Massimo Guidolin
Massimo.Guidolin@unibocconi.it
Dept. of Finance

Università Commerciale
Luigi Bocconi

STATISTICS/ECONOMETRICS PREP COURSE – PROF. MASSIMO GUIDOLIN

SECOND PART, LECTURE 3: HYPOTHESIS TESTING

OVERVIEW

1) Null vs. alternative hypotheses and rejection region
2) Likelihood ratio tests
3) Type I and II errors: power functions
4) Size of a test
5) Finding a test for given size
6) Uniformly Most Powerful tests and Neyman-Pearson’s lemma
7) The concept of p-value of a test

Null vs. Alternative Hypotheses

- A hypothesis is a statement about a population parameter
- The goal of a test is to decide, based on a sample from the population, which of two complementary hypotheses is true
- Definition: The two complementary hypotheses in a testing problem are called the null hypothesis and the alternative hypothesis; they are denoted by H_o and H_1 , respectively
- If θ is a population parameter, the general format of the null and alternative hypotheses is H_o: $\theta \in \Theta_o$ and H_1: $\theta \in \Theta^c_o$ where Θ_o is a subset of the parameter space and Θ^c_o its complement
  - For example, H_o: $\theta = 0$ versus H_1 : $\theta \neq 0$
  - Usually the null hypothesis implies no effect from a treatment and it tends to be the hypothesis against the conjecture of a researcher
  - Other example: H_o: $\theta < 0$ versus H_1 : $\theta \geq 0$, θ = increase in GPA from attending this prep course

NULL VS. ALTERNATIVE HYPOTHESES

- In a hypothesis testing problem, after observing a random sample, experimenter decides whether to reject H_o as false
- The subset of the sample space for which H_o will be rejected is called the rejection or critical region; the complement of the rejection region is called the non-rejection region
- Typically, a hypothesis test is specified in terms of a test statistic $W(X_1, ..., X_n) = W(\mathbf{X})$, a function of the sample
  - E.g., a test might specify that H_o is to be rejected if the sample mean, is greater than 3; in this case $W(X)$ = sample mean is the test statistic and the rejection region is $\{x_1, ..., x_n)$: sample mean > 3}
- Four methods of finding a test:

  ❶ Likelihood ratio tests

  ❸ Bayesian methods (not here)

  Not here, not explicitly!

  ❷ Union-Intersection Tests

  ❹ Intersection-Union Tests
- The likelihood ratio method is related to ML estimators and

LIKELIHOOD RATIO TESTS

is as widely applicable

- Recall that if X$_1$,... , X$_n$ is a random sample from a population with pdf or pmf f(x; θ) (θ may be a vector), the likelihood function is defined as

$$L(\boldsymbol{\theta};\mathbf{x}) \equiv L(\theta_1,\theta_2,\ldots,\theta_K;x_1,x_2,\ldots,x_n) = f(\mathbf{x};\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i;\theta_1,\theta_2,\ldots,\theta_K)$$

- Definition [LIKELIHOOD RATIO TEST]: A likelihood ratio test (LRT) is any test that has a rejection region of the form {**x**: LRT(**x**) ≤ c}, where c ∈ [0, 1] and $LRT(\mathbf{x}) = \dfrac{\sup_{\Theta_0} L(\theta;\mathbf{x})}{\sup_{\Theta} L(\theta;\mathbf{x})}$
  - The numerator is the maximum probability of the observed sample, the maximum being computed over parameters under the null hypothesis; the denominator is the maximum probability of the observed sample over all possible parameters
  - The ratio of these two maxima is small if there are parameter points in the alternative hypothesis for which the observed sample is much

LIKELIHOOD RATIO TESTS

more likely than for any parameter point under the null hypothesis: in this situation, the LRT criterion says H_o should be rejected

- If you think of doing the maximization over both the entire parameter space (unrestricted maximization)—obtaining $\sup_{\Theta_0} L(\theta;\mathbf{x})$ -- and a subset of the parameter space (restricted maximization) obtaining $\sup_{\Theta} L(\theta;\mathbf{x})$ then the correspondence between LRTs and MLEs is clear
- How do you pick c? Formally, any number in [0, 1] will do; practically, we shall see how to optimize such a choice
- <u>Example</u> (Normal LRT): Let $X_1, \ldots , X_n$ be a random sample from a $N(\theta, 1)$ population. Consider H_o: $\theta = \theta_o$ versus H_1: $\theta \neq \theta_o$, where θ_o is fixed
- Since there is only one value of θ specified by H_o, the numerator of the LRT is $L(\theta_o ;\mathbf{x})$; as for the denominator, we know that the MLE of the mean is the sample mean. So the LRT statistic is

$$LRT(\mathbf{x}) = \frac{L(\theta_0;\mathbf{x})}{\sup_{\Theta} L(\theta;\mathbf{x})} = \frac{L(\theta_0;\mathbf{x})}{L(\hat{\theta}^{ML};\mathbf{x})} = \frac{(2\pi)^{-n/2}\exp[(-1/2)\sum_{i=1}^{n}(x_i-\theta_0)^2]}{(2\pi)^{-n/2}\exp[(-1/2)\sum_{i=1}^{n}(x_i-\bar{x})^2]}$$

Likelihood Ratio Tests

$$= \exp[-(1/2)\sum_{i=1}^{n}(x_i-\theta_0)^2 + (1/2)\sum_{i=1}^{n}(x_i-\bar{x})^2]$$

$$= \exp[-(1/2)\sum_{i=1}^{n}(x_i-\bar{x})^2 - (1/2)n(\bar{x}-\theta_0)^2 + (1/2)\sum_{i=1}^{n}(x_i-\bar{x})^2]$$

$$= \exp[-(1/2)n(\bar{x}-\theta_0)^2]$$

which yields the following rejection region:

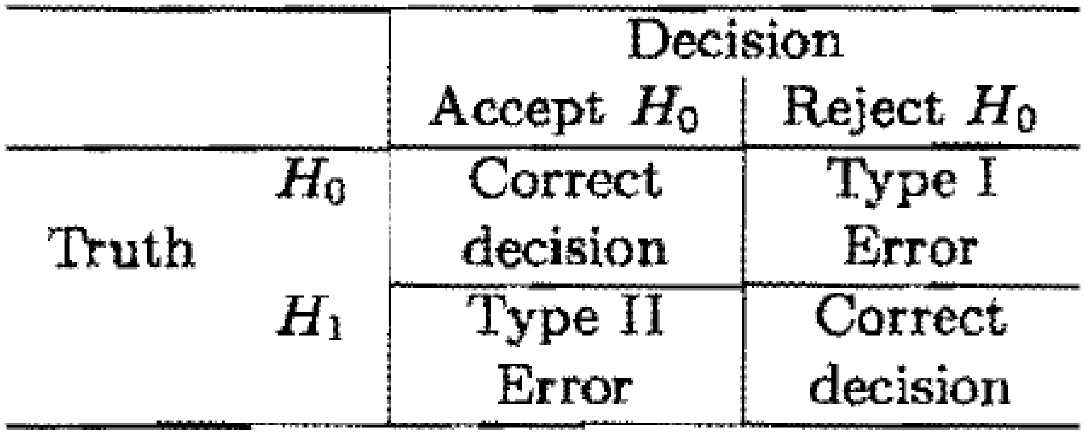
$$LRT(\mathbf{x}) \ = \ \exp[-(1/2)n(\bar{x}-\theta_0)^2] \le c \Longrightarrow \frac{1}{2}n(\bar{x}-\theta_0)^2 \ge -\ln c$$

$$\Longrightarrow (\bar{x}-\theta_0)^2 \ge -\frac{2}{n}\ln c \Longrightarrow \boxed{\left\{\mathbf{x} \ni -\sqrt{-\frac{2}{n}\ln c} \ge \bar{x}-\theta_0\right\} \cup \left\{\mathbf{x} \ni \bar{x}-\theta_0 \ge \sqrt{-\frac{2}{n}\ln c}\right\}}$$

– The LRTs are just those tests that reject H_o: $\theta = \theta_o$ if sample mean differs from the hypothesized value θ_o by more than a specified amount

- In performing a test, an experimenter might be making a mistake: hypothesis tests are evaluated and compared through their probabilities of making mistakes
- There are two possible types of errors

POWER FUNCTION

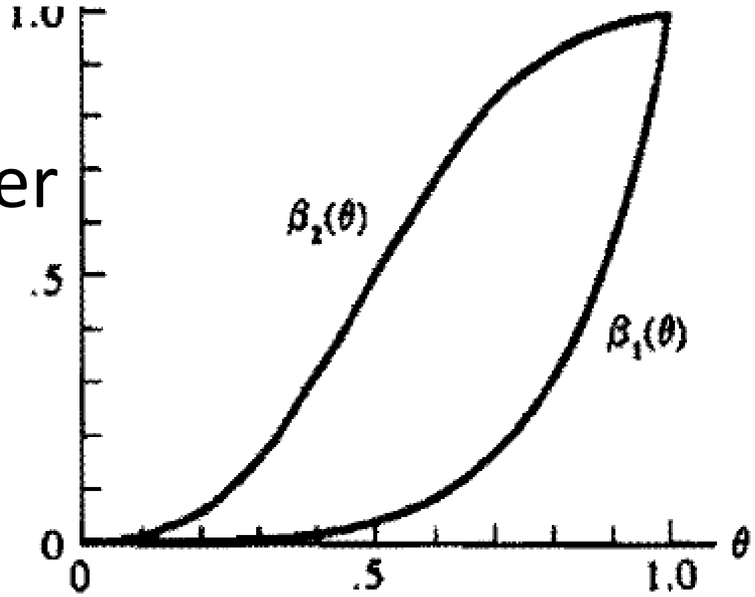


		Decision	
		Accept H_0	Reject H_0
Truth	H_0	Correct decision	Type I Error
	H_1	Type II Error	Correct decision

- Definition [TYPE I ERROR]: If $\theta \in \Theta_o$ but the test incorrectly decides to reject H_o, then the test has made a Type I Error
- Definition [TYPE II ERROR]: If $\theta \in \Theta^c_o$ but the test decides not to reject H_o, a Type II Error has been made

$$P_\theta(\mathbf{X} \in R) = \begin{cases} \text{probability of a Type I Error} & \text{if } \theta \in \Theta_0 \\ \boxed{\text{one minus}}\text{ the probability of a Type II Error} & \text{if } \theta \in \Theta_0^c \end{cases}$$

- This information is summarized by the:
- Definition [POWER FUNCTION]: The power function of a hypothesis test with rejection region R is the function of θ defined by $\beta(\theta) \equiv \Pr(\mathbf{X} \in R)$
  - The ideal power function is 0 for all $\theta \in \Theta_o$



Power Function

and 1 for all $\theta \notin \Theta_o$

- Except in trivial situations, this ideal cannot be attained; therefore qualitatively, a good test has power function near 1 for most $\theta \notin \Theta_o$ and near 0 for most $\theta \in \Theta_o$
- Think of picture in the previous page: would you pick $\beta_1(\theta)$ or $\beta_2(\theta)$?
  - $\beta_1(\theta)$ has very good type I error probability but it goes up 1 very late, which means the type II probability remains large; $\beta_2(\theta)$ has large type I error probability but it goes up 1 rather fast, which means the type II probability declines soon

- For a fixed sample size, n, it is usually impossible to make both types of error probabilities arbitrarily small
- In searching for a good test, we restrict consideration to tests that control the Type I Error probability at a specified level; within this class of tests we then search for tests that have Type II Error probability that is as small as possible

SIZE AND LEVEL

- The following is useful:
- Definition [SIZE OF A TEST]: For $0 \leq \alpha \leq 1$, a test with power function $\beta(\theta)$ is a size-α test if $\sup_{\theta \in \Theta o} \beta(\theta) = \alpha$
  - We commonly specify the level of the test at $\alpha = 0.01$, 0.05, and 0.10
  - Remember that, in fixing the level, you are controlling only the Type I Error probabilities, not the Type II Error
  - In this approach, you should specify the null and alternative hypotheses so that it is most important to control the Type I Error
  - E.g., you want to give support to a particular hypothesis—your thesis—but you do not wish to make the assertion unless the data give convincing support, the test can be set up so that the alternative hypothesis is the one that you "would like" the data to support
    - The alternative hypothesis is sometimes called the research hypothesis
- In this case it is appropriate to set a size α that is small

Setting the Parameter c

- So far, c was unspecified, so not one but an entire class of LRTs was defined, one for each value of c

- The restriction to size α tests may now lead to choice of one out of the class of tests, pinning down the optimal value of c
  - Example: In our example with $X_1, \ldots, X_n$ a random sample from a N(θ, 1) population, we set H_o: $\theta = \theta_o$ versus H_1: $\theta \neq \theta_o$, where θ_o is fixed and concluded that the rejection region was:

$$\Pr{}_{\theta \in \Theta_0}(\bar{x} \ \in \ R) = \Pr\left(|Z| \geq \sqrt{-2\ln c}\right) = \alpha \Longrightarrow \underbrace{\Pr\left(Z \leq -\sqrt{-2\ln c}\right)}_{=\alpha/2} + \underbrace{\Pr\left(Z \geq \sqrt{-2\ln c}\right)}_{=\alpha/2} = \alpha$$

  where the sum of two terms derives from symmetry
  - At this point, the following steps allow is to derive the optimal c and to see that in this case this corresponds to a normal critical value:

$$\Pr{}_{\theta \in \Theta_0}(\bar{x} \ \in \ R) = \Pr\left(|Z| \geq \sqrt{-2\ln c}\right) = \alpha \Longrightarrow \underbrace{\Pr\left(Z \leq -\sqrt{-2\ln c}\right)}_{=\alpha/2} + \underbrace{\Pr\left(Z \geq \sqrt{-2\ln c}\right)}_{=\alpha/2} = \alpha$$

$$\Longrightarrow \Pr\left(Z \geq \sqrt{\text{-}2\ln c}\right) = \Pr\left(Z \geq z_{\alpha/2}\right) = \frac{\alpha}{2} \Longrightarrow z_{\alpha/2} = \sqrt{\text{-}2\ln c} \Longrightarrow \boxed{c^* = \exp(-z^2_{\alpha/2}/2)}$$

Uniformly Most Powerful Tests

- In the above example we used the notation $z_{\alpha/2}$ to denote the point having probability $\alpha/2$ to the right of it for a standard normal PDF
- This standard, general notation: z_α is such that $\Pr(Z > z_\alpha) = \alpha$ where $Z \sim N(0,1)$; $t_{n-1,\alpha/2}$ is such that $\Pr(T_{n-1} > t_{n-1,\alpha/2}) = \alpha/2$; $\chi^2_{p,1-\alpha}$ is such that $\Pr(\chi^2_p > \chi^2_{p,1-\alpha}) = 1 - \alpha$; all these points are known as cutoff points

- Not always fixing the size of a test delivers a unique choice for the parameter c; in this case a criterion to rank tests is needed
  - The logic is however clear: for fixed size, we want to maximize the power of a test because type II error pr. = $1 - \beta(\theta)$ and so large $\beta(\theta) \Rightarrow$ low type II error
- Definition [UNIFORMLY MOST POWERFUL TEST]: Let $\mathbb{C}$ be a class of tests for H_o: $\theta \in \Theta_o$ vs. H_1: $\theta \in \Theta^c_o$. A test in $\mathbb{C}$, with power function $\beta(\theta)$ is a uniformly most powerful (UMP) class C test if $\beta(\theta) \geq \beta'(\theta)$ for every $\theta \in \Theta^c_o$ and every $\beta'(\theta)$ that is a power function of a test in class $\mathbb{C}$

UMP Tests: Neymann-Pearson Lemma

- What is $\mathbb{C}$? For instance, all tests with a fixed size
- At this point, there are good and bad news; bad news, for most applications, UMP tests do not exist and are always hard to find
- Good news: when the hypotheses are simple (i.e., they both consist of only one probability distribution for the sample), then we know quite a lot about the structure of UMP tests, as shown by the famous:

- <u>Key Result 12</u> [NEYMAN-PEARSON LEMMA]: Consider H_o: $\theta = \theta_0$ vs. versus H_1: $\theta = \theta_1$ where the pdf or pmf corresponding to θ_i is $f(x; \theta_i)$, $i = 0, 1$, using a rejection region R that satisfies

$$\begin{cases} \mathbf{x} \in R \text{ if } f(\mathbf{x};\theta_1) > kf(\mathbf{x};\theta_0) \\ \mathbf{x} \in R^c \text{ if } f(\mathbf{x};\theta_1) \leq kf(\mathbf{x};\theta_0) \end{cases} \quad \text{for some } k \geq 0 \text{ and } \alpha = \Pr_{\theta\in\Theta_o}(\mathbf{x} \in R)$$

(a, sufficiency) Any test that satisfies these conditions is a UMP level-α test; (b, necessity) if there exists a test satisfying these

The Concept of P-Value

conditions with $k > 0$, then every UMP sized-α test satisfies them almost surely

- Since we have mentioned the simple ones, let's add that hypotheses that assert that a univariate parameter is large, for example, H: $\theta \geq \theta_0$, or small, for example, H: $\theta < \theta_0$, are called **one-sided hypotheses**
- Hypotheses that assert that a parameter is either large or small, for example, H: $\theta \neq \theta_0$ are called **two-sided hypotheses**

- After a hypothesis test is done, the conclusions must be reported in some meaningful way
  - One method of reporting the results of a hypothesis test is to report the size, α, of the test used and the decision to reject H_o or not
  - The size of the test carries important information. If α is small, the decision to reject H_0 is fairly convincing; but if α is large, the decision to reject H_o is not very convincing because the test has a large probability of incorrectly making that decision

The Concept of P-Value

- Another way of reporting the results of a test is to report the value of a certain kind of test statistic called a p-value
- Definition [P-VALUE]: A p-value p(**X**) is a sample statistic satisfying p(**x**) $\in$ [0,1] for every sample point **x** that captures the largest possibile size of all tests that reject the null hypothesis
  - Formally, let W(**X**) be a test statistic such that large values of W give evidence that H_1 is true; for each sample point **x**, a p-value p(**x**) is such that p(**x**) = $\sup_{\theta \in \Theta o}$Pr(W(**X**) > W(**x**)) (note the “sup” under null)
  - An advantage to reporting a test result via a p-value is that each reader can choose the α she considers appropriate and then can compare the reported p(**x**) to α
  - The smaller the p-value, the stronger the evidence for rejecting H_0
  - A p-value reports the test result on a more continuous scale, rather than just the dichotomous decision "Reject H_o“, “Not Reject H_o"

USEFUL NOTIONS REVIEWED IN THIS LECTURE

- Let me give you a list to follow up to:
- Logical structure of a statistical test of hypothesis
- Rejection vs. non-rejection region
- Likelihood ratio tests
- Type I, type II errors and power function of a test
- Size of a test
- Uniformly most powerful tests
- The Lehman-Pearson’s lemma
- Concept of p-value from a test