# LECTURE 10: HYPOTHESIS TESTING

MECO 7312.
INSTRUCTOR: DR. KHAI CHIONG
NOVEMBER 12, 2020

**Hypothesis:** statement about an unknown population parameter

Examples: 1.) The average age of viewers for a TV program is 27 (statement about population mean).
2.) A company's promotional policy has zero effect on sales (statement about the population regression coefficient).
3.) Portfolio A is less volatile than Portfolio B (statement about variances of stocks and portfolios).

In hypothesis testing, we are interested in testing between two mutually exclusive hypotheses, called the **null hypothesis** (denoted $H_0$) and the **alternative hypothesis** (denoted $H_1$).

$H_0$ and $H_1$ are complementary hypotheses, in the following sense:

If the parameter space is $S$, then the null and alternative hypotheses form a partition of $S$. That is,

$H_0$: $\theta \in S_0 \subset S$
$H_1$: $\theta \in S_0^c$ (the complement of $S_0$ in $S$).

**Examples:**

(i) $H_0: \theta = 0$ vs. $H_1: \theta \neq 0$, where the parameter space is $\mathbb{R}$.

(ii) $H_0: \theta \leq 0$ vs. $H_1: \theta > 0$, where the parameter space is $\mathbb{R}$.

(iii) $H_0: \theta = 1$ vs. $H_1: \theta = -1$, where the parameter space is $\{-1, 1\}$.

(iv) $H_0: \theta \in [-1, 1]$ vs. $H_1: \theta \notin [-1, 1]$, where the parameter space is $\mathbb{R}$.

## 1. Test statistics

There are two main ingredients in a hypothesis test. One is a test statistic, the other is a decision rule.

A **test statistic**, similar to an estimator, is just some real-valued function $T_n \equiv T(X_1, \ldots, X_n)$ of your data sample $X_1, \ldots, X_n$. Clearly, a test statistic is a random variable.

A **decision rule** is a function mapping values of the test statistic into $\{0, 1\}$, where

- "0" implies that you accept the null hypothesis $H_0 \Leftrightarrow$ reject the alternative hypothesis $H_1$.
- "1" implies that you reject the null hypothesis $H_0 \Leftrightarrow$ accept the alternative hypothesis $H_1$.

*Example:*

Let $\mu$ denote the (unknown) average age of the viewers for a TV show.

You want to test: $H_0: \ \mu = 27$ vs. $H_1: \ \mu \neq 27$

Let your test statistic be $\bar{X}_{100} = \frac{1}{100}\sum_{i=1}^{100} X_i$, the average age of 100 randomly-drawn viewers of the TV show.

There are many different possible decision rules. Consider the following decision rules:

(i) $\mathbb{1}(\bar{X}_{100} \neq 27)$

(ii) $\mathbb{1}(\bar{X}_{100} \notin [24, 30])$

(iii) $\mathbb{1}(\bar{X}_{100} \notin [26, 28])$

Also, there are many possible test statistics, such as: (i) $med_{100}$ (sample median); (ii) $\max(X_1, \ldots, X_{100})$ (sample maximum); (iii) $mode_{100}$ (sample mode)

Which ones make the most sense?

If our proposed test is $\mathbb{1}(\bar{X}_{100} \notin [24, 30])$, then the **rejection region** or the **critical region** is $\bar{X}_{100} \notin [24, 30]$, i.e. the region of the real line where we will reject the null hypothesis.

Next we consider some common types of hypothesis tests.

## 2. Likelihood Ratio Test

Let: $X_1, \ldots, X_n \sim \ i.i.d \, f(x|\theta)$, and likelihood function $L(\theta|\boldsymbol{x}) = \prod_{i=1}^{n} f(x_i|\theta)$.

Define: the **likelihood ratio test statistic** for testing $H_0: \theta \in S_0$ vs. $H_1: \theta \in S_0^c$ as

$$\lambda(\boldsymbol{x}) \equiv \frac{\max_{\theta \in S_0} L(\theta|\boldsymbol{x})}{\max_{\theta \in S} L(\theta|\boldsymbol{x})}$$

Where the parameter space is $S$ and $S_0^c \equiv S \setminus S_0$. $\boldsymbol{x} = (x_1, \ldots, x_n)$ is the realized sample. The numerator of $\lambda(\boldsymbol{x})$ is the "restricted" likelihood function, and the denominator is the unrestricted likelihood function.

The support of the LR test statistic is $[0, 1]$.

Intuitively speaking, if $H_0$ is true (i.e., $S \in S_0$), then $\lambda(\boldsymbol{x})$ will be close to 1 (since the restriction of $S \in S_0$ will not bind). However, if $H_0$ is false, then $\lambda(\boldsymbol{x})$ can be small (close to zero).

So an LR test should be one which rejects $H_0$ when $\lambda(\boldsymbol{x})$ is small enough.

A **Likelihood Ratio Test (LRT)** is a test where we reject the null hypothesis if $\lambda(\boldsymbol{x}) \leq c$, where $c$ is any number satisfying $0 \leq c \leq 1$. In another words, a Likelihood Ratio Test consists of the test statistic $\lambda(\boldsymbol{x})$, as well as the decision rule that we reject the null hypothesis whenever $\lambda(\boldsymbol{x}) \leq c$.

### 2.1. Example: Normal LRT

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

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

Here, $S_0 = \{2\}$ and $S = \mathbb{R}$.

$$\begin{aligned}\lambda(\boldsymbol{x}) &= \frac{\max_{\theta \in S_0} L(\theta|\boldsymbol{x})}{\max_{\theta \in S} L(\theta|\boldsymbol{x})} \\ &= \frac{L(2|\boldsymbol{x})}{L(\hat{\theta}_{MLE}|\boldsymbol{x})}\end{aligned}$$

Maximizing the unrestricted likelihood is exactly the Maximum Likelihood Estimator (MLE). Therefore $\hat{\theta}_{MLE} = \bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the MLE for $\theta$.

$$\begin{aligned}\lambda(\boldsymbol{x}) &= \frac{L(2|\boldsymbol{x})}{L(\hat{\theta}_{MLE}|\boldsymbol{x})} \\ &= \frac{(2\pi)^{-n/2}\exp\left(-\frac{1}{2}\sum_i (x_i-2)^2\right)}{(2\pi)^{-n/2}\exp\left(-\frac{1}{2}\sum_i (x_i-\bar{x}_n)^2\right)} \\ &= \exp\left(-\frac{1}{2}\sum_{i=1}^n (x_i-2)^2 + \frac{1}{2}\sum_{i=1}^n (x_i-\bar{x}_n)^2\right) \\ &= \exp\left(-\frac{n}{2}(\bar{x}_n-2)^2\right)\end{aligned}$$

More generally, test $H_0: \ \theta = \theta_0$ vs. $H_1: \ \theta \neq \theta_0$. The likelihood ratio test statistic is:

$$\lambda(\boldsymbol{x}) = \exp\left(-\frac{n}{2}(\bar{x}_n-\theta_0)^2\right)$$

For this to be a test, we need to specify the decision rule: $\mathbb{1}(\lambda(\boldsymbol{x}) \leq c)$, which we will do so later.

### 2.2. Example: Uniform LRT

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

#### 2.2.1. Null hypothesis is a point, $S_0$ is a singleton

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

Here, $S_0 = \{2\}$ and $S = (0, \infty)$.

The likelihood function $L(\theta|\boldsymbol{x})$ is:

$$L(\theta|\boldsymbol{x}) = \begin{cases} \left(\frac{1}{\theta}\right)^n & \text{if } \max(x_1, \ldots, x_n) \leq \theta \\ 0 & \text{if } \max(x_1, \ldots, x_n) > \theta \end{cases}$$

The denominator of the LRT statistic is the unrestricted likelihood, $\max_{\theta \in S} L(\theta|\boldsymbol{x})$, which is maximized at $\hat{\theta}_{MLE} = \max(x_1, \ldots, x_n)$. Hence the denominator of the LR statistic is $L(\hat{\theta}_{MLE}|\boldsymbol{x}) = \left(\frac{1}{\max(x_1,\ldots,x_n)}\right)^n$.

The numerator of the LRT statistic is the restricted likelihood, $\max_{\theta \in S_0} L(\theta|\boldsymbol{x})$:

$$L(2|\boldsymbol{x}) = \begin{cases} \left(\frac{1}{2}\right)^n & \text{if } \max(x_1, \ldots, x_n) \leq 2 \\ 0 & \text{if } \max(x_1, \ldots, x_n) > 2. \end{cases}$$

Putting them together,

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

To complete the LR test: we have to specify the decision rule, which is to reject the null if $\lambda(\boldsymbol{x})$ is small enough, say $\mathbb{1}(\lambda(\boldsymbol{x}) \leq c)$. We see that the dritical region depends on the data only through $\max(x_1, \ldots, x_n)$.

Plot the graph depicting the rejection region (it will consist of two disconnected parts). We will reject the null if either $\max(x_1, \ldots, x_n) > 2$, or $\max(x_1, \ldots, x_n) \leq 2c^{1/n}$.

### 2.2.2. Null hypothesis is an interval, $S_0$ is an interval

Test $H_0: \ \theta \in [0, 2]$ vs. $H_1: \ \theta > 2$.

Here, $S_0 = (0, 2]$ and $S = (0, \infty)$.

The unrestricted likelihood is the same as before. But the restricted likelihood is

$$\max_{\theta \in (0,2]} L(\theta|\boldsymbol{x}) = \begin{cases} \left(\frac{1}{\max(x_1, \ldots, x_n)}\right)^n & \text{if } \max(x_1, \ldots, x_n) \leq 2 \\ 0 & \text{otherwise.} \end{cases}$$

so

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

The LR test is $\mathbb{1}(\lambda(\boldsymbol{x}) \leq c)$. Therefore for $0 < c < 1$, we reject the null if $\max(x_1, \ldots, x_n) > 2$. If $c = 1$, then we will always reject the null, regardless of what data we observe. If $c = 0$, then we will never reject the null. Later, we will talk about how to set $c$, but in this example, the only sensible choice is $c \in (0, 1)$, but all $c \in (0, 1)$ leads to the same decision rule.

Therefore the test of $H_0 : \theta \in [0, \theta_0]$ vs. $H_1 : \theta > \theta_0$ has a very simple form, which is to reject the null hypothesis whenever $\max(x_1, \ldots, x_n) > \theta_0$.

### 2.3. Exponential LRT

Let $X_1, \ldots, X_n$ be a random sample from an exponential population with pdf:

$$f(x|\theta) = \begin{cases} e^{-(x-\theta)} & x \geq \theta \\ 0 & x < \theta \end{cases}$$

where $-\infty < \theta < \infty$

The likelihood function is:

$$L(\theta|\boldsymbol{x}) = \begin{cases} e^{n\theta - \sum x_i} & \min(x_1, \ldots, x_n) \geq \theta \\ 0 & \text{otherwise} \end{cases}$$

Consider testing $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$.

The unrestricted maximum of $L(\theta|\boldsymbol{x})$ is achieved at $\theta = \min(x_1, \ldots, x_n)$. Therefore $\max_{\theta \in (-\infty, \infty)} L(\theta|\boldsymbol{x}) = e^{n \min(x_1, \ldots, x_n) - \sum x_i}$

Maximizing $L(\theta|\boldsymbol{x})$ with respect to the parameter space $\theta \in (-\infty, \theta_0]$,

$$\max_{\theta \in (-\infty, \theta_0]} L(\theta|\boldsymbol{x}) = \begin{cases} e^{n\theta_0 - \sum x_i} & \min(x_1, \ldots, x_n) \geq \theta_0 \\ e^{n \min(x_1, \ldots, x_n) - \sum x_i} & \min(x_1, \ldots, x_n) < \theta_0 \end{cases}$$

Therefore,

$$\lambda(\boldsymbol{x}) = \begin{cases} e^{n(\theta_0 - \min(x_1, \ldots, x_n))} & \min(x_1, \ldots, x_n) > \theta_0 \\ 1 & \min(x_1, \ldots, x_n) \leq \theta \end{cases}$$

Try plotting the LR test statistic $\lambda(\boldsymbol{x})$ as a function of $\min(x_1, \ldots, x_n)$. We reject the null hypothesis when $e^{n(\theta_0 - \min(x_1, \ldots, x_n))} \leq c$, that is, when $\min(x_1, \ldots, x_n) \geq$

$\theta_0 - \log c/n$, i.e. when $\min(x_1, \ldots, x_n)$ is sufficiently larger than $\theta_0$. Note that $\log c$ is a negative number because $0 \leq c \leq 1$.

## 3. Wald Tests ($t$-test)

Another common way to generate test statistics is to focus on statistics which are either normally distributed or asymptotically normal distributed, under $H_0$. For example, regression coefficients, all Maximum Likelihood estimators, sample mean, sample variances, etc.

Suppose that the population parameter of interest is $\theta$, and that we have an estimator $\hat{\theta}_n$ for $\theta$ that is consistent and asymptotically Normal, with some asymptotic variance $V$. That is, $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, V)$ as $n \to \infty$. We want to test $H_0: \theta = \theta_0$. Then, if the null were true:

$$Z(\boldsymbol{x}) \equiv \frac{\sqrt{n}(\hat{\theta}_n - \theta_0)}{\sqrt{V}} \xrightarrow{d} \mathcal{N}(0, 1). \tag{2}$$

The quantity $\frac{\sqrt{n}(\hat{\theta}_n - \theta_0)}{\sqrt{V}}$ is called the **$t$-test statistic**, which is approximately Normal when $n$ is large.

To fix idea, take $\theta \equiv \mathbb{E}[X]$ to be the (unknown) population mean, and the estimator for $\theta$ is the sample mean $\hat{\theta}_n = \frac{1}{n} \sum_{i=1} X_i$. Then, the Central-Limit Theorem implies that $\sqrt{n}(\bar{X} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$, and the $t$-test statistic becomes $\frac{\sqrt{n}(\bar{X} - \theta_0)}{\sigma}$ or $\frac{\bar{X} - \theta_0}{\sigma/\sqrt{n}}$.

Know that the $t$-test statistic here can be applied more generally to any asymptotically Normal estimator $\hat{\theta}_n$ of $\theta$ such that $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, V)$ as $n \to \infty$.

In most cases, the asymptotic variance $V$ will not be known, and will also need to be estimated. However, if we have an estimator $\hat{V}_n$ such that $\hat{V}_n \xrightarrow{p} V$, then the statement

$$\frac{\sqrt{n}(\hat{\theta}_n - \theta_0)}{\sqrt{\hat{V}}} \xrightarrow{d} \mathcal{N}(0, 1)$$

still holds (using the continuous mapping theorem and the Slutsky theorem). For hypothesis tests involving the population mean, the $t$-test statistic becomes $\frac{\bar{X} - \theta_0}{\sqrt{S^2/n}}$, where $S^2$ is the sample variance.

To see how the $t$ statistic can be used for hypothesis testing, we consider two cases:

(i) Two-sided (two-tailed) test: $H_0: \ \theta = \theta_0$ vs. $H_1: \ \theta \neq \theta_0$.

Under $H_0$: the $t$-test statistic is approximately (asymptotically) $\mathcal{N}(0,1)$
Under $H_1$: assume that the true value is some $\theta_1 \neq \theta_0$. Then the $t$-statistic can be written as:

$$t = \frac{\sqrt{n}(\hat{\theta}_n - \theta_0)}{\sigma} = \frac{\sqrt{n}(\hat{\theta}_n - \theta_1)}{\sigma} + \frac{\sqrt{n}(\theta_1 - \theta_0)}{\sigma}.$$

The first term $\xrightarrow{d} \mathcal{N}(0,1)$, but the second (non-stochastic) term diverges to $\infty$ or $-\infty$, depending on whether the true $\theta_1$ exceeds or is less than $\theta_0$. Hence the $t$-statistic diverges to $-\infty$ or $\infty$ with probability 1.

Hence, in this case, your test should be $\mathbb{1}(|t| > c)$, where $c$ should be some number in the tails of the $\mathcal{N}(0,1)$ distribution. Later, we will discuss how to choose $c$.

(ii) One-sided test: $H_0: \ \theta \leq \theta_0$ vs. $H_1: \ \theta > \theta_0$.

Here the null hypothesis is $\theta \in (-\infty, \theta_0]$

Just as for the two-sided test, let's consider the test statistic $t \equiv \frac{\sqrt{n}(\hat{\theta}_n - \theta_0)}{\sigma}$.

Suppose $H_0$ is true and $\theta < \theta_0$, then $t$ diverges to $-\infty$ with probability 1 as $n \to \infty$.

Suppose $H_0$ is true and $\theta = \theta_0$, then $t$ is approximately $\mathcal{N}(0,1)$.

Suppose $H_1$ is true, $t$ diverges to $\infty$ with probability 1 as $n \to \infty$.

Hence, we should reject the null when the test statistic is reasonably large. That is, your test should be $\mathbb{1}(t > c)$, for some $c$.

Both LR and Wald tests are applicable in some scenarios, for example, in the example at the beginning where we test the population mean of a Normal distribution, we could either use the LRT or the Wald test. Another family of test statistics is the Lagrange Multiplier test or the Score test.

LRT requires both the restricted and unrestricted models to be estimated, which is more complicated than the Wald test, especially for a null hypothesis like $H_0: \theta \leq \theta_0$. It also requires that we correctly specify the likelihood function. Wald test seems simpler but it does require the estimator to be asymptotically Normal, and having a consistent estimate of the asymptotic variance.

### 3.1. Wald test for MLE

(*Optional reading)

Suppose that $\hat{\theta}_{MLE}$ is a MLE of $\theta$ given the data $x_1, \dots, x_n$ generated from $f(x_1, \dots, x_n|\theta)$.

$$\sqrt{n}(\hat{\theta}_{MLE} - \theta) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}(\theta)^{-1}) \quad \text{as } n \to \infty$$

Where $\mathcal{I}(\theta) = \mathbb{E}\left[\left(\frac{\partial \log f(X_1,\dots,X_n|\theta)}{\partial\theta}\right)^2\right]$ is the Fisher's information number. For $x_1, \dots, x_n$ generated i.i.d from $f(x|\theta)$, we can further simplify: $\mathcal{I}(\theta) = n\,\mathbb{E}\left[\left(\frac{\partial \log f(X|\theta)}{\partial\theta}\right)^2\right]$. A consistent estimator of $\mathcal{I}(\theta) = n\,\mathbb{E}\left[\left(\frac{\partial \log f(X|\theta)}{\partial\theta}\right)^2\right]$ is $\hat{I}(\theta) = \sum_{i=1}^n \left(\frac{\partial \log f(x_i|\theta)}{\partial\theta}\right)^2$. The $t$-test statistic becomes $\sqrt{n}(\hat{\theta}_{MLE} - \theta)\hat{I}(\theta)$.

In general, MLE is obtained numerically (for example, in Probit regressions), so we do not know the (finite-sample) sampling distribution of the estimator. We rely on asymptotic approximations and take $\frac{1}{n}\mathcal{I}(\theta)^{-1}$ to be the variance of the estimator. Of course $\frac{1}{n}\mathcal{I}(\theta)^{-1}$ has no closed-form either, so we estimate $\mathbb{E}\left[\left(\frac{\partial \log f(X|\theta)}{\partial\theta}\right)^2\right]$ using its sample moment: $\frac{1}{n}\sum_{i=1}^n \left(\frac{\partial \log f(x_i|\theta)}{\partial\theta}\right)^2$, or we estimate $\mathbb{E}\left[\left(\frac{\partial \log f(X_1,\dots,X_n|\theta)}{\partial\theta}\right)^2\right]$ using a single draw: $\left(\frac{\partial \log f(x_1,\dots,x_n|\theta)}{\partial\theta}\right)^2$

*Multivariate version:* The Wald test can be used to test a hypothesis on multiple parameters. Let $\vec{\theta}$ be a $k$-dimensional estimator that is asymptotically Normal:

$$\sqrt{n}(\vec{\theta}_n - \vec{\theta}) \xrightarrow{d} \mathcal{N}(\vec{0}, \Sigma).$$

The MLE with multiple parameters satisfies this. Under $H_0 : \vec{\theta} = \vec{\theta}_0$, then we have

$$\sqrt{n}(\vec{\theta}_n - \vec{\theta}_0) \xrightarrow{d} \mathcal{N}(0, \Sigma)$$

The multivariate version of the $t$-test statistic is the following quadratic form:

$$Z_n \equiv n \cdot (\vec{\theta}_n - \vec{\theta}_0)^T \Sigma^{-1} (\vec{\theta}_n - \vec{\theta}_0)$$

Under the null hypothesis, $Z_n \xrightarrow{d} \chi^2_k$. Intuitively, sum of squares of $k$ Normally distributed variables have a $\chi^2_k$ distribution. Since $\chi^2$ takes only positive values, the rejection region of the test would take the form: $\mathbb{1}(Z_n > c)$.