# 6 Lecture 6: Hypothesis Testing

## 6.1 Motivation

Assume $\mathbf{X} = (X_1, X_2, .., X_n)$ are i.i.d. from $f(x, \theta), \theta \in \Theta$. Point estimation of $\theta$ will give an estimated value of $\theta$ which will be in general different from the true $\theta$. In fact, if $\Theta$ was not a finite set but an interval (as it often happens) then the estimator and the true value will coincide with probability zero! This observation alone is convincing enough to claim that it is not enough just to give a single estimated value of the parameter. The problems of constructing confidence intervals (if the parameter was one-dimensional), or confidence sets (if the parameter was multi-dimensional), and the problems of testing hypotheses about $\theta$ naturally arise.

We will mainly constrain ourselves to hypothesis testing and will avoid the thorough discussion of confidence sets. Besides lack of time, the following argument can be put forward to defend this decision. In your introductory Statistics courses you have studied the interrelationship between Hypothesis testing and the construction of Confidence intervals. In particular, given certain $\alpha$ size test of a hypothesis $H_0 : \theta = \theta_0$ , and having obtained the sample, we can take the set of the parameter values for which the test "answers" with an acceptance when the sample is substituted in the test statistic. This set of parameter values is a confidence set at level $1 - \alpha$. Symbolically, we can say that the subset in $\Theta$ defined via

$$\{\theta' | H_0 : \theta = \theta' \text{ is accepted given realization } \mathbf{X} = \mathbf{x} \text{ of the sample } \}$$

represents a confidence set at level $(1 - \alpha)$ for the unknown parameter $\theta$.

In other words, knowing how to construct tests, we basically also know how to construct confidence sets. Moreover, the usefulness of the relationship between testing hypotheses and confidence sets is further exemplified by the fact that some optimality results carry over. It can be shown quite generally that the above procedure of constructing confidence sets leads to confidence sets with optimality properties if the hypothesis test used in the construction was optimally designed.

## 6.2 General terminology in relation to hypothesis testing.

Let us start with the case of testing a *simple* hypothesis against a *simple* alternative. This is the easiest case to discuss. Besides, the technique that is being used in this simple case (the **Neyman-Pearson lemma** below) is indeed fundamental and serves as a basis to deal with the more difficult cases, too.

Assume that the unknown parameter $\theta$ can be one of the two values $\{\theta_0, \theta_1\}$ only. In other words, we are testing a *simple hypothesis* $H_0 : \theta = \theta_0$ versus a *single alternative* $H_1 : \theta = \theta_1$. A *test* $\varphi(\mathbf{x})$ is defined as

$$\varphi(\mathbf{x}) = P(\text{ reject } H_0 \mid \mathbf{X} = \mathbf{x}).$$

Generally, we would prefer deterministic decisions, i.e. we would like $\varphi(\mathbf{x})$ to be equal to either zero or one. Based on the observations we calculate $\varphi(\mathbf{x})$ and according to its value, we reject $H_0$ (if it happens that $\varphi(\mathbf{x}) = 1$) or do not reject it (if $\varphi(\mathbf{x}) = 0$). This means that we need to decompose the sample space $\mathcal{X}$ into two regions A and $S$ :

$$\mathcal{X} = A \cup S; A \cap S = \oslash$$

so that depending on $\mathbf{x} \in A$ or $\mathbf{x} \in S$ we decide for either $H_0$ or $H_1$. We shall be looking for a definition of $\varphi(\mathbf{x})$(or equivalently, to a decomposition of $\mathcal{X}$ into A and $S$) in some sort of optimal way. To define reasonably what optimality could mean in this setting, we need to examine the sorts of errors that we can encounter when deciding to reject or accept $H_0$. As is well-known from your introductory Statistics course, we can commit 2 types of errors:

- to reject $H_0$ given that $H_0$ is correct (first type of error)
- to accept $H_0$ given that $H_1$ is correct (second type of error).

The corresponding probabilities are denoted as follows:

$$P(\text{ reject } H_0 \mid H_0 \text{ correct }) = \text{level of the test (significance)}$$

$$P(\text{ accept } H_0 \mid H_1 \text{ correct }) = 1 - \text{(power of the test)}$$

It is also well known that minimizing level and maximizing power simultaneously is **not possible**. Because of this well known fact, one possible way out is to formulate a **constrained optimization** problem as follows: we decide to fix certain (small) value $\alpha$(e.g. $\alpha = 0.005, 0.01, 0.05, 0.10$) (level of significance) that is not allowed to be exceeded for the first type error and in the set of all tests having first type error equal to $\alpha$, we are looking for the one with a smallest possible second type error (or equivalently the highest possible power).

In the signal processing literature, the first type error is called, not unreasonably, the "false alarm". This is because if the null hypothesis is about no (enemy) signal in the generally recorded noise level then the rejection of the null implies that a false alarm has been raised. You can imagine now that choosing the probability level for a false alarm ($\alpha$) too high means that too many false alarms could be raised; on the other hand, choosing it too low implies that we are increasing the chance for a signal to be missed. This is why choosing the suitable level $\alpha$ represents a compromise. There is no unique recommendation for the choice of $\alpha$ and this choice is often related to the particular field of study. Indeed, very often we have an idea what highest first type error we could tolerate. Despite the above comments, the value $\alpha = 0.05$ is often considered as a "default".

Once having decided on the level of significance, we would like to "perform optimally". Of course, after having constructed the optimal test, we would still like to examine its power to see if it is not too low for the purpose of our analysis. If this turns out to be the case, we might need to increase the **sample size** in order to improve the power.

Unfortunately, even this constrained optimization problem turns out sometimes not to have a solution in cases of **discrete** observations. Sometime, in the discrete case, it is not possible to decompose $\mathcal{X}$ in such a way that the first type error is exactly equal to $\alpha$(we can not "exhaust the level"). To be able to do this, one needs to introduce *randomized* *tests* by allowing $\varphi(\mathbf{x})$ to take *any value* in [0,1]. With this extension of $\varphi$, the two types of errors discussed above have the following interpretation:

$$P(\text{reject } H_0 \mid H_0 \text{ correct }) = \int .. \int P(\text{ reject } H_0 \mid \mathbf{X} = \mathbf{x}) L(\mathbf{x}, \theta_0) d\mathbf{x} =$$

$$\int .. \int \varphi(\mathbf{x}) L(\mathbf{x}, \theta_0) d\mathbf{x} = E_{\theta_0}\varphi$$

(where $d\mathbf{x}$ is a shorthand notation for $d\mathbf{x} = dx_1 dx_2 .. dx_n$)

$$P(\text{ accept } H_0 \mid H_1 \text{ true }) = \text{ similarly to the argument above} = 1 - E_{\theta_1}\varphi$$

These definitions of the two types of errors can be easily interpreted also in cases of composite hypotheses and will be used from now on.

## 6.3 Fundamental Lemma of Neyman- Pearson

You must have heard about it from your earlier statistics courses.

**Lemma 6.1.** *i) For every $\alpha \in (0,1)$ there exists a constant $C$ and a test*

$$\varphi^* = \begin{cases} 1 \textit{ if } x \in S = \{x : L(x,\theta_1)/L(x,\theta_0) > C\}, \\ \gamma \textit{ if } x \in R = \{x : L(x,\theta_1)/L(x,\theta_0) = C\}, \\ 0 \textit{ if } x \in A = \{x : L(x,\theta_1)/L(x,\theta_0) < C\} \end{cases}$$

*with $E_{\theta_0}\varphi^* = \alpha$. The constant $\gamma \in (0,1)$ in the definition of the test is equal to $\gamma = \frac{\alpha - P_{\theta_0}(S)}{P_{\theta_0}(R)}$;*

*ii)$\varphi^*$ is the best $\alpha$-test, i.e. $E_{\theta_1}\varphi^*$ is maximal among all tests $\varphi \in \Phi_\alpha = \{\varphi \mid E_{\theta_0}\varphi \leq \alpha\}$.*

*iii) $\varphi^*$ is essentially unique, i.e. all other "best" $\alpha$-tests in the sense of ii) must coincide with $\varphi^*$ on $S$ and $A$.*

**Proof** (sketch, details at lecture):

i) Given $\alpha$, we define $C$ to be the smallest value on the real line for which $P_{\theta_0}\{\frac{L(X,\theta_1)}{L(X,\theta_0)} > C\}$ is still $\leq \alpha$ (in the continuous case we would actually have precisely $P_{\theta_0}\{\frac{L(X,\theta_1)}{L(X,\theta_0)} > C\} = \alpha$ but equality might not be possible in the discrete case). The constant $C$ which we choose in this manner has a specific name-it is called the **upper** $\alpha * 100\%-$**point** of the distribution of $\frac{L(X,\theta_1)}{L(X,\theta_0)}$ when $\theta_0$ is the true parameter. Then, looking at the definition of $\varphi^*$ and using the definition of $\gamma$ we see that

$$E_{\theta_0}\varphi^* = 1 * P_{\theta_0}(\mathbf{X} \in \mathbf{S}) + \gamma * \mathbf{P}_{\boldsymbol{\theta_0}}(\mathbf{X} \in \mathbf{R}) = \cdots = \alpha$$

ii) Take *any other* $\alpha-$test $\varphi$ and divide the sample space $\mathcal{X}$ into $\mathcal{X} = \mathcal{X}^+ \bigcup \mathcal{X}^- \bigcup \mathcal{X}^=$ with:

$$\mathcal{X}^+ = \{\mathbf{X} : \varphi^*(\mathbf{X}) - \varphi(\mathbf{X}) > 0\}$$

$$\mathcal{X}^- = \{\mathbf{X} : \varphi^*(\mathbf{X}) - \varphi(\mathbf{X}) < 0\}$$
$$\mathcal{X}^= = \{\mathbf{X} : \varphi^*(\mathbf{X}) - \varphi(\mathbf{X}) = 0\}$$

Analyzing the expression $Z(X) = (\varphi^*(X) - \varphi(X))(L(X, \theta_1) - CL(X, \theta_0))$ separately for values of $X \in \mathcal{X}^+, X \in \mathcal{X}^-$ and $X \in \mathcal{X}^=$, we see that always $Z(X) \geq 0$ holds! (Why (!)). But then, of course,

$$\int_{\mathcal{X}} Z(X)dX \geq 0 \tag{12}$$

holds. Substituting back the value of $Z(X)$ in (12) we get: $E_{\theta_1}\varphi^* \geq E_{\theta_1}\varphi$. Since $\varphi$ was arbitrarily chosen in the set of $\alpha-$tests, this implies that $\varphi^*$ can not be improved with respect to power, that is, it is the best $\alpha-$ size test.

iii) If $\bar{\varphi}$ is another "best" $\alpha$ test (that is $E_{\theta_1}\varphi^* = E_{\theta_1}\bar{\varphi}$ holds) then according to our discussion in ii), we necessarily need to have $Z(X) \equiv 0$. Since $Z(X)$ is a product of two factors, we either have one of the factors being zero: $\varphi^*(X) = \bar{\varphi}(X)$ or, if not then the other one must be zero (which means $X \in R$). Hence, always when $X$ is not in $R$ but in $S$ or in $A$, we must have $\varphi^*(X) = \bar{\varphi}(X)$.

## 6.4 Comments related to the Neyman-Pearson Lemma

Any test $\varphi$ with $E_{\theta_0}\varphi = \alpha$ is called an *$\alpha$-test* (equivalently, an $\alpha-$size test. The region $S$ is called a *rejection region (critical region).* The optimal test $\varphi^*$ has the highest power among all tests of size $\leq \alpha$. Looking at the structure of $\varphi^*$ given in Part i) of Lemma 6.1 we see that it has a very simple and intuitively appealing interpretation: we look at the ratio of likelihoods for the sample under the alternative and under the null hypothesis. When this ratio is large enough, the alternative is more likely to have generated the sample and we "vote" for the alternative. If the ratio is small enough, the hypothesis is more likely to have generated the sample and we "vote" for the hypothesis. In the "intermediate case" of the ratio being equal to $C$ , we are in doubt and this is why our decision is random (we decide for the alternative with a probability $\gamma \in (0, 1)$). The choice of $C$ and $\gamma$ is tailored to make the test have exactly a size equal to $\alpha$, since by exhausting the level given in advance to us we are hoping to maximize the power. All this simple reasoning finds its rigorous support in the Fundamental Lemma given above.

## 6.5 Simple $\mathrm{H}_0$ versus composite $\mathrm{H}_1$-the "simple case"

We consider now the more realistic situation where we have a *collection* of alternatives instead of a simple one. This is a new situation, not covered by the Neyman-Pearson Lemma. One simple case can be handled immediately. If we obtain the same size $\alpha$ best critical region for all $\theta$-values in the alternative set then the optimal Neyman-Pearson test $\varphi^*$ that one can construct for *one concrete alternative value* $\theta_1$ will be *uniformly most powerful* (UMP) $\alpha$-size test for the simple hypothesis versus the *collection* of alternatives.

*Example:* $\mathbf{X} = (X_1, X_2, .., X_n)$: i.i.d. $N(\theta, 1)$. Consider $H_0 : \theta = \theta_0 \in R^1$ versus $H_1 : \theta > \theta_0$. We want to find the UMP $\alpha$-test of $H_0$ versus $H_1$, i.e. we want to find a test $\varphi^*$ which is such that for any other test $\varphi \in \Phi_\alpha : E_\theta\varphi^* \geq E_\theta\varphi$ *for all* $\theta > \theta_0$ holds. Take and fix *any* $\theta_1 > \theta_0$ and consider testing $H_0$ versus $\bar{H}_1 : \theta = \theta_1$. Then the Neyman-Pearson (NP) lemma can be applied and after simple transformations, we get the rejection region $S$ of the best (NP test) for $H_0$ versus $\bar{H}_1$ in the form $S = \{\mathbf{x} : \bar{x} \geq \theta_0 + (z_\alpha/\sqrt{n})\}$( which obviously does not depend on the specific $\theta_1$ in the alternative. Then $\varphi^* = \begin{cases} 1 \text{ if } \bar{x} \geq \theta_0 + (z_\alpha/\sqrt{n}), \\ 0 \text{ if } \bar{x} < \theta_0 + (z_\alpha/\sqrt{n}) \end{cases}$ will be the uniformly most powerful $\alpha$-test of $H_0$ versus $H_1$.

## 6.6 Composite $\mathbf{H}_0$ versus composite $\mathbf{H}_1$

In general, for such type of hypothesis testing problems, there is no UMP $\alpha$-size test. But for *some types of distributions* and specific hypotheses/alternatives, like intervals on the real line, one can find UMP $\alpha$-tests:

### 6.6.1 MLR family of distributions

The family $L(\mathbf{x}, \theta), \theta \in R$ has a *monotone likelihood ratio* (MLR) in the statistic $T(\mathbf{X})$ if for any fixed $\theta'$ and $\theta''$ such that $\theta' < \theta''$, it holds that $\frac{L(\mathbf{x},\theta'')}{L(\mathbf{x},\theta')}$ is a *non-decreasing function* of $T(\mathbf{x}) = T(x_1, x_2, .., x_n)$.

**Note:** typical examples are from the one-parameter exponential family: if $f(x, \theta) = a(\theta)b(x)\exp(c(\theta)d(x))$ and $c(\theta)$ is strictly monotone increasing then

$$\frac{L(\mathbf{x}, \theta'')}{L(\mathbf{x}, \theta')} = \frac{a^n(\theta'')}{a^n(\theta')}.\exp\{[(c(\theta'') - c(\theta')]\sum_{i=1}^{n} d(x_i)\}$$

and clearly, this family has a MLR in $T(\mathbf{X})$=$\sum_{i=1}^{n} d(X_i)$.

### 6.6.2 Theorem of Blackwell & Girshick

Suppose $\mathbf{X} \sim L(\mathbf{x}, \theta)$ and the family is with MLR in $T(\mathbf{X})$. Then for testing $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$, the $\alpha$-test $\varphi^*$ with the structure:

$$\varphi^*(\mathbf{x}) = \begin{cases} 1 \text{ if } T(\mathbf{x}) > k \\ \gamma \text{ if } T(\mathbf{x}) = k \\ 0 \text{ if } T(\mathbf{x}) < k \end{cases}$$

($k$ being the upper $\alpha$.100% point of the $P_{\theta_0}$-distribution of $T(\mathbf{X})$) has an increasing power function $E_\theta\varphi^*$( i.e. its power as a function of $\theta$ is increasing) and the test is UMP $\alpha$-test.

**Note:** There is an obvious variant of the theorem: under the same conditions on the family, the test that rejects $H_0 : \theta \geq \theta_0$ in favour of $H_1 : \theta < \theta_0$ when $T(\mathbf{x}) < k$ ($\alpha = P_{\theta_0}(T < k) + \gamma P_{\theta_0}(T = k)$) is the UMP $\alpha$-test.

### 6.6.3 Examples

(see lectures):

## 6.7 Unbiasedness. UMPU $\alpha$-tests.

### 6.7.1 General discussion and definition.

We have already seen that for a one parameter exponential family, for example, a UMP $\alpha$-test for composite alternatives exists. If no UMP test exists, we should think about another criterion to make the optimal choice among possible tests (i.e. a further *restriction* of the set of $\alpha$-tests is necessary in order to find an optimal solution in a *smaller* set of competing tests). For example, one can:

-choose a 'typical' alternative in the set of alternatives and use the most powerful test for that alternative;

-maximize the power locally, by considering only the $\theta$-values from the alternative set that are close to the hypothetical $\theta$-values. This leads to the notion of a *locally most powerful test* (not to be discussed in this course);

-maximize some weighted average of power for the different alternatives.

One more solution is mathematically very attractive and leads to very reasonable tests. This is to restrict ourselves to the set of *unbiased tests.*

**Definition 1.** A test $\varphi$ of $H_0 : \theta \in \Theta_0 (\Theta_0 \subset \Theta)$ versus $H_1 : \theta \in \Theta \backslash \Theta_0$ is an *unbiased* size $\alpha$-test if $E_\theta \varphi \leq \alpha$ for all $\theta \in \Theta_0$ and $E_\theta \varphi \geq \alpha$ for all $\theta \in \Theta \backslash \Theta_0$.

Basically, the above definition of unbiasedness ensures that there exist no alternatives for which acceptance of the hypothesis is more probable than in cases when the null hypothesis is true. This is a very reasonable requirement (don'*t* you think) so that asking in addition for it to be satisfied would not restrict too seriously the set of tests of interest (the majority of reasonable tests of size $\alpha$ would still be allowed to compete).

### 6.7.2 Basic results

**Theorem 6.2.** *Suppose* $\mathbf{X} \sim L(\mathbf{x}, \theta)$ *with* $L(\mathbf{x}, \theta) = (a(\theta))^n \prod_{i=1}^n b(x_i) \exp[c(\theta). \sum_{i=1}^n d(x_i)]$ *and* $T(\mathbf{X}) = \sum_{i=1}^n d(x_i)$. *Then, for testing* $H_0 : \theta_1 \leq \theta \leq \theta_2$ *versus* $H_1 : \theta < \theta_1 or\ \theta > \theta_2$, *the test* $\varphi^*$ *is* :$\varphi^*(\mathbf{x}) = \begin{cases} 1 \ \textit{if}\ T(\mathbf{x}) \notin [c_1, c_2] \\ \gamma_i \ \textit{if}\ T(\mathbf{x}) = c_i, i = 1, 2 \\ 0 \ \textit{if}\ c_1 < T < c_2 \end{cases}$ *where* $c_1, \gamma_1, c_2, \gamma_2$ *are determined by the conditions* $E_{\theta_1} \varphi^* = E_{\theta_2} \varphi^* = \alpha$.

*Moreover,the power function has a minimum somewhere within* $(\theta_1, \theta_2)$ *and is monotone outside* $(\theta_1, \theta_2)$.

Example: negative exponential (to be considered at lecture)

**Theorem 6.3.** *Consider the same family like in the previous Theorem 6.2. Then , for testing $H_0 : \theta = \theta_0$ versus $H_2 : \theta \neq \theta_0$ an UMPU $\alpha$-test exists with the structure:* $\varphi^*(\mathbf{x}) = \begin{cases} 1 \text{ if } T(\mathbf{x}) < c_1 \text{ or } T(\mathbf{x}) > c_2 \\ \gamma_i \text{ if } T(\mathbf{x}) = c_i, i = 1, 2 \\ 0 \text{ if } c_1 < T < c_2 \end{cases}$ *. The constants $c_i, \gamma_i$ satisfy: Power$(\theta_0) = \alpha = E_{\theta_0}\varphi^*$ and $\frac{\partial}{\partial\theta}$Power$(\theta_0) = 0 = \frac{\partial}{\partial\theta} E_\theta \varphi^* |_{\theta=\theta_0}$ .*

**Note:** The latter Theorem 6.3 can be justified as a limiting case of Theorem 6.2 when the interval $[\theta_1, \theta_2]$ collapses to a single point $\theta_0$.

## 6.8 Examples

a) Assume, a "sample" of one observation ($n = 1$) from an exponential family with density $f(x, \theta) = \frac{1}{\theta} \exp(-\frac{x}{\theta}), x > 0$ is available. The parameter $\theta > 0$ is to be tested. One would like to test $H_0 : \theta = 1$ versus $H_1 : \theta \neq 1$.

According to Theorem 6.3, the UMPU $\alpha$-test $\varphi^*$ has the structure:

$\varphi^*(x) = \begin{cases} 1 \text{ if } T < C_1 \text{ or } T > C_2 \\ 0 \text{ if } C_1 \leq T \leq C_2 \end{cases}$ where $T(x) = x$ in this case (one-parameter exponential family, $d(x) = x, n = 1$). We only need to find $C_1$ and $C_2$ in order to uniquely specify the above test. Note that $E_\theta\varphi^* = P_\theta(x \notin (C_1, C_2)) = 1 - \exp(-C_1/\theta) + \exp(-C_2/\theta)$(since the cdf is $F(x, \theta) = 1 - \exp(-x/\theta), x > 0$). The two conditions on $E_\theta\varphi^*$ are:

- $E_\theta\varphi^* |_{\theta=1} = \alpha = 1 - \exp(-C_1) + \exp(-C_2)$
- $\frac{\partial}{\partial\theta} E_\theta\varphi^* |_{\theta=1} = -\frac{C_1}{\theta^2} \exp(-\frac{C_1}{\theta}) + \frac{C_2}{\theta^2} \exp(-\frac{C_2}{\theta}) |_{\theta=1} = -C_1 \exp(-C_1) + C_2 \exp(-C_2) = 0$

We get a system of two equations with respect to $C_1$ and $C_2$. It can be solved numerically (iteratively) given the level $\alpha$ and hence the UMPU $\alpha$-test will be completely specified.

b) If $X_1, X_2, \ldots, X_n$ are i.i.d. $N(\theta, 1)$ then for testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ there exists an UMPU $\alpha$-test (according to Theorem 6.3). We want to show that it coincides with the well-known test $\varphi^*(\mathbf{x}) = \begin{cases} 1 \text{ if } \sqrt{n}|\bar{\mathbf{x}} - \theta_0| \geq z_{\alpha/2} \\ 0 \text{ if } \sqrt{n}|\bar{\mathbf{x}} - \theta_0| < z_{\alpha/2} \end{cases}$

$E_\theta\varphi^* = P_\theta(\bar{\mathbf{x}} \leq C_1 \text{ or } \bar{\mathbf{x}} \geq C_2) = P_\theta\{\sqrt{n}(\bar{\mathbf{x}}-\theta) \leq \sqrt{n}(C_1-\theta) \text{ or } \sqrt{n}(\bar{\mathbf{x}}-\theta) \geq \sqrt{n}(C_2-\theta)\} =$

$P\{N(0, 1) \leq \sqrt{n}(C_1-\theta)\}+P\{N(0, 1) \geq \sqrt{n}(C_2-\theta)\} = \Phi(\sqrt{n}(C_1-\theta))+1-\Phi(\sqrt{n}(C_2-\theta)).$

Here $\Phi$ is the cdf of the standard normal distribution. The two equations that have to be satisfied, are:

$$E_{\theta_0}\varphi^* = \alpha$$

and

$$\frac{\partial}{\partial\theta}E_\theta\varphi^*|_{\theta=\theta_0}=0.$$

This leads us to the following two equations:

$$\Phi(\sqrt{n}(C_1-\theta_0))+1-\Phi(\sqrt{n}(C_2-\theta_0)=\alpha$$
$$\Phi'(\sqrt{n}(C_1-\theta_0))=\Phi'(\sqrt{n}(C_2-\theta_0))$$

From the second equation we get (since $C_1\neq C_2$ and the standard normal density is symmetric around zero) : $C_1+C_2=2\theta_0$. Substituting into the first equation, we get: $2[1-\Phi(\sqrt{n}(C_2-\theta_0))]=\alpha$. The latter relation means that $C_2=\theta_0+\frac{z_{\alpha/2}}{\sqrt{n}}$ and $C_1=2\theta_0-C_2=\theta_0-\frac{z_{\alpha/2}}{\sqrt{n}}$

Hence the form of $\varphi^*$ is indeed $\varphi^*(\mathbf{x})=\begin{cases}1 \text{ if } \sqrt{n}|\bar{\mathbf{x}}-\theta_0|\geq z_{\alpha/2}\\ 0 \text{ if } \sqrt{n}|\bar{\mathbf{x}}-\theta_0|< z_{\alpha/2}\end{cases}$

## 6.9 Locally most powerful tests

Another way to handle the situation in which no UMP test exists is to restrict attention to values of the alternative parameter, i.e. we look at tests which have high power at some *particular alternatives*. In most cases, we consider the behaviour of the power for alternative parameter values that are close to the null hypothesis. One reckons that deriving a test that works well in *such "difficult" situations* when the hypothesis and alternative are close to each other, is most essential for the applications (when hypothesis and alternative are relatively far from each other, hopefully many tests would do a good job).

**Definition 2** (locally most powerful test).

Test $\varphi^*$ with power function $E_\theta\varphi^*$ is (at its size) locally most powerful (LMP) for $H_0:\theta\leq\theta_0$ versus $H_1:\theta>\theta_0$ if for any other test $\varphi$ with $E_{\theta_0}\varphi=E_{\theta_0}\varphi^*$, there exists $\Delta>0$ such that $E_\theta\varphi^*\geq E_\theta\varphi$ for every $\theta\in(\theta_0,\theta_0+\Delta)$.

**Note:** In most practical situations, the tests we consider have differentiable power functions. In such cases, a LMP test will obviously maximize:

$$\frac{\partial}{\partial\theta}E_\theta\varphi_{|\theta=\theta_0}=\int..\int\varphi(\mathbf{x})\frac{\partial}{\partial\theta}L(\mathbf{x},\theta)_{|\theta=\theta_0}d\mathbf{x}\text{ under the constraint }\int\varphi(\mathbf{x})L(\mathbf{x},\theta_0)d\mathbf{x}=\alpha=E_{\theta_0}\varphi.$$

But note that the structure of this optimization problem is the same as in the NP Lemma. So, proceeding along the same lines (please, reread the proof of the NP Lemma) we arrive at following optimal solution:

$$\varphi^*(\mathbf{x})=\begin{cases}1 \text{ if } (\partial/\partial\theta)L(\mathbf{x},\theta)_{|\theta=\theta_0}>kL(\mathbf{x},\theta_0)\\ 0 \text{ if } (\partial/\partial\theta)L(\mathbf{x},\theta)_{|\theta=\theta_0}\leq kL(\mathbf{x},\theta_0)\end{cases}$$

and the above optimal test is unique. If we denote, as usual, $V(\mathbf{x},\theta)$ to be the score function then obviously $\varphi^*$ has the following simple and appealing structure:

$$\varphi^*(\mathbf{x}) = \begin{cases} 1 \text{ if } V(\mathbf{x}, \theta_0) > k \\ 0 \text{ if } V(\mathbf{x}, \theta_0) \leq k \end{cases}$$

**Example.** $\mathbf{X}$=$(X_1, X_2, .., X_n)$ be i.i.d. Cauchy$(\theta, 1)$ variables with density $f(x, \theta) = \frac{1}{\pi} \cdot \frac{1}{1+(x-\theta)^2}, x \in R, \theta$ being an unknown parameter. Test $H_0 : \theta \leq 0$ versus $H_1 : \theta > 0$ (see discussion at lecture).

## 6.10 Likelihood ratio tests

We have now looked at a variety of *ad hoc* criteria for optimal tests, including unbiasedness, locally most powerful etc. Other criteria, not discussed here, include similarity, invariance etc. The reason to consider so many different criteria comes from the lack of universal criterion for comparing sets of models. But if we agree with the strong likelihood principle and believe the set of models to be best represented by its most likely member given the observed data, we can arrive at a relatively simple and universal procedure also in a testing context.

When no points in the parameter space specified by $H_0$ are preferred to others, the likelihood function can be maximized under the null and alternative hypotheses:

### 6.10.1 General formulation

Assume, for example, that $\Theta \subseteq R^{r+s} = R^k; \Theta_0 = \{\theta \in \Theta \mid \theta_1 = \theta_{10}, .., \theta_r = \theta_{r0}\}, H_0 : \theta \in \Theta_0$and $H_1 : \theta \in \Theta \backslash \Theta_0$. Let us define the statistic

$$\lambda(\mathbf{X}) = \frac{\sup_{\theta \in \Theta_0} L(\mathbf{X}, \theta)}{\sup_{\theta \in \Theta} L(\mathbf{X}, \theta)}$$

which is obviously in the interval [0,1]. Intuitively, it makes sense to define the rejection region as $S = \{\mathbf{x} \mid \lambda(\mathbf{x}) \leq C\}$ for a certain constant C. However, the optimum properties of likelihood ratios for simple hypotheses, as discussed in the NP lemma, no longer apply, except asymptotically. In addition, the exact distribution of $\lambda(\mathbf{X})$(that is needed to determine the constant $C$) is also difficult to obtain without using asymptotic approximations. Thus, the fact that the *deviance statistic* has an asymptotic distribution which is well known is generally used to obtain significance levels or to get the constant $C$ for $\alpha$ given in advance.

Bearing in mind the derivations related to the deviance in Lecture 5, we can formulate the following, slightly more general, statement:

**Theorem 6.4.** *Let $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ be a random sample from $f(x, \theta), \theta \in R^{r+s}$. Suppose the regularity conditions for consistency and asymptotic normality of MLE under $H_0$ and $H_1$ hold. Then under $H_0 : -2\ln\lambda(\mathbf{X}) \to^d \chi^2_r$ (r is interpreted as the difference between the number of free parameters specified by $\theta \in \Theta$ and the number of free parameters specified by $\theta \in \Theta_0$).*

**Examples:**

i) For testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ for a sample of $n$ i.i.d. $N(\mu, \sigma^2)$, ($\sigma^2$ known) one has $-2 \ln \lambda(\mathbf{x}) = \frac{n(\bar{\mathbf{x}}-\mu_0)^2}{\sigma^2} = D(\mu_0) \sim \chi^2_1$ (and the result is *exact).*

ii) Normal sample with both $\mu$ and $\sigma^2$ unknown. Testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$. The MLR test is equivalent to the classical $t$-test with a rejection region

$$S = \{\mathbf{x} || \frac{(\bar{\mathbf{x}} - \mu_0)\sqrt{n}}{s} |\geq k\}$$

Here $s = (s^2)^{1/2} = [\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{\mathbf{x}})^2]^{1/2}$ and to make the size equal to $\alpha$, we must choose $k = t_{\alpha/2,(n-1)}$(the upper $(\alpha/2).100\%$ point of the $t$ distribution with $(n-1)$ degrees of freedom.)

Thus, a popular test (known to be also the UMPU $\alpha$-test for the above problem) could be constructed using the Likelihood Ratio Test construction.

We shall only note here that the asymptotic distribution result in Theorem 6.4 is formulated under the hypothesis only but more sophisticated reasoning can be used to derive the asymptotic distribution under alternative parameter values, too. This distribution is a non-central $\chi^2$ and can be used for (approximate asymptotic) power computations. We omit the details.

## 6.11 Alternatives to the GLRT.

The GLRT is widely used but there are circumstances where other test procedures may be preferred. We define two of these test procedures for the case $s = 0$, that is, $k = r$.

### 6.11.1 Score test.

It uses $S = V(\mathbf{X}, \theta_0)' I^{-1}_{\mathbf{X}}(\theta_0) V(\mathbf{X}, \theta_0)$ instead of $-2 \log \lambda(\mathbf{X})$.

### 6.11.2 Wald test.

It uses $(\hat{\theta} - \theta_0)' I_{\mathbf{X}}(\hat{\theta})(\hat{\theta} - \theta_0)$ instead of $-2 \log \lambda(\mathbf{X})$ where $\theta_0$ is the hypothetical vector and $\hat{\theta}$ is the MLE.

For both the Score test, and the Wald test, the *asymptotic* distribution of the test statistic under $H_0$ is the *same* as the distribution of the GLRT statistic (that is, chi-square with $r = k$ degrees of freedom). Score tests have a numerical advantage in comparison to GLRT and the Wald test, that they do *not* require the MLE to be calculated! Specifically in the econometrics literature, the Score test is known as **Lagrange Multiplier Test**. The name comes from its alternative derivation in which the Likelihood function is maximized subject to the restrictions of $H_0$ and the maximization method uses Lagrange multipliers.

Much research has been devoted to selecting one of the three tests as a preferred test in a particular situation for relatively small sample sizes. We will only say that this is a difficult task and will not be discussed in our course.

**Exercise.** Let X be the number of successes in a binomial experiment with a probability of success $p$. We wish to test $H_0 : p = p_0$ versus $H_1 : p \neq p_0$. Denote $\hat{p} = \frac{X}{n}$. Then the Score statistic is

$$S = \frac{(X - np_0)^2}{np_0(1 - p_0)},$$

the Wald statistic is

$$W = \frac{(X - np_0)^2}{n\hat{p}(1 - \hat{p})}$$

and

$$-2 \ln \lambda = 2\{X \ln \frac{\hat{p}(1 - p_0)}{p_0(1 - \hat{p})} + n \ln \frac{1 - \hat{p}}{1 - p_0}\}.$$