Hypothesis Testing

We want to assess the validity of a claim about the population against a counter-claim using sample data. The two claims are:

Null Hypothesis H0 is the claim of "no difference."
Alternative Hypothesis H1 is the claim we are trying to prove.

Example (Vaccine Trial)
Let P1 be the fraction of the control group population who get polio.
P2 is the fraction of the treatment population who get polio.

H0 is the claim that the vaccine is not effective. P1=P2.
H1 is the claim that the vaccine is effective. P1>P2.

The main idea: assume H0 is true. Does the data contradict the assumption beyond a reasonable doubt?
If yes: accept H1 and reject H0.
If no: H0 can't be ruled out as an explanation for the data. *In this case we can't accept any hypothesis.*

Analogy: Presumed innocent until proven guilty
H0: not guilty
H1: guilty
Does the data contradict H0? If yes, then rule as guilty. If no, it doesn't mean the person is innocent but evidence is insufficient to establish guilt, it gives the person the benefit of the doubt.

We can think of a hypothesis test as a method to weigh evidence against H0 (rather than a decision procedure between H0 and H1).

Usually H0 is chosen to represent a hypothesis of "no difference" between the new and existing methods, just chance alone. If we can reject this sort of hypothesis then we have a **statistically significant** proof of claim H1. In this case, the hypothesis test is a **significance test.**

Again, the hypothesis test is not a decision procedure between H0 and H1. If H1 is not accepted, it does not mean we can accept H0. I never want to hear you say: "therefore we accept the null hypothesis".

A **type I error** is made when a test rejects H0 in favor of H1 when H0 is actually true (false positive).

E.g. person does not have the disease but test is positive.

A **type II error** is made when the test fails to reject H0 when H1 is true. (false negative).

E.g. person has the disease but test is negative.

	Decision	
	Do not reject H0	**Reject H0**
H0 is true	Correct Decision	Type I Error
H0 is false	Type II Error	Correct Decision

The probability of a type I error is called the **α-risk**.
The probability of a type II error is called the **β-risk**.

That is:

$$\alpha = P(\text{type I error}) = P(\text{reject H0} \mid \text{H0})$$
$$\beta = P(\text{type II error}) = P(\text{fail to reject H0} \mid \text{H1}).$$

Type I error is sometimes more serious than Type II, e.g., Type I is like convicting an innocent person and Type II is letting a guilty person go free due to lack of evidence.

Here we create a test that is required to satisfy P(type I error) $\leq \alpha$. In this case, α is called the **level of significance**, and the test is a **α-level test**.

Usually we choose $\alpha = 0.05$ (or 0.10 or 0.01). This says that most of the time, we accept less than 5% probability of Type I error.

E.g., if we are checking whether: P(type I error) = P(test rejects H0 | H0) ≤ 0.10, then the test is a 0.10-level test.

The **p-value** is the probability of observing a sample statistic *as extreme or more extreme* than the one observed under the assumption that the null hypothesis is true. It is the "observed level of significance."

When testing a hypothesis, state α. Calculate the p-value and if the p-value $\leq \alpha$, then reject H0. Otherwise do not reject H0.

Note: in order to compute the α-risk, you need H0 and a decision rule (such as "reject if $X>a$"). For computing β-risk, you need H1 and a decision rule.

Example: Let's say we reject a shipment from a vendor if we find more than 1 defective item in a shipment of 100. (That's our decision rule for determining whether the probability of error is 1%). The number of defective items in a shipment obeys the binomial distribution. If the true probability of error really is 1%, then:

$$\alpha = P(\text{type I error}) = P(\text{reject H0} \mid \text{H0})$$
$$= P(\text{2 or more defective} \mid p=0.01) = 0.264 .$$

(We can compute this using the first 2 terms of the binomial distribution.)

Define $\pi = 1-\beta$ to be the **power** of the test,

$$\pi = 1-\beta = P(\text{reject H0} \mid \text{H1}).$$

The higher the power, the better the test. The power is useful for assessing whether the test has sufficiently high probability to reject H0 when H1 is true.

Misuse of Hypothesis Tests

- Data in practice are not always a random sample from a distribution.

- It is possible to have a highly statistically significant result that is practically insignificant
    - E.g. an SAT coaching program that has an average improvement of 15.5 points with p-value < 0.001, but the average retest gain is already 15 points.

- "If you torture data long enough it will confess" – testing so many hypotheses that you will likely find at least one significant result – and then you only report that one. This is commonly done yet highly incorrect! There are methods (e.g., Bonferroni) that lower the acceptable significance levels when there are multiple tests.

15.075J / ESD.07J Statistical Thinking and Data Analysis
Fall 2011