This version: November 2, 2016

Notation: $\mathbf{X}$ denotes a random variable or random vector. $\mathbf{x}$ is its realization.

1 Hypothesis Testing

A *hypothesis* is a statement about the parameter space Θ. The *null hypothesis* Θ_0 is a subset of Θ of interest, typically suggested by some scientific theory. The *alternative hypothesis* $\Theta_1 = \Theta \backslash \Theta_0$ is the complement of Θ_0. *Hypothesis testing* is a decision whether to accept the null hypothesis or to reject it according to the observed evidence.

A *test function* is a mapping

$$\phi : \mathcal{X}^n \mapsto \{0, 1\},$$

where $\mathcal{X}$ is the sample space. We accept the null hypothesis if $\phi(\mathbf{x}) = 0$, or reject it if $\phi(\mathbf{x}) = 1$. The *acceptance region* is defined as $A_\phi = \{\mathbf{x} \in \mathcal{X}^n : \phi(\mathbf{x}) = 0\}$, and the *rejection region* is $R_\phi = \{\mathbf{x} \in \mathcal{X}^n : \phi(\mathbf{x}) = 1\}$. The *power function* of the test ϕ is

$$\beta_\phi(\theta) = P_\theta(\phi(\mathbf{X}) = 1) = E_\theta(\phi(\mathbf{X})).$$

The power function measures, at a given point, the probability that the test function rejects the null. The *size* of the test ϕ is

$$\alpha = \sup_{\theta \in \Theta_0} \beta_\phi(\theta).$$

The *level* of the test ϕ is a value $\alpha \in (0, 1)$ such that $\alpha \geq \sup_{\theta \in \Theta_0} \beta_\phi(\theta)$, which is often used when it is difficult to attain the exact supremum. The *probability of committing Type I error* is $\beta_\phi(\theta)$ for some $\theta \in \Theta_0$. The *probability of committing Type II error* is $1 - \beta_\phi(\theta)$ for $\theta \in \Theta_1$.

There has been a philosophical debate for decades about the hypothesis testing framework. At present the prevailing framework in statistics education is the frequentist perspective. A frequentist views the parameter as a

Table 1: Decisions and Status

	accept H_0 (reject H_1)	reject H_0 (accept H_1)
H_0 true (H_1 false)	correct decision	Type I error
H_0 false (H_1 true)	Type II error	correct decision

size $= P(\text{reject } H_0 | H_0 \text{ true})$
power $= P(\text{reject } H_0 | H_0 \text{ false})$

fixed constant, and they are conservative about the Type I error. Only if overwhelming evidence is demonstrated should a researcher reject the null.

Under the philosophy of protecting the null hypothesis, a desirable test should have a small level. Conventionally we take $\alpha = 0.01$, 0.05 or 0.1. There can be many tests of the correct size.

Example 1. A trivial test function, $\phi(\mathbf{X}) = 1\{0 \leq U \leq \alpha\}$, where U is a random variable from a uniform distribution on $[0,1]$, has correct size but no power. Another trivial test function $\phi(\mathbf{X}) = 1$ has the biggest power but incorrect size.

Usually, we design a test by proposing a test statistic $T_n : \mathcal{X}^n \mapsto \mathbb{R}^+$ and a critical value c_α, and then define

$$\phi(\mathbf{X}) = 1\{T_n(\mathbf{X}) > c_\alpha\}.$$

To ensure such a $\phi(\mathbf{x})$ has correct size, we figure out the distribution of T_n under the null hypothesis (called the *null distribution*), and choose c_α according to the null distribution and the desirable size or level α.

Example 2. The concept of *level* is useful if we do not have information to derive the exact size of a test. If $(X_{1i}, X_{2i})_{i=1}^n$ are randomly drawn from some unknown joint distribution, but we only know that the marginal distribution is $X_{ji} \sim N(\theta_j, 1)$, for $j = 1, 2$. In order to test the joint hypothesis $\theta_1 = \theta_2 = 0$, we can construct a test function

$$\phi(\mathbf{X}_1, \mathbf{X}_2) = 1\left\{\left\{\left|\overline{X}_1\right| \geq c_{1-\alpha/4}\right\} \cup \left\{\left|\overline{X}_2\right| \geq c_{1-\alpha/4}\right\}\right\},$$

where $c_{1-\alpha/4}$ is the $(1-\alpha/4)$-th quantile of the standard normal distribution. The level of this test is

$$\begin{aligned} P_{\theta_1=\theta_2=0}\left(\phi\left(\mathbf{X}_1,\mathbf{X}_2\right)\right) &\leq P_{\theta_1=0}\left(\left|\overline{X}_1\right|\geq c_{1-\alpha/4}\right)+P_{\theta_2=0}\left(\left|\overline{X}_2\right|\geq c_{1-\alpha/4}\right) \\ &= \alpha/2+\alpha/2=\alpha. \end{aligned}$$

where the inequality follows by the Bonferroni inequality $P\left(A\cup B\right)\leq P\left(A\right)+P\left(B\right)$. Therefore, the level of $\phi\left(\mathbf{X}_1,\mathbf{X}_2\right)$ is α, but the exact size is unknown without the knowledge of the joint distribution. (Even if we know the correlation of X_{1i} and X_{2i}, putting two marginally normal distributions together does not make a jointly normal vector in general.)

There can be many tests of a correct level. Denote the class of test functions of level smaller than α as $\Psi_\alpha=\left\{\phi:\sup_{\theta\in\Theta_0}\beta_\phi\left(\theta\right)\leq\alpha\right\}$. A *uniformly most powerful test* $\phi^*\in\Psi_\alpha$ is a test function such that, for every $\phi\in\Psi_\alpha$,

$$\beta_{\phi^*}\left(\theta\right)\geq\beta_\phi\left(\theta\right)$$

uniformly over $\theta\in\Theta_1$.

Example 3. Suppose a random sample of size 6 is generated from

$$\left(X_1,\ldots,X_6\right)\sim\text{i.i.d.}N\left(\theta,1\right),$$

where θ is unknown. We want to infer the population mean of the normal distribution. The null hypothesis is H_0: $\theta\leq0$ and the alternative is H_1: $\theta>0$. All tests in

$$\Psi=\left\{1\left\{\bar{X}\geq c/\sqrt{6}\right\}:c\geq1.64\right\}$$

has the correct level. Since $\bar{X}=N\left(\theta,1/\sqrt{6}\right)$, the power function for those in Ψ is

$$\beta_\phi\left(\theta\right)=P\left(\bar{X}\geq\frac{c}{\sqrt{6}}\right)=P\left(\sqrt{6}\left(\bar{X}-\theta\right)\geq c-\sqrt{6}\theta\right)=1-\Phi\left(c-\sqrt{6}\theta\right).$$

The test function
$$\phi(\mathbf{X}) = 1\left\{\bar{X} \geq 1.64/\sqrt{6}\right\}$$
is the most powerful test in Ψ.

Another commonly used indicator in hypothesis testing is p-value:
$$\sup_{\theta\in\Theta_0} P_\theta\left(T_n(\mathbf{x}) \leq T_n(\mathbf{X})\right).$$
In the above expression, $T_n(\mathbf{x})$ is the realized value of the test statistic T_n, while $T_n(\mathbf{X})$ is the random variable generated by $\mathbf{X}$ under the null $\theta \in \Theta_0$. p-value is closely related to the corresponding test. When p-value is smaller than the specified test size α, the test rejects the null hypothesis.

p-value is a measure whether the data is consistent with the null hypothesis, or whether the evidence from the data is compatible with the null hypothesis. p-value is *not* the probability that the null hypothesis is true. Under the frequentist perspective, the null hypothesis is either true or false, with certainty. The randomness of a test comes only from sampling, not from the hypothesis.

2 Confidence Interval

An *interval estimate* is a function $C : \mathcal{X}^n \mapsto \{\Theta' : \Theta' \subseteq \Theta\}$ that maps a point in the sample space to a subset of the parameter space. The *coverage probability* of an *interval estimator* $C(\mathbf{X})$ is defined as $P_\theta(\theta \in C(\mathbf{X}))$. The coverage probability is the frequency that the interval estimator captures the true parameter that generates the sample (From the frequentist perspective, the parameter is fixed while the region is random). It is *not* the probability that θ is inside the given region (From the Bayesian perspective, the parameter is random while the region is fixed conditional on $\mathbf{X}$.)

Exercise 1. Suppose a random sample of size 6 is generated from
$$(X_1, \ldots, X_6) \sim \text{i.i.d. } N(\theta, 1).$$

Find the coverage probability of the random interval

$$\left[\bar{X} - 1.96/\sqrt{6}, \bar{X} + 1.96/\sqrt{6}\right].$$

Hypothesis testing and confidence interval are closely related. Sometimes it is difficult to directly construct the confidence interval, but easier to test a hypothesis. One way to construct confidence interval is by *inverting a corresponding test.* Suppose ϕ is a test of size α. If $C(\mathbf{X})$ is constructed as

$$C(\mathbf{x}) = \{\theta \in \Theta : \phi_\theta(\mathbf{x}) = 0\},$$

then its coverage probability

$$P_\theta(\theta \in C(\mathbf{X})) = 1 - P_\theta(\phi_\theta(\mathbf{X}) = 1) = 1 - \alpha.$$

3 Application in OLS

3.1 Wald Test

Suppose the OLS estimator $\widehat{\beta}$ is asymptotic normal, i.e.

$$\sqrt{n}\left(\widehat{\beta} - \beta\right) \xrightarrow{d} N(0, \Omega)$$

where Ω is a $K \times K$ positive definite covariance matrix and R is a $q \times K$ constant matrix, then $R\sqrt{n}\left(\widehat{\beta} - \beta\right) \xrightarrow{d} N(0, R\Omega R')$. Moreover, if $\operatorname{rank}(R) = q$, then

$$n\left(\widehat{\beta} - \beta\right)' R' \left(R\Omega R'\right)^{-1} R\left(\widehat{\beta} - \beta\right) \xrightarrow{d} \chi^2_q.$$

Now we intend to test the null hypothesis $R\beta = r$. Under the null, the Wald statistic

$$W_n = n\left(R\widehat{\beta} - r\right)' \left(R\widehat{\Omega}R'\right)^{-1} \left(R\widehat{\beta} - r\right) \xrightarrow{d} \chi^2_q$$

where $\widehat{\Omega}$ is a consistent estimator of Ω.

Example 4. In a linear regression

$$\begin{aligned} y &= x_i'\beta + e_i = \sum_{k=1}^{5} \beta_k x_{ik} + e_i. \\ E[e_i x_i] &= \mathbf{0}_5, \end{aligned} \tag{1}$$

where y is wage and

$$x = \left(\text{edu}, \text{age}, \text{experience}, \text{experience}^2, 1\right)'.$$

To test whether *education* affects *wage*, we specify the null hypothesis $\beta_1 = 0$. Let $R = (1, 0, 0, 0, 0)$.

$$\sqrt{n}\widehat{\beta}_1 = \sqrt{n}\left(\widehat{\beta}_1 - \beta_1\right) = \sqrt{n}R\left(\widehat{\beta} - \beta\right) \xrightarrow{d} N\left(0, R\Omega R'\right) \sim N\left(0, \Omega_{11}\right), \tag{2}$$

where Ω_{11} is the $(1,1)$ (scalar) element of Ω. Therefore,

$$\sqrt{n}\frac{\widehat{\beta}_1}{\widehat{\Omega}_{11}^{1/2}} = \sqrt{\frac{\Omega_{11}}{\widehat{\Omega}_{11}}}\sqrt{n}\frac{\widehat{\beta}_1}{\Omega_{11}^{1/2}}$$

If $\widehat{\Omega} \xrightarrow{p} \Omega$, then $\left(\Omega_{11}/\widehat{\Omega}_{11}\right)^{1/2} \xrightarrow{p} 1$ by the continuous mapping theorem. As $\sqrt{n}\widehat{\beta}_1/\Omega_{11}^{1/2} \xrightarrow{d} N(0,1)$, we conclude $\sqrt{n}\widehat{\beta}_1/\widehat{\Omega}_{11}^{1/2} \xrightarrow{d} N(0,1)$.

Example 4 is a test about a single coefficient, and the test statistic is essentially a t-statistic. Example 5 gives a test about a joint hypothesis.

Example 5. We want to simultaneously test $\beta_1 = 1$ and $\beta_3 + \beta_4 = 2$ in (1). The null hypothesis can be expressed in the general form $R\beta = r$, where the restriction matrix R is

$$R = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \end{pmatrix}$$

and $r = (1, 2)'$.

Example 4 and 5 are linear restrictions. In order to test a nonlinear regression, we need the so-called *delta method.*

Theorem 1 (delta method)**.** *If* $\sqrt{n}\left(\widehat{\theta}-\theta_0\right) \xrightarrow{d} N\left(0, \Omega_{K\times K}\right)$*, and* $f : \mathbb{R}^K \mapsto \mathbb{R}^q$ *is a continuously differentiable function for some* $q \leq K$*, then*

$$\sqrt{n}\left(f\left(\widehat{\theta}\right)-f\left(\theta_0\right)\right) \xrightarrow{d} N\left(0, \frac{\partial f}{\partial \theta}\left(\theta_0\right)\Omega\frac{\partial f}{\partial \theta}\left(\theta_0\right)'\right).$$

Example 6. In the regression (1), the optimal experience level can be found by setting the first order condition with respective to experience to set, $\beta_3 + 2\beta_4\text{experience}^* = 0$. We test the hypothesis that the optimal experience level is 20 years; in other words,

$$\text{experience}^* = -\frac{\beta_3}{2\beta_4} = 20.$$

This is a nonlinear hypothesis. According to Theorem 1, if rank $\left(\frac{\partial f}{\partial \theta}\left(\theta_0\right)\right) = q \leq K$, we have

$$n\left(f\left(\widehat{\theta}\right)-f\left(\theta_0\right)\right)'\left(\frac{\partial f}{\partial \theta}\left(\theta_0\right)\Omega\frac{\partial f}{\partial \theta}\left(\theta_0\right)'\right)^{-1}\left(f\left(\widehat{\theta}\right)-f\left(\theta_0\right)\right) \xrightarrow{d} \chi^2_q,$$

where in this example, $\theta = \beta$, $f\left(\beta\right) = -\beta_3/\left(2\beta_4\right)$. The gradient

$$\frac{\partial f}{\partial \beta}\left(\beta\right) = \left(0, 0, -\frac{1}{2\beta_4}, \frac{\beta_3}{2\beta_4^2}\right)$$

Since $\widehat{\beta} \xrightarrow{p} \beta_0$, by the continuous mapping theorem theorem, if $\beta_{0,4} \neq 0$, we have $\frac{\partial}{\partial \beta}f\left(\widehat{\beta}\right) \xrightarrow{p} \frac{\partial}{\partial \beta}f\left(\beta_0\right)$. Therefore, the (nonlinear) Wald test is

$$W_n = n\left(f\left(\widehat{\beta}\right)-20\right)'\left(\frac{\partial f}{\partial \beta}\left(\widehat{\beta}\right)\widehat{\Omega}\frac{\partial f}{\partial \beta}\left(\widehat{\beta}\right)'\right)^{-1}\left(f\left(\widehat{\beta}\right)-20\right) \xrightarrow{d} \chi^2_1.$$

3.2 Lagrangian Multiplier Test*

Restricted least square

$$\min_{\beta} (y - X\beta)' (y - X\beta) \text{ s.t. } R\beta = r.$$

Turn it into an unrestricted problem

$$L(\beta, \lambda) = \frac{1}{2n} (y - X\beta)' (y - X\beta) + \lambda' (R\beta - r).$$

The first-order condition

$$\begin{aligned}
\frac{\partial}{\partial \beta} L &= -\frac{1}{n} X' \left(y - X\tilde{\beta} \right) + \tilde{\lambda} R = -\frac{1}{n} X' e + \frac{1}{n} X' X \left(\tilde{\beta} - \beta^* \right) + R' \tilde{\lambda} = 0. \\
\frac{\partial}{\partial \beta} L &= R\tilde{\beta} - r = R \left(\tilde{\beta} - \beta^* \right) = 0
\end{aligned}$$

Combine these two equations into a linear system,

$$\begin{pmatrix} \widehat{Q} & R' \\ R & 0 \end{pmatrix} \begin{pmatrix} \tilde{\beta} - \beta^* \\ \tilde{\lambda} \end{pmatrix} = \begin{pmatrix} \frac{1}{n} X' e \\ 0 \end{pmatrix}.$$

$$\begin{aligned}
\begin{pmatrix} \tilde{\beta} - \beta^* \\ \tilde{\lambda} \end{pmatrix} &= \begin{pmatrix} \widehat{Q} & R' \\ R & 0 \end{pmatrix}^{-1} \begin{pmatrix} \frac{1}{n} X' e \\ 0 \end{pmatrix} \\
&= \begin{pmatrix} \widehat{Q}^{-1} - \widehat{Q}^{-1} R' \left(R\widehat{Q}^{-1} R' \right)^{-1} R\widehat{Q}^{-1} & \widehat{Q}^{-1} R' \left(R\widehat{Q}^{-1} R' \right)^{-1} \\ \left(R\widehat{Q}^{-1} R' \right)^{-1} R\widehat{Q}^{-1} & -\left(R\widehat{Q}^{-1} R' \right)^{-1} \end{pmatrix} \begin{pmatrix} \frac{1}{n} X' e \\ 0 \end{pmatrix}.
\end{aligned}$$

We conclude that

$$\sqrt{n}\tilde{\lambda} = \left(R\widehat{Q}^{-1} R' \right)^{-1} R\widehat{Q}^{-1} \frac{1}{\sqrt{n}} X' e$$

$$\sqrt{n}\tilde{\lambda} \Rightarrow N \left(0, \left(RQ^{-1} R' \right)^{-1} RQ^{-1} \Omega Q^{-1} R' \left(RQ^{-1} R' \right)^{-1} \right).$$

Let $W = \left(RQ^{-1}R'\right)^{-1} RQ^{-1}\Omega Q^{-1}R' \left(RQ^{-1}R'\right)^{-1}$, we have

$$n\tilde{\lambda}'W^{-1}\tilde{\lambda} \Rightarrow \chi_q^2.$$

If homoskedastic, then $W = \sigma^2 \left(RQ^{-1}R'\right)^{-1} RQ^{-1}QQ^{-1}R' \left(RQ^{-1}R'\right)^{-1} = \sigma^2 \left(RQ^{-1}R'\right)^{-1}$.

$$\begin{aligned} \frac{n\tilde{\lambda}'RQ^{-1}R'\tilde{\lambda}}{\sigma^2} &= \frac{1}{n\sigma^2}\left(y - X\tilde{\beta}\right)' XQ^{-1}X' \left(y - X\tilde{\beta}\right) \\ &= \frac{1}{n\sigma^2}\left(y - X\tilde{\beta}\right)' P_X \left(y - X\tilde{\beta}\right). \end{aligned}$$

3.3 Likelihood-Ratio test*

For likelihood ratio test, the starting point can be a criterion function $L(\beta) = (y - X\beta)'(y - X\beta)$. It does not have to be the likelihood function.

$$\begin{aligned} L\left(\tilde{\beta}\right) - L\left(\widehat{\beta}\right) &= \frac{\partial L}{\partial \beta}\left(\widehat{\beta}\right) + \frac{1}{2}\left(\tilde{\beta} - \widehat{\beta}\right)' \frac{\partial L}{\partial\beta\partial\beta}\left(\dot{\beta}\right)\left(\tilde{\beta} - \widehat{\beta}\right) \\ &= 0 + \frac{1}{2}\left(\tilde{\beta} - \widehat{\beta}\right)' \widehat{Q}\left(\tilde{\beta} - \widehat{\beta}\right). \end{aligned}$$

From the derivation of LM test, we have

$$\begin{aligned} \sqrt{n}\left(\tilde{\beta} - \beta^*\right) &= \left(\widehat{Q}^{-1} - \widehat{Q}^{-1}R'\left(R\widehat{Q}^{-1}R'\right)^{-1} R\widehat{Q}^{-1}\right)\frac{1}{\sqrt{n}}X'e \\ &= \frac{1}{\sqrt{n}}\left(X'X\right)X'e - \widehat{Q}^{-1}R'\left(R\widehat{Q}^{-1}R'\right)^{-1} R\widehat{Q}^{-1}\frac{1}{\sqrt{n}}X'e \\ &= \sqrt{n}\left(\widehat{\beta} - \beta^*\right) - \widehat{Q}^{-1}R'\left(R\widehat{Q}^{-1}R'\right)^{-1} R\widehat{Q}^{-1}\frac{1}{\sqrt{n}}X'e \end{aligned}$$

Therefore

$$\sqrt{n}\left(\tilde{\beta} - \widehat{\beta}\right) = -\widehat{Q}^{-1}R'\left(R\widehat{Q}^{-1}R'\right)^{-1} R\widehat{Q}^{-1}\frac{1}{\sqrt{n}}X'e$$

and

$$
\begin{aligned}
& n\left(\tilde{\beta}-\beta\right)'\widehat{Q}\left(\tilde{\beta}-\widehat{\beta}\right) \\
= & \frac{1}{\sqrt{n}}e'X\widehat{Q}^{-1}R'\left(R\widehat{Q}^{-1}R'\right)^{-1}R\widehat{Q}^{-1}\widehat{Q}\widehat{Q}^{-1}R'\left(R\widehat{Q}^{-1}R'\right)^{-1}R\widehat{Q}^{-1}\frac{1}{\sqrt{n}}X'e \\
= & \frac{1}{\sqrt{n}}e'X\widehat{Q}^{-1}R'\left(R\widehat{Q}^{-1}R'\right)^{-1}R\widehat{Q}^{-1}\frac{1}{\sqrt{n}}X'e
\end{aligned}
$$

In general, it is a quadratic form of normal distributions. If homoskedastic, then

$$
\left(R\widehat{Q}^{-1}R'\right)^{-1/2}R\widehat{Q}^{-1}\frac{1}{\sqrt{n}}X'e
$$

has variance

$$
\sigma^2\left(RQ^{-1}R'\right)^{-1/2}RQ^{-1}QQ^{-1}R'\left(RQ^{-1}R'\right)^{-1/2}=\sigma^2 I_q.
$$

We can view the optimization of the log-likelihood as a two-step optimization with the inner step $\sigma=\sigma\left(\beta\right)$. By the envelop theorem, when we take derivative with respect to β, we can ignore the indirect effect of $\partial\sigma\left(\beta\right)/\partial\beta$.