CHAPTER

HYPOTHESIS TESTING 10

CHAPTER CONTENTS

When tossing a coin 20 times, heads are obtained 17 times. Can we then conclude that the coin is biased? The framework of *hypothesis testing* allows us to answer this question statistically. In this chapter, the basic idea of hypothesis testing and standard tests is introduced.

10.1 FUNDAMENTALS OF HYPOTHESIS TESTING

The hypothesis that we want to test is called a *null hypothesis*, while the opposite is called the *alternative hypothesis*. In the above coin-toss example, the null hypothesis is that the coin is not biased (i.e., the probability of obtaining heads is $1/2$), while the alternative hypothesis is that the coin is biased (i.e., the probability of obtaining heads is not $1/2$). The null hypothesis and the alternative hypothesis are often denoted as H_0 and H_1, respectively.

In hypothesis testing, probability that the current samples are obtained under the null hypothesis is computed. If the probability, called the *p-value*, is less than the pre-specified *significance level* α, then the null hypothesis is *rejected*; otherwise the null hypothesis is *accepted*. Conventionally, significance level α is set at either 5% or 1%.

As shown in Section 3.2, the probability of obtaining heads in coin tossing follows the binomial distribution. Thus, if the coin is not biased (i.e., the probability of obtaining heads is $1/2$), the probability that heads are obtained more than or equal to

An Introduction to Statistical Machine Learning. DOI: 10.1016/B978-0-12-802121-7.00021-2

17 times for 20 trials is given by

$$\left(\binom{20}{17}+\binom{20}{18}+\binom{20}{19}+\binom{20}{20}\right)\times\left(\frac{1}{2}\right)^{20}\approx 0.0013.$$

If significance level α is set at 0.01, 0.0013 is less than the significance level. Thus, the null hypothesis is rejected and the alternative hypothesis is accepted, and the coin is concluded to be biased. Note that the probabilities of observing heads more than or equal to 16, 15, and 14 times are 0.0059, 0.0207, and 0.0577, respectively. Thus, if heads is observed no more than 15 times, the null hypothesis is accepted under significance level $\alpha = 0.01$ and the coin is concluded not to be biased.

As illustrated above, when rejecting a null hypothesis by hypothesis testing, the null hypothesis is shown to seldom occur based on samples. On the other hand, when a null hypothesis is accepted, its validity is not actively proved—there is no strong enough evidence that the null hypothesis is wrong and thus the null hypothesis is accepted inevitably. Such a logic is called *proof by contradiction*.

A *two-sided test* is aimed at testing whether the observed value is equal to a target value. For example, if a computationally efficient algorithm of a machine learning method is developed, a two-sided test is used to confirm whether the same performance can still be obtained by the new algorithm. On the other hand, a *one-sided test* is aimed at testing whether the observed value is larger (or smaller) than a target value. If a new machine learning method is developed, a one-sided test is used to see whether superior performance can be obtained by the new method.

In hypothesis testing, a *test statistic* z that can be computed from samples is considered, and its probability distribution is computed under the null hypothesis. If the value $\widehat{z}$ of the test statistic computed from the current samples can occur only with low probability, then the null hypothesis is rejected; otherwise the null hypothesis is accepted. The region to which rejected $\widehat{z}$ belongs is called a *critical region*, and its threshold is called the *critical value*. The critical region in a two-sided test is the left and right tails of the probability mass/density of test statistic z, while that in a one-sided test is the left (or right) tail (see Fig. 10.1).

10.2 TEST FOR EXPECTATION OF NORMAL SAMPLES

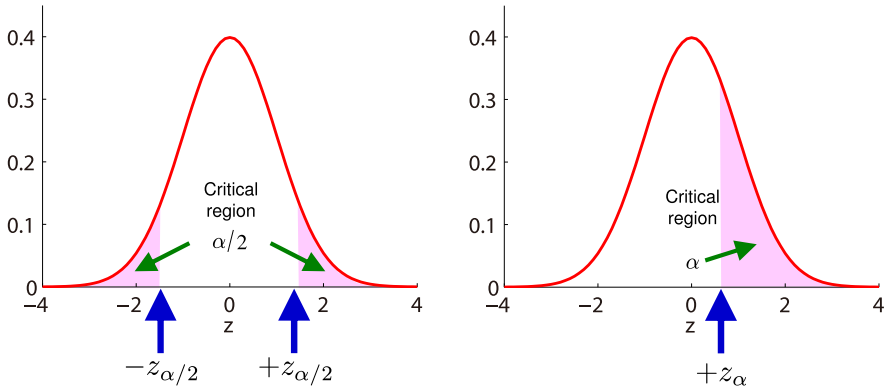
For one-dimensional i.i.d. normal samples $x_1,\ldots,x_n$ with variance σ^2, a test for the null hypothesis that its expectation is μ is introduced.

Since the sample average,

$$\widehat{\mu}=\frac{1}{n}\sum_{i=1}^{n}x_i,$$

follows normal distribution $N(\mu,\sigma^2/n)$ under the null hypothesis that the expectation is μ, its standardization,

$$z=\frac{\widehat{\mu}-\mu}{\sqrt{\sigma^2/n}},$$



(a) Two-sided test (critical value $z_{\alpha/2}$) (b) One-sided test (critical value z_α)

FIGURE 10.1

Critical region and critical value.

follows the standard normal distribution $N(0,1)$. The hypothesis test that uses the above z as a test statistic is called a *z-test*. The critical region and critical values are set in the same way as in Fig. 10.1.

When the variance σ^2 is unknown, it is replaced with an unbiased estimator,

$$\widehat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \widehat{\mu})^2.$$

Then the test statistic,

$$t = \frac{\widehat{\mu} - \mu}{\sqrt{\widehat{\sigma}^2/n}},$$

follows the t-distribution with $n-1$ degrees of freedom under the null hypothesis that the expectation is μ [7]. This is called a *t-test*.

10.3 NEYMAN-PEARSON LEMMA

The error that the correct null hypothesis is rejected is called the *type-I error* or *false positive*. In the coin-toss example, the type-I error corresponds to concluding that an unbiased coin is biased. On the other hand, the error that the incorrect null hypothesis is accepted is called the *type-II error* or *false negative*, which corresponds to concluding that a biased coin is unbiased (Table 10.1). The type-II error is denoted by β, and $1-\beta$ is called the *power* of the test. In the framework of hypothesis testing, the type-I error is set at α, and the type-II error is reduced (equivalently the power is increased) as much as possible.

Table 10.1 Type-I Error α (False Positive) and Type-II Error β (False Negative)

	Truth	
Test Result	**Null Hypothesis is Correct**	**Alternative Hypothesis is Correct**
Null hypothesis is accepted	OK	Type-II error (false negative)
Alternative hypothesis is accepted	Type-I error (false positive)	OK

Suppose that null hypothesis $\theta = \theta_0$ and alternative hypothesis $\theta = \theta_1$ are tested using samples $\mathcal{D} = \{x_1, \ldots, x_n\}$. Under the null hypothesis $\theta = \theta_0$, let η_α be a scalar such that

$$\Pr\left(\frac{L(\theta_0)}{L(\theta_1)} \leq \eta_\alpha\right) = \alpha,$$

where $L(\theta)$ is the likelihood. Then setting the critical value at η_α and the critical region as

$$\frac{L(\theta_0)}{L(\theta_1)} \leq \eta_\alpha$$

minimizes the type-II error (equivalently maximizes the power) subject to the constraint that the type-I error is fixed at α. This is called the *Neyman-Pearson lemma* and a test based on the ratio of likelihoods is called a *likelihood-ratio test*.

10.4 TEST FOR CONTINGENCY TABLES

In this section, a *goodness-of-fit test* and an *independence test* for contingency tables (see Table 10.2) are introduced. For discrete random variables $x \in \{1, \ldots, \ell\}$ and $y \in \{1, \ldots, m\}$, the *Pearson divergence* from $p_{x,y}$ to $q_{x,y}$ is considered below:

$$\sum_{x=1}^{\ell} \sum_{y=1}^{m} \frac{(p_{x,y} - q_{x,y})^2}{q_{x,y}}. \tag{10.1}$$

In the goodness-of-fit test, the null hypothesis that the sample joint probability mass function $\widehat{f}(x, y) = c_{x,y}/n$ is equivalent to a target value $f(x, y)$ is tested. More specifically, Pearson divergence (10.1) for

$$p_{x,y} = \widehat{f}(x, y) \quad \text{and} \quad q_{x,y} = f(x, y)$$

Table 10.2 **Contingency Table for $x \in \{1,\dots,\ell\}$ and $y \in \{1,\dots,m\}$. $c_{x,y}$ Denotes the Frequency of (x,y), $d_x = \sum_{y=1}^{m} c_{x,y}$, $e_y = \sum_{x=1}^{\ell} c_{x,y}$, and $n = \sum_{x=1}^{\ell}\sum_{y=1}^{m} c_{x,y}$**

$x \setminus y$	1	$\cdots$	m	Total
1	$c_{1,1}$	$\cdots$	$c_{1,m}$	d_1
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
ℓ	$c_{\ell,1}$	$\cdots$	$c_{\ell,m}$	d_ℓ
Total	e_1	$\cdots$	e_m	n

is used as a test statistic, and the critical region is computed based on the fact that the Pearson divergence follows the chi-squared distribution with $\ell m - 1$ degrees of freedom [7].

In the independence test, statistical independence between random variables x and y is tested by considering the null hypothesis that the sample joint probability mass function $\widehat{f}(x,y) = c_{x,y}/n$ is equivalent to the product of marginals $\widehat{g}(x)\widehat{h}(y)$, where

$$\widehat{g}(x) = \frac{d_x}{n} = \frac{1}{n}\sum_{y=1}^{m} c_{x,y},$$

$$\widehat{h}(y) = \frac{e_y}{n} = \frac{1}{n}\sum_{x=1}^{\ell} c_{x,y}.$$

More specifically, Pearson divergence (10.1) for

$$p_{x,y} = \widehat{f}(x,y) \quad \text{and} \quad q_{x,y} = \widehat{g}(x)\widehat{h}(y)$$

is used as a test statistic, and the critical region is computed based on the fact that the Pearson divergence follows the chi-squared distribution with $(\ell - 1)(m - 1)$ degrees of freedom when x and y follow multinomial distributions.

The test that uses the *Kullback-Leibler (KL) divergence* (see Section 14.2),

$$\sum_{x=1}^{\ell}\sum_{y=1}^{m} p_{x,y} \log \frac{p_{x,y}}{q_{x,y}},$$

instead of the Pearson divergence is called a *G-test*. This is a likelihood-ratio test and the KL divergence approximately follows the chi-squared distribution with $(\ell - 1)(m - 1)$ degrees of freedom.

A test whose test statistic (approximately) follows the chi-squared distribution under the null hypothesis is called a *chi-square test* [7].

10.5 TEST FOR DIFFERENCE IN EXPECTATIONS OF NORMAL SAMPLES

Let $\mathcal{D} = \{x_1, \ldots, x_n\}$ and $\mathcal{D}' = \{x'_1, \ldots, x'_{n'}\}$ be the i.i.d. samples with normal distributions $N(\mu, \sigma^2)$ and $N(\mu', \sigma^2)$, where the variance is common but the expectations can be different. In this section, a test for the difference in expectation, $\mu - \mu'$, is introduced.

10.5.1 TWO SAMPLES WITHOUT CORRESPONDENCE

Let $\widehat{\mu}$ and $\widehat{\mu}'$ be the sample averages for $\mathcal{D} = \{x_1, \ldots, x_n\}$ and $\mathcal{D}' = \{x'_1, \ldots, x'_{n'}\}$:

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and} \quad \widehat{\mu}' = \frac{1}{n'}\sum_{i=1}^{n'} x'_i.$$

Since $\mathcal{D} = \{x_1, \ldots, x_n\}$ and $\mathcal{D}' = \{x'_1, \ldots, x'_{n'}\}$ are statistically independent of each other, the variance of the difference in sample average, $\widehat{\mu} - \widehat{\mu}'$, is given by $\sigma^2(1/n + 1/n')$ under the null hypothesis $\mu = \mu'$. Thus, its standardization,

$$z_{\mathrm{u}} = \frac{\widehat{\mu} - \widehat{\mu}'}{\sqrt{\sigma^2(1/n + 1/n')}},$$

follows the standard normal distribution $N(0, 1)$. The test that uses the above z_{u} as a test statistic is called an *unpaired z-test*. The critical region and critical values are set in the same way as in Fig. 10.1.

When the variance σ^2 is unknown, it is replaced with an unbiased estimator:

$$\widehat{\sigma}^2_{\mathrm{u}} = \frac{\sum_{i=1}^{n}(x_i - \widehat{\mu})^2 + \sum_{i=1}^{n'}(x'_i - \widehat{\mu}')^2}{n + n' - 2}.$$

Then the test statistic,

$$t_{\mathrm{u}} = \frac{\widehat{\mu} - \widehat{\mu}'}{\sqrt{\widehat{\sigma}^2_{\mathrm{u}}(1/n + 1/n')}},$$

follows the t-distribution with $n + n' - 2$ degrees of freedom under the null hypothesis $\mu = \mu'$. This is called an *unpaired t-test*.

If the variances of $\mathcal{D} = \{x_1, \ldots, x_n\}$ and $\mathcal{D}' = \{x'_1, \ldots, x'_{n'}\}$ can be different, the variances, say σ^2 and σ'^2, are replaced with their unbiased estimators:

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i - \widehat{\mu})^2}{n - 1},$$
$$\widehat{\sigma}'^2 = \frac{\sum_{i=1}^{n'}(x'_i - \widehat{\mu}')^2}{n' - 1}.$$

Then, under the null hypothesis $\mu = \mu'$, the test statistic,

$$t_{\mathrm{W}} = \frac{\widehat{\mu} - \widehat{\mu}'}{\sqrt{\widehat{\sigma}^2/n + \widehat{\sigma}'^2/n'}},$$

approximately follows the t-distribution with round(k) degrees of freedom, where

$$k = \frac{\left(\widehat{\sigma}^2/n + \widehat{\sigma}'^2/n'\right)^2}{\widehat{\sigma}^4/(n^2(n-1)) + \widehat{\sigma}'^4/(n'^2(n'-1))},$$

and round($\cdot$) rounds off the value to the nearest integer. This is called *Welch's t-test*.

Under the null hypothesis $\sigma^2 = \sigma'^2$, the test statistic,

$$F = \frac{\widehat{\sigma}^2}{\widehat{\sigma}'^2},$$

follows the F-distribution explained in Section 4.6. This is called the *F-test*, which allows us to test the equality of the variance.

10.5.2 TWO SAMPLES WITH CORRESPONDENCE

Suppose that two sets of samples, $\mathcal{D} = \{x_1, \ldots, x_n\}$ and $\mathcal{D}' = \{x'_1, \ldots, x'_{n'}\}$, have correspondence, i.e., for $n = n'$, the samples are paired as

$$\{(x_1, x'_1), \ldots, (x_n, x'_n)\}.$$

Then, whether the expectations of $\mathcal{D}$ and $\mathcal{D}'$ are equivalent can be tested by the unpaired z-test:

$$z_{\mathrm{u}} = \frac{\Delta\widehat{\mu}}{\sqrt{2\sigma^2/n}},$$

where $\Delta\widehat{\mu}$ is the average of the difference between the paired samples:

$$\Delta\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n}(x_i - x'_i) = \widehat{\mu} - \widehat{\mu}'.$$

In this situation, the power of the test (see Section 10.3) can be improved if positive correlation exists between the two samples. More specifically, under the null hypothesis $\mu - \mu' = 0$, the variance of $\Delta\widehat{\mu}$ is given by $2\sigma^2(1-\rho)/n$, where ρ is the correlation coefficient:

$$\rho = \frac{\mathrm{Cov}[x, x']}{\sqrt{V[x]}\sqrt{V[x']}}.$$

Then its standardization,

$$z_{\mathrm{p}} = \frac{\Delta\widehat{\mu}}{\sqrt{2\sigma^2(1-\rho)/n}},$$

follows the standard normal distribution $N(0,1)$. The test that uses the above z_{p} as a test statistic is called a *paired z-test*. If $\rho > 0$,

$$|z_{\mathrm{p}}| > |z_{\mathrm{u}}|$$

holds and thus the power of the test can be improved.

When the variance σ^2 is unknown, it is replaced with

$$\widehat{\sigma}_{\mathrm{p}}^2 = \frac{\sum_{i=1}^{n}(x_i - x_i' - \Delta\widehat{\mu})^2}{2(n-1)}. \tag{10.2}$$

Then the test statistic,

$$t_{\mathrm{p}} = \frac{\Delta\widehat{\mu}}{\sqrt{2\widehat{\sigma}_{\mathrm{p}}^2(1-\rho)/n}},$$

follows the t-distribution with $n-1$ degrees of freedom under the null hypothesis $\mu - \mu' = 0$. This is called a *paired t-test*.

The test statistic of the unpaired t-test for $n = n'$ can be expressed as

$$t_{\mathrm{u}} = \frac{\Delta\widehat{\mu}}{\sqrt{2\widehat{\sigma}_{\mathrm{u}}^2/n}},$$

where

$$\widehat{\sigma}_{\mathrm{u}}^2 = \frac{\sum_{i=1}^{n}\left((x_i - \widehat{\mu})^2 + (x_i' - \widehat{\mu}')^2\right)}{2(n-1)}.$$

$\widehat{\sigma}_{\mathrm{u}}^2$ and $\widehat{\sigma}_{\mathrm{p}}^2$ defined in Eq. (10.2) can be expressed by using $\widehat{\sigma}_{\mathrm{u}}^2$ as

$$\begin{aligned}
\widehat{\sigma}_{\mathrm{p}}^2 &= \frac{\sum_{i=1}^{n}(x_i - x_i' - \Delta\widehat{\mu})^2}{2(n-1)} \\
&= \frac{\sum_{i=1}^{n}((x_i - \mu) - (x_i' - \mu'))^2}{2(n-1)} \\
&= \frac{\sum_{i=1}^{n}(x_i - \mu)^2 + \sum_{i=1}^{n}(x_i' - \mu')^2 - 2\sum_{i=1}^{n}(x_i - \mu)(x_i' - \mu')}{2(n-1)} \\
&= \widehat{\sigma}_{\mathrm{u}}^2 - \widehat{\mathrm{Cov}}[x, x'],
\end{aligned}$$

where $\widehat{\mathrm{Cov}}[x, x']$ is the sample covariance given by

$$\widehat{\mathrm{Cov}}[x, x'] = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)(x_i' - \mu').$$

If $\widehat{\mathrm{Cov}}[x, x'] > 0$,

$$|t_{\mathrm{p}}| > |t_{\mathrm{u}}|$$

holds and thus the power of the test can be improved.

Table 10.3 **Wilcoxon Rank-Sum Test. In this Example,** $r_1 = 3$, $r_2 = 5.5$, $r_3 = 1$, **and the Rank-Sum is** $r = 9.5$

$\mathcal{D}$	x_3		x_1		x_2		
$\mathcal{D}'$		x'_2		x'_4		x'_3	x'_1
Sample value	−2	0	1	3.5	7	7	7.1
Rank	1	2	3	4	5.5	5.5	7

10.6 NONPARAMETRIC TEST FOR RANKS

In the previous section, samples $\mathcal{D} = \{x_1, \ldots, x_n\}$ and $\mathcal{D}' = \{x'_1, \ldots, x'_{n'}\}$ were assumed to follow normal distributions. When samples do not follow the normal distributions, particularly in the presence of outliers, tests based on the normality may not be reliable. In this section, *nonparametric tests* that do not require parametric assumptions on the probability distributions are introduced.

10.6.1 TWO SAMPLES WITHOUT CORRESPONDENCE

Without loss of generality, assume $n \leq n'$ below (if $n > n'$, just $\mathcal{D}$ and $\mathcal{D}'$ are swapped to satisfy $n \leq n'$).

Let us merge all samples $x_1, \ldots, x_n$ and $x'_1, \ldots, x'_{n'}$ together and sort them in the ascending order. Let us denote the ranks of $x_1, \ldots, x_n$ in the set of $n + n'$ samples by $r_1, \ldots, r_n$. If there are ties, the mean rank is used. For example, if x_i the third smallest sample in $x_1, \ldots, x_n$ and $x'_1, \ldots, x'_{n'}$, r_i is set at 3; if the fifth and sixth smallest samples share the same value, their ranks are 5.5 (Table 10.3). The *Wilcoxon rank-sum test* uses the sum of the ranks of $x_1, \ldots, x_n$,

$$r = \sum_{i=1}^{n} r_i,$$

a test statistic.

Under the null hypothesis that $\mathcal{D}$ and $\mathcal{D}'$ follow the same probability distribution, the above test statistic r approximately follows the normal distribution with expectation and variance given by

$$\mu = \frac{n(n + n' + 1)}{2},$$
$$\sigma^2 = \frac{nn'(n + n' + 1)}{12}.$$

Since the standardized statistic $(r - \mu)/\sigma$ follows the standard normal distribution $N(0, 1)$, setting the critical region and critical values as in Fig. 10.1 allows us to perform hypothesis testing.

The *Mann-Whitney U-test* is essentially the same as the Wilcoxon rank-sum test.

10.6.2 TWO SAMPLES WITH CORRESPONDENCE

For $n = n'$, suppose that two sets of samples $\mathcal{D} = \{x_1, \ldots, x_n\}$ and $\mathcal{D}' = \{x'_1, \ldots, x'_{n'}\}$ have correspondence as

$$\{(x_1, x'_1), \ldots, (x_n, x'_n)\}.$$

Let us remove pairs such that $x_i = x'_i$ and reduce the value of n accordingly (i.e., n denotes the number of sample pairs such that $x_i \neq x'_i$ below). Let us sort the sample pairs in the ascending order of $|x_i - x'_i|$, and let r_i be its rank. If there are ties, the mean rank is used in the same way as in the Wilcoxon rank-sum test. Then the *Wilcoxon signed-rank test* uses the sum of the ranks of samples such that $x_i - x'_i > 0$,

$$s = \sum_{i: x_i - x'_i > 0} r_i,$$

a test statistic.

Under the null hypothesis that $\mathcal{D}$ and $\mathcal{D}'$ follow the same probability distribution, the above test statistic s approximately follows the normal distribution with expectation and variance given by

$$\mu = \frac{n(n+1)}{4},$$
$$\sigma^2 = \frac{n(n+1)(2n+1)}{24}.$$

Since the standardized statistic $(s - \mu)/\sigma$ follows the standard normal distribution $N(0, 1)$, setting the critical region and critical values as in Fig. 10.1 allows us to perform hypothesis testing.

10.7 MONTE CARLO TEST

The test statistics introduced above all (approximately) follow the normal distribution, t-distribution, and chi-squared distribution under the null hypothesis. However, if a test statistic is more complicated, its distribution cannot be analytically derived even approximately. In such a situation, computing the value of a test statistic using samples generated by a *Monte Carlo method* and numerically obtaining the critical region are practically useful. The Monte Carlo method is a generic name of algorithms that use random numbers, and its name stems from the Monte Carlo Casino in Monaco. A test based on the Monte Carlo method is called a *Monte Carlo test*.

For testing whether the expectation of samples $\mathcal{D} = \{x_1, \ldots, x_n\}$ is μ (which was discussed in Section 10.2) by a Monte Carlo test, the bootstrap method introduced in Section 9.3.2 is used. More specifically, bootstrap resampling of $\mathcal{D} = \{x_1, \ldots, x_n\}$ and computing their average are repeated many times, and a histogram of the average is constructed. Then hypothesis testing can be approximately performed by verifying

whether the target value μ is included in the critical region (see Fig. 10.1). The p-value can also be approximated from the histogram. As illustrated above, the bootstrap-based hypothesis test is highly general and can be applied to any statistic computed from samples following any probability distribution.

In the contingency table explained in Section 10.4, enumeration of all possible combinations allows us to obtain the probability distribution of any test statistic. For 2×2 contingency tables, computing the p-value by such an exhaustive way is called *Fisher's exact test*. In the Monte Carlo test, combinations are randomly generated and the test is performed approximately. Since this allows us to numerically approximate the p-value for contingency tables of arbitrary size and for arbitrary test statistic, it can be regarded as generalization of Fisher's exact test.

For testing the null hypothesis that $\mathcal{D} = \{x_1, \ldots, x_n\}$ and $\mathcal{D}' = \{x'_1, \ldots, x'_{n'}\}$ follow the same probability distributions, let us merge all samples $x_1, \ldots, x_n$ and $x'_1, \ldots, x'_{n'}$ together and partition them into two sets with sizes n and n'. Then enumeration of all possible partitions allows us to obtain the probability distribution of any test statistic. This is called a *permutation test*, and the Monte Carlo test can be regarded as its approximate implementation with a limited number of repetitions.