Hypothesis testing: review

Suppose that there is a hypothesis/theory that you don't want to reject unless there is evidence that it is false. If the data you collect contains evidence that the hypothesis is false, you will reject it. If not, you won't. A classic example is a court trial: it is assumed that you are not guilty unless proven otherwise. Another example is testing a new drug: being conservative, we assume that the new drug is ineffective (or even detrimental) unless proven otherwise.

Our mathematical framework will be the following: we have data (which in this course we will always assume to be independent and identically distributed) $X_1, X_2, \dots, X_n$ coming from a model with PMF/PDF $f_\theta(x)$, where θ is unknown. The hypothesis that we assume "true" unless proven otherwise is called the **null hypothesis** (H_0), which is tested against the **alternative hypothesis** (H_1). The **alternative hypothesis** is not accepted unless there is evidence that supports it.

Example: We want to know if we want to release a new drug. Assume that we measure an outcome that is zero if there is no treatment effect, positive if the treatment works well, and negative values if the treatment is harmful. The measurements are modeled with a Normal(μ, σ^2) distribution, so μ represents the expected treatment effect. We might want to test $H_0 : \mu \leq 0$ against the alternative $H_1 : \mu > 0$.

We can make two types of error: rejecting H_0 when it is true (false positive, also known as type I error) and not rejecting H_0 when H_1 is true (false negative, also known as type II error). Most textbooks include the following table to illustrate this:

	Don't Reject H_0	Reject H_0
H_0 is true	OK	Type I error (false positive)
H_1 is true	Type II error (false negative)	OK

The basic logic behind all the tests that we will see is the following:

1. Assume that H_0 is true.
2. Compute the probability that a statistic T is as "extreme" or more "extreme" (in the direction of H_1) than the observed value t_{obs} assuming that H_0 is true.
3. We reject H_0 if the probability that we computed in step 3 is low. If it isn't low, we don't reject.

One group

Let $X_1, X_2, \dots, X_n$ be independent and identically distributed as Normal(μ, σ^2). Let $H_0 : \mu = \mu_0$ be the null hypothesis. If H_0 is true, we have

$$T = \frac{\sqrt{n}(\overline{X} - \mu_0)}{S} \sim \text{Student}-t(n-1),$$

where $\overline{X}$ is the sample mean and S is the sample standard deviation. Suppose that we observe data $x = (x_1, x_2, \dots, x_n)$ and compute $t = \sqrt{n}(\overline{x} - \mu_0)/s$.

The t**-test** at a significance level α proceeds as follows:

1. If the alternative hypothesis is $H_1 : \mu > \mu_0$, the t-test rejects H_0 if the observed t-value is unusually high. That is, if $P_{H_0}(T \geq t) < \alpha$.

2. If the alternative hypothesis is $H_1 : \mu < \mu_0$, the t-test rejects H_0 if the observed t-value is unusually low. That is, if $P_{H_0}(T \leq t) < \alpha$.

3. If the alternative hypothesis is $H_1 : \mu \neq \mu_0$, the t-test rejects H_0 if the observed t-value is unusually high or low under H_0. That is, if $P_{H_0}(|T| \geq |t|) = 2P_{H_0}(T \geq |t|) < \alpha$.

If n is large enough, the assumption of normality isn't very important provided the data are independent, identically distributed, and come from a distribution with finite variance. This is a consequence of the central limit theorem.

Example: Let -1.46, 0.02, -0.07, 0.15, -1.75 be iid data coming from Normal(μ, σ^2) and suppose that the null hypothesis is $\mu = 1$, which we want to test a significance level $\alpha = 0.05$. The t-value is $t = \sqrt{n}(\overline{X} - \mu_0)/s = \sqrt{5}(-0.622 - 1)/0.90 \approx -4.02$. The degree of freedom of the t-distribution is $n - 1 = 4$.

- If $H_1 : \mu > 1$, the p-value is $P_{H_0}(T \geq t) = P(T \geq -4.02) > 0.99$, so we don't reject H_0.
- If the alternative hypothesis is $H_1 : \mu < 1$, the p-value is $P_{H_0}(T \leq -4.02) < 0.01 < \alpha$, so we reject H_0.
- If the alternative hypothesis is $H_1 : \mu \neq 1$, the p-value is $P_{H_0}(T \geq 4.02) + P_{H_0}(T \leq -4.02) < 0.02 < \alpha$, so we reject H_0 at a significance level $\alpha = 0.05$.

Example: In 2007, the online dating platform OKCupid ran some experiments. In one of them, they worked with a group of 40 women. In the sample, the women under-reported their weight on average by 8.48 lb, with a standard deviation of 8.87 lb. We want to know if there is evidence to claim that this observed difference in the sample generalizes to the population.

We can use the t-test for this task. If μ is the expected difference between the reported and the actual age, we can test $H_0 : \mu = 0$ against $H_0 : \mu \neq 0$. The t-value is $t = \sqrt{40}(8.48 - 0)/8.87 \approx 6.05$ and the degrees of freedom are 39. If you compute the p-value, you will see that it is significant.

Exercise: OKcupid also ran a study with 40 men, and they recorded the outcome "reported age - actual age' '. The sample mean difference was -0.51 years with a standard deviation of 1.61 years. Test the hypothesis that the reported age by men in online dating websites is accurate against the hypothesis that it isn't at the 0.01 significance level.