs}$ and then for the rest of the set $\mathcal{D}$, $\mathbb{I}((\#D_i \geq D_{obs}) \leq k) = 0$.

□

Appendix

Here we briefly review some of the results that you should remember from your inference course.

As we have seen many times in class, if $X_1, \ldots, X_n \overset{iid}{\sim} N(0,1)$, then $\sum_{i=1}^{n} X_i^2$ has a χ^2 distribution with n degrees of freedom.

Theorem

Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, 0)$. Define $\bar{X} \triangleq \frac{1}{n}\sum_{i=1}^{n} X_i$. Then $\sum_{i=1}^{n}(X_i - \bar{X})^2$ is χ^2 with n-1 degrees of freedom.

As you can see above the degrees of freedom has reduced from n to $n-1$. The main reason is that $\sum_{i=1}^{n}(X_i - \bar{X}) = 0$, i.e., they are linearly dependent. The proof of this lemma is optional (you do not have to study that). But, I mention it here for the interested readers.

Appendix

First we prove the following lemma

Lemma

Suppose that $Z \sim N(0, I_n)$, where I_n is the $n \times n$ identity matrix. Also, let P be a symmetric matrix that satisfies $P^2 = P$. Let r denote the rank of P. Then $Z^T PZ \sim \chi^2_r$, i.e., a χ^2 with r degrees of freedom.

Appendix

Proof.

- Let z be an eigenvector of P that corresponds to eigenvalue λ. We have
$$P^2 = P \Rightarrow P^2 z = Pz = \lambda z.$$
- However $P^2 z = P(Pz) = P(\lambda z) = \lambda Pz = \lambda^2 z$.
- Hence, $\lambda^2 z = \lambda z$ for every eigenvector. Since $z \neq 0$, we have $\lambda^2 = \lambda$.
- In other words, all the eigenvalues of P are either 0 or 1.
- This means that $Tr(P)$ is equal to the number of nonzero eigenvalues of P.
- Since, P is symmetric, there exists an orthonormal matrix Q, i.e., $Q^T Q = QQ^T = I$, (constructed by the eigenvectors of P) and a diagonal matrix Λ constructed by the eigenvalues of P such that:
$$P = Q \Lambda Q^T.$$

□

Appendix

Proof.

- Hence, $Z^T P Z = Z^T Q \Lambda Q^T Z$. Define $X \triangleq Q^T Z$.
- It is clear that X is a Gaussian vector with zero mean and identity covariance matrix. Why?
- Finally, we have $Z^T P Z = \sum_{i=1}^{n} \lambda_i X_i^2$, where λ_i is the i^{th} eigenvalue.
- If the first eigenvalues are non-zero, then we can simplify $Z^T P Z = \sum_{i=1}^{Tr(P)} \lambda_i X_i^2$, which proves the result, since all the nonzero eigenvalues are equal to 1.

□

Appendix

Now we get back to the proof of the Theorem:

Proof.

- Define the matrix $P = \frac{1}{n}\mathbf{1}_{n\times n}$, where $\mathbf{1}$ denotes the matrix whose elements are all equal to 1.
- Also define the random variables $Z_i = X_i - \bar{X}$.
- Our goal is to convert this problem to Lemma.
- Note that if Z represents the vector of Z_i it can be written as $Z = (I - P)X$.
- The matrix $P = \frac{1}{n}\mathbf{1}_{n\times n}$ is symmetric and idempotent.
- Then so is $I_n - P$:

$$(I_n - P)(I_n - P) = I_n - 2P + P^2 = I_n - P.$$

- Then, by the last Lemma,

$$W = ||Y - PY||^2 = Y^\top (I_n - P)Y \sim \chi^2_{n-1}.$$

□

Appendix

The next lemma that is very useful in the T-test is the following lemma.

Lemma

Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, 0)$*. Define* $\bar{X} \triangleq \frac{1}{n}\sum_{i=1}^{n} X_i$*. Then* $\sum_{i=1}^{n}(X_i - \bar{X})^2$ *is independent of* $\bar{X}$*.*

Appendix

Proof.

We only write the proof sketch here and you should be able to complete the proof. First, note that $\bar{X}$ and $X_i - \bar{X}$ are jointly Gaussian.[a] Why? Therefore, they are independent if and only if the covariance of these two random variables is zero, i.e.

$$E[(\bar{X} - \mu)(X_i - \bar{X})] = 0. \tag{6}$$

Can you prove that (6) holds. Therefore, $\bar{X}$ is independent of $X_1 - \bar{X}, X_2 - \bar{X}, \ldots, X_n - \bar{X}$. Hence, $\bar{X}$ is independent of any function of $X_1 - \bar{X}, X_2 - \bar{X}, \ldots, X_n - \bar{X}$. This implies that $\bar{X}$ is independent of $\sum_{i=1}^{n}(X_i - \bar{X})^2$. □

[a] It is so distributed because every linear combination $a_1(X_i - \bar{X}) + a_2\bar{X}$ a 1-dimensional normal distribution.

Appendix

Lemma

If $X_1 \sim \chi^2(k)$ and $X_2 \sim \chi^2(m)$ independent of X_1, then $X_1 + X_2 \sim \chi^2(k+m)$.

Note that if X_1 is $\chi^2(k)$, it means that it can be written as $Z_1^2, \ldots, Z_k^2$, where $Z_i \overset{iid}{\sim} N(0,1)$. X_2 is the summation of square of m iid Gaussians. Hence, the summation of the two is the summation of the squares of $m + k$ iid Gaussians and hence it is $\chi^2(m+k)$.