Lecture Note 10 *
Hypothesis Testing

MIT 14.30 Spring 2006
Herman Bennett

Hypothesis testing: Given a random sample from a certain population, Is the sample evidence ***enough*** *to disregard a particular belief about that population? (E.g.: the value of a parameter.)*

25 Definitions

25.1 Hypothesis Testing

A (parametric) hypothesis is a statement about one or more population parameters [1]. This hypothesis can be tested using a *hypothesis test*.

A hypothesis test consists of:

1. Two complementary hypotheses: the null hypothesis and the alternative hypothesis, denoted H_0 and H_1 respectively.
2. A decision rule that specifies for which sample values the *null hypothesis* is not rejected ('accepted'), and for which sample values the *null hypothesis* is rejected in favor of the *alternative hypothesis*.

The set of sample values for which H_0 is rejected is called the rejection or critical region. The complement of the critical region is called the acceptance region (where H_0 is accepted).

*Caution: These notes are not necessarily self-explanatory notes. They are to be used as a complement to (and not as a substitute for) the lectures.

[1] An example of a nonparametric hypothesis would be to make a statement about the distribution of the RV X. E.g.: $X \sim N(\ )$.

A hypothesis that if true completely specifies the population distribution, is called a simple hypothesis; one that does not is called a composite hypothesis.

25.2 General Setting for Hypothesis Testing

Let $X_1, ..., X_n$ be a random sample from a population with pmf/pdf $f(x|\theta)$. Define the following hypothesis test about the parameter $\theta \in \Omega$,:

$$H_0: \quad \theta \in \Omega_0$$
$$H_1: \quad \theta \in \Omega_1,$$

where $\Omega_0 \cup \Omega_1 = \Omega$ and $\Omega_0 \cap \Omega_1 = \emptyset$. H_0 is rejected if the random sample $X_1, ..., X_n$ lies in the $n-$dimensional space C. The space C is the *critical region* defined in terms of $\mathbf{x}$, the $n-$dimensional vector that contains the random sample.

The two complementary hypotheses, H_0 and H_1, usually take one of the following five structures:

1.- Singleton H_0 and singleton H_1 :
$$H_0: \quad \theta = \theta_0$$
$$H_1: \quad \theta = \theta_1 \tag{76}$$

2.- Singleton H_0 and composite 2-sided H_1 :
$$H_0: \quad \theta = \theta_0$$
$$H_1: \quad \theta \neq \theta_0 \tag{77}$$

3.- Singleton H_0 and composite 1-sided H_1 :
$$H_0: \quad \theta = \theta_0$$
$$H_1: \quad \theta < \theta_0 \text{ (or } \theta > \theta_0) \tag{78}$$

4.- Composite 1-sided H_0 and composite 1-sided H_1 :
$$H_0: \quad \theta \leq \theta_0 \text{ (or } \theta \geq \theta_0)$$
$$H_1: \quad \theta > \theta_0 \text{ (or } \theta < \theta_0) \tag{79}$$

5.- Composite 2-sidedH_0 and composite 2-sided H_1 :
$$H_0: \quad \theta_{n_1} \leq \theta \leq \theta_{n_2}$$
$$H_1: \quad \theta < \theta_{n_1} \text{ and } \theta > \theta_{n_2} \tag{80}$$

25.3 Type of Errors in Hypothesis Testing

A type I error occurs when H_0 is rejected when indeed is true. The probability that this error occurs, denoted α_θ, is defined as follows:

$$\alpha_\theta = P(\text{type I error}) = P\big(\text{rejecting } H_0 \big|\ \theta \in \Omega_0\big) \tag{81}$$

A type II error occurs when H_0 is not rejected when indeed H_1 is true. The probability that this error occurs, denoted β_θ, is defined as follows:

$$\beta_\theta = P(\text{type II error}) = P\big(\text{accepting } H_0 \big|\ \theta \in \Omega_1\big) \tag{82}$$

- Wrap up:

25.3.1 Level of Significance and Optimal Tests

The level of significance, or size, of a hypothesis test is the highest type I error. The level of significance is denoted by α. [2] Formally:

$$\alpha = \sup_{\theta \in \Omega_0} \alpha_\theta. \tag{83}$$

If Ω_0 is singleton: $\alpha = \alpha_\theta$.

For a given pair of *null* and *alternative* hypotheses, and a given level of α, an optimal hypothesis test is defined as a test that minimizes $\beta_\theta\ \forall\theta$. Note that optimal tests do not exist for many hypothesis test structures (more on this later).

[2] There is a technical difference between the *level* and the *size*, which in practice becomes only relevant in complicated testing situations. For the purpose of this course we will use them interchangeably.

Example 25.1. Assume a random sample of size n from a normal population $N(\mu, 4)$. i) Use the statistic $\bar{X}$ to construct a hypothesis test with $H_0 : \mu = 0$, $H_1 : \mu = 1$, and a decision rule of the form "reject H_0 when $\bar{x} > k$", such that the probability of type I error is 5%. ii) Compute the probability of type II error. What is the size of the test? iii) What happens to α and β as $k \uparrow$ or $k \downarrow$? Which is the trade-off? iv) What happens if the sample size n increases? v) How would the answers change if we redefine the hypotheses as $H_0 : \mu = 0$ and $H_1 : \mu \neq 0$.

- Be careful when interpreting the results of a hypothesis test: accepting v/s failing to reject H_0.

25.4 Power Function

Let's denote the characteristics of a hypothesis test (the null hypothesis, the alternative hypothesis, and the decision rule) by the letter δ.

The power function of a hypothesis test δ is the probability of rejecting H_0 given that the true value of the parameter is $\theta \in \Omega$.

$$\pi(\theta|\delta) = P\big(\text{rejecting } H_0 \big| \ \theta \in \Omega\big) = P(\mathbf{X} \in C|\theta) \qquad \text{for all } \theta \in \Omega. \tag{84}$$

Thus,

$$\begin{aligned} \pi(\theta|\delta) &= \alpha_\theta(\delta) \qquad \text{if } \theta \in \Omega_0 \\ 1 - \pi(\theta|\delta) &= \beta_\theta(\delta) \qquad \text{if } \theta \in \Omega_1 \end{aligned} \tag{85}$$

Example 25.2. Ideal power function...a=? b=?

$$\pi(\theta|\delta) = \begin{cases} a & \text{if } \theta \in \Omega_0 \\ b & \text{if } \theta \in \Omega_1. \end{cases}$$

• If Ω_0 is singleton: $\alpha = \pi(\theta|\delta)$.

• For a given pair of *null* and *alternative* hypotheses, and a given level of α, an optimal hypothesis test, δ^*, is a test that minimizes $\beta(\delta)$ for all $\theta \in \Omega_1$. In other words, δ^* maximizes the power function for all $\theta \in \Omega_1$.

Example 25.3. Assume a random sample of size n from a U$[0, \theta]$, where θ is unknown. Suppose the following hypothesis test δ:

$$\begin{aligned} H_0 : &\quad 3 \leq \theta \leq 4 \\ H_1 : &\quad \theta < 3 \text{ or } \theta > 4 \end{aligned}$$

Decision rule: Accept H_0 if $\hat{\theta}_{MLE} \in [2.9, 4.1]$, and reject H_0 otherwise.

Find the power function $\pi(\theta|\delta)$ (note: $\forall\theta$). Which is the size of this test?

25.5 p-value

The p-value describes the minimum level of significance α that would have implied, given the particular realization of the random sample ($\mathbf{x}$), a rejection of H_0. Thus, the p-value, as well as whether H_0 is rejected or not, are *ex-post* calculations.

26 (Four) Most Common Hypothesis Tests Structures

26.1 Likelihood Ratio Test (LRT):

$$H_0: \quad \theta = \theta_0$$
$$H_1: \quad \theta = \theta_1$$

$$\text{Decision Rule form: "Reject } H_0 \text{ if } f_1(\mathbf{x})/f_0(\mathbf{x}) > k\text{".} \tag{86}$$

Where $k > 0$ is a constant chosen according to the size of the test (α_0), such that $P(f_1(\mathbf{x})/f_0(\mathbf{x}) > k|\theta_0) = \alpha_0$. The statistic $f_1(\mathbf{x})/f_0(\mathbf{x})$ is given by:

$$f_i(\mathbf{x}) = f(x_1, x_2, ..., x_n|\theta_i) = f(x_1|\theta_i)f(x_2|\theta_i)...f(x_n|\theta_i) \qquad (iid \text{ sample}) \tag{87}$$

- The ratio $f_1(\mathbf{x})/f_0(\mathbf{x})$ is called the likelihood ratio of the sample.

Optimality of the LRT

Minimize the probability of type II error given the probability of type I error:

$$\min_{\delta} \beta \; ; \quad \text{given} \;\; \alpha_0.$$

(α_0 is the size imposed on the test.)

(Neyman-Pearson lemma) Let δ^* be a hypothesis test where H_0 and H_1 are simple hypotheses, and where H_0 is accepted if $kf_0(\mathbf{x}) > f_1(\mathbf{x})$ ($k > 0$). Otherwise, H_1 is accepted, except if $kf_0(\mathbf{x}) = f_1(\mathbf{x})$ where both H_0 and H_1 may be accepted. Then, for every other hypothesis test δ:

$$\beta(\delta) < \beta(\delta^*) \longleftrightarrow \alpha(\delta) > \alpha(\delta^*) \tag{88}$$

Example 26.1. Assume a random sample of size $n = 20$ from a Bernoulli distribution, where p is unknown. Suppose the following hypotheses:

$$H_0 : p = 0.2$$
$$H_1 : p = 0.4$$

Find the optimal test procedure δ^* with $\alpha(\delta^*) = 0.05$.

• For the case of a normal random sample, the hypothesis test (86) implies the following decision rule:

- "Reject H_0 if $\bar{x} > k'$ " when $\theta_1 > \theta_0$.
- "Reject H_0 if $\bar{x} < k'$ " when $\theta_1 < \theta_0$.

(For the derivation of this result check DeGroot and Schervish (2002) page 465.)

26.2 One-sided Test:

$$H_0: \quad \mu = \mu_0$$
$$H_1: \quad \mu > \mu_0 \qquad \text{or} \qquad H_1: \mu < \mu_0$$

$$\text{Decision Rule form: "Reject } H_0 \text{ if } \bar{x} > c\text{" or "Reject } H_0 \text{ if } \bar{x} < c\text{".} \tag{89}$$

Where c is a constant chosen according to the size of the test (α_0), such that $P(\bar{X} > c|\mu_0) = \alpha_0$ or $P(\bar{X} < c|\mu_0) = \alpha_0$.

Optimality of One-sided tests

What does it mean to be optimal in these cases? Should we use non optimal tests?

A generalization of the relevant optimality results for these cases is out of the scope of this course. [3] However, we can handily state the following result:

- Assume a random sample from a binomial or normal distribution, a *null* and an *alternative* hypotheses given by (89), and a level of significance α_0. Then, the **optimal** test δ^*, which minimizes $\beta(\delta)$ for all $\theta \in \Omega_1$, is given by test (89).

• The decision rule of test (89) is widely used for cases (78) **and** (79), even if it is not an optimal test.

[3] An excellent reference is DeGroot and Schervish (2002) Ch. 8.3.

Example 26.2. Assume a random sample of size $n = 100$ from a $\mathrm{N}(\mu, 1)$, where μ is unknown and $\bar{x} = 1.13$. Suppose the following hypotheses:

$$H_0 : \mu = 1$$
$$H_1 : \mu > 1$$

Construct a one-sided hypothesis test of size 0.05. Test the null hypothesis and find the p-value.

26.3 Two-sided Test:

$$H_0 : \quad \mu = \mu_0$$
$$H_1 : \quad \mu \neq \mu_0$$

$$\text{Decision Rule form: "Reject } H_0 \text{ if } \bar{x} \text{ is outside the interval } [c_1, c_2]. \tag{90}$$

Where c_1 and c_2 are a constants chosen according to the size of the test (α_0), such that $P(\bar{x} \notin [c_1, c_2]\ |\mu_0) = \alpha_0$. Usually, hypothesis tests are constructed in a symmetric way, which means that $P(\bar{X} < c_1|\mu_0) = \alpha_0/2$ and $P(\bar{X} > c_2|\mu_0) = \alpha_0/2$.

Optimality of Two-sided tests

Unfortunately, there is no robust result regarding optimality in this case. No test procedure δ^* will minimize $\beta_\theta(\delta)$ for all $\theta \in \Omega_1$. However, the optimality results for the 1-sided hypothesis test case suggest that a reasonable decision rule for the hypotheses described in (90), could be given by the decision rule of test (90). [4]

- The decision rule of test (90) is widely used for cases (77) and (80).

Example 26.3. A candle producer company claims that their candles last for 60 minutes on average. One consumer, curious about this claim, bought 40 candles and tested them. He found that on average they last for 65.22 minutes. With the data collected he also computed the statistic $s^2 = 225$. Can the consumer say, with 99% of significance, that the company is wrong in its claim? (Assume the sample is *iid*.) Also, compute the p-value and the limiting n such that H_0 is rejected at $\alpha = 0.01$ (assume s^2 and $\bar{x}$ keep their value).

[4] In fact, this is what most researchers do.

26.4 Generalized Likelihood Ratio Test (GLRT):

H_0 and H_1 : any composite or simple hypothesis

$$\text{Decision Rule form: "Reject } H_0 \text{ if } W > k\text{".} \tag{91}$$

Where $k > 0$ is a constant chosen according to the size of the test (α_0), such that $P(W > k|H_0) = \alpha_0$. The statistic W is given by:

$$W = \frac{\sup_{\theta \in \Omega_1} L(\theta_1, ..., \theta_k | x_1, ..., x_n)}{\sup_{\theta \in \Omega_0} L(\theta_1, ..., \theta_k | x_1, ..., x_n)} = \frac{\sup_{\theta \in \Omega_1} f(\mathbf{x}|\theta \in \Omega_1)}{\sup_{\theta \in \Omega_0} f(\mathbf{x}|\theta \in \Omega_0)}. \tag{92}$$

• As with previous tests, the constant k will depend on the distribution of the statistic W and α_0. If computing the distribution of W becomes a nightmare, it is possible to use an equivalent definition of the GLRT, (93), which has a known limiting distribution.

$$T = \frac{\sup_{\theta \in \Omega_0} L(\theta_1, ..., \theta_k | x_1, ..., x_n)}{\sup_{\theta \in \Omega} L(\theta_1, ..., \theta_k | x_1, ..., x_n)} = \frac{\sup_{\theta \in \Omega_0} f(\mathbf{x}|\theta \in \Omega_0)}{\sup_{\theta \in \Omega} f(\mathbf{x}|\theta \in \Omega)} \tag{93}$$

Decision Rule form: "Reject H_0 if $T < d$"; where $d > 0$ is a constant chosen according to the test size (α_0), such that $P(T < d|H_0) = \alpha_0$. The limiting distribution of $-2lnT$ is known:

$$-2lnT \overset{n \to \infty}{\sim} \chi^2_{(r)} \; ; \tag{94}$$

where r is the # of free parameters in Ω minus the # of free parameters in Ω_0. Reject H_0 if $-2lnT > \chi^2_{(r),\alpha}$. [5]

If it is possible to compute directly the distribution of W or T, then it is better to use that distribution instead of the limiting χ^2.

Optimality of GLRT

The GLRT is a generalization of the LRT; it works for any case where either H_0 or/and H_1 are composite hypotheses. However, GLRT is **not necessarily optimal**, as the LRT is. In particular, it will depend on the case at hand (further details on this issue are out of the scope of this course [6]).

[5] The technical result says that the distribution is a $\chi^2_{(r)}$ with degrees of freedom $r = \text{dim}\Omega - \text{dim}\Omega_0$.

[6] An excellent reference is DeGroot and Schervish (2002) Ch. 8.

27 Hypothesis Testing Based on Two Normal Samples

Example 27.1. Assume 2 random samples:

$$X_i \sim N(\mu_X, \sigma_X^2) \text{ of sample size } n_X$$
$$Y_i \sim N(\mu_Y, \sigma_Y^2) \text{ of sample size } n_Y,$$

and the following hypotheses to be tested:

$$a) \quad H_0: \ \mu_X = \mu_Y \qquad\qquad b) \quad H_0: \ \sigma_X^2 = \sigma_Y^2$$
$$H_1: \ \mu_X \neq \mu_Y \qquad\qquad H_1: \ \sigma_X^2 \neq \sigma_Y^2$$

For each case, construct a hypothesis test of size 95%. In part a) assume that you know σ_X^2 and σ_Y^2.

That's all Folks!