LECTURE 11: HYPOTHESIS TESTING

MECO 7312.
INSTRUCTOR: DR. KHAI CHIONG
NOVEMBER 15, 2023

1. Methods of evaluating tests

Suppose you want to test that the population mean is 2.

Test $H_0 : \mu = 2$ versus $H_1 : \ \mu \neq 2$.

Why are the following good or bad tests?

(i) $\mathbb{1}(\bar{X}_n \neq 2)$

(ii) $\mathbb{1}(\bar{X}_n \notin [1.8, 2.2])$

(iii) $\mathbb{1}(\bar{X}_n \notin [-10, 30])$

Test 1 "rejects too often" (in fact, for every n, you reject with probability 1). Test 2 seems ok, Test 3 seems to accept too often.

Since the outcome of a test itself is a random variable, even if the null hypothesis is correct, we could just reject it wrongly by chance. This is called the **Type 1 error**. Moreover, when the alternative hypothesis is correct (the null is wrong), we might also fail to reject the null hypothesis by chance. This is called the **Type 2 error**.

There are two types of mistakes that we are worried about:

- **Type-I error:** Rejecting H_0 when it is true. (This is the problem with test 1.)
- **Type-II error:** Accepting H_0 when it is false. (This is the problem with test 3.)

	Accept H_0	Reject H_0
Truth is H_0	Correct decision	Type 1 error
Truth is H_1	Type 2 error	Correct decision

Let $T(\vec{X}) \equiv T(X_1, \ldots, X_n)$ denote the test statistic, and let $\mathbb{1}(T(\vec{X}) \in R)$ be the test, where R is the rejection region. The null hypothesis is $H_0 : \theta \in S_0$, and the alternative hypothesis is $H_1 : \theta \notin S_0$. Then:

$$P(\text{type I error}|\theta) = P(T(\vec{X}) \in R \,|\theta) \quad \text{for } \theta \in S_0$$
$$P(\text{type II error}|\theta) = P(T(\vec{X}) \notin R \,|\theta) \quad \text{for } \theta \in S_0^c$$

1.1. Example

$X_1, X_2 \sim^{\text{i.i.d}}$ Bernoulli with probability p.

- Test $H_0 : \; p = \frac{3}{4}$ vs. $H_1 : \; p \neq \frac{3}{4}$.
- Consider the test $\mathbb{1}\left(\frac{X_1+X_2}{2} \neq 1\right)$ or equivalently, $\mathbb{1}\left(\frac{X_1+X_2}{2} \in \{0, \frac{1}{2}\}\right)$.
- Type I error: Rejecting H_0 when $p = \frac{3}{4}$.

  P(Type I error) $=$ P$(\frac{X_1+X_2}{2} \neq 1|p = \frac{3}{4}) =$ P$(\frac{X_1+X_2}{2} = 0|p = \frac{3}{4})+$ P$(\frac{X_1+X_2}{2} = \frac{1}{2}|p = \frac{3}{4}) = \frac{1}{16} + \frac{3}{8} = \frac{7}{16} = 0.4375$.
- Type II error: Accepting H_0 when $p \neq \frac{3}{4}$

$$\frac{X_1 + X_2}{2} = \begin{cases} 0 & \text{with prob } (1-p)^2 \\ \frac{1}{2} & \text{with prob } 2(1-p)p \\ 1 & \text{with prob } p^2 \end{cases}$$

  Therefore, P(Type II error) $=$ P$(\frac{X_1+X_2}{2} = 1|p \neq \frac{3}{4}) = p^2$, where $p \neq \frac{3}{4}$.

Type-2 error here depends on what the true value of p is. Therefore when the true value of p is near zero, type-2 error is small.

2. Power function

More generally, Type-I and Type-II errors are summarized in the **power function**.

Definition: Suppose $H_0 : \theta \in S_0$ and $H_1 : \theta \in S_0^c$. Let the test statistic be $T(X_1, \ldots, X_n)$, and the test be $\mathbb{1}(T(\vec{X}) \in R)$, where R is the rejection region. The *power function* of the test is defined by $\beta(\theta) = P(T(\vec{X}) \in R|\theta)$.

We can interpret the power function as follows. Suppose $H_0 : \theta \in S_0$, then $\max_{\theta \in S_0} \beta(\theta)$ is the maximum Type-1 error. The Type-2 error can also be obtained from the power function as $1 - \beta(\theta)$ for $\theta \notin S_0$.

The name power function comes from the fact that a test that has a high Type-2 error is said to have low power (the test accepts too often and cannot discriminate the null from the alternative).

Example: Consider the previous example where $X_1, X_2 \sim^{\text{i.i.d}}$ Bernoulli with probability p. Let $H_0 : p = \frac{3}{4}$ vs. $H_1 : p \neq \frac{3}{4}$. The test statistic is $T = \frac{X_1+X_2}{2}$, while the rejection region is $R = \{0, \frac{1}{2}\}$.

The Power Function of this test is $\beta(p) = P\left(\frac{X_1+X_2}{2} \in \{0, \frac{1}{2}\}|p\right) = 1 - p^2$. Graph the power function as a function of p.

What can we say from this power function?

- $\beta(\frac{3}{4})$ is the Type I error, and $1 - \beta(p)$ for all $p \neq \frac{3}{4}$ is the Type-II error.
- A good test should have both low Type I and Type II errors.
- In general for $H_0 : \theta \in S_0$ vs. $H_1 : \theta \notin S_0$, a good statistical test has a power function that is low for $\theta \in S_0$ and high for $\theta \notin S_0$.

Consider a different test: $\mathbb{1}\left(\frac{X_1+X_2}{2} = 0\right)$? We can use power function to compare the two tests. For this test, the power function is $\beta_2(p) = (1-p)^2$. When we plot this power function alongside the previous one, we see that $\beta(p) = 1 - p^2 > \beta_2(p) = (1-p)^2$ for $p \in (0, 1)$. As such, the power function $\beta_2(p) = (1-p)^2$ lies below the power function $\beta(p) = 1 - p^2$.

This means that the second test has a lower Type-1 error, but the Type-2 error is higher. Thus, we say that the first test is a higher powered, more discriminating test.

2.1. Example: Binomial power function

Let $X \sim$ Binomial$(5, \theta)$. Consider testing $H_0 : \theta \leq \frac{1}{2}$ versus $H_1 : \theta > \frac{1}{2}$. Consider the test $\mathbb{1}(X = 5)$, i.e. we reject H_0 if and only if all successes are observed. The power function for this test is:

$$\beta_1(\theta) = P(X = 5|\theta) = \theta^5$$

Plotting the power function for this test, the maximum Type 1 error is 0.0312, and occurs at $\theta = \frac{1}{2}$.[1] However the Type 2 error is large: at $\theta = 0.75$, the Type-2 error is $1 - \beta(0.75) = 1 - 0.24 = 0.76$. This test appears to reject too infrequently.

[1] $\frac{1}{2^5} = 0.0312$

Consider another test that rejects H_0 if and only if $X = 3, 4, 5$. The power function for this test is:

$$\beta_2(\theta) = P(X = \{3,4,5\}|\theta) = \binom{5}{3}\theta^3(1-\theta)^2 + \binom{5}{4}\theta^4(1-\theta) + \theta^5$$

Plotting this power function, the Type 2 error is now lower, but at the expense of a larger Type 1 error. This test has a much higher power than the first test, but the Type-1 error is high.

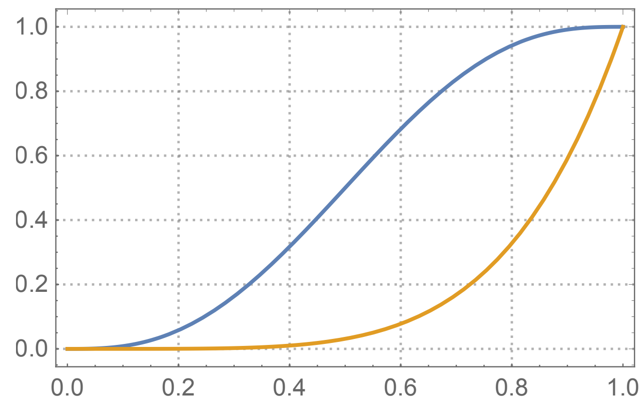

FIGURE 1. The top line is the power function $\theta^5 + 5(1-\theta)\theta^4 + 10(1-\theta)^2\theta^3$, while the bottom line is the function θ^5. The horizontal axis is θ.

By varying the rejection region, we obtain different magnitudes of Type-1 errors. Depending on the desired Type-1 or Type-2 errors, we then choose the appropriate rejection region.

2.2. Example: Uniform power function

$X_1, \ldots, X_n \sim U[0,\theta]$.

Test $H_0 : \theta \leq 2$ versus $H_1 : \theta > 2$. Derive $\beta(\theta)$ for the Likelihood Ratio Test $\mathbb{1}(\lambda(\boldsymbol{x}) < c)$. Recall previously that:

$$\lambda(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \max(x_1, \ldots, x_n) \leq 2 \\ 0 & \text{otherwise.} \end{cases}$$

Hence, for $0 < c < 1$:

$$\begin{aligned}\beta(\theta) &= P(\lambda(\boldsymbol{X}) < c|\theta) \\ &= P(\max(X_1, \ldots, X_n) > 2|\theta) \\ &= 1 - P(\max(X_1, \ldots, X_n) \leq 2|\theta)\end{aligned}$$

Since $X_i \sim U[0, \theta]$,

$$P(X_i \leq 2|\theta) = \begin{cases} \frac{2}{\theta} & \text{for } \theta \geq 2 \\ 1 & \text{for } \theta < 2 \end{cases}$$

$$\begin{aligned}\beta(\theta) &= 1 - P(\max(X_1, \ldots, X_n) \leq 2|\theta) \\ &= \begin{cases} 1 - \left(\frac{2}{\theta}\right)^n & \text{for } \theta \geq 2 \\ 0 & \text{for } \theta < 2 \end{cases}\end{aligned}$$

Graph the power function. The power function shows that this is a good test, especially when the sample size is large.

3. Level and size of a test

Researchers are often more concerned with Type-I error (i.e. not rejecting the null hypothesis unless overwhelming evidence against it). Type-2 error is a secondary concern.

This motivates the definition of *size* and *level* of a test.

- A test with power function $\beta(\theta)$ is a **size** α test if $\max_{\theta \in S_0} \beta(\theta) = \alpha$.
- A test with power function $\beta(\theta)$ is a **level** α test if $\max_{\theta \in S_0} \beta(\theta) \leq \alpha$.
- For $0 \leq \alpha \leq 1$.
- Level α tests consist of tests that have size α or less.

Reflecting perhaps the "conservative" bias, researcher often use tests of size $\alpha =$ 0.05, or 0.10.

A size α test means that you will never commit a Type-1 error greater than α. Of course this says nothing about the power of the test.

3.1. Size of Likelihood Ratio tests

For a Likelihood Ratio Test $\mathbb{1}(\lambda(\boldsymbol{x}) < c)$, the desired size can be controlled and achieved by manipulating c. That is, if we desire a $\alpha = 0.05$ Likelihood Ratio Test, then choose c such that $\max_{\theta \in S_0} P(\lambda(\boldsymbol{X}) < c|\theta) = \alpha$.

3.1.1. Example 1

Recall previously that $X_1, \ldots, X_n \sim^{\text{i.i.d}} \mathcal{N}(\theta, 1)$, and we are testing $H_0: \theta = \theta_0$ versus $H_1: \theta \neq \theta_0$. The Likelihood Ratio Test is $\mathbb{1}(\lambda(\boldsymbol{X}) < c)$, where $\lambda(\boldsymbol{X}) = \exp\left(-\frac{n}{2}(\bar{X}_n - \theta_0)^2\right)$ and $\bar{X}_n$ is the sample mean.

$$\max_{\theta \in S_0} P\left(\exp\left(-\frac{n}{2}(\bar{X}_n - \theta_0)^2\right) < c \,\middle|\, \theta\right) = 0.05$$

$$\max_{\theta \in S_0} P\left(|\bar{X}_n - \theta_0| > \sqrt{-\frac{2}{n}\log c} \,\middle|\, \theta\right) = 0.05$$

$$P\left(|\bar{X}_n - \theta_0| > \sqrt{-\frac{2}{n}\log c} \,\middle|\, \theta = \theta_0\right) = 0.05$$

Conditional on $\theta = \theta_0$, we have $\bar{X}_n \sim \mathcal{N}(\theta_0, \frac{1}{n})$, and $\sqrt{n}(\bar{X}_n - \theta_0) \sim \mathcal{N}(0, 1)$. Therefore we can then find c such that: $P\left(|\sqrt{n}(\bar{X}_n - \theta_0)| > \sqrt{-2\log c}\right) = 0.05$ according to the standard Normal distribution. It follows that $\Phi(-\sqrt{-2\log c}) = 0.05/2$, and hence, $c = \exp(-\frac{1}{2}\Phi^{-1}(0.025)^2) = 0.1465$, where Φ is the cdf of the standard Normal distribution, and Φ^{-1} is its inverse cdf.

For general α, the critical value is determined as $c = \exp(-\frac{1}{2}\Phi^{-1}(\frac{\alpha}{2})^2)$.

3.1.2. Example 2

$X_1, \ldots, X_n \sim^{\text{i.i.d}} U[0, \theta]$.

Test $H_0: \theta = 2$ vs. $H_1: \theta \neq 2$.

The Likelihood Ratio Test Statistics is:

$$\lambda(\boldsymbol{x}) = \begin{cases} 0 & \text{if } \max(x_1, \ldots, x_n) > 2 \\ \left(\frac{\max(x_1, \ldots, x_n)}{2}\right)^n & \text{otherwise} \end{cases}$$

For a Likelihood Ratio test, $\mathbb{1}(\lambda(\boldsymbol{x}) \leq c)$, the number c will determine the size of the test.

$$\begin{aligned}\alpha =& P(\lambda(\boldsymbol{X}) \leq c|\theta = 2)\\ \alpha =& P(\left(\frac{\max(X_1,\ldots,X_n)}{2}\right)^n \leq c|\theta = 2)\\ \alpha =& P(\max(X_1,\ldots,X_n) \leq 2c^{1/n}|\theta = 2)\\ \alpha =& \left(c^{1/n}\right)^n\\ c =& \alpha\end{aligned}$$

3.2. Size of t-tests

$X_1,\ldots,X_n \sim^{\text{i.i.d}} f(x|\mu)$, where $\mu \equiv \mathbb{E}[X]$ is the population mean.

$H_0:\ \mu \leq \mu_0$ versus $H_1:\ \mu > \mu_0$

Recall the t-test, where $Z_n = \frac{\sqrt{n}(\bar{X}_n - \mu_0)}{S_n}$, $\bar{X}_n$ is the sample mean, and S_n is the sample standard deviation. The (one-sided) t-test is $\mathbb{1}(Z_n > c)$ for some c.

The power function,

$$\begin{aligned}\beta_n(\mu) &= P(Z_n > c|\mu)\\ \beta_n(\mu) &= P\left(\frac{\sqrt{n}(\bar{X}_n - \mu_0)}{S_n} > c\,\middle|\,\mu\right)\\ &= P\left(\frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} - \frac{\sqrt{n}(\mu_0 - \mu)}{S_n} > c\,\middle|\,\mu\right)\end{aligned}$$

Then, consider the asymptotic power function,

$$\lim_{n\to\infty} \beta_n(\mu) = \begin{cases} 0 & \text{for } \mu < \mu_0\\ 1 - \Phi(c) & \text{for } \mu = \mu_0\end{cases}$$

To see this, when $\mu < \mu_0$, $\frac{\sqrt{n}(\mu_0-\mu)}{S_n}$ will diverge to ∞ with probability 1, hence $\frac{\sqrt{n}(\bar{X}_n-\mu)}{S_n} - \frac{\sqrt{n}(\mu_0-\mu)}{S_n}$ will diverge to $-\infty$. When $\mu = \mu_0$, we have $\frac{\sqrt{n}(\bar{X}-\mu)}{S_n} \to_d \mathcal{N}(0,1)$ as $n \to \infty$, by the Central Limit Theorem.

Therefore, asymptotically, the maximum type-1 error occurs at $\mu = \mu_0$. Then, the (asymptotic) critical value c is obtained as $\alpha = \max_{\mu\leq\mu_0} \lim_{n\to\infty} \beta_n(\mu) = 1 - \Phi(c)$.

$$\begin{aligned}\alpha &= 1 - \Phi(c)\\ c &= \Phi^{-1}(1-\alpha)\end{aligned}$$

c is the $(1-\alpha)$-th quantile of the standard normal distribution. You can get these from the usual tables.

For $\alpha = 0.025$, then $c^* = 1.96$. For $\alpha = 0.05$, then $c^* = 1.64$.

3.2.1. p-values

For the t-tests, the smaller the size of a test, the more conservative the test is (since the Type-1 error is smaller), the harder it is to reject the null. If the size of a test is zero, then we would never reject the null hypothesis. Thus, rejecting the null at size $\alpha = 0.01$ constitutes a stronger evidence against the null, compared to rejecting the null at size $\alpha = 0.05$.The p-value of a test is the smallest size such that the null would still be rejected.

The notion of p-values applies mainly to t-tests. Let α denote the size of a test. The *outcome* of a test has a p-value given by p^* if p^* is the smallest size such that the null is still rejected. That is, we would reject the null hypothesis under all corresponding tests that have size $\alpha \geq p^*$. While the size (and the critical region) determines when to reject the null, the p-values can tell us "how much" you reject the null. The smaller the p-value, the greater the evidence against the null.

Consider a size α one-sided test before, $H_0 : \mu \leq \mu_0$ vs. $H_1 : \ \mu > \mu_0$. Let Z be the one-sided t-test statistic . The null is rejected when $z \geq \Phi^{-1}(1-\alpha)$, where z is the realized test statistic. When α is smaller, $\Phi^{-1}(1-\alpha)$ becomes larger. Therefore, finding the smallest α such that the null is still rejected is equivalent to solving for α such that $z = \Phi^{-1}(1-\alpha)$. Therefore, the p-value, denoted as p^*, is defined as $p^* = 1 - \Phi(z)$, where Φ is the cdf of the standard Normal distribution. The larger the t-test statistic value z is, the smaller the p-value.

4. Size of Likelihood Ratio tests using asymptotic approximation

We can use asymptotic approximation in order to determine the approximate critical regions for many common test statistics.

For the Likelihood Ratio test statistics, it can be difficult to derive its sampling distribution. We use the following result.

Wilks' Theorem: Let $X_1, \ldots, X_n \sim^{\text{i.i.d}} f(x|\theta)$. Hypothesis test: $H_0 : \ \theta \in S_0$ vs. $H_1 : \ \theta \notin S_0$. Let $\lambda(X_1, \ldots, X_n)$ be the Likelihood Ratio Test statistics. Then under H_0, as $n \to \infty$:

$$-2 \log \lambda(X_1, \ldots, X_n) \xrightarrow{d} \chi^2_1.$$

Note: χ_1^2 denotes a random variable from the Chi-squared distribution with 1 degree of freedom. If $Z \sim N(0,1)$, then $Z^2 \sim \chi_1^2$. Clearly, χ^2 random variables only have positive support.

Wilks Theorem holds true under some assumptions. The theorem assumes i.i.d. data generating process. Moreover, the theorem will not work when the unrestricted likelihood function is maximized at a corner, and not at an interior solution. In another words, the MLE is not obtained through first order conditions, such as in the Uniform distribution examples above.

4.1. Example 1

If the data-generating process is Normal, then this asymptotic approximation holds exactly with finite n.

$X_1, \ldots, X_n \sim^{\text{i.i.d}} \mathcal{N}(\theta, 1)$

Test $H_0: \theta = \theta_0$ vs. $H_1: \theta \neq \theta_0$.

The likelihood ratio test statistic is:

$$\begin{aligned}\lambda(X_1, \ldots, X_n) &= \exp\left(-\frac{n}{2}(\bar{X} - \theta_0)^2\right) \\ -2\log\lambda(X_1, \ldots, X_n) &= n(\bar{X} - \theta_0)^2\end{aligned}$$

Under the null hypothesis that $\theta = \theta_0$, we have $\sqrt{n}(\bar{X} - \theta_0) \sim \mathcal{N}(0,1)$, it follows that $-2\log\lambda(X_1, \ldots, X_n) = n(\bar{X} - \theta_0)^2 \sim \chi_1^2$.

4.2. Example 2

$X_1, \ldots, X_n \sim$ *i.i.d.* Bernoulli with probability p. Test $H_0: p = p_0$ vs. $H_1: p \neq p_0$.

The likelihood function is $L(p|x_1, \ldots, x_n) = \prod_{i=1}^n p^{x_i}(1-p)^{1-x_i} = p^{\sum_i x_i}(1-p)^{n-\sum_i x_i}$. $Y = \sum_{i=1}^n X_i$.

$$\lambda(X_1, \ldots, X_n) = \frac{(p_0)^Y (1-p_0)^{n-Y}}{\left(\frac{Y}{n}\right)^Y \left(\frac{n-Y}{n}\right)^{n-Y}}.$$

The sampling distribution of this LR test statistic is not analytically tractable, so we appeal to its asymptotic distribution. For an asymptotic size α:

$$\begin{aligned}
\alpha &= P(\lambda(X_1, \ldots, X_n) \leq c \,|\, p = p_0) \\
&= P(-2\log\lambda(X_1, \ldots, X_n) \geq -2\log c \,|\, p = p_0) \\
&= P(\chi_1^2 \geq -2\log c) \\
&= 1 - F_{\chi_1^2}(-2\log c) \\
\Rightarrow c &= \exp\left(-\frac{1}{2}F_{\chi_1^2}^{-1}(1-\alpha)\right).
\end{aligned}$$

For instance, for $\alpha = 0.05$, then $F_{\chi_1^2}^{-1}(1-\alpha) = 3.841$ and $c^* = 0.1465$. For $\alpha = 0.10$, then $F_{\chi_1^2}^{-1}(1-\alpha) = 2.706$ and $c^* = 0.2584$. Note, these (asymptotic) critical values do not depend on p_0. Regardless of what our null/alternative hypotheses are, we always have these critical values.

Let's verify in R and Python using simulations that the following

$$\lambda(\vec{X}) = \left(\frac{p_0}{\bar{X}}\right)^{\sum X_i}\left(\frac{1-p_0}{1-\bar{X}}\right)^{n-\sum X_i}$$

has the asymptotic distribution $-2\log\lambda(\vec{X}) \to \chi_1^2$ under the null hypothesis.

5. Uniformly Most Powerful test

(*Optional reading)

The Likelihood Ratio Test is one of the most commonly used test because under some conditions, it is optimal.

Let $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \notin \Theta_0$.

Consider all level α tests, that is, the Type-1 error (with respect to the null hypothesis above) is at most α.

Let $\beta(\theta)$ be the power function of a level-α test that is called the Uniformly Most Powerful test, then $\beta(\theta) \geq \beta'(\theta)$ for all $\theta \notin \Theta_0$, where $\beta'(\theta)$ are any other power functions that are level-α.

5.1. Neyman-Pearson Lemma

For simple hypothesis tests, the Neyman-Pearson Lemma says that the Likelihood Ratio Test is the Uniformly Most Powerful test.

$H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$.

$X_1, \ldots, X_n \sim f(x_1, \ldots, x_n|\theta)$ (need not be i.i.d)

Consider the test statistics

$$\lambda(x_1, \ldots, x_n) = \frac{f(x_1, \ldots, x_n|\theta_0)}{f(x_1, \ldots, x_n|\theta_1)} \tag{1}$$

We reject the null if $\lambda(x_1, \ldots, x_n) < c$, for some $c > 0$. Here c can be greater than 1. Suppose c is such that $P(\lambda(x_1, \ldots, x_n) < c|\theta = \theta_0) = \alpha$.

Neyman-Pearson Lemma says that any test that satisfies the above is a Uniformly Most Powerful level α test. Conversely, every Uniformly Most Powerful satisfy the above, except for some pathological cases.