Multiple Hypothesis Testing

Tianhang Zheng

https://tianzheng4.github.io

Multiple Hypothesis Testing

A single null hypothesis might look like H_0: the expected blood pressures of mice in the control and treatment groups are the same.

We will now consider testing m null hypotheses, H_{01},…, H_{0m}, where e.g. H_{0j} : the expected values of the jth biomarker among mice in the control and treatment groups are equal.

A Quick Review of Hypothesis Testing

Hypothesis tests allow us to answer simple “yes-or-no” questions, such as:

Is the true coefficient in a linear regression equal to zero?

Does the expected blood pressure among mice in the treatment group equal the expected blood pressure among mice in the control group?

Process of Hypothesis Testing

Hypothesis testing proceeds as follows:

1. Define the null and alternative hypotheses

2. Construct the test statistic (t-statistics, F-statistics)

3. Compute the p-value

4. Decide whether to reject the null hypothesis

Define the Null Hypotheses

We divide the world into null and alternative hypotheses.

The null hypothesis, H_0, is the default state of belief about the world. For instance:

The true coefficient equals zero.

There is no difference in the expected blood pressures.

Define the Alternative Hypotheses

The alternative hypothesis, H_a, represents something different and unexpected. For instance:

The true coefficient is non-zero.

There is a difference in the expected blood pressures.

Construct the Test Statistic

The test statistic summarizes the extent to which our data are consistent with H_0.

Let $\hat{\mu}_t$ / $\hat{\mu}_c$ respectively denote the average blood pressure for the n_t / n_c mice in the treatment and control groups.

To test $H_0 : \mu_t = \mu_c$, we use a two-sample t-statistic

$$T = \frac{\hat{\mu}_t - \hat{\mu}_c}{s\sqrt{\frac{1}{n_t} + \frac{1}{n_c}}}$$

P-Value

The p-value is the probability of observing a test statistic at least as extreme as the observed statistic, under the assumption that H_0 is true.

A small p-value provides evidence against H_0 .

A large p-value indicates that H_0 is likely to be true.

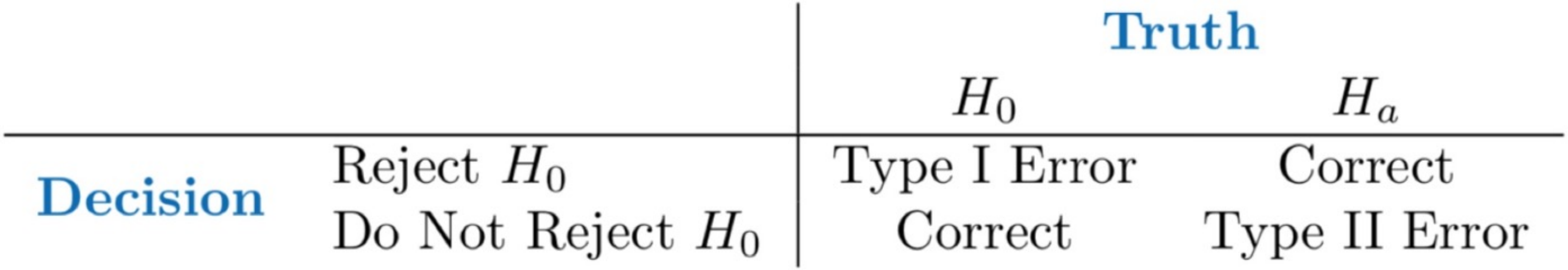
Decide Whether to Reject $\boldsymbol{H_0}$

A small p-value indicates that such a large value of the test statistic is unlikely to occur under H_0.

A small p-value provides evidence against H_0 .

If the p-value is sufficiently small, then we will want to reject H0 (and, therefore, make a potential “discovery”).

Type I Error and Type II Error



		Truth	
		H_0	H_a
Decision	Reject H_0	Type I Error	Correct
	Do Not Reject H_0	Correct	Type II Error

Type I Error Rate

The Type I error rate is the probability of making a Type I error.

We want to ensure a small Type I error rate.

If we only reject H_0 when the p-value is less than α, then the Type I error rate will be at most α.

So, we reject H_0 when the p-value falls below some α: often we choose α to equal 0.05 or 0.01 or 0.001.

Multiple Testing

Now suppose that we wish to test m null hypotheses

Can we simply reject all null hypotheses for which the corresponding p-value falls below (say) 0.01?

If we reject all null hypotheses for which the p-value falls below 0.01, then how many Type I errors will we make?

The Challenge of Multiple Testing

Suppose we test H_{01},…, H_{0m}, all of which are true, and reject any null hypothesis with a p-value below 0.01.

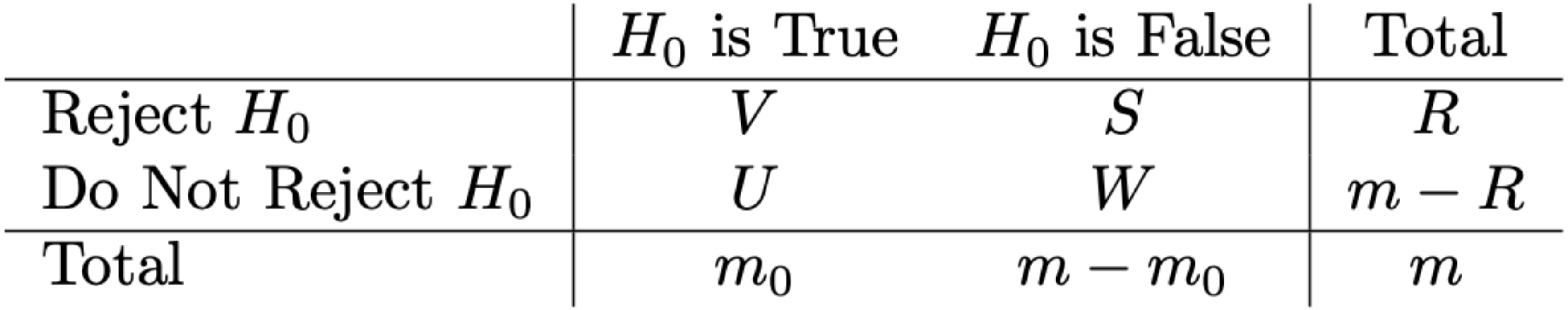
Then we expect to falsely reject approximately 0.01 × m null hypotheses.

If m = 10,000, then we expect to falsely reject 100 null hypotheses by chance!

The Family-Wise Error Rate

The family-wise error rate (FWER) is the probability of making at least one Type I error when conducting m hypothesis tests.

FWER=Pr(V ≥1)



	H_0 is True	H_0 is False	Total
Reject H_0	V	S	R
Do Not Reject H_0	U	W	$m - R$
Total	m_0	$m - m_0$	m

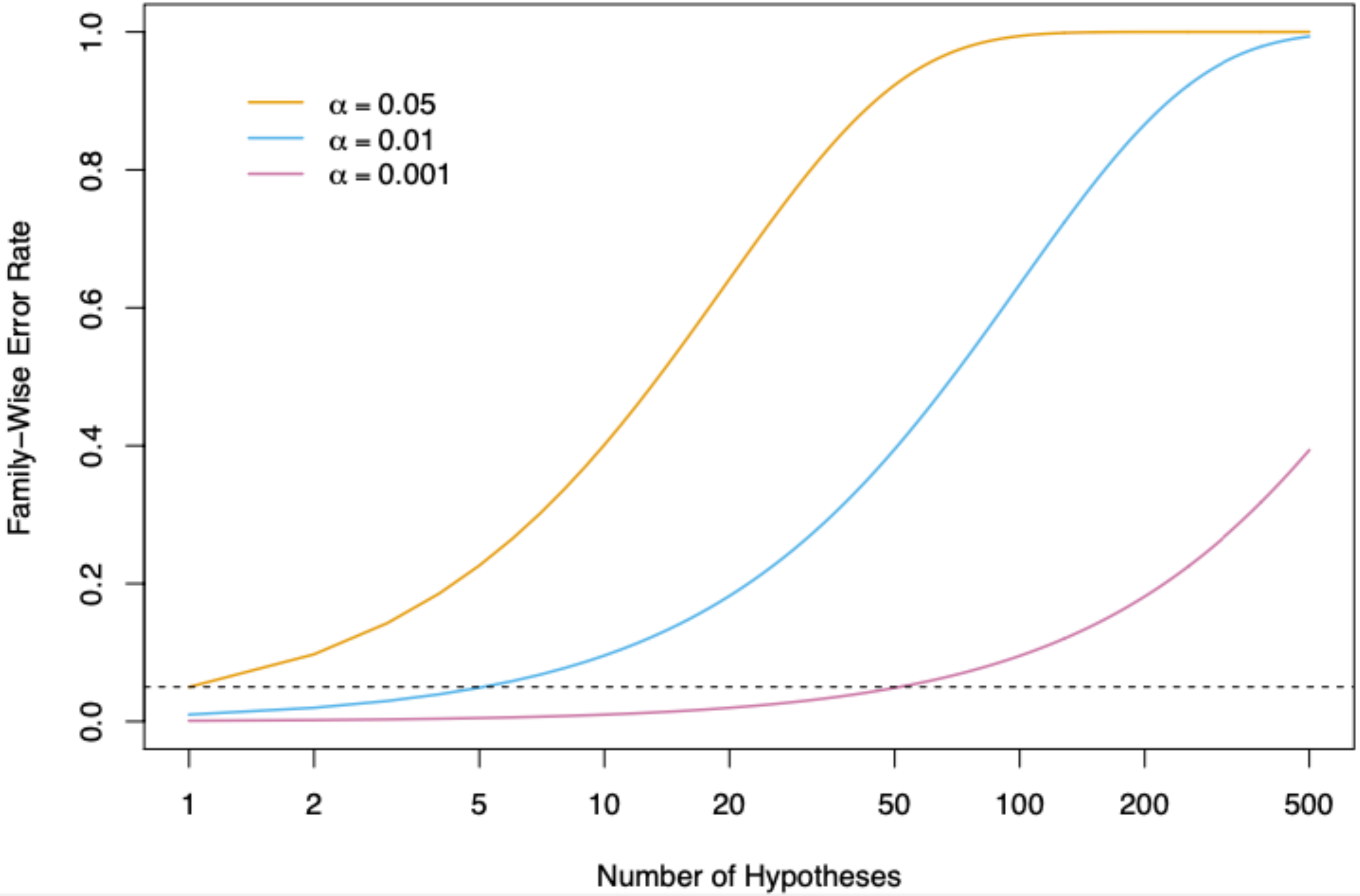
The Family-Wise Error Rate

$$\begin{aligned}\text{FWER} &= 1 - \Pr(\text{do not falsely reject any null hypotheses}) \\ &= 1 - \Pr\left(\textstyle\bigcap_{j=1}^{m} \{\text{do not falsely reject } H_{0j}\}\right).\end{aligned}$$

If the tests are independent and all H_{0j} are true then

$$\text{FWER} = 1 - \prod_{j=1}^{m}(1-\alpha) = 1 - (1-\alpha)^m.$$

The Family-Wise Error Rate



Holm’s Method

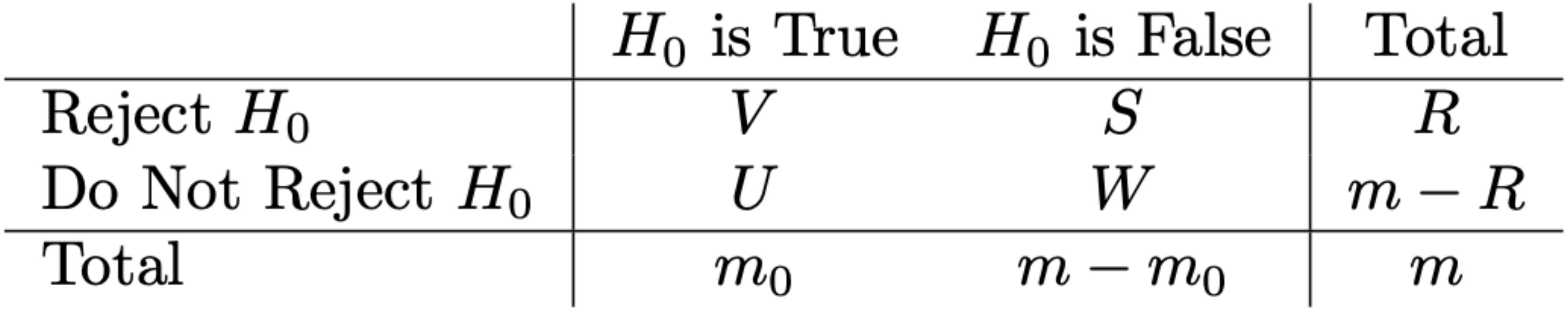
Compute p-values, $p_1, \ldots, p_m$, for the m null hypotheses $H_{01}, \ldots, H_{0m}$.

Order the m p-values so that $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$.

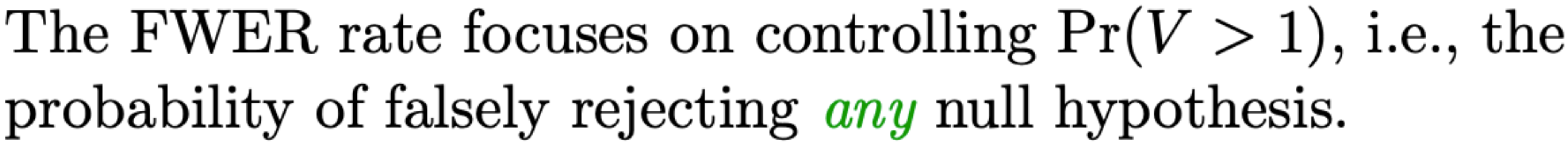
Define

$$L = \min \left\{ j : p_{(j)} > \frac{\alpha}{m+1-j} \right\}.$$

False Discovery Rate



	H_0 is True	H_0 is False	Total
Reject H_0	V	S	R
Do Not Reject H_0	U	W	$m - R$
Total	m_0	$m - m_0$	m



The FWER rate focuses on controlling $\Pr(V > 1)$, i.e., the probability of falsely rejecting *any* null hypothesis.

False Discovery Rate

This is a tough ask when m is large! It will cause us to be super conservative (i.e. to very rarely reject).

Instead, we can control the *false discovery rate*:

$$\mathrm{FDR} = \mathrm{E}\left(\frac{V}{R}\right) = \mathrm{E}\left(\frac{\text{number of false rejections}}{\text{total number of rejections}}\right)$$

Benjamini-Hochberg Procedure to Control FDR

Specify q, the level at which to control the FDR.

Compute p-values $p_1, \ldots, p_m$ for the null hypotheses $H_{01}, \ldots, H_{0m}$.

Order the p-values so that $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$.

Define $L = \max\left\{j : p_{(j)} < qj/m\right\}$.

Re-Sampling

So far, we have assumed that we want to test some null hypothesis H0 with some test statistic T, and that we know (or can assume) the distribution of T under H0.

This allows us to compute the p-value.

What if this theoretical null distribution is unknown?

Two-Sample t-Test

Suppose we want to test $H_0 : E(X) = E(Y)$ versus $H_a : E(X) \neq E(Y)$, using n_X independent observations from X and n_Y independent observations from Y.

The two-sample t-statistic takes the form

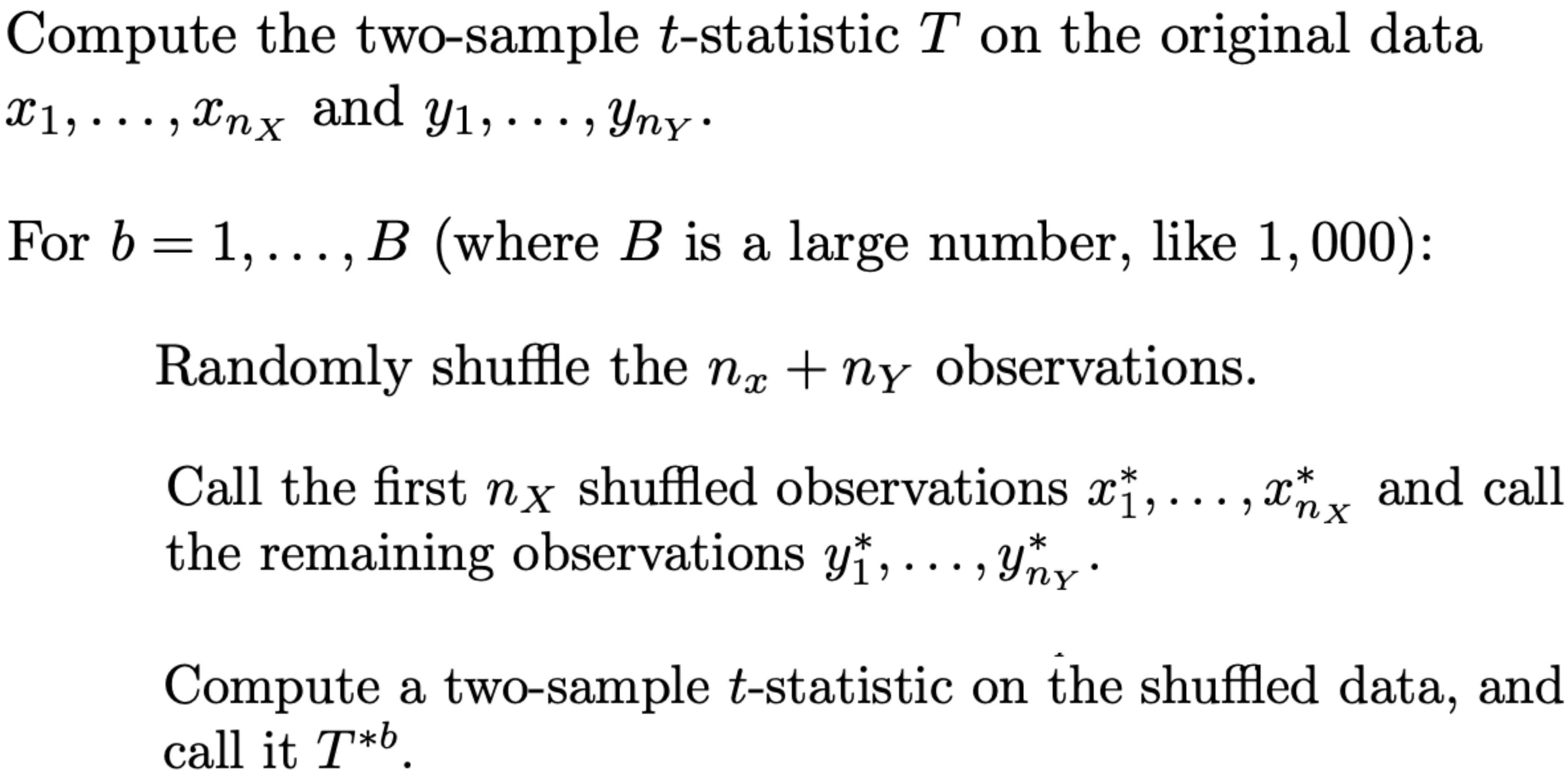
$$T = \frac{\hat{\mu}_X - \hat{\mu}_Y}{s\sqrt{1/n_X + 1/n_Y}}.$$

Two-Sample t-Test

If n_X and n_Y are large, then T approximately follows a $N(0,1)$ distribution under H_0.

If n_X and n_Y are small, then we don't know the theoretical null distribution of T.

Resampling



Compute the two-sample t-statistic T on the original data $x_1, \ldots, x_{n_X}$ and $y_1, \ldots, y_{n_Y}$.

For $b = 1, \ldots, B$ (where B is a large number, like $1,000$):

Randomly shuffle the $n_x + n_Y$ observations.

Call the first n_X shuffled observations $x_1^*, \ldots, x_{n_X}^*$ and call the remaining observations $y_1^*, \ldots, y_{n_Y}^*$.

Compute a two-sample t-statistic on the shuffled data, and call it T^{*b}.

P-Value

The *p*-value is given by

$$\frac{\sum_{b=1}^{B} 1_{(|T^{*b}| \geq |T|)}}{B}.$$

Re-sampling approaches are useful if the theoretical null distribution is unavailable, or requires stringent assumptions.