Hypothesis Testing

CB: chapter 8; section 10.3

Hypothesis: statement about an unknown population parameter

Examples: The average age of males in Sweden is 27. (statement about population mean)
The lowest time it takes to run 30 miles is 2 hours. (statement about population max)
Stocks are more volatile than bonds. (statement about variances of stock and bond returns)

In hypothesis testing, you 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 being hypothesized about is θ, and the parameter space (i.e., possible values for θ) is Θ, then the null and alternative hypotheses form a partition of Θ:

H_0: $\theta \in \Theta_0 \subset \Theta$
H_1: $\theta \in \Theta_0^c$ (the complement of Θ_0 in Θ).

Examples:

1. $H_0: \ \theta = 0$ vs. $H_1: \ \theta \neq 0$
2. $H_0: \ \theta \leq 0$ vs. $H_1: \ \theta > 0$

1 Definitions of test statistics

A **test statistic**, similarly 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 **test** 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.

The subset of the real line $\mathcal{R}$ for which the test is equal to 1 is called the rejection (or "critical") region. The complement of the critical region (in the support of the test statistic) is the acceptance region.

Example: let μ denote the (unknown) mean male age in Sweden.

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

Let your test statistic be $\bar{X}_{100}$, the average age of 100 randomly-drawn Swedish males.

Just as with estimators, there are many different possible tests, for a given pair of hypotheses H_0, H_1. Consider the following four tests:

1. Test 1: $\mathbf{1}\left(\bar{X}_{100} \notin [25, 29]\right)$
2. Test 2: $\mathbf{1}\left(\bar{X}_{100} \not\leq 29\right)$
3. Test 3: $\mathbf{1}\left(\bar{X}_{100} \not> 35\right)$
4. Test 4: $\mathbf{1}\left(\bar{X}_{100} \neq 27\right)$

Which ones make the most sense?

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); (iv) $\sin X_1$ (the sine of the first observation).

In what follows, we refer to a test as a combination of both (i) a test statistic; and (ii) the mapping from realizations of the test statistic to $\{0, 1\}$.

Next we consider some common types of tests.

■■■

1.1 Likelihood Ratio Test

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

Define: the **likelihood ratio test statistic** for testing $H_0: \ \theta \in \Theta_0$ vs. $H_1: \ \theta \in \Theta_0^c$ is

$$\lambda(\vec{X}) \equiv \frac{\sup_{\theta \in \Theta_0} L(\theta|\vec{X})}{\sup_{\theta \in \Theta} L(\theta|\vec{X})}.$$

The numerator of $\lambda(\vec{X})$ 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., $\theta \in \Theta_0$), then $\lambda(\vec{X}) = 1$ (since the restriction of $\theta \in \Theta_0$ will not bind). However, if H_0 is false, then $\lambda(\vec{X})$ can be small (close to zero).

So an LR test should be one which rejects H_0 when $\lambda(\vec{X})$ is small; for example, $\mathbf{1}(\lambda(\vec{X}) < 0.75)$

$\blacksquare\blacksquare\blacksquare$

Example: $X_1, \ldots, X_n \sim \;$ *i.i.d.* $N(\theta, 1)$

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

Then

$$\lambda(\vec{X}) = \frac{\exp\left(-\frac{1}{2}\sum_i (X_i - 2)^2\right)}{\exp\left(-\frac{1}{2}\sum_i (X_i - \bar{X}_n)^2\right)}$$

(the denominator arises because $\bar{X}_n$ is the unrestricted MLE estimator for θ.

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

(i) Test $H_0 : \; \theta = 2$ vs. $H_1 : \; \theta \neq 2$.

Restricted likelihood function

$$L(\vec{X}|2) = \begin{cases} \left(\frac{1}{2}\right)^n & \text{if } \max(X_1, \ldots, X_n) \leq 2 \\ 0 & \text{if } \max(X_1, \ldots, X_n) > 2. \end{cases}$$

Unrestricted likelihood function

$$L(\vec{X}|\theta) = \begin{cases} \left(\frac{1}{\theta}\right)^n & \text{if } \max(X_1, \ldots, X_n) \leq \theta \\ 0 & \text{if } \max(X_1, \ldots, X_n) > \theta \end{cases}$$

which is maximized at $\theta_n^{MLE} = \max(X_1, \ldots, X_n)$.

Hence the denominator of the LR statistic is $\left(\frac{1}{\max(X_1,\ldots,X_n)}\right)^n$, so that

$$\lambda(\vec{X}) = \begin{cases} 0 & \text{if } \max(X_1, \ldots, X_n) > 2 \\ \left(\frac{\max(X_1,\ldots,X_n)}{2}\right)^n & \text{otherwise} \end{cases}$$

LR test would say: $\mathbf{1}(\lambda(\vec{X}) \leq c)$. Critical region consists of two disconnected parts (graph).

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

In this case: the restricted likelihood is

$$\sup_{\theta \in [0,2]} L(\vec{X}|\theta) = \begin{cases} \left(\frac{1}{\max(X_1,\ldots,X_n)}\right)^n & \text{if } \max(X_1,\ldots,X_n) \leq 2 \\ 0 & \text{otherwise.} \end{cases}$$

so

$$\lambda(\vec{X}) = \begin{cases} 1 & \text{if } \max(X_1,\ldots,X_n) \leq 2 \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

(graph)

So now LR test is $\mathbf{1}(\lambda(\vec{X}) \leq c) = \mathbf{1}(\max(X_1,\ldots,X_n) > 2)$.

■■■

1.2 Wald Tests

Another common way to generate test statistics is to focus on statistics which are asymptotically normal distributed, under H_0 (i.e., if H_0 were true).

A common situation is when the estimator for θ, call it $\hat{\theta}_n$, is asymptotically normal, with some asymptotic variance V (eg. MLE). Let the null be $H_0: \ \theta = \theta_0$. Then, if the null were true:

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

The quantity on the LHS is the **T-test statistic**.

Note: 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 $\hat{V} \xrightarrow{p} V$, then the statement

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

still holds (using the plim operator and Slutsky theorems). In what follows, therefore, we assume for simplicity that we know V.

We consider two cases:

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

Under H_0: the CLT holds, and the t-stat is $N(0,1)$
Under H_1: assume that the true value is some $\theta_1 \neq \theta_0$. Then the t-stat can be written as

$$\frac{(\hat{\theta}_n - \theta_0)}{\sqrt{\frac{1}{n}V}} = \frac{(\hat{\theta}_n - \theta_1)}{\sqrt{\frac{1}{n}V}} + \frac{(\theta_1 - \theta_0)}{\sqrt{\frac{1}{n}V}}.$$

The first term $\xrightarrow{d} N(0,1)$, but the second (non-stochastic) term diverges to ∞ or $-\infty$, depending on whether the true θ_1 exceeds or is less than θ_0. Hence the t-stat diverges to $-\infty$ or ∞ with probability 1.

Hence, in this case, your test should be $\mathbf{1}(|Z_n| > c)$, where c should be some number in the tails of the $N(0,1)$ distribution.

Multivariate version: θ is K-dimensional, and asymptotic normal, so that under H_0, we have

$$\sqrt{n}(\theta_n - \theta_0) \xrightarrow{d} N(0, \Sigma).$$

Then we can test $H_0: \ \theta = \theta_0$ vs. $H_1: \ \theta \neq \theta_0$ using the quadratic form

$$Z_n \equiv n \cdot (\theta_n - \theta_0)' \Sigma^{-1} (\theta_n - \theta_0) \xrightarrow{d} \chi^2_k.$$

Test takes the form: $\mathbf{1}(Z_n > c)$.

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

Here the null hypothesis specifies a whole range of true θ ($\Theta_0 = (-\infty, \theta_0]$), whereas the t-test statistic is evaluated at just one value of θ.

Just as for the two-sided test, the one-sided t-stat is evaluated at θ_0, so that $Z_n = \frac{\hat{\theta}_n - \theta_0}{\sqrt{\frac{1}{n}V}}$.

Under H_0 and $\theta < \theta_0$: Z_n diverges to $-\infty$ with probability 1. Under H_0 and $\theta = \theta_0$: the CLT holds, and the t-stat is $N(0,1)$.
Under H_1: Z_n diverges to ∞ with probability 1.

Hence, in this case, you will reject the null only for very large values of Z_n. Correspondingly, your test should be $\mathbf{1}(Z_n > c)$, where c should be some number in the right tail of the $N(0,1)$ distribution.

Later, we will discuss how to choose c.

■■■

1.3 Score test

Consider a model with log-likelihood function $\log L(\theta|\vec{X}) = \frac{1}{n}\sum_i \log f(x_i|\theta)$.

Let $H_0 : \ \theta = \theta_0$. The *sample score* function evaluated at θ_0 is

$$S(\theta_0) \equiv \frac{\partial}{\partial \theta} \log L(\theta|\vec{X})\big|_{\theta=\theta_0} = \frac{1}{n}\sum_i \frac{\partial}{\partial \theta} \log f(x_i|\theta)\big|_{\theta=\theta_0}.$$

Define $W_i = \frac{\partial}{\partial \theta} \log f(x_i|\theta)\big|_{\theta=\theta_0}$. Under the null hypothesis, $S(\theta_0)$ converges to

$$\begin{aligned} E_{\theta_0} W_i &= \int \frac{\frac{d}{d\theta} f(x|\theta)|_{\theta=\theta_0}}{f(x|\theta_0)} f(x|\theta_0) dx \\ &= \frac{\partial}{\partial \theta} \int f(x|\theta) dx \\ &= \frac{\partial}{\partial \theta} \cdot 1 = 0 \end{aligned}$$

(the information inequality). Hence,

$$V_{\theta_0} W_i = E_{\theta_0} W_i^2 = E_{\theta_0} \left(\frac{\partial}{\partial \theta} \log f(X|\theta)\big|_{\theta=\theta_0} \right)^2 \equiv V_0.$$

(Note that $\frac{1}{V_0}$ is the usual variance matrix for MLE, which is the CRLB.)

Therefore, you can apply the CLT to get that, under H_0,

$$\frac{S(\theta_0)}{\sqrt{\frac{1}{n} V_0}} \xrightarrow{d} N(0,1).$$

(If we don't know V_0, we can use some consistent estimator $\hat{V}_0$ of it.)

So a test of $H_0 : \theta = \theta_0$ could be formulated as $\mathbf{1}\left(|\frac{\frac{1}{n}S(\theta_0)}{\sqrt{\frac{1}{n}V_0}}| > c\right)$, where c is in the right tail of the $N(0,1)$ distribution.

Multivariate version: if θ is K-dimensional:

$$S_n \equiv nS(\theta_0)'V_0^{-1}S(\theta_0) \xrightarrow{d} \chi^2_k.$$

■■■

Recall that, in the previous lecture notes, we derived the following asymptotic equality for the MLE:

$$\begin{aligned}\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) &\stackrel{a}{=} \sqrt{n}\frac{-\frac{1}{n}\sum_i \frac{\partial \log f(x_i|\theta)}{\partial \theta}\Big|_{\theta=\theta_0}}{\frac{1}{n}\sum_i \frac{\partial^2 \log f(x_i|\theta)}{\partial \theta^2}\Big|_{\theta=\theta_0}} \\ &= \sqrt{n}\frac{S(\theta_0)}{V_0}.\end{aligned} \tag{3}$$

(The notation $\stackrel{a}{=}$ means that the LHS and RHS differ by some quantity which is $o_p(1)$.)

Hence, the above implies that (applying the information inequality, as we did before)

$$\underbrace{\sqrt{n}\left(\hat{\theta}_n - \theta_0\right)}_{(1)} \stackrel{a}{=} \underbrace{\sqrt{n}\frac{S(\theta_0)}{V_0}}_{(2)} \xrightarrow{d} N(0, \frac{1}{V_0})$$

The Wald statistic is based on (1), while the Score statistic is based on (2). In this sense, these two tests are asymptotically equivalent. Note: the asymptotic variance of the Wald statistic (2) equal the reciprocal of V_0.

(Later, you will also see that the Likelihood Ratio Test statistic is also asymptotically equivalent to these two.)

■■■

The LR, Wald, and Score tests (the "trinity" of test statistics) require different models to be estimated.

- LR test requires both the restricted and unrestricted models to be estimated
- Wald test requires only the unrestricted model to be estimated.

- Score test requires only the restricted model to be estimated.

Applicability of each test then depends on the nature of the hypotheses. For $H_0 : \theta = \theta_0$, the restricted model is trivial to estimate, and so LR or Score test might be preferred. For $H_0 : \theta \leq \theta_0$, restricted model is a constrained maximization problem, so Wald test might be preferred.

$\blacksquare\blacksquare\blacksquare$

2 Methods of evaluating tests

Consider $X_1, \ldots, X_n \sim$ *i.i.d.* (μ, σ^2). Test statistic $\bar{X}_n$.

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

Why are the following good or bad tests?

1. $\mathbf{1}(\bar{X}_n \neq 2)$
2. $\mathbf{1}(\bar{X}_n \geq 1.2)$
3. $\mathbf{1}(\bar{X}_n \notin [1.8, 2.2])$
4. $\mathbf{1}(\bar{X}_n \notin [-10, 30])$

Test 1 "rejects too often" (in fact, for every n, you reject with probability 1). Test 2 is even worse, since it rejects even when $\bar{X}_n$ is close to 2. Test 3 is not so bad, Test 4 accepts too often.

Basically, we are worried when a test is wrong. Since the test itself is a random variable, we cannot guarantee that a test is never wrong, but we can characterize how often it would be wrong.

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

- **Type I error:** rejecting H_0 when it is true. (This is the problem with tests 1 and 2.)
- **Type II error:** Accepting H_0 when it is false. (This is the problem with test 4.)

Let $T_n \equiv T(X_1, \ldots, X_n)$ denote the sample test statistic. Consider a test with rejection region R (i.e., test is $\mathbf{1}(T_n \in R)$). Then:

$$
\begin{aligned}
P(\text{type I error}) &= P(T_n \in R \mid \theta \in \Theta_0) \\
P(\text{type II error}) &= P(T_n \notin R \mid \theta \in \Theta_0^c)
\end{aligned}
$$

■■■

Example: $X_1, X_2 \sim$ *i.i.d.* Bernoulli, with probability p.

- Test $H_0: \ p = \frac{1}{2}$ vs. $H_1: \ p \neq \frac{1}{2}$.
- Consider the test $\mathbf{1}\left(\frac{X_1+X_2}{2} \neq 1\right)$.
- Type I error: rejecting H_0 when $p = \frac{1}{2}$.
  P(Type I error) = P($\frac{X_1+X_2}{2} \neq 1 | p = \frac{1}{2}$) = P($\frac{X_1+X_2}{2} = 0 | p = \frac{1}{2}$)+ P($\frac{X_1+X_2}{2} = \frac{1}{2} | p = \frac{1}{2}$) $= \frac{1}{4} + \frac{1}{2} = \frac{3}{4}$.
- Type II error: Accepting H_0 when $p \neq \frac{1}{2}$.
  For $p \neq \frac{1}{2}$:

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

  So P($\frac{X_1+X_2}{2} = 1 | p$)=$p^2$. Graph.

■■■

Power function

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

Definition: the *power function* of a hypothesis test with a rejection region R is the function of θ defined by $\beta(\theta) = P(T_n \in R|\theta)$.

Example: For above example,

$\beta(p) = P\left(\frac{X_1+X_2}{2} \neq 1|p\right) = 1 - p^2$. Graph

- Power function gives the Type I error probabilities, for any singleton null hypothesis $H_0: \ \theta = \theta_0$.

  From $\beta(p)$, see that if you are worried about Type I error, then you should only use this test when your null is that p is close to 1 (because only for p_0 close to 1 is the power function low).

- $1 - \beta(p)$ gives you the Type II error probabilities, for any point alternative hypothesis

  So if you are worried about Type II error, then $\beta(p)$ tells you that you should use this test when your alternative hypothesis postulates that p is low (close to zero). We say that this test *has good power* against alternative values of p close to zero.

- **Important:** power function is specific to a given test $\mathbf{1}(T_n \in R)$, regardless of the specific hypotheses that the test may be used for.

■■■

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

Test $H_0: \ \theta \leq 2$ vs. $H_1: \ \theta > 2$. Derive $\beta(\theta)$ for the LR test $\mathbf{1}(\lambda(\vec{X}) < c)$.

Recall our earlier derivation of LR test in Eq. (1). Hence,

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

Graph.

■■■

In practice, researchers often concerned about Type I error (i.e., don't want to reject H_0 unless evidence overwhelming against it): "conservative bias"?

But if this is so, then you want a test with a power function $\beta(\theta)$ which is low for $\theta \in \Theta_0$, but high elsewhere:

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

- For $0 \leq \alpha \leq 1$, a test with power function $\beta(\theta)$ is a **size** α test if $\sup_{\theta \in \Theta_0} \beta(\theta) = \alpha$.
- For $0 \leq \alpha \leq 1$, a test with power function $\beta(\theta)$ is a **level** α test if $\sup_{\theta \in \Theta_0} \beta(\theta) \leq \alpha$.
- The $\theta \in \Theta_0$ at which the "sup" is achieved is called the "least favorable value" of θ under the null, for this test. It is the value of $\theta \in \Theta_0$ for which the null holds, but which is most difficult to distinguish (in the sense of having the highest rejection probability) from any alternative parameter $\theta \notin \Theta_0$.

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

Example: $X_1, \ldots, X_n \sim$ *i.i.d.* $N(\mu, 1)$. Then $\bar{X}_n \sim N(\mu, 1/n)$, and $Z_n(\mu) \equiv \sqrt{n}(\bar{X}_n - \mu) \sim N(0, 1)$.

- Consider the test $\mathbf{1}\left(Z_n(2) > c\right)$, for the hypotheses
  $H_0: \ \mu \leq 2$ vs. $H_1: \ \mu > 2$.

- The power function

$$\begin{aligned}\beta(\mu) &= P(\sqrt{n}(\bar{X}_n - 2) > c|\mu) \\ &= P(\sqrt{n}(\bar{X}_n - \mu) > c + \sqrt{n}(2 - \mu)|\mu) \\ &= 1 - \Phi(c + \sqrt{n}(2 - \mu))\end{aligned}$$

  where $\Phi(\cdot)$ is the standard normal CDF. Note that $\beta(\mu)$ is increasing in μ.

- Size of test= $\sup_{\mu \leq 2} 1 - \Phi(c + \sqrt{n}(2 - \mu))$. Since $\beta(\mu)$ is increasing in μ, the supremum occurs at $\mu = 2$, so that size is $\beta(2) = 1 - \Phi(c)$.

- Assume you want a test with size α. Then you want to set c such that

$$1 - \Phi(c) = \alpha \Leftrightarrow c = \Phi^{-1}(1 - \alpha).$$

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

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

■■■

Now consider the above test, with $c^* = 1.96$, but change the hypotheses to $H_0: \ \mu = 2$ vs. $H_1: \ \mu \neq 2$.

Test still has size $\alpha = 0.05$.

But there is something intuitively wrong about this test. You are less likely to reject when $\mu < 2$. So the Type II error is very high for the alternatives $\mu < 2$.

We wish to rule out such tests

Definition: a test with power function $\beta(\theta)$ is **unbiased** if $\beta(\theta') \geq \beta(\theta'')$ for every pair (θ', θ'') where $\theta' \in \Theta_0^c$ and $\theta'' \in \Theta_0$.

Clearly, the test above is biased for the stated hypotheses. What would be an unbiased test, with the same size $\alpha = 0.05$?

Definition: Let C be a class of tests for testing $H_0: \ \theta \in \Theta_0$ vs. $H_1: \ \theta \in \Theta_0^c$. A test in class C, with power function $\beta(\theta)$, is **uniformly most powerful (UMP)** in class C if, for every other test in class C with power function $\tilde{\beta}(\theta)$

$$\beta(\theta) \geq \tilde{\beta}(\theta), \text{ for every } \theta \in \Theta_0^c.$$

Often, the classes you consider are test of a given size α. Graphically, the power function for a UMP test lies above the upper envelope of power functions for all other tests in the class, for $\theta \in \Theta_0^c$:

(graph)

2.1 UMP tests in special case: two simple hypotheses

It can be difficult in general to see what form an UMP test for any given size α is. But in the simple case where both the null and alternative hypotheses are simple hypotheses, we can appeal to the following result.

Theorem 8.3.12 (Neyman-Pearson Lemma): Consider testing

$$H_0: \ \theta = \theta_0 \text{ vs. } H_1: \ \theta = \theta_1$$

(so both the null and alternative are singletons). The pdf or pmf corresponding to θ_i is $f(\vec{X}|\theta_i)$, for $i = 0, 1$. The test has a rejection region R which satisfies:

$$\begin{aligned} \vec{X} \in R \qquad & \text{if } f(\vec{X}|\theta_1) > k \cdot f(\vec{X}|\theta_0) \quad (*) \\ \vec{X} \in R^c \qquad & \text{if } f(\vec{X}|\theta_1) < k \cdot f(\vec{X}|\theta_0) \end{aligned}$$

and

$$\alpha = P(\vec{X} \in R|\theta_0). \tag{4}$$

Then:

- Any test satisfying (*) is a UMP test with level α.
- If there exists a test satisfying (*) with $k > 0$, then every UMP level α test is a size α test and every UMP level α test satisfies (*).

In other words, for simple hypotheses, a likelihood ratio test of the sort

$$\mathbf{1}\left(\frac{f(\vec{X}|\theta_1)}{f(\vec{X}|\theta_0)} > k\right)$$

is UMP-size α (where k is chosen so that $P(\frac{f(\vec{X}|\theta_1)}{f(\vec{X}|\theta_0)} > k|\theta_0) = \alpha$).

Note that this LR statistic is different than $\lambda(\vec{X})$, which for these hypotheses is $\frac{f(\vec{X}|\theta_0)}{max\left[f(\vec{X}|\theta_0), f(\vec{X}|\theta_1)\right]}$.

Example: return to 2-coin toss again. Test $H_0: \; p = \frac{1}{2}$ vs. $H_1: \; p = \frac{3}{4}$.

The likelihood ratios for the three possible outcomes $\sum_i X_i = 0, 1, 2$ are:

$$\begin{aligned} \frac{f(0|p=\frac{3}{4})}{f(0|p=\frac{1}{2})} &= \frac{1}{4} \\ \frac{f(1|p=\frac{3}{4})}{f(1|p=\frac{1}{2})} &= \frac{3}{4} \\ \frac{f(2|p=\frac{3}{4})}{f(2|p=\frac{1}{2})} &= \frac{9}{4}. \end{aligned}$$

Let $l(\vec{X}) = \frac{f(\sum_i X_i|p=\frac{3}{4})}{f(\sum_i X_i|p=\frac{1}{2})}$. Hence, there are 4 possible rejection regions, for values of $l(\vec{X})$:

(i) $(\frac{9}{4}, +\infty)$ (size $\alpha = 0$)
(ii) $(\frac{3}{4}, +\infty)$ (size $\alpha = \frac{1}{4}$)
(iii) $(\frac{1}{4}, +\infty)$ (size $\alpha = \frac{3}{4}$)
(iv) $(-\infty, +\infty)$ (size $\alpha = 1$).

At all other values for α, no value of k satisfies (*) exactly, and so the test above are level-α tests. This is due to the discreteness of this example.

Example A: $X_1, \ldots, X_n \sim N(\mu, 1)$. Test $H_0 : \mu = \mu_0$ vs. $H_1 : \mu = \mu_1$. (Assume $\mu_1 > \mu_0$.)

By the NP-Lemma, the UMP test has rejection region characterized by

$$\frac{\exp\left(-\frac{1}{2}\sum_i (X_i - \mu_1)^2\right)}{\exp\left(-\frac{1}{2}\sum_i (X_i - \mu_0)^2\right)} > k.$$

Taking logs of both sides:

$$\begin{aligned}
(\mu_1 - \mu_0)\sum_i X_i &> \log k + \frac{n}{2}(\mu_1 - \mu_0)^2 \\
\Leftrightarrow \frac{1}{n}\sum_i X_i &> \frac{\frac{1}{n}\log k + \frac{1}{2}(\mu_1 - \mu_0)^2}{(\mu_1 - \mu_0)} \equiv d.
\end{aligned}$$

where d is determined such that $P(\bar{X}_n > d) = \alpha$, where α is the desired size. This makes intuitive sense: you reject the null when the sample mean is too large (because $\mu_1 > \mu_0$).

Under the null, $\sqrt{n}(\bar{X}_n - \mu_0) \sim N(0, 1)$, so for $\alpha = 0.05$, you want to set $d = \mu_0 + 1.64/\sqrt{n}$.

2.1.1 Discussion of NP Lemma

(From Amemiya pp. 190-192) Rather than present a proof of NP Lemma (you can find one in CB), let's consider some intuition for the NP test. In doing so, we will derive an interesting duality between Bayesian and classical approaches to hypothesis testing (the NP Lemma being a seminal result for the latter).

Start by considering a Bayesian approach to this testing problem. Assume that decisionmaker incurs loss γ_1 if he mistakenly chooses H_1 when H_0 is true, and γ_2 if he mistakenly chooses H_0 when H_1 is true. Then, given data observations $\vec{x}$, he will

$$\text{Reject } H_0 \text{ (=accept } H_1) \Leftrightarrow \gamma_1 P(\theta_0|\vec{x}) < \gamma_2 P(\theta_1|\vec{x}) \quad (*)$$

where $P(\theta_0|\vec{x})$ denotes the posterior probability of the null hypothesis given data $\vec{x}$. In other words, this Bayesian's rejection region R_0 is given by

$$R_0 = \left\{ \vec{x} : \frac{P(\theta_1|\vec{x})}{P(\theta_0|\vec{x})} > \frac{\gamma_1}{\gamma_2} \right\}.$$

If we multiply and divide the ratio of posterior probabilities by $f(\vec{x})$, the (marginal) joint density of $\vec{x}$, and use the laws of probability, we get:

$$\begin{aligned} R_0 &= \left\{ \vec{x} : \frac{P(\theta_1|\vec{x})f(\vec{x})}{P(\theta_0|\vec{x})f(\vec{x})} > \frac{\gamma_1}{\gamma_2} \right\} \\ &= \left\{ \vec{x} : \frac{L(\vec{x}|\theta_1)P(\theta_1)}{L(\vec{x}|\theta_0)P(\theta_0)} > \frac{\gamma_1}{\gamma_2} \right\} \\ &= \left\{ \vec{x} : \frac{L(\vec{x}|\theta_1)}{L(\vec{x}|\theta_0)} > \frac{\gamma_1 P(\theta_0)}{\gamma_2 P(\theta_1)} \equiv c \right\}. \end{aligned}$$

In the above, $P(\theta_0)$ and $P(\theta_1)$ denote the prior probabilities for, respectively, the null and alternative hypotheses. This provides some intuition for the "likelihood ratio" form of the NP rejection region.

Next we show a duality between the Bayesian and classical approaches to finding the "best" test. In the case of two simple hypotheses, a best test should be such that for a given size α, it should have the smallest type 2 error β (these are called "admissible" tests). For any testing scenario, the frontier (in (β, α) space) of tests is convex. (Why?) Both the Bayesian and the classical statistician want to choose a test that is on the frontier, but they might differ in the one they choose.

First consider the Bayesian. She wishes to employ a test, equivalently, choose a rejection region R, to minimize expected loss

$$\min_R \phi(R) \equiv \gamma_1 P(\theta_0|\vec{X} \in R)P(\vec{X} \in R) + \gamma_2 P(\theta_1|\vec{X} \in \bar{R})P(\vec{X} \in \bar{R}).$$

We will show that the region R_0 defined above optimizes this problem. Consider any other region R_1. Recall that $R_0 = (R_0 \cap R_1) \cup (R_0 \cap \bar{R}_1)$. Then we can rewrite

$$\begin{aligned}\phi(R_0) =& \gamma_1 P(\theta_0|R_0 \cap R_1)P(R_0 \cap R_1) + \gamma_1 P(\theta_0|R_0 \cap \bar{R}_1)P(R_0 \cap \bar{R}_1)\\ &+ \gamma_2 P(\theta_1|\bar{R}_0 \cap R_1)P(\bar{R}_0 \cap R_1) + \gamma_2 P(\theta_1|\bar{R}_0 \cap \bar{R}_1)P(\bar{R}_0 \cap \bar{R}_1).\end{aligned}$$

$$\begin{aligned}\phi(R_1) =& \gamma_1 P(\theta_0|R_1 \cap R_0)P(R_1 \cap R_0) + \gamma_1 P(\theta_0|R_1 \cap \bar{R}_0)P(R_1 \cap \bar{R}_0)\\ &+ \gamma_2 P(\theta_1|\bar{R}_1 \cap R_0)P(\bar{R}_1 \cap R_0) + \gamma_2 P(\theta_1|\bar{R}_1 \cap \bar{R}_0)P(\bar{R}_1 \cap \bar{R}_0).\end{aligned}$$

First and fourth terms of equations above are identical. From the definition of R_0 above, we know that $\phi(R_0)_{ii} < \phi(R_1)_{iii}$. That is, for all $\vec{x} \in R_0 \supseteq R_0 \cap \bar{R}_1$, we know from (*) that $\gamma_1 P(\theta_0|\vec{x}) < \gamma_2 P(\theta_1|\vec{x})$. Similarly, we know that $\phi(R_0)_{iii} < \phi(R_1)_{ii}$, implying that $\phi(R_0) < \phi(R_1)$.

Moreover, using the laws of probability, we can re-express

$$\phi(R) = \gamma_1 P(\theta_0)P(R|\theta_0) + \gamma_2 P(\theta_1)P(\bar{R}|\theta_1) \equiv \eta_0 \alpha_R + \eta_1 \beta_R$$

with $\eta_0 = \gamma_1 P(\theta_0)$ and $\eta_1 = \gamma_2 P(\theta_1)$. In (β, α) space, the "iso-loss" curves are parallel lines with slope $-\eta_0/\eta_1$, decreasing towards the origin. Hence, the optimal test (with region R_0) lies at tangency of (α, β) frontier with a line with slope $-\eta_0/\eta_1$. Hence, the optimal Bayesian test (R_0).

Now consider the classical statistician. He doesn't want to specify priors $P(\theta_0)$, $P(\theta_1)$, so η_0, η_1 are not fixed. But he is willing to specify a desired size α^*. Hence the optimal test is the one with a rejection region characterized by the slope of the line tangent at α^* in the (β, α) space. This is in fact the NP test. From the above calculations, we know that this slope is equivalent to $-c$. That is, the NP test corresponds to an optimal Bayesian test with $\eta_0/\eta_1 = c$.

2.2 Extension of NP Lemma: models with monotone likelihood ratio

The case covered by NP Lemma– that both null and alternative hypotheses are simple – is somewhat artificial. For instance, we may be interested in the *one-sided* hypotheses $H_0:\ \theta \leq \theta_0$ vs. $H_1:\ \theta \geq \theta_0$, where θ is scalar. It turns out UMP for one-sided hypotheses exists under an additional assumption on the family of density functions $\left\{f(\vec{X}|\theta)\right\}_\theta$:

Definition: the family of densities $f(\vec{X}|\theta)$ has *monotone likelihood ratio in* $T(\vec{X})$ is there exists a function $T(\vec{X})$ such that for any pair $\theta < \theta'$ the densities $f(\vec{X}|\theta)$

and $f(\vec{X}|\theta')$ are distinct and the ratio $f(\vec{X}|\theta')/f(\vec{X}|\theta)$ is a nondecreasing function of $T(\vec{X})$. That is, there exists a nondecreasing function $g(\cdot)$ and function $T(\vec{X})$ such that $f(\vec{X}|\theta')/f(\vec{X}|\theta) = g(T(\vec{X}))$.

For instance: let $\vec{X} = x$ and $T(x) = x$, then MLR in x means that $f(x|\theta')/f(x|\theta)$ is nondecreasing in x. Roughly speaking: larger x's are more "likely" under larger θ's.

Theorem: Let θ be scalar, and let $f(\vec{X}|\theta)$ have MLR in $T(\vec{X})$. Then:

(i) For testing $H_0: \theta \leq \theta_0$ vs. $H_1: \theta \geq \theta_0$, there exists a UMP test $\phi(\vec{X})$, which is given by

$$\phi(\vec{X}) = \begin{cases} 1 & \text{when } T(\vec{X}) > C \\ \gamma & \text{when } T(\vec{X}) = C \\ 0 & \text{when } T(\vec{X}) < C \end{cases}$$

where (γ, C) are chosen to satisfy $E_{\theta_0}\phi(\vec{X}) = \alpha$.

(ii) The power function $\beta(\theta) = E_\theta\phi(\vec{X})$ is strictly increasing for all points θ for which $0 < \beta(\theta) < 1$.

(iii) For all θ', the test described in part i is UMP for testing $H_0': \theta \leq \theta'$ vs. $H_1': \theta \geq \theta'$ at level $\alpha = \beta(\theta')$. ■

Sketch of Proof:[1] Consider testing θ_0 vs. $\theta_1 > \theta_0$. By NP Lemma, this depends on ratio $f(\vec{X}|\theta_1)/f(\vec{X}|\theta_0)$. Given MLR condition, this ratio can be written $g(T(\vec{X}))$ where $g(\cdot)$ is nondecreasing. Then UMP test rejects when $f(\vec{X}|\theta_1)/f(\vec{X}|\theta_0)$ is large, which is when $T(X)$ is large; this is test $\phi(\vec{X})$. Since this test does not depend on θ_1, $\phi(\vec{X})$ is also UMP-size α for testing $H_0: \theta_0$ vs. $H_1: \theta_1 > \theta_0$ (composite alternative).

Now since MLR holds for all $(\theta', \theta''; \theta'' > \theta')$, the test $\phi(\vec{X})$ is also UMP-size $E_{\theta'}\phi(\vec{X})$ for testing θ' vs. θ''. Hence $\beta(\theta'') \geq \beta(\theta')$, otherwise $\phi(\vec{X})$ cannot be UMP (why?[2]). Furthermore, the distinctiveness of $f(\vec{X}|\theta'')$ and $f(\vec{X}|\theta')$ rules out $\beta(\theta'') = \beta(\theta')$. Hence we get (ii).

Then since the power function is monotonic increasing in θ, the UMP-size α feature of $\phi(\vec{X})$ for testing θ_0 vs. $H_1: \theta > \theta_0$ extends to the composite null hypothesis $H_0: \theta \leq \theta_0$ vs. $H_1: \theta > \theta_0$, which is (i). (iii) follows immediately.

■■■

Example A cont'd: From above, we see that for $\mu_1 > \mu_0$, the likelihood ratio simplifies to $\exp\left((\mu_1 - \mu_0)\sum_i X_i - \frac{n}{2}(\mu_1^2 - \mu_0^2)\right)$ which is increasing in $\bar{X}_n$. Hence,

[1] Lehmann and Romano, Testing Statistical Hypotheses, pg. 65.

[2] Consider the purely random test $\phi(\vec{X}) = 1$ with probability α.

this satisfies MLR with $T(\vec{X}) = \bar{X}_n$.

Using the theorem above, the one-sided T-test which rejects when

$$\bar{X}_n > \mu_0 + \frac{1.64}{\sqrt{n}}$$

is also UMP for size $\alpha = 0.05$ for the one-sided hypotheses $H_0 : \mu \leq \mu_0$ vs. $H_1 : \mu > \mu_0$. Call this "test 1".

Taking this example further, we have that for the one-sided hypotheses $H_0 : \mu > \mu_0$ vs. $H_1 : \mu < \mu_0$, the one-sided T-test which rejects when $\bar{X}_n < \mu_0 - \frac{1.64}{\sqrt{n}}$ will be UMP for size $\alpha = 0.05$. Call this "test 2".

Now consider testing

$$H_0 : \mu = \mu_0 \ vs. \ H_1 : \mu \neq \mu_0. \quad (*)$$

The alternative hypothesis is equivalent to "$\mu < \mu_0$ or $\mu > \mu_0$". Can we find an UMP size-α test?

Note that both Test 1 and Test 2 are size-α tests for hypotheses (*). So we consider each in turn.

For any alternative point $\mu_1 > \mu_0$, test 1 is UMP, implying that $\beta_1(\mu_1)$ is maximal among all size-α tests. For $\mu_2 < \mu_0$, however, test 2 is UMP, implying that $\beta_2(\mu_2)$ is maximal. Furthermore, $\beta_1(\mu_2) < \alpha < \beta_2(\mu_2)$ from part (ii) of theorem above, so neither test can be uniformly most powerful for all $\mu \neq \mu_0$. And indeed there is no UMP size-α test for problem (*).

But note that both Test 1 and Test 2 are biased for the hypotheses (*). It turns out that the two-sided T-test which rejects when $\bar{X}_n > \mu_0 + 1.96/\sqrt{n}$ or $\bar{X}_n < \mu_0 - 1.96/\sqrt{n}$ is UMP among size-α *unbiased* tests. See discussion in CB, pp. 392-395.

3 Large-sample properties of tests

In practice, we use large-sample theory — that is, LLN's and CLT's — in order to determine the approximate critical regions for the most common test statistics.

Why? Because finite-sample properties can be difficult to determine:

Example: $X_1, \ldots, X_n \sim$ *i.i.d.* Bernoulli with prob. p.

Want to test $H_0: \ p \leq \frac{1}{2}$ vs. $H_1: \ p > \frac{1}{2}$, using the test stat $\bar{X}_n = \frac{1}{n}\sum_i X_i$.

n is finite. The exact finite-sample distribution for $\bar{X}_n$ is the distribution of $\frac{1}{n}$ times a $B(n,p)$ random variable, which is:

$$\begin{cases} 0 & \text{with prob } \binom{n}{0}(1-p)^n \\ \frac{1}{n} & \text{with prob } \binom{n}{1}p(1-p)^{n-1} \\ \frac{2}{n} & \text{with prob } \binom{n}{2}p^2(1-p)^{n-2} \\ \ldots & \ldots \\ 1 & \text{with prob } p^n \end{cases}$$

Assume your test is of the following form: $\mathbf{1}(\bar{X}_n > c)$, where the critical value c is to be determined such that the size $\sup_{p \leq \frac{1}{2}} P(\bar{X}_n > c|p) = \alpha$, for some specified α.[3] This equation is difficult to solve for!

On the other hand, by the CLT, we know that $\frac{\sqrt{n}(\bar{X}_n - p)}{\sqrt{p(1-p)}} \xrightarrow{d} N(0,1)$. Hence, consider the T-test statistic $Z_n \equiv \sqrt{n}\left(\bar{X}_n - \frac{1}{2}\right)/\sqrt{\frac{1}{4}} = 2\sqrt{n}\left(\bar{X}_n - \frac{1}{2}\right)$.

Under any $p \leq \frac{1}{2}$ in the null hypothesis,

$$P(Z_n > \Phi^{-1}(1-\alpha)) \leq \alpha,$$

for n large enough. (In fact, this equation holds with equality for $p = \frac{1}{2}$, and holds with strict inequality for $p < \frac{1}{2}$.)

Corresponding to this test, you can derive the *asymptotic* power function, which is $\beta^a(p) \equiv \lim_{n\to\infty} P(Z_n > c)$, for $c = \Phi^{-1}(1-\alpha)$:

(Graph)

[3]Indeed, the Clopper-Pearson (1934) confidence intervals for p are based on inverting this exact finite-sample test.

- Note that the asymptotic power function is equal to 1 at all values under the alternative. This is the notion for **consistency** for a test: that it has asymptotic power 1 under every alternative.

- Note also that asymptotic power (rejection probability) is zero under every p of the null, except $p = \frac{1}{2}$.

- (skip) Accordingly, we see that asymptotic power, vs. *fixed alternatives*, is not a sufficiently discerning criterion for distinguishing between tests. We can deal with this by considering *local alternatives* of the sort $\tilde{p} = \frac{1}{2} + h/\sqrt{n}$. Under additional smoothness assumptions on the distributional convergence of $\frac{\sqrt{n}(\bar{X}_n - p)}{\sqrt{p(1-p)}}$ around $p = \frac{1}{2}$, we can obtain asymptotic power functions under these local alternatives.

3.1 Likelihood Ratio Test Statistic: asymptotic distribution

Theorem 10.3.1 (Wilks Theorem): For testing $H_0: \theta = \theta_0$ vs. $H_1: \theta \neq \theta_0$, if $X_1, \dots, X_n \sim$ *i.i.d.* $f(X|\theta)$, and $f(X|\theta)$ satisfies the regularity conditions in Section 10.6.2. Then under H_0, as $n \to \infty$:

$$-2 \log \lambda(\vec{X}) \xrightarrow{d} \chi_1^2.$$

Note: χ_1^2 denotes a random variable from the Chi-squared distribution with 1 degree of freedom. By Lemma 5.3.2 in CB, if $Z \sim N(0,1)$, then $Z^2 \sim \chi_1^2$. Clearly, χ^2 random variables only have positive support.

Proof: Assume null holds. Use Taylor-series expansion of log-likelihood function around the MLE estimator $\hat{\theta}_n$:

$$\begin{aligned}
\sum_i \log f(x_i|\theta_0) =& \sum_i \log f(X_i|\hat{\theta}_n) + \sum_i \frac{\partial}{\partial \theta} \log f(X_i|\theta)\big|_{\theta=\hat{\theta}_n} \cdot (\theta_0 - \hat{\theta}_n) \\
&+ \frac{1}{2} \sum_i \frac{\partial^2}{\partial \theta^2} \log f(X_i|\theta)\big|_{\theta=\hat{\theta}_n} \cdot (\theta_0 - \hat{\theta}_n)^2 + \dots \qquad (5) \\
=& \sum_i \log f(X_i|\hat{\theta}_n) + \frac{1}{2} \sum_i \frac{\partial^2}{\partial \theta^2} \log f(X_i|\theta)\big|_{\theta=\hat{\theta}_n} \cdot (\theta_0 - \hat{\theta}_n)^2 + \dots
\end{aligned}$$

where (i) the second term disappeared because the MLE $\hat{\theta}_n$ sets the first-order condition $\sum_i \frac{\partial}{\partial \theta} \log f(X_i|\theta)\big|_{\theta=\hat{\theta}_n} = 0$; (ii) the remainder term is $o_p(1)$. This is a second-order Taylor expansion.

Rewrite the above as:

$$-2\sum_i \log\left(\frac{f(x_i|\theta_0)}{f(X_i|\hat{\theta}_n)}\right) = -\frac{1}{n}\sum_i \frac{\partial^2}{\partial\theta^2}\log f(X_i|\theta)\big|_{\theta=\hat{\theta}_n}\cdot\left[\sqrt{n}(\theta_0-\hat{\theta}_n)\right]^2 + o_p(1).$$

Now

$$-\frac{1}{n}\sum_i \frac{\partial^2}{\partial\theta^2}\log f(X_i|\theta)\big|_{\theta=\hat{\theta}_n} \xrightarrow{p} -E_{\theta_0}\frac{\partial^2}{\partial\theta^2}\log f(X_i|\theta_0) = \frac{1}{V_0(\theta_0)},$$

where $V_0(\theta_0)$ denotes the asymptotic variance of the MLE estimator (and is equal to the familiar information bound $1/I(\theta_0)$).

Finally, we note that $\frac{\sqrt{n}(\hat{\theta}_n-\theta_0)}{\sqrt{V_0(\theta_0)}} \xrightarrow{d} N(0,1)$. Hence, by the definition of the χ_1^2 random variable:

$$-2\log\lambda(\vec{X}) \xrightarrow{d} \chi_1^2.$$

■■■

In the multivariate case (θ being k-dimensional), the above says that

$$-2\log(\lambda(\vec{X})) \stackrel{a}{=} n(\theta_0-\theta_n)I(\theta_0)^{-1}(\theta_0-\theta_n) \sim \chi_k^2.$$

Hence, LR-statistic is asymptotically equivalent to Wald and Score tests.

■■■

Example: $X_1,\ldots,X_N \sim$ *i.i.d.* Bernoulli with prob. p. Test $H_0: \ p=\frac{1}{2}$ vs. $H_1: \ p\neq\frac{1}{2}$. Let Y_n denote the number of 1's.

$$\lambda(\vec{X}) = \frac{\binom{n}{y_n}\left(\frac{1}{2}\right)^{y_n}\left(\frac{1}{2}\right)^{n-y_n}}{\binom{n}{y_n}\left(\frac{y_n}{n}\right)^{y_n}\left(\frac{n-y_n}{n}\right)^{n-y_n}}.$$

For test with asymptotic size α:

$$\begin{aligned}\alpha &= P(\lambda(\vec{X})\le c) \quad (c<1)\\ &= P(-2\log\lambda(\vec{X})\ge -2\log c)\\ &= P(\chi_1^2\ge -2\log c)\\ &= 1-F_{\chi_1^2}(-2\log c)\\ \Rightarrow c &= \exp\left(-\frac{1}{2}F^{-1}_{\chi_x^2}(1-\alpha)\right).\end{aligned}$$

For instance, for $\alpha=0.05$, then $F^{-1}_{\chi_x^2}(1-\alpha)=3.841$; $\alpha=0.10$, then $F^{-1}_{\chi_x^2}(1-\alpha)=2.706$.