A discussion of the coin tossing experiment: In a series of coin tosses, how likely THH will appear before HHH occurs? Three students reported their results from computer simulations.

Student 1: $P(HHH\ first) = 0.37$ based on 10,000 trials;

Student 2: the probability is 0.36 based on 100 trials;

Student 3: the probability is 0.13xxx based on 10,000 trials.

The exact probability is 1/8. Statistically we can tell that the first two students made mistakes in their simulations. Why?

Recall the margin of error is $2\sqrt{p(1-p)/n} \leq 1/\sqrt{n}$. For $n = 100$, it is 0.1, and for $n = 10,000$, it is 0.01. So the answers obtained by the first two students are far off; that is, it is very unlikely that with correct computer simulation they will get such answers. For the third student, the answer is possible based on the margin of error calculations.

Back to hypothesis testing:

Example: The reported numbers of regular residents in Shanghai in the last five years show year to year increases. Those numbers have measurement errors. Do we have confidence that the real size of the regular residents in Shanghai is increasing? What is an appropriate statistical analysis for this question?

We may use hypothesis testing. The first step is to identify the parameter of interest. For example, here we can use the slope of the regression line, or the correlation θ. If we take the correlation θ between the number of residents and the year (1,2,3,4,5), we can formulate the problem:

The null hypothesis is $\theta = 0$;

The alternative hypothesis is $\theta > 0$.

In the alternative, we use " $>$ " since our goal is to confirm that the trend is "increas-

ing".

Let's use the p-value to do the job. If we have the p-value, we can compare it with the significance level $\alpha = 0.05$. If the p-value is less than the significance level, we will reject the null hypothesis. The logic here is that, a very small p-value means that if the null hypothesis is true, then the observed one (and more extreme cases) is unlikely to happen, so the assumption is very likely wrong, therefore, we reject the null hypothesis with high confidence.

Is there a unique answer for the p-value? The answer is no. It depends on the testing procedure or the test statistic used.

The sign test: For the Shanghai population example, the data show that every year, the number of residents is higher than that in the previous year. If we take the four yearly changes, they are all positive (++++). Under the null hypothesis that there is no trend, each of the the four signs should be positive or negative with probability 0.5. Then, roughly we should have two "+" and two "-". In this scenario, "++++" is the most extreme case. So the p-value P("++++" or more extreme)=$P(\text{"}++++\text{"}) = 1/16 = 0.0625 > 0.05$.

If one more data point from next year is available and if the increment is also positive, then the p-value would be $1/32 < 0.05$.

In this problem, we may use the the t test based on the correlation r. We have that, under the null hypothesis,

$$t = \frac{\sqrt{n-2}r}{1-r^2} \sim t_{n-2},$$

and the p-value is $P(t_{n-2} >$ observed t). If we use the t test here, then the (one-sided) p-value is less than 0.05. If we use the t-test, we can reject the null hypothesis at level 0.05.

You may say that in this case the t-test is more powerful than the sign test. But this

is not always the case. More importantly, the t-test used here assumes that the variables are normally distributed. This assumption is questionable for at least one of the variables (time).

For the t test, the assumptions include that 1) The errors are normally distributed if the sample size is small. (If normal, the t test is valid and the best; if not normal, and if n is large, then the t test is approximately valid). 2) The errors are independent. For the sign test, the assumptions include that from year to year, the increments are independent.

The t test needs more assumptions, and the sign test loses some information (because it does not use the exact value of each data point). We introduce a new test: **permutation test**.

Consider all the permutations of the data. Under the null hypothesis, all the permutations are equally likely. For $n = 5$, we have 120 permutations. If we calculate all the correlations from the 120 permutations, we get the distribution of r under the null hypothesis, which is called the reference distribution. The p-value is then the proportion of those r's that are equal to or greater than the observed correlation from the original data. The permutation test makes no assumption on the distribution of the variables under consideration.

Example: We ask if the mean income of new college graduates in Beijing is higher than in Shanghai. Let μ_1 be the mean income in Beijing, and μ_2 in Shanghai. We get a sample of 10 people for each city with the following income data: $B_1, \cdots, B_10$ for Beijing, and $S_1, \cdots, S_10$ for Shanghai.

To use the permutation test, we pool the 20 data points together, shuffle them, and use the first 10 for a permuted sample for Beijing, and the rest for Shanghai. Here we have

20! permutations, although many permutations lead to the same samples.

We can use the difference of two sample means $\bar{B} - \bar{S}$ to order the outcome, and then compute the p-value.

The assumptions of the permutation test include 1) independence. 2) Under the null, the two distributions are the same, which is more than the null hypothesis that the two means are equal. But this assumption is weaker than the assumption of "normal, equal variances" for the t test.

For example, from the given data, we get the sample mean difference 10; then we do permutations, and get the sample mean differences based on each permuted data: 2,-1,-3,4,12,3,-1,... From the first 7 permutations only, the p-value is = 2/8=0.25; because two of the values out of 8 are 10 or higher. In reality, we compute the p-value based on a large number of permutations.

Traps in hypothesis testing

Example: A research team reported ten factors that affect how longer people live. Those factors include the habit of eating breakfast, regular exercise, smoking, and getting married. For these ten factors, the statistical tests all have p-values less than 0.01. Usually the scientists would not tell us how many factors are not significant in their studies. For this particular study, the research team considered 500 factors and tested their associations with how long a person lives.

The significance level 0.01 for each individual test means that if none of the 500 factors are relevant, on average, we would find $500 \times 0.01 = 5$ factors to be significant. This means that we do not have much confidence on the reported significance of 10 factors.

Example: A financial advisor claims that one can make a lot of money by following

the trend in a stock, where the trend is identified based on the stock price of the past 50 days. The 50 days form a moving window in the strategy. How does the advisor find this window size of 50 days? The advisor actually used past data and examined the hypothetical performance of an investment strategy based on many different window sizes from 10 to 200 days. The window size of 50 days worked the best! Because it was selected from the past data after so many "tests", the strategy is unlikely to work well in the future. This phenomenon is similar the multiple testing trap of the previous example.

We can see that, if we perform many many tests, we can have false discoveries.

If we have m tests with type-1 error α for each of them, then the familywise type-1 error rate is

$$P(\text{at least making one type-1 error}) = 1 - (1 - \alpha)^m.$$

For $\alpha = 0.05$ and $m = 2$, this probability is about 0.1. But it goes up quickly with m.

By observing that

$$1 - (1 - \alpha)^m \leq m\alpha,$$

if we set the individual significance level to be α/m, then the familywise type-1 error is controlled to be less than α. This adjustment is called Bonferroni correction. For very large m, α/m is very small, then it is hard to detect any factor, which means, the test loses power.

Example: if someone with a headache goes to the hospital and the doctor scans the brain to detect possible problems, the machine would scan many small areas, and a statistical test is preformed on each area to see if something abnormal shows up there. After Bonferroni adjustment, we are unlikely to detect early problems. Without any adjustment, there might be too many false positives. Neither result is desirable.

Table 1:

	test results	
	Positive	Negative
Truth: Positive	True Positives (TP)	False Negatives (FN)
Truth: Negative	False Positives (FP)	True Negatives (TN)
	Total Discoveries (TD)	

In confirmatory analysis, such as testing whether a new drug is effective or not, we do not want to use the new drug due to higher costs and side effects unless it is really useful. So for this case, it is important to control the type-1 error. But for some other scenarios, the goal might be to discover. Then it is more important to see if the discoveries are real. The familywise type-1 error rate applies only if all the null hypotheses are true. This might not be the case in most applications.

When a large number of tests are performed, the relationship between the truth and the statistical results is summarized in the Table 1.

In such studies, it is important to remain confident that what we discover is truly significant. That is, we want to control the **false discovery rate**

$$\frac{\text{FP}}{\text{TD}}.$$

If the FDR is small, then we have confidence that most of what we have discovered is important. The following procedure ensures that the false discovery rate (on average over many experiments) is no more than α.

Benjamini-Hochberg procedure: With m independent tests, we order the hypotheses by their p-values such that $p_1 \leq p_2 \leq \cdots \leq p_m$, then we compare p_i with $i\alpha/m$ for $i = 1, 2, \cdots, m$. Let k be the largest i such that $p_i < i\alpha/m$. We reject the first k null hypotheses corresponding to k smallest p-values.

For example, if $m = 5$ and $\alpha = 0.05$, if we get p-values 0.001, 0.012, 0.08, 0.25, and 0.66, then k=2, and the first two are significant. If we get p-values 0.001, 0.012, 0.034, 0.035, and 0.66, then k=4, and the first four are significant. Recall that the false discovery rate of FP/4 is controlled to be no more than 0.05 on average, so most of the time FP would be zero.

The validity of the Benjamini-Hochberg procedure is based on the fact that each p-value is uniformly distributed on (0,1) if the null hypothesis is true.