Ch. 5 Hypothesis Testing

(May 30, 2016)

1 Introduction

Inference, so far as we have seen, often take the form of numerical estimates, either as single points as confidence intervals. But not always. In many experimental situations the conclusion to be drawn in *not* numerical and is more aptly phrased as a choice between two conflicting theories, or *hypotheses*. Thus a court psychiatrist is called upon to pronounce an accused murderer either "sane" or "insane"; a stockbroker "buy" or "sell" a certain stock. In this chapter we examine the statistical methodology involved in making decisions of this sort.

The current framework of hypothesis testing is largely due to the work of Neyman and Pearson in the late 1920s, early 30s, complementing Fisher's work on estimation. As in estimation, we begin by postulating a statistical model but instead of seeking an estimator of $\boldsymbol{\theta}$ in $\boldsymbol{\Theta}$ we consider the question whether $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0 \subset \boldsymbol{\Theta}$ or $\boldsymbol{\theta} \in \boldsymbol{\Theta}_1 = \boldsymbol{\Theta} - \boldsymbol{\Theta}_0$ is most supported by the observed data. The discussion which follows will proceed in a similar way, though less systematically and formally, to the discussion of estimation. This is due to the complexity of the topic which arises mainly because one is asked to assimilate too many concepts too quickly just to be able to define the problem properly. This difficulty, however, is inherent in testing, if any proper understanding of the topic is to be attempted, and thus unavoidable.

2 Testing: Definition and Concepts

2.1 The Decision Rule

Let X be a random variables defined on the probability space $(\mathcal{S}, \mathcal{F}, \mathcal{P}(\cdot))$ and consider the statistical model associated with X:
(a) $\Phi = \{f(x;\boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$;
(b) $\mathbf{x} = (X_1, X_2, ..., X_n)$' is a random sample from $f(x;\boldsymbol{\theta})$.

Definition.
The problem of hypothesis testing is one of deciding whether or not some conjectures about $\boldsymbol{\theta}$ of the form $\boldsymbol{\theta}$ belongs to some subset $\boldsymbol{\Theta}_0$ of $\boldsymbol{\Theta}$ is supported[1] by the data $\boldsymbol{x} = (x_1, x_2, ..., x_n)'$. We call such a conjecture the *null hypothesis* and denoted it by

$$H_0 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0, \quad \boldsymbol{\Theta}_0 \subseteq \boldsymbol{\Theta}.$$

The null hypothesis is generally assumed to be true until evidence indicates otherwise. Against the null hypothesis H_0 we postulate the *alternative* H_1 which take the form:

$$H_1 : \boldsymbol{\theta} \notin \boldsymbol{\Theta}_0$$

or, equivalently

$$H_1 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_1 = \boldsymbol{\Theta} - \boldsymbol{\Theta}_0.$$ ■

It is important to note that H_0 and H_1 are in effect hypotheses about the distribution of the sample $f(\mathbf{x};\boldsymbol{\theta})$, i.e.

$$H_0 : f(\mathbf{x};\boldsymbol{\theta}),\ \boldsymbol{\theta} \in \boldsymbol{\Theta}_0,\ v.s.\ \ H_1 : f(\mathbf{x};\boldsymbol{\theta})\ \boldsymbol{\theta} \in \boldsymbol{\Theta}_1.$$

[1]The description of testable implications suggests the subset of values in the null hypothesis is contained within the unrestricted set, i.e. $\boldsymbol{\Theta}_0 \in \boldsymbol{\Theta}$. In this way, the models are said to be *nested*. Now consider an alternative pairs of models: $H_0 : f(\boldsymbol{x}, \boldsymbol{\theta})$ *v.s.* $H_1 : g(\boldsymbol{x}, \boldsymbol{\vartheta})$. These two models are *non-nested*. we are concerned only with nested models in this chapter.

Definition.
A hypothesis H_0 or H_1 is called *simple* if knowing $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ or $\boldsymbol{\theta} \in \boldsymbol{\Theta}_1$ specifies $f(\mathbf{x};\boldsymbol{\theta})$ completely, otherwise it is called a *composite* hypothesis. That is, if $f(\mathbf{x};\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ or $f(\mathbf{x};\boldsymbol{\theta})$ $\boldsymbol{\theta} \in \boldsymbol{\Theta}_1$ contain only one density function we say that H_0 or H_1 are simple hypothesis, respectively; otherwise they are said to be composite. ■

In testing a null hypothesis H_0 against an alternative H_1 the issue is to decide whether the sample realization $\boldsymbol{x}$ "support" H_0 or H_1. In the former case we say that H_0 is *accepted*, in the latter H_0 is *rejected*. In order to be able to make such a decision we need to formulate a mapping which related $\boldsymbol{\Theta}_0$ to some subset of the observation space $\mathcal{X}$, say C_0, we call an *acceptance region*, and its complement C_1 $(C_0 \cup C_1 = \mathcal{X}, C_0 \cap C_1 = \varnothing)$ we call the *rejection region*. If the sample realization $\boldsymbol{x} \in C_0$ we accept H_0, if $\boldsymbol{x} \notin C_0$ we reject it.

Since the observation space $\mathcal{X} \in \mathbb{R}^n$, but both the acceptance region $C_0 \in \mathbb{R}^1$ and rejection region $\{\mathbb{R}^1 - C_0\}$, we need a mapping from $\mathbb{R}^n$ to $\mathbb{R}^1$.

Definition. (The Test Statistics)
The mapping $\tau(\cdot)$ which enables us to define C_0 and C_1 from observation space $\mathcal{X}$ is called a *test statistic*, i.e.

$$\tau(\mathbf{x}) : \mathcal{X} \to \mathbb{R}^1. \qquad \blacksquare$$

Example. (Simple Null against Composite Alternative)
Let X be the random variables representing the marks achieved by students in an econometric theory paper and let the statistical model be:
(a). $\Phi = \left\{ f(x;\theta) = \frac{1}{8\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\theta}{8}\right)^2 \right] \right\}$, $\theta \in \Theta \equiv [0, 100]$;
(b). $\mathbf{x} = (X_1, X_2, ..., X_n)'$, n=40, is random sample from Φ.

The hypothesis to be tested is

$$H_0 : \theta = 60 \ \ (i.e.\ X \sim N(60, 64)),\ \Theta_0 = \{60\}$$

against

$$H_1 : \theta \neq 60 \ \ (i.e.\ X \sim N(\mu, 64), \mu \neq 60),\ \Theta_1 = [0, 100] - \{60\}.$$

Common sense suggests that if some "good" estimator of θ, say $\bar{X}_n = (1/n)\sum_{i=1}^{n} x_i$ for the sample realization $\boldsymbol{x}$ takes a value "around" 60 then we will be inclined to accept H_0. Let us formalise this argument:

"The *acceptance region* takes the form: $60 - \varepsilon \leq \bar{X}_n \leq 60 + \varepsilon,\ \varepsilon > 0$, or

$$C_0 = \{\boldsymbol{x} : |\bar{X}_n - 60| \leq \varepsilon\}$$

and

$$C_1 = \{\boldsymbol{x} : |\bar{X}_n - 60| \geq \varepsilon\}, \quad \text{is} \quad \text{the } rejection\ region."$$ ■

2.2 Type I and Type II Errors

In the above example, the next question is "how do we choose ε ?" If ε is too small we run the risk of *rejecting H_0 when it is true*; we call this *type I error*. On the other hand, if ε is too large we run the risk of *accepting H_0 when it is false*; we call this *type II error*.

Definition.
Formally, if $\boldsymbol{x} \in C_1$ (reject H_0) and $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ (H_0 is true)–we call it *type I error*, α; if $\boldsymbol{x} \in C_0$ (accept H_0) and $\boldsymbol{\theta} \in \boldsymbol{\Theta}_1$ (H_0 is false)–we call it *type II error*, β. ■

That is, if we were to choose ε too small we run a higher risk of committing a type I error than of committing a type II error and vice versa. That is, there is a trade-off between the probability of type I error, i.e.

$$Pr(\boldsymbol{x} \in C_1; \boldsymbol{\theta} \in \boldsymbol{\Theta}_0) = \alpha,$$

and the probability β of type II error, i.e.

$$Pr(\boldsymbol{x} \in C_0; \boldsymbol{\theta} \in \boldsymbol{\Theta}_1) = \beta.$$

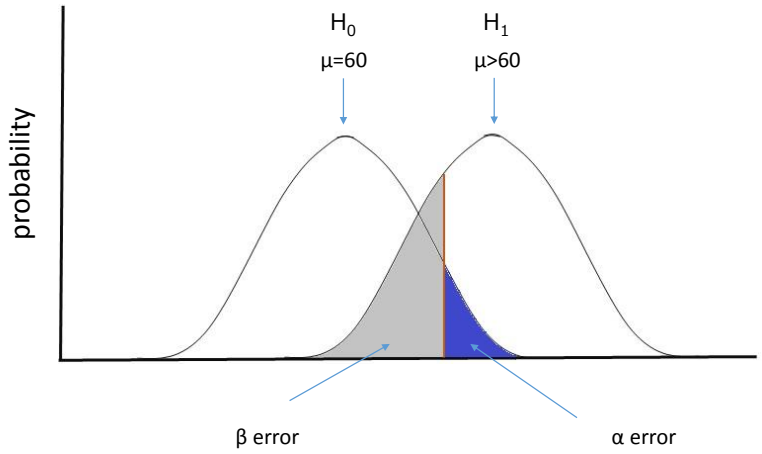

Figure (5-1a). Type I and Type II Error

Ideally we would like $\alpha = \beta = 0$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ which is not possible for a fixed n. Moreover we cannot control both simultaneously because of the trade-off between them. The strategy adopted in hypothesis testing where a small value of α is chosen and for a given α, β is minimized. Formally, this amounts to choose α^* such that

$$Pr(\boldsymbol{x} \in C_1; \boldsymbol{\theta} \in \boldsymbol{\Theta}_0) = \alpha(\boldsymbol{\theta}) \leq \alpha^* \quad for\ \boldsymbol{\theta} \in \boldsymbol{\Theta}_0,$$

and

$$Pr(\boldsymbol{x} \in C_0; \boldsymbol{\theta} \in \boldsymbol{\Theta}_1) = \beta(\boldsymbol{\theta}), \quad is\ minimized\ for\ \boldsymbol{\theta} \in \boldsymbol{\Theta}_1$$

by choosing C_1 or C_0 appropriately.

In the case of the above example if we were to choose α, say $\alpha^* = 0.05$, then

$$Pr(|\bar{X}_n - 60| > \varepsilon; \theta = 60) = 0.05.$$

"How do we determine ε, then ?" The only random variable involved in the statement is $\bar{X}$ and hence it has to be its sampling distribution. For the above probabilistic statement to have any operational meaning to enable us to determine ε, the distribution of $\bar{X}_n$ must be known. In the present case we know that

$$\bar{X}_n \sim N\left(\theta, \frac{\sigma^2}{n}\right) \quad where\ \frac{\sigma^2}{n} = \frac{64}{40} = 1.6,$$

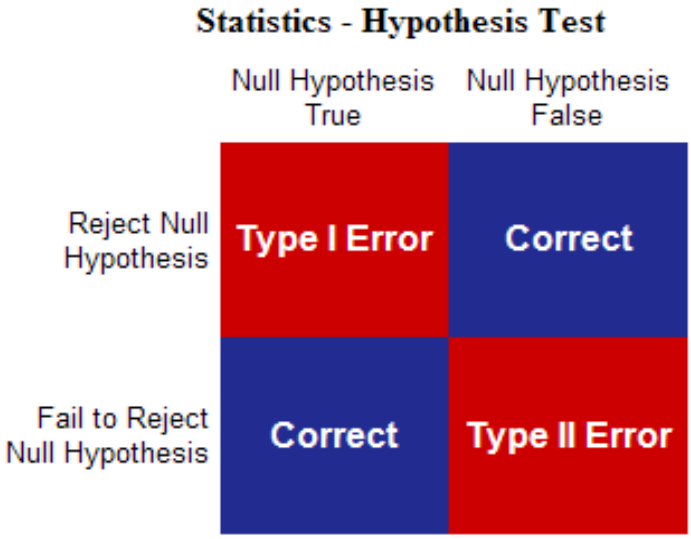

Figure (5-1b). Type I and Type II Error

which implies that for $\theta = 60$, (i.e. when H_0 is true) we can "construct" a *test statistic* $\tau(\mathbf{x})$ from sample $\mathbf{x}$ such that

$$\tau(\mathbf{x}) = \left(\frac{\bar{X}_n - \theta}{\sqrt{1.6}}\right) = \left(\frac{\bar{X}_n - 60}{\sqrt{1.6}}\right) = \left(\frac{\bar{X}_n - 60}{1.265}\right) \sim N(0,1),$$

and thus the distribution of $\tau(\cdot)$ is known completely (*no unknown parameters*). When this is the case this distribution can be used in conjunction with the above probabilistic statement to determine ε. In order to do this we need to relate $|\bar{X}_n - 60|$ to $\tau(\mathbf{x})$ (a statistics) for which the distribution is known. The obvious way is to standardize the former. This suggests changing the above probabilistic statement to the equivalent statement

$$Pr\left(\frac{|\bar{X}_n - 60|}{1.265} \geq c_\alpha; \theta = 60\right) = 0.05 \quad where \;\; c_\alpha = \frac{\varepsilon}{1.265}.$$

The value of c_α given from the $N(0,1)$ table is $c_\alpha = 1.96$. This in turn implies that the rejection region for the test is

$$C_1 = \left\{\mathbf{x} : \frac{|\bar{X}_n - 60|}{1.265} \geq 1.96\right\} = \{\mathbf{x} : |\tau(\mathbf{x})| \geq 1.96\}$$

or

$$C_1 = \{\mathbf{x} : |\bar{X}_n - 60| \geq 2.48\}.$$

That is, for sample realization $\mathbf{x}$ which give rises to $\bar{X}_n$ falling outside the interval (57.52, 62.48) we reject H_0.

Let us summarize the argument so far.

> "We set out to construct a size-α test for $H_0 : \theta = 60$ against $H_1 : \theta \neq 60$ and intuition suggested the rejection region $(|\bar{X}_n - 60| \geq \varepsilon)$. In order to determine ε we have to
> (a). Choose an α; and then
> (b). define the rejection region in term of some statistic $\tau(\mathbf{x})$."

The latter is necessary to enable us to determine ε via some known distribution. This is the distribution of the *test statistic* $\tau(\mathbf{x})$ *under* H_0 (i.e. when H_0 is true).

Definition.
A *pivotal test statistics* for $\tau(\cdot)$ is a function of sample whose distribution does not depend on $\boldsymbol{\theta}$ under the null hypothesis. ■

Example.
Assume a random sample of size 11 is drawn from a normal distribution $N(\mu, 400)$. In particular, $y_1 = 62, y_2 = 52, y_3 = 68, y_4 = 23, y_5 = 34, y_6 = 45, y_7 = 27, y_8 = 42, y_9 = 83, y_{10} = 56$ and $y_{11} = 40$. Test the null hypothesis that $H_0 : \mu = 55$ versus $H_1 : \mu \neq 55$.

Since σ^2 is known, the sample mean will distributed as

$$\bar{Y} \sim N(\mu, \sigma^2/n) \equiv N(\mu, 400/11),$$

therefore under $H_0 : \mu = 55$,

$$\bar{Y} \sim N(55, 36.36)$$

or

$$\frac{\bar{Y} - 55}{\sqrt{36.36}} \sim N(0, 1) \equiv Z.$$

We accept H_0 when the test statistics $\tau(\mathbf{x}) = (\bar{Y} - 55)/\sqrt{36.36}$ lying in the interval $C_0 = [Z_{-.025}, Z_{0.975}] = [-1.96, 1.96]$ under the size of the test $\alpha = 0.05$.

We now have

$$\sum_{i=1}^{11} y_i = 532 \ \textit{and} \ \bar{y} = \frac{532}{11} = 48.4.$$

Then

$$\frac{48.4 - 55}{\sqrt{36.36}} = -1.01$$

which is in the accept region. Therefore we accept the null hypothesis that $H_0 : \mu = 55$. ■

Example.

Assume a random sample of size 11 is drawn from a normal distribution $N(\mu, \sigma^2)$. In particular, $y_1 = 62, y_2 = 52, y_3 = 68, y_4 = 23, y_5 = 34, y_6 = 45, y_7 = 27, y_8 = 42, y_9 = 83, y_{10} = 56$ and $y_{11} = 40$. Test the null hypothesis that $H_0 : \mu = 55$ versus $H_1 : \mu \neq 55$.

Since σ^2 is unknown, the sample mean distributed as

$$\bar{Y} \sim N(\mu, \sigma^2/n),$$

therefore under $H_0 : \mu = 55$

$$\bar{Y} \sim N(55, \sigma^2/n)$$

or

$$\frac{\bar{Y} - 55}{\sqrt{\sigma^2/n}} \sim N(0, 1),$$

however, it is not a pivotal test statistics since an unknown parameter σ^2.

From the fact that $\sum(Y_i - \bar{Y})^2/\sigma^2 \sim \chi^2_{n-1}$ or $s^2(n-1)/\sigma^2 \sim \chi^2_{n-1}$ where $s^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2/(n-1)$ is an unbiased estimator of σ^2,[2] we have

$$\frac{(\bar{Y} - 55)/(\sqrt{\sigma^2/n})}{\sqrt{(n-1)s^2/(n-1)\sigma^2}} = \frac{\bar{Y} - 55}{\sqrt{s^2/n}} \sim t_{n-1}.$$

We accept H_0 when the test statistics $\tau(\mathbf{x}) = \frac{\bar{Y}-55}{\sqrt{s^2/n}}$ lying in the interval $C_0 = [t(10)_{0.025}, t(10)_{0.975}] = [-2.23, 2.23]$. We now have

$$\begin{aligned} \sum_{i=1}^{11} y_i &= 532, \quad \sum_{i=1}^{11} y_i^2 = 29000 \ \text{ and} \\ s^2 &= \frac{\sum y_i^2 - n\bar{y}^2}{10} = \frac{29000 - 11(48.4)^2}{10} = 323.19. \end{aligned}$$

[2] See p. 27 of Chapter 3.

Then

$$\frac{48.4 - 55}{\sqrt{323.19/11}} = -1.2$$

which is also in the accept region C_0. Therefore we accept the null hypothesis that $H_0 : \mu = 55$. ■

The next question which naturally arises is: "What do we need the probability of type II error β for ?" The answer is that we need β to decide whether the test defined in terms of C_1(of course C_0) is a "good" or a "bad" test. As we mentioned at the outset, the way we decided to solve the problem of the trade-off between α and β was to choose a given small value of α and define C_1 so as to minimize β. At this stage we do not know whether the test defined above is a "good" test or not. Let us consider setting up the apparatus to enable us to consider the question of optimality.

3 Optimal Tests

In the last section, we developed $(1-\alpha)$ confidence intervals and size-α tests using pivotal quantities. However, there may be many pivotal test statistic, each leading to its own confidence intervals and test. In choosing which of many size α test to use, in this section we try to find size-α procedures which have small probability of falsely accepting an untrue hull hypothesis, .i.e. minimizing of $Pr(\mathbf{x} \in C_0)$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}_1$ or maximizing $Pr(\mathbf{x} \in C_1)$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}_1$.

Definition. (Power of the Test)
The probability of reject H_0 when false at some point $\boldsymbol{\theta}_1 \in \boldsymbol{\Theta}_1$, i.e. $Pr(\mathbf{x} \in C_1; \boldsymbol{\theta} = \boldsymbol{\Theta}_1)$ is called the *power of the test* at $\boldsymbol{\theta} = \boldsymbol{\theta}_1$. ■

Note that

$$Pr(\mathbf{x} \in C_1; \boldsymbol{\theta} = \boldsymbol{\theta}_1) = 1 - Pr(\mathbf{x} \in C_0; \boldsymbol{\theta} = \boldsymbol{\theta}_1) = 1 - \beta(\boldsymbol{\theta}_1).$$

Example.
In the last example we can define the power of the test at some $\theta_1 \in \Theta_1$, say $\theta = 54$, to be $Pr[(|\bar{X}_n - 60|)/1.265 \geq 1.96;\ \theta = 54]$.

Under the alternative hypothesis that $\theta = 54$, then it is true that $\frac{\bar{X}_n - 54}{1.265} \sim N(0,1)$. We would like to know that the probability of the statistics constructed *under the null hypothesis* that $\frac{\bar{X}_n - 60}{1.265}$ would fall in the rejection region; that is, the power of the test at $\theta = 54$ to be

$$Pr\left(\frac{|\bar{X}_n - 60|}{1.265} \geq 1.96;\ \theta = 54\right) = Pr\left(\frac{|\bar{X}_n - 54|}{1.265} \leq -1.96 - \frac{(54-60)}{1.265}\right) + Pr\left(\frac{|\bar{X}_n - 54|}{1.265} \geq 1.96 - \frac{(54-60)}{1.265}\right) = 0.993.$$

Hence, the power of the test defined by C_1 above is indeed very high for $\theta = 54$. From this we know that to calculate the power of a test we need to know the distribution of the test statistics $\tau(\mathbf{x})$ under the alternative hypothesis. In this case it is the distribution of $\frac{\bar{X}_n - 54}{1.265}$.[3]

[3] In the example above, the test statistic $\tau(\mathbf{x})$ have a standard normal distribution under both the

Following the same procedure the power of the test defined by C_1 is as following for all $\theta \in \Theta_1$:

$$Pr(|\tau(\mathbf{x})| \geq 1.96; \theta = 56) = 0.8849;$$
$$Pr(|\tau(\mathbf{x})| \geq 1.96; \theta = 58) = 0.3520;$$
$$Pr(|\tau(\mathbf{x})| \geq 1.96; \theta = 60) = 0.0500;$$
$$Pr(|\tau(\mathbf{x})| \geq 1.96; \theta = 62) = 0.3520;$$
$$Pr(|\tau(\mathbf{x})| \geq 1.96; \theta = 64) = 0.8849;$$
$$Pr(|\tau(\mathbf{x})| \geq 1.96; \theta = 66) = 0.9973.$$

As we can see, the power of the test increases as we go further away from $\theta = 60$ (i.e. the null hypothesis H_0) and the power at $\theta = 60$ equals the probability of type I error. This prompts us to define the power function as follows.

Definition. (Power Function)
$\mathrm{P}(\theta) = Pr(\mathbf{x} \in C_1),\ \theta \in \Theta$ is called the *power function of the test* defined by the rejection region C_1. ■

Definition. (Size of a Test)
$\alpha = max_{\theta \in \Theta_0}\mathrm{P}(\theta)$ is defined to be the *size* (or the significance level) of the test.[4] In the case where H_0 is simple, say $\theta = \theta_0$, then $\alpha = \mathrm{P}(\theta_0)$. ■

These definitions enable us to define a criterion for a "best" test of a given α to be the one whose power function $\mathrm{P}(\theta),\ \theta \in \Theta_1$ is maximum at every θ.

Definition. (Uniformly Most Powerful Test)
A test of $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$ as defined by some rejection region C_1 is said to be uniformly most powerful (UMP) test of size α if

null and the alternative hypothesis. However, it is quite often the case when it happen that a test statistics have a different distribution under the null and the alternative hypotheses. For example, the unit root test. See Chapter 21.

[4]That, the probability that $Pr(\mathbf{x} \in C_1)$, but H_0 is correct, i.e. $\theta \in \Theta_0$.

(a). $max_{\theta\in\Theta_0} \mathrm{P}(\theta) = \alpha$;

(b). $\mathrm{P}(\theta) \geq \mathrm{P}^*(\theta)$ for all $\theta \in \Theta_1$, where $\mathrm{P}^*(\theta)$ is the power function of any other test of size α. ■

Example.

Suppose we observe $X \sim N(\theta, 1)$, independent. We assume for simplicity that $n = 1$. Then

$$\tau(\mathbf{x}) = \tau(X_1) = Q \sim N(\theta, 1).$$

The 0.05 one-side test for testing that $\theta = 0$ against $\theta > 0$ reject the null hypothesis if $(Q - 0)/1 = Q > 1.645$. To find the power of the test, we calculate

$$\begin{aligned} \mathrm{P}^*(\theta) = Pr(Q > 1.645; \theta) &= Pr((Q - \theta + \theta) > 1.645) \\ &= Pr(Z > 1.645 - \theta) \\ &= 1 - N(1.645 - \theta). \end{aligned}$$

Now the 0.05 one-side test one the other side is to testing that $\theta = 0$ against $\theta < 0$. We reject the null hypothesis if $(Q - 0)/1 < 1.645$. To find the power of the test, we calculate

$$\begin{aligned} \mathrm{P}_*(\theta) = Pr(Q < 1.645; \theta) &= Pr((Q - \theta + \theta) < 1.645) \\ &= Pr(Z < 1.645 - \theta) \\ &= N(1.645 - \theta). \end{aligned}$$

Finally, the two side size-0.05 test which reject the null hypothesis if $Q > 1.96$ or $Q < -1.96$. Then the power function is given by

$$\begin{aligned} \mathrm{P}(\theta) = 1 - Pr(-1.96 < Q < 1.96; \theta) &= 1 - Pr(-1.96 < (Q - \theta) + \theta < 1.96) \\ &= 1 - Pr(-1.96 - \theta < Z < 1.96 - \theta) \\ &= 1 - N(1.96 - \theta) + N(-1.96 - \theta). \quad ■ \end{aligned}$$

The three power functions are drawn in the following.

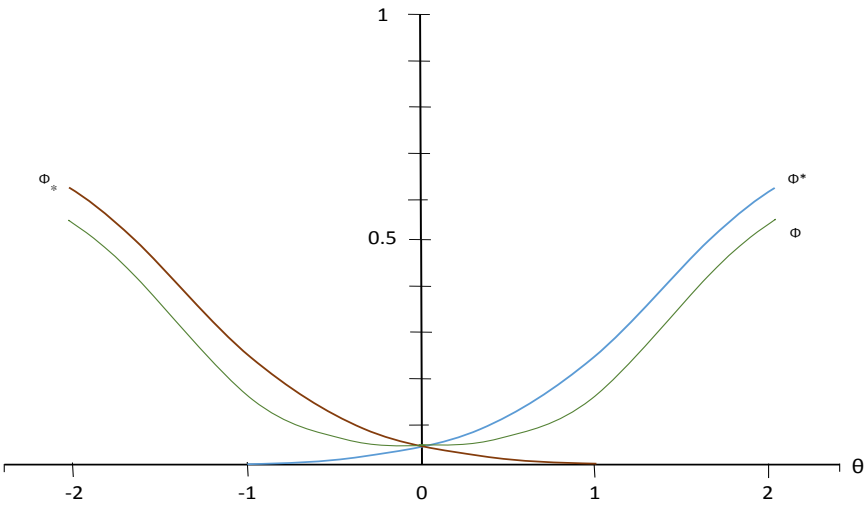

Figure (5-2). Power functions of $\mathrm{P}(\theta)$, $\mathrm{P}^*(\theta)$, and $\mathrm{P}_*(\theta)$

As will be seen in the example above, no UMP tests exists in most situations of interest in practice. The procedure adopted in such cases is to reduce the class of all tests to some subclass by imposing some more criteria and consider the question of UMP tests within the subclass.

Definition.
A test of $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$ is said to be *unbiased* if

$$max_{\theta\in\Theta_0}\mathrm{P}(\theta) \leq max_{\theta\in\Theta_1}\mathrm{P}(\theta).$$

In other word, a test is unbiased if it reject H_0 more often when it is false than when it is true. ■

Collecting all the above concepts together we say that a test has been defined when the following components have been specified:
(a). a test statistic $\tau(\mathbf{x})$;
(b). the size of the test α;
(c). the distribution of $\tau(\mathbf{x})$ under H_0 and H_1;
(d). the rejection region C_1 (or, equivalently, C_0).

4 Asymptotic Test Procedures

As discussed in the last section, the main problem in hypothesis testing is to construct a test statistics $\tau(\mathbf{x})$ whose distribution we know under both the null hypothesis H_0 and the alternative H_1 and it does not depend on the unknown parameters $\boldsymbol{\theta}$. The first part of the problem, that of constructing $\tau(\mathbf{x})$, can be handled relatively easy using various methods (Neyman-Pearson likelihood ratio) when certain condition are satisfied. The second part of the problem, that of determining the distribution of $\tau(\mathbf{x})$ under both H_0 and H_1, is much more difficult to solve and often we have to resort to asymptotic theory. This amount to deriving the asymptotic distribution of $\tau(\mathbf{x})$ and using that to determine the rejection region C_1 (or C_0) and associated probabilities.

4.1 Asymptotic Properties

For a given sample size n, if the distribution of $\tau_n(\mathbf{x})$ is not known (otherwise we use that), we do not know how "good" the asymptotic distribution of $\tau_n(\mathbf{x})$ is an accurate approximation of its finite sample distribution. This suggest that when asymptotic results are used we should be aware of their limitations and the inaccuracies they can lead to.

Consider the test defined by the rejection region

$$C_1^n = \{\mathbf{x} : |\tau_n(\mathbf{x})| \geq c_n\},$$

and whose power function is

$$\mathrm{P}_n(\boldsymbol{\theta}) = Pr(\mathbf{x} \in C_1^n), \;\; \theta \in \Theta.$$

Since the distribution of $\tau_n(\mathbf{x})$ is not known we cannot determine c_n or P_n. If the asymptotic distribution of $\tau_n(\mathbf{x})$ is available, however we can use that instead to define c_n from some fixed α and the asymptotic power function

$$\lim_{n\to\infty} \pi_n(\boldsymbol{\theta}) = Pr(\mathbf{x} \in C_1^\infty), \;\; \theta \in \Theta.$$

In this sense we can think of $\{\tau_n(\mathbf{x}),\ n \geq 1\}$ as a sequence of test statistics defining a sequence of rejection region $\{C_1^n, n \geq 1\}$ with power function $\{\mathrm{P}_n(\theta), n \geq 1,\ \theta \in \Theta\}$.

Definition. (Consistence of a Test)
The sequence of tests for $H_0 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ against $H_1 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_1$ defined by $\{C_1^n, n \geq 1\}$ is said to be *consistent of size* α if

$$\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} \pi(\boldsymbol{\theta}) = \alpha$$

and

$$\pi(\boldsymbol{\theta}) = 1, \quad \boldsymbol{\theta} \in \boldsymbol{\Theta}_1. \qquad \blacksquare$$

It is often happened that use of conventional asymptotic critical asymptotic critical values (C_1^∞) for many tests may cause *size distortions* in a finite sample, however.

Definition. (Finite Sample Size Distortion)
The size distortion of a test is defined as

$$\left| \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} \pi_n(\boldsymbol{\theta}) - \alpha \right| \quad \text{for finite } n. \qquad \blacksquare$$

Definition. (Unbias of a Test):
A sequence of test as defined above is said to be *asymptotically unbiased of size* α if

$$\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} \pi(\boldsymbol{\theta}) = \alpha$$

and

$$\alpha < \pi(\boldsymbol{\theta}) < 1, \quad \boldsymbol{\theta} \in \boldsymbol{\Theta}_1. \qquad \blacksquare$$

4.2 Three Asymptotically Equivalent Test Procedures

The use of increasingly complex statistical models has led to heavy reliance on maximum likelihood methods for both estimation and testing. In such a setting only asymptotic properties can be expected for estimators or tests. In this section three general test procedures–"Holy Trinity"–which gives rise to asymptotically optimal tests will be considered: the *likelihood ratio, Wald* and *Lagrange multiplier* test. All three test procedures can be interpreted as utilizing the information incorporated in the log likelihood function in different but asymptotically equivalent ways.

For expositional purpose the test procedures will be considered in the context of the simplest statistical model where
(a). $\Phi = \{f(x;\boldsymbol{\theta}),\ \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ is the probability model; and
(b). $\mathbf{x} \equiv (X_1, X_2, ..., X_n)'$ is a random sample.

We consider maximum likelihood estimation of parameters $\boldsymbol{\theta}$ and a test of the simple null hypothesis that $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0,\ \boldsymbol{\theta} \in \boldsymbol{\Theta} \equiv \mathbb{R}^m$ against $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$.

4.2.1 The Likelihood Ratio Test

We first discuss the likelihood ratio test which is a test statistics calculated under both the null and the alternative hypothesis.

Definition. (Likelihood Ratio Test)
If the restriction $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ is valid, then imposing it should not lead to a large reduction in the log-likelihood function. Therefore, we base the test on the difference, $\ln L - \ln L_R$, where L is the value of the likelihood function at the unrestricted value of $\boldsymbol{\theta}$ and L_R is the value of the likelihood function at the restricted estimate. ■

The likelihood ratio test statistics takes the form

$$\lambda(\mathbf{x}) = \frac{L(\boldsymbol{\theta}_0;\mathbf{x})}{\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} L(\boldsymbol{\theta};\mathbf{x})} = \frac{L(\boldsymbol{\theta}_0;\mathbf{x})}{L(\hat{\boldsymbol{\theta}};\mathbf{x})},$$

where $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$.

Under certain regularity conditions which include RC1-RC3 (see Chapter 3), $\log L(\boldsymbol{\theta};\mathbf{x})$ can be expanded in a Taylor series at $\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}$:

$$\log L(\boldsymbol{\theta};\mathbf{x}) \simeq \log L(\hat{\boldsymbol{\theta}};\mathbf{x}) + \left.\frac{\partial \log L}{\partial \boldsymbol{\theta}'}\right|_{\hat{\boldsymbol{\theta}}} (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}) + \frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})' \left.\frac{\partial^2 \log L}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right|_{\hat{\boldsymbol{\theta}}} (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}) + o(1).$$

Since

$$\left.\frac{\partial \log L}{\partial \boldsymbol{\theta}'}\right|_{\hat{\boldsymbol{\theta}}} = \mathbf{0},$$

being the FOC of the MLE, and

$$\frac{1}{n}\left(\frac{\partial^2}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\log L(\hat{\boldsymbol{\theta}};\mathbf{x})\right) \xrightarrow{p} \mathbf{I}(\boldsymbol{\theta}),$$

the above expansion can be simplified to

$$\log L(\boldsymbol{\theta};\mathbf{x}) = \log L(\hat{\boldsymbol{\theta}};\mathbf{x}) + \frac{1}{2}n(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})'\mathbf{I}(\boldsymbol{\theta})(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}) + o(1).$$

This implies that under the null hypothesis that $H_0 : \boldsymbol{\theta}=\boldsymbol{\theta}_0$

$$-2\log\lambda(\mathbf{x}) = 2[\log L(\hat{\boldsymbol{\theta}};\mathbf{x}) - L(\boldsymbol{\theta}_0;\mathbf{x})] = n(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)'\mathbf{I}(\boldsymbol{\theta})(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) + o(1).$$

From the asymptotic properties of the MLE's it is known[5] that under certain regularity conditions

$$\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) \sim N(0, \mathbf{I}^{-1}(\boldsymbol{\theta})).$$

Using this we can deduce that

$$\begin{aligned} LR = -2\log\lambda(\mathbf{x}) &= 2[\log L(\hat{\boldsymbol{\theta}};\mathbf{x}) - L(\boldsymbol{\theta}_0;\mathbf{x})] \\ &\simeq n(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)'\mathbf{I}(\boldsymbol{\theta})(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) \\ &\overset{H_0}{\sim} \chi^2(m), \end{aligned} \tag{5-1}$$

being a quadratic form in asymptotically normal random variables.

[5] See equation (3-9) of Chapter 3

4.2.2 Wald test

To compute the LR test statistics, both likelihood function under he null and the alternative hypotheses needed to be calculated. Wald (1943), using the above approximation of $-2\log\lambda(\mathbf{x})$ in (1), proposed an alternative test that is computed under H_1 only.

Definition. (Wald Test)
Replacing $\mathbf{I}(\boldsymbol{\theta})$ with $\mathbf{I}(\hat{\boldsymbol{\theta}})$ in the above approximation of $-2\log\lambda(\mathbf{x})$ in (5-1), we obtain that

$$W = n(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'\mathbf{I}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \overset{H_0}{\sim} \chi^2(m),$$

given that $\mathbf{I}(\hat{\boldsymbol{\theta}}) \xrightarrow{p} \mathbf{I}(\boldsymbol{\theta})$.[6]

4.2.3 The Lagrange Multiplier Test

Rao (1947) using the asymptotic distribution of the score function to propose the LM test which is computed under H_0 only.

Definition. (The LM Test)
Under the null hypothesis, the test statistics

$$LM = \frac{1}{n}\left(\frac{\partial \log L(\boldsymbol{\theta}_0;\mathbf{x})}{\partial\boldsymbol{\theta}}\right)' \mathbf{I}^{-1}(\boldsymbol{\theta}_0)\left(\frac{\partial \log L(\boldsymbol{\theta}_0;\mathbf{x})}{\partial\boldsymbol{\theta}}\right) \overset{H_0}{\sim} \chi^2(m).$$

Proof.
Expanding score function of $\log L(\boldsymbol{\theta};\mathbf{x})$ (i.e. $\frac{\partial \log L(\boldsymbol{\theta};\mathbf{x})}{\partial\boldsymbol{\theta}}$) around $\hat{\boldsymbol{\theta}}$ we have:

$$\frac{\partial \log L(\boldsymbol{\theta};\mathbf{x})}{\partial\boldsymbol{\theta}} \simeq \frac{\partial \log L(\hat{\boldsymbol{\theta}};\mathbf{x})}{\partial\boldsymbol{\theta}} + \frac{\partial^2 \log L(\hat{\boldsymbol{\theta}};\mathbf{x})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}).$$

As in the LR test, $\frac{\partial \log L(\hat{\boldsymbol{\theta}};\mathbf{x})}{\partial\boldsymbol{\theta}} = 0$, and the above equation reduces to

$$\frac{\partial \log L(\boldsymbol{\theta};\mathbf{x})}{\partial\boldsymbol{\theta}} \simeq \frac{\partial^2 \log L(\hat{\boldsymbol{\theta}};\mathbf{x})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}).$$

[6] $\mathbf{I}(\hat{\boldsymbol{\theta}})$ can be anyone of the three estimators which estimate the asymptotic variance of the MLE. See section 3.3.6 of Chapter 3.

Now we consider the test statistics under the null hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$

$$LM = \frac{1}{n}\left(\frac{\partial \log L(\boldsymbol{\theta}_0; \mathbf{x})}{\partial \boldsymbol{\theta}}\right)' \mathbf{I}^{-1}(\boldsymbol{\theta}_0)\left(\frac{\partial \log L(\boldsymbol{\theta}_0; \mathbf{x})}{\partial \boldsymbol{\theta}}\right),$$

which is equivalent to

$$LM = \frac{1}{n}(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})' \frac{\partial^2 \log L(\hat{\boldsymbol{\theta}}; \mathbf{x})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \mathbf{I}^{-1}(\boldsymbol{\theta}_0) \frac{\partial^2 \log L(\hat{\boldsymbol{\theta}}; \mathbf{x})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} (\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}).$$

Given that $\frac{1}{n}\left(\frac{\partial^2 \log L(\hat{\boldsymbol{\theta}}; \mathbf{x})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right) \xrightarrow{p} \mathbf{I}(\boldsymbol{\theta}) = \mathbf{I}(\boldsymbol{\theta}_0)$ (under H_0), we have

$$\begin{aligned} LM &= \frac{1}{n}\left(\frac{\partial \log L(\boldsymbol{\theta}_0; \mathbf{x})}{\partial \boldsymbol{\theta}}\right)' \mathbf{I}^{-1}(\boldsymbol{\theta}_0)\left(\frac{\partial \log L(\boldsymbol{\theta}_0; \mathbf{x})}{\partial \boldsymbol{\theta}}\right) \\ &= n(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' \mathbf{I}(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \sim \chi^2(m) \end{aligned}$$

as in the proof of LR test. ■

Example.

To test the hypothesis $H_0 : c(\theta) = 0$ from these three tests. The logic of these tests can be seen in the following figure. The figure plot the log-likelihood function $\ln(\theta)$, its derivative with respect to θ, $d\ln(\theta)/d\theta$, and the constrain $c(\theta)$.

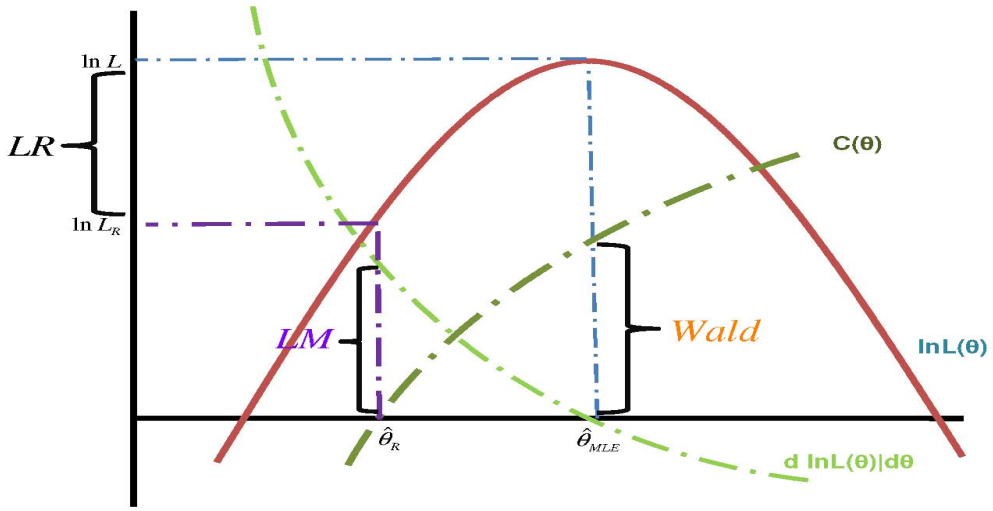

Figure (5-3). The Holy Trinity in asymptotic test statistics

■

Example.

Consider a set of T *i.i.d.* observations on a Bernoulli random variable which takes on the value:

$$y_t = \begin{cases} 1, & \text{with probability } \theta, \\ 0, & \text{with probability } 1-\theta. \end{cases}$$

We wish to test $\theta = \theta_0$ against $\theta \neq \theta_0$ for $\theta \in (0,1)$.

The log-likelihood function is given by:

$$\ln L(\theta, \mathbf{y}) = \sum_t [y_t \ln \theta + (1-y_t)\ln(1-\theta)],$$

with the MLE, $\hat{\theta} = \bar{y}$. The first derivative (score) is

$$s(\theta, \mathbf{y}) = \frac{1}{\theta(1-\theta)} \sum_t (y_t - \theta).$$

The information is

$$\begin{aligned} \mathrm{I}(\theta) &= E\left[\frac{T\theta(1-\theta) + (1-2\theta)\sum(y_y - \theta)}{\theta^2(1-\theta)^2}\right] \Big/ T \\ &= \frac{1}{\theta(1-\theta)}. \end{aligned}$$

The Wald test is given by

$$\begin{aligned} W &= T(\bar{y}-\theta_0)' \frac{1}{\bar{y}(1-\bar{y})}(\bar{y}-\theta_0) \\ &= T(\bar{y}-\theta_0)^2/\bar{y}(1-\bar{y}). \end{aligned}$$

The LM test is

$$\begin{aligned} LM &= \frac{1}{T}\frac{\sum(y_t-\theta_0)}{\theta_0(1-\theta_0)} \left[\frac{1}{\theta_0(1-\theta_0)}\right]^{-1} \frac{\sum(y_t-\theta_0)}{\theta_0(1-\theta_0)} \\ &= \left[\frac{\sum(y_t-\theta_0)}{\theta_0(1-\theta_0)}\right]^2 \frac{\theta_0(1-\theta_0)}{T} \\ &= T(\bar{y}-\theta_0)^2/\theta_0(1-\theta_0). \end{aligned}$$

The likelihood Ratio test statistics is given by

$$LR = 2T\left[\bar{y}\ln \bar{y}/\theta_0 + (1-\bar{y})\ln(1-\bar{y})/(1-\theta_0)\right]. \qquad \blacksquare$$

Example.

Of particular interest in practice is the case where $\boldsymbol{\theta} \equiv (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')'$ and $H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_1^0$ against $H_1 : \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_1^0$, $\boldsymbol{\theta}_1$ is $r \times 1$ with $\boldsymbol{\theta}_2$ $(m-r) \times 1$ left unrestricted. In this case the three test statistics take the form

$$\begin{aligned} LR &= 2(\ln L(\hat{\boldsymbol{\theta}}; \mathbf{x}) - \ln L(\tilde{\boldsymbol{\theta}}; \mathbf{x})); \\ W &= n(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1^0)'[\mathbf{I}_{11}(\hat{\boldsymbol{\theta}}) - \mathbf{I}_{12}(\hat{\boldsymbol{\theta}})\mathbf{I}_{22}^{-1}(\hat{\boldsymbol{\theta}})\mathbf{I}_{21}(\hat{\boldsymbol{\theta}})](\hat{\boldsymbol{\theta}_1} - \boldsymbol{\theta}_1^0) \\ LM &= \frac{1}{n}\boldsymbol{\mu}(\tilde{\boldsymbol{\theta}})'[\mathbf{I}_{11}(\tilde{\boldsymbol{\theta}}) - \mathbf{I}_{12}(\tilde{\boldsymbol{\theta}})\mathbf{I}_{22}^{-1}(\tilde{\boldsymbol{\theta}})\mathbf{I}_{21}(\tilde{\boldsymbol{\theta}})]^{-1}\boldsymbol{\mu}(\tilde{\boldsymbol{\theta}}), \end{aligned}$$

where

$$\hat{\boldsymbol{\theta}} \equiv (\hat{\boldsymbol{\theta}}_1', \hat{\boldsymbol{\theta}}_2')', \quad \tilde{\boldsymbol{\theta}}' \equiv (\boldsymbol{\theta}_1^{0\prime}, \tilde{\boldsymbol{\theta}}_2')', \quad \mathbf{I}(\boldsymbol{\theta}) = \begin{bmatrix} \mathbf{I}_{11}(\boldsymbol{\theta}) & \mathbf{I}_{12}(\boldsymbol{\theta}) \\ \mathbf{I}_{21}(\boldsymbol{\theta}) & \mathbf{I}_{22}(\boldsymbol{\theta}) \end{bmatrix},$$

and $\tilde{\boldsymbol{\theta}}_2$ is the solution to

$$\left.\frac{\partial \ln(\boldsymbol{\theta}; \mathbf{x})}{\partial \boldsymbol{\theta}_2}\right|_{\boldsymbol{\theta_1} = \boldsymbol{\theta}_1^0} = \mathbf{0}, \quad \textit{and} \quad \boldsymbol{\mu}(\tilde{\boldsymbol{\theta}}) = \left.\frac{\partial \ln(\boldsymbol{\theta}; \mathbf{x})}{\partial \boldsymbol{\theta}_1}\right|_{\boldsymbol{\theta_1} = \boldsymbol{\theta}_1^0}.$$

Proof.

Since

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \sim N\ (\mathbf{0},\ \mathbf{I}^{-1}(\boldsymbol{\theta})),$$

under $H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_1^0$, and the fact that $\mathbf{I}(\hat{\boldsymbol{\theta}}) - \mathbf{I}(\hat{\boldsymbol{\theta}}) \xrightarrow{p} \mathbf{0}$, we have the results from partitioned inverse rule[7]

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1^0) \sim N(\mathbf{0},\ [\mathbf{I}_{11}(\hat{\boldsymbol{\theta}}) - \mathbf{I}_{12}(\hat{\boldsymbol{\theta}})\mathbf{I}_{22}^{-1}(\hat{\boldsymbol{\theta}})\mathbf{I}_{21}(\hat{\boldsymbol{\theta}})]^{-1})$$

The Wald test statistics therefore is

$$W = (\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1^0)'[\mathbf{I}_{11}(\hat{\boldsymbol{\theta}}) - \mathbf{I}_{12}(\hat{\boldsymbol{\theta}})\mathbf{I}_{22}^{-1}(\hat{\boldsymbol{\theta}})\mathbf{I}_{21}(\hat{\boldsymbol{\theta}})](\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1^0).$$

By definition the LM test statistics is

$$LM = \frac{1}{n}\left(\frac{\partial \log L(\tilde{\boldsymbol{\theta}}; \mathbf{x})}{\partial \boldsymbol{\theta}}\right)' \mathbf{I}^{-1}(\tilde{\boldsymbol{\theta}}) \left(\frac{\partial \log L(\tilde{\boldsymbol{\theta}}; \mathbf{x})}{\partial \boldsymbol{\theta}}\right),$$

[7] Reminder: For a general 2×2 partitioned matrix, $\begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{F}_1 & -\mathbf{F}_1\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ -\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{F}_1 & \mathbf{A}_{22}^{-1}(\mathbf{I} + \mathbf{A}_{21}\mathbf{F}_1\mathbf{A}_{12}\mathbf{A}_{22}^{-1}) \end{bmatrix}$, where $\mathbf{F}_1 = (\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})^{-1}$.

but

$$\left(\frac{\partial \log L(\tilde{\boldsymbol{\theta}};\mathbf{x})}{\partial \boldsymbol{\theta}}\right) = \begin{pmatrix} \frac{\partial \log L}{\partial \boldsymbol{\theta}_1}\Big|_{\boldsymbol{\theta}_1=\boldsymbol{\theta}_1^0} \\ \frac{\partial \log L}{\partial \boldsymbol{\theta}_2}\Big|_{\boldsymbol{\theta}_2=\tilde{\boldsymbol{\theta}}_2} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu}(\tilde{\boldsymbol{\theta}}) \\ \mathbf{0} \end{pmatrix},$$

Using partitioned inverse rule we have

$$LM = \boldsymbol{\mu}(\tilde{\boldsymbol{\theta}})'[\mathbf{I}_{11}(\tilde{\boldsymbol{\theta}}) - \mathbf{I}_{12}(\tilde{\boldsymbol{\theta}})\mathbf{I}_{22}^{-1}(\tilde{\boldsymbol{\theta}})\mathbf{I}_{21}(\tilde{\boldsymbol{\theta}})]^{-1}\boldsymbol{\mu}(\tilde{\boldsymbol{\theta}}).$$

The proof of LR test statistics is straightforward. ■

Exercise 1.

Reproduce the results of Table 17.1 on p.490 of Greene 5th edition. ■

5 Monte Carlo Simulation in Statistics

Monte Carlo methods (or Monte Carlo experiments) are a broad class of computational algorithms that rely on repeated random sampling to obtain numerical results. In applied statistics, Monte Carlo methods are generally used for two purposes:

(a). To provide implementations of hypothesis tests such as when the critical values for the asymptotic distribution are often impossible or hard to compute.

(b). To compare the finite small properties for competing test statistics under realistic data conditions. Although Type I error and power properties of statistics can be calculated for data drawn from classical theoretical distributions for asymptotic conditions, real finite sample data often do not have such distributions.

5.1 With Data Generated from Pseudo-Random Number

5.1.1 Simulation of Critical Values

Nowadays, modern Econometrics-Package (such as *Gauss*) has built-in Pseudo-Random Number generator. The critical value of the statistics we are interested in (always the test statistics $\tau(\mathbf{x})$) can be easily obtained with precision by the Monte-Carlo simulation rather by the integration of the probability density.

Example.

The distribution of a standard Cauchy random variable Y, that is the ratio of two independent standard normal variables and has the probability density function

$$f(y;0,1) = \frac{1}{\pi(1+y^2)}.$$

If under certain null hypothesis that the test statistics $\tau(\mathbf{x},\theta_0) \sim Y$. Under one right one-side alternative hypothesis, we reject the null hypothesis when the realization of the sample is such that $\tau(\boldsymbol{x}) > F_{Y_{0.95}}(y)$. We can obtain the "critical value a^*" either by theoretical integration:

$$\int_{\infty}^{a^*} \frac{1}{\pi(1+y^2)} dy = 0.95,$$

or simply by the 95% quantile of (sorting) repeated independent draws from

$$\mathcal{Y}^{(i)} = \frac{\mathcal{Z}_1^{(i)}}{\mathcal{Z}_2^{(i)}},\ i = 1, ..., m,$$

where $\mathcal{Z}_1$ and $\mathcal{Z}_2$ are independent standard normal distributions. The larger the m, the more precise the Monte-Carlo experiment. ■

Example.

In many applications, we may need a non-random (univariate) normal sample of size T. We can view this sample as a one draw from a multivariate normal distribution, $\mathbf{x} \sim N(\boldsymbol{\mu},\ \boldsymbol{\Sigma})$, where the $\boldsymbol{\Sigma}$ is to take into account of serial correlations. It can be generated as follow. We begin with a draw, $\mathbf{z}$, from the T-variate standard normal distribution just by stacking T independent draws from the univariate standard normal distribution. Let $\mathbf{P}$ be the square root of $\boldsymbol{\Sigma}$ such that $\mathbf{PP}' = \boldsymbol{\Sigma}$. The desired draw is then just $\mathbf{x} = \boldsymbol{\mu} + \mathbf{Pz}$. ■

5.1.2 Finite Sample Size Distortion and Power Performance

The following example is taken from Lee *et al.* (2014) to illustrate use of Monte-Carlo simulation in comparisons of finite sample size distortion and power performance of a test statistic.

> "Lee at al. (2012) used multivariate Beveridge-Nelson decomposition to show that these two cointegrating relations $\boldsymbol{\alpha}_x'\mathbf{x}_t \sim I(0)$ and $\boldsymbol{\alpha}_y'\mathbf{y}_t \sim I(0)$ can be represented by
>
> $$\begin{aligned} S_{TW,t} - \beta_x S_{1t} - \mu_x - \delta_x t &= \nu_t \sim I(0), \\ S_{TW,t} - \beta_y S_{2t} - \mu_y - \delta_y t &= \varsigma_t \sim I(0), \end{aligned} \tag{5-2}$$
>
> where μ_x, δ_x, μ_y, and δ_y are constants and ν_t and ς_t are de-meaned and de-trended stationary cointegrating equilibrium errors.
>
> Given that $\mathbf{x}_t$ and $\mathbf{y}_t$ are conintegrated, we are interested in discriminating the closer relationship from the cointegrating relationships. To this end, we compare the magnitude of the variances of cointegrating equilibrium errors by developing the equal variance test of the null hypothesis,
>
> $$H_0 : \sigma_\nu^2 = \sigma_\varsigma^2,\ \ v.s.\ \ H_1 : \sigma_\nu^2 \gtrless \sigma_\varsigma^2,$$

in which a smaller variance of cointegrating equilibrium error is associated with a closer linkage between the variables in the cointegrating regression model. Hence, the rejection of the null hypothesis can discriminate the closer relationship from the cointegrating relationships.

To derive the asymptotic distributions of the proposed equal variance test, we need the following assumptions on the cointegrating equilibrium errors and their squared processes.

Assumption 1: The cointegrating equilibrium error process $\mathbf{u}_t = (\nu_t, \varsigma_t)'$ is stationary and ergodic with $E(\nu_t, \varsigma_t)' = (0,0)'$, $E(\nu_t^2, \varsigma_t^2)' = (\sigma_\nu^2, \sigma_\varsigma^2)'$.

Assumption 2: The squared cointegrating equilibrium error process $\mathbf{z}_t = (\nu_t^2 - \sigma_\nu^2, \varsigma_t^2 - \sigma_\varsigma^2)'$ is L_2-NED of size$-\frac{1}{2}$ on a process $\{e_t\}_{t=-\infty}^{\infty}$, where e_s is an α-mixing of size $-r/(r-2)$ for $r > 2$ or ϕ-mixing of size $-r/(2r-2)$ for $r \geq 2$. In addition, $\sup_t \|\mathbf{z}_t\|_r < \infty$.

Theorem 3: Under the null hypothesis of equal variance, $H_0 : \sigma_\nu^2 = \sigma_\varsigma^2$, the test statistic $\mathcal{Z} = \widehat{\mathcal{M}}^{-1} T^{1/2} \left(\hat{\sigma}_\nu^2 - \hat{\sigma}_\varsigma^2\right)$ is asymptotically distributed as functionals of Brownian motion:

$$\mathcal{Z} = \widehat{\mathcal{M}}^{-1} T^{1/2} \left(\hat{\sigma}_\nu^2 - \hat{\sigma}_\varsigma^2\right) \xrightarrow{d} \frac{W(1)}{\left\{\int_0^1 [W(r) - rW(1)]^2 dr\right\}^{1/2}}.$$

The cointegrating equilibrium errors processes must be stationary and the aim of the proposed variance test is to determine the degree of cointegration by comparing the variances of the cointegrating equilibrium errors. Here for simplicity, we restrict our experiment to examine the finite sample performance of the new variance test and the test of Lee et al. (2012) by assuming that the cointegrating equilibrium errors are known in the simulation. In fact, the cointegrating equilibrium errors are unknown and must be estimated. However, the simulations in this setup are helpful and insightful to understand the performance of the variance tests. We assume that the data generating processes (DGP) of cointegrating equilibrium errors follow a stationary $AR(1)$ process with an exogenous common factor:

$$\nu_t = a f_t + b\nu_{t-1} + \varepsilon_{1t}, \tag{5-3}$$

$$\varsigma_t = c f_t + d\varsigma_{t-1} + \varepsilon_{2t}, \tag{5-4}$$

and we employ the variance tests to the hypothesis,

$$H_0 : \sigma_\nu^2 = \sigma_\varsigma^2, \text{ against } H_1 : \sigma_\nu^2 < \sigma_\varsigma^2,$$

where f_t, ε_{1t} and ε_{2t} are mutually independent and assumed to be $i.i.d.N(0,1)$. f_t is the common effect of the process ν_t and ς_t. To compare the finite sample properties of the variance tests, we focus on two scenarios: the cross-sectionally independent and the cross-sectionally dependent case. In this setup, a and c in Eqs. (5-3) and Eq. (5-4) are used to control the cross-sectionally dependence, and if $a = c = 0$, the DGP degenerates to the cross-sectionally independent, but serially correlated case. b and d in Eqs. (5-3) and (5-4) are used to describe the serial correlation, where $b, d \in [0, 1)$. It is also worthy to note that, for other sample size, the finite sample properties of the tests are generally similar although they are not reported here to save space. The finite sample properties for $T = 100$ are reported in Table 7.

Table 7: Size and power of the variance tests

	Cross-sectional independent case: size						Cross-sectional independent case: power					
(b,d)	(0.1,0.1)	(0.5,0.5)	(0.75,0.75)	(0.9,0.9)	(0.95,0.95)	(0.98,98)	(0.1,0.2)	(0.1,0.4)	(0.1,0.5)	(0.1,0.6)	(0.1,0.7)	(0.1,0.9)
Lee	0.000	0.004	0.018	0.079	0.137	0.229	0.001	0.007	0.037	0.109	0.329	0.955
New	0.049	0.056	0.051	0.064	0.059	0.096	0.076	0.152	0.246	0.361	0.525	0.708
	Cross-sectional dependent case: size											
(a,b)	(1,0.1)	(1,0.75)	(1,0.9)	(1,0.95)	(1,0.975)	(1,0.99)	(5,0.1)	(5,0.75)	(5,0.9)	(5,0.95)	(5,0.975)	(5,0.99)
(c,d)	(1,0.1)	(1,0.75)	(1,0.9)	(1,0.95)	(1,0.975)	(1,0.99)	(5,0.1)	(5,0.75)	(5,0.9)	(5,0.95)	(5,0.975)	(5,0.99)
Lee	0.000	0.012	0.048	0.101	0.167	0.210	0.000	0.000	0.001	0.003	0.008	0.010
New	0.047	0.048	0.056	0.084	0.098	0.118	0.055	0.051	0.048	0.078	0.085	0.096
	Cross-sectional dependent case: power											
(a,b)	(1,0.1)	(1,0.1)	(1,0.1)	(1,0.1)	(1,0.1)	(1,0.1)	(5,0.1)	(5,0.1)	(5,0.1)	(10,0.1)	(10,0.1)	(10,0.1)
(c,d)	(1,0.2)	(1,0.4)	(1,0.6)	(1,0.7)	(1,0.8)	(1,0.98)	(5,0.3)	(5,0.5)	(5,0.7)	(10,0.3)	(10,0.5)	(10,0.7)
Lee	0.000	0.002	0.088	0.343	0.951	0.995	0.000	0.000	0.261	0.000	0.000	0.259
New	0.061	0.161	0.398	0.581	0.681	0.701	0.237	0.561	0.766	0.306	0.646	0.738

Notes: The tests were one-sided with the nominal size set at 5%, and were conducted for sample size $T = 100$ using 1000 replications. "Lee" denotes the test of Lee et al. (2012), and "New" denotes the variance test proposed in this paper.

The simulations indicate that: (a) the size control of the variance test of Lee et al. (2012) depends strongly on the serial correlation, thus a strong serial correlation can induce a spurious rejection of the null, while the proposed variance test has relatively good size; (b) a violation of the cross-sectionally

uncorrelated squared cointegrating errors can invalidate the variance test of Lee et al. (2012), especially in the cases where the squared cointegrating errors are strongly cross-sectionally dependent. However, the proposed variance test can be used to achieve satisfactory performance."

■

5.2 With Data Generated from the Real Data, Bootstrapping

When we are doing quantitative research such as testing and estimation, it is required to have information about the population distribution. However, if the usual assumptions of the population cannot be made (normality assumptions, or small samples), then the traditional approach does not work The bootstrap technique developed by Efron (1979) is to estimates the population distribution by using the information based on a number of resamples from the sample. The basic idea of bootstrapping is that inference about a population from sample data (sample → population) can be modeled by resampling the sample data and performing inference on (resample → sample). As the population is unknown, the true error in a sample statistic against its population value is unknowable. In bootstrap-resamples, the "population" is in fact the sample, and this is known; hence the quality of inference from resample data → "true" sample is measurable

5.2.1 Bootstrap Sample

For a univariate random variable, suppose a random sample of size n, $\mathbf{x} = (X_1, X_2, ..., X_n)'$ is observed, $\boldsymbol{x} = (x_1, x_2, ..., x_n)'$ from which we compute a statistic (an estimate) of interest $\hat{\theta}(\boldsymbol{x})$ traditionally.

Definition. (Bootstrap Sample)
A *bootstrap sample* $\boldsymbol{x}^* = (x_1^*, x_2^*, ..., x_m^*)$ is obtained by randomly sampling m times, with replacement, from the original observed data point $x_1, x_2, ..., x_n$. (The bootstrap sample size, m, may be larger or smaller than n.) For instance, with $n = 7, m = 8$

we might obtain $\boldsymbol{x}^* = (x_5, x_7, x_5, x_4, x_1, x_2, x_1, x_3)$. The bootstrap algorithm begins by generating a large number (i.e. B times) of independent bootstrap samples $\boldsymbol{x}^*_{(1)}$, $\boldsymbol{x}^*_{(2)}$,...,$\boldsymbol{x}^*_{(B)}$. ■

After a total of B times, the desired sampling characteristic the statistic $\hat{\theta}(\mathbf{x})$ can be computed from the sampling distribution of $\hat{\theta}(\boldsymbol{x}_b)$, (w.r.t. all possible bootstrap samples $\hat{\theta}(\boldsymbol{x}^*_{(1)})$,$\hat{\theta}(\boldsymbol{x}^*_{(2)})$,...,$\hat{\theta}(\boldsymbol{x}^*_{(B)})$. Bickel and Freedman (1981) and Singh (1981) provided large sample answers for most of the commonly used statistics. In limit, as ($n \to \infty$), the sampling distribution of $\hat{\theta}(\boldsymbol{x}^*)$ is also bell shaped with $\hat{\theta}(\mathbf{x})$. Thus, bootstrap distribution (after suitable standardization) of $(\hat{\theta}(\boldsymbol{x}^*) - \hat{\theta}(\boldsymbol{x}))$ approximates (fairly well) the sampling distribution of $(\hat{\theta}(\mathbf{x}) - \theta)$.

5.2.2 Method of Bootstrapping

(a). Independent Data:
One can resample an observation from the discrete uniform distribution $[1, 2, ..., n]$ by multiplying n by a draw from the continuous $U[0, 1]$. distribution.

(b). Dependent Data: Block bootstrap methods.
It divides the quantities that are being resampled, $[\underbrace{x_1, x_2, x_3, x_4}_{1st\ block}, \underbrace{x_5, x_6, x_7, x_8}_{2nd\ block}, \cdots, x_n]$ into blocks of b consecutive observations.[8] We then resample the blocks. Blocks may be either overlapping or nonoverlapping; overlapping seems to be better.

5.2.3 Primary Applications of Bootstrap

(a). Approximating standard error of a sample estimate.
If $\hat{\theta}(\mathbf{x})$ is consistent, then one might approximate the variance of the estimator $\hat{\theta}(\mathbf{x})$ by using

$$\text{Est } Var\left(\hat{\theta}(\mathbf{x})\right) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}(\boldsymbol{x}^*_{(b)}) - \hat{\theta}(\boldsymbol{x})\right)^2. \tag{5-5}$$

[8] This is an example of nonoverlapping blocks.

(b). Bootstrap percentile method.

Suppose one settles for $1,000$ bootstrap replications of $\hat{\theta}(\boldsymbol{x}^*)$, denoted by $\hat{\theta}(\boldsymbol{x}^*_{(1)})$, $\hat{\theta}(\boldsymbol{x}^*_{(2)})$,..., $\hat{\theta}(\boldsymbol{x}^*_{(1000)})$. After ranking from bottom to top, let us denote these bootstrap values as $\hat{\theta}(\boldsymbol{x}^*_{\underline{(1)}})$, $\hat{\theta}(\boldsymbol{x}^*_{\underline{(2)}})$,..., $\hat{\theta}(\boldsymbol{x}^*_{\underline{(1000)}})$. Then the bootstrap percentile confidence interval for $\left(\hat{\theta}(\mathbf{x})\right)$ at 95% level of confidence would be $\left[\hat{\theta}(\boldsymbol{x}^*_{\underline{(25)}}),\ \hat{\theta}(\boldsymbol{x}^*_{\underline{(975)}})\right]$. Turning to the theoretical aspects of this method, it should be pointed out that the method requires the symmetry of the sampling distribution of $\hat{\theta}(\mathbf{x})$ around θ.

(c). Bootstrap-t Methods.

Suppose we form a standard t-statistics to test the null hypothesis $H_0 : \theta = theta_0$ from $t_{\hat{\theta}} = \frac{\hat{\theta}(\mathbf{x})-\theta}{SE}$, where SE is a sample estimate of the standard error of $\hat{\theta}(\mathbf{x})$. The distribution of $t_{\hat{\theta}}$ depends on the probability model and sample assumptions on $\mathbf{x}$. This distribution may sometimes hard to derive. However, we can approximate the distribution of $t_{\hat{\theta}}$ by bootstrap. The bootstrap counterpart of such a function t is $t_{\hat{\theta}_b}$,

$$t_{\hat{\theta}_b} = \frac{\hat{\theta}(\boldsymbol{x}^*_{(b)}) - \hat{\theta}(\boldsymbol{x})}{SE_b},$$

where SE_b is exactly like SE but computed on a bootstrap sample as the square root of eq. (5-5). Denote the $100s-th$ bootstrap percentile of $t_{\hat{\theta}_b}$ by $t_{\hat{\theta}_{b,s}}$ and consider the statement: $t_{\hat{\theta}}$ lies within $(t_{\hat{\theta}_{b,0.025}},\ t_{\hat{\theta}_{b,0.975}})$. After the substitution $t_{\hat{\theta}} = \frac{\hat{\theta}(\mathbf{x})-\theta}{SE}$, the above statement translates to θ lies within

$$\left(\hat{\theta} - SE \cdot t_{\hat{\theta}_{b,0.975}},\ \hat{\theta} - SE \cdot t_{\hat{\theta}_{b,0.025}}\right).$$

This range for θ is called bootstrap-t based confidence interval for θ at coverage level 95%.

Marshall-Adams Hall, MSU. (Economics Department).

End of this Chapter