Theory of Hypothesis Testing

Inference is divided into two broad categories:

- Estimation
- Testing

Chapter 7 devoted to point estimation. Will discuss interval estimation in Chapter 9.

Testing is the subject of Chapter 8.

Definition 14 A hypothesis is a statement about population parameters.

There are usually two hypotheses, called the null and alternative hypotheses.
The null hypothesis is denoted $H_0$, and the alternative is denoted $H_1$.

Usually, $H_1$ is the negation, or complement, of $H_0$. Often the two hypotheses have the form

$$H_0 : \theta \in \Theta_0 \quad H_1 : \theta \in \Theta^c_0.$$ 

$\Theta_0 \subset \Theta$ and $\Theta^c_0$ is the complement of $\Theta_0$.

A test of the hypotheses is carried out as follows:

We are to observe a value of random vector $X$ having distribution $f(\cdot | \theta)$. Before observing the data, we formulate a decision rule of the form:

Reject $H_0$ if the observed value $x$ of $X$ is in some set $R$.

Don’t reject $H_0$ if $x \in R^c$. 

126
$R$ is called the rejection region or critical region.

We could also phrase the decision rule in terms of a sufficient statistic.

Any statistic whose values are used to define the rejection region is called a test statistic.

Suppose that $\Theta_0$ contains only a single element. Then $H_0 : \theta \in \Theta_0$ is called a simple hypothesis. Otherwise it is called a composite hypothesis.

We have two hypotheses $H_0$ and $H_1$. After observing the data we either take action $a_0$ or $a_1$. 127
\[ a_0 = "\text{accept, or fail to reject, } H_0" \]

\[ a_1 = "\text{accept } H_1" \]

**Consequences of actions**

<table>
<thead>
<tr>
<th>True state of nature</th>
<th>( \theta \in \Theta_0 )</th>
<th>( \theta \in \Theta_0^c )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Action</td>
<td>( a_0 )</td>
<td>( a_1 )</td>
</tr>
<tr>
<td>Correct</td>
<td>Correct</td>
<td>Type I Error</td>
</tr>
<tr>
<td>Type II Error</td>
<td>Type II Error</td>
<td>Correct</td>
</tr>
</tbody>
</table>

128
Example 29  Suppose $X_1, \ldots, X_n$ is a random sample from $N(\theta, 1)$. We're interested in testing the hypotheses

$H_0 : \theta \leq 10 \quad H_1 : \theta > 10.$

$\Theta = (-\infty, \infty) \quad \Theta_0 = (-\infty, 10]$ 

$\Theta_0^c = (10, \infty)$

For example:

- $X_i$ might be a measure of product quality when a new process is used.

- The average quality measure using the old process is 10.

- $H_0$ says that the new process is no better than the old.

- $H_1$ says the new process is better than the old.
A sufficient statistic in this model is

\[ \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i. \]

Also, we know that \( \bar{X} \) is both the MLE and the UMVUE of \( \theta \). A sensible test would have the following form:

**Take action** \( a_1 \) if \( \bar{x} \geq c_n \), and

**take action** \( a_0 \) if \( \bar{x} < c_n \),

where \( \bar{x} \) is the observed value of \( \bar{X} \) and \( c_n \) is some constant larger than 10.

**Type I error:** Conclude new process is better when it isn’t.

**Type II error:** Conclude new process is no better than the old when in fact it is better.
A general approach to testing

When \( \theta \) is the true parameter value, let

\[
\beta(\theta) = P_{\theta}(\text{rejecting } H_0) = P_{\theta}(X \in R),
\]

where \( R \) is the rejection region of the test. We have

\[
\beta(\theta) = \begin{cases} 
P_{\theta}(\text{Type I error}), & \text{if } \theta \in \Theta_0 \\
1 - P_{\theta}(\text{Type II error}), & \text{if } \theta \in \Theta_0^c.
\end{cases}
\]

Note that for \( \theta \in \Theta_0^c \),

\[
\beta(\theta) = P_{\theta}(\text{making correct decision}).
\]

The function \( \beta \) restricted to \( \Theta_0^c \) is called the power function of the test, and \( \beta(\theta) \) is called the power at \( \theta \).
Define

\[ \alpha = \sup_{\theta \in \Theta_0} \beta(\theta). \]

\( \alpha \) is called the **size** of the test. Casella and Berger say that the test is of **level** \( \eta \) if \( \alpha \leq \eta \).

**Notation** A test can be characterized by a **test function** \( \phi \).

\[ \phi(x) = \begin{cases} 1, & \text{if } x \in R \\ 0, & \text{if } x \in R^c \end{cases} \]

Note that \( \beta(\theta) = E_{\theta}[\phi(X)] \).

**Example 29 (continued)** Suppose we use a test function \( \phi \) as follows:

\[ \phi(x) = \begin{cases} 1, & \text{if } \bar{x} \geq 10 + \frac{1.645}{\sqrt{n}} \\ 0, & \text{otherwise.} \end{cases} \]
\[ \beta(\theta) = P_\theta \left( \bar{X} \geq 10 + \frac{1.645}{\sqrt{n}} \right) \]
\[ = P_\theta \left( \frac{\bar{X} - \theta}{1/\sqrt{n}} \geq \sqrt{n}(10 - \theta) + 1.645 \right) \]
\[ = P(Z \geq \sqrt{n}(10 - \theta) + 1.645), \]

where \( Z \sim N(0, 1) \).

Remarks

- \( \beta(10) = 0.05 \) for each \( n \).

- \( \beta(\theta) \) increases monotonically, from 0 at \( \theta = -\infty \) to 1 at \( \theta = \infty \).

- So, the size of the test is 0.05, no matter the value of \( n \).
The numbers beside the curves indicate sample size, $n$. 
Goal in constructing a test

Make $\alpha$ as small as possible while making the power as large as possible.

Clearly we can construct a test with $\alpha = 0$ by taking $R = \text{empty set}$! Also, we can make the power equal to 1 for all $\theta \in \Theta^c$ by using a test with $R = \text{sample space}$.

However, neither of these tests attains the goal of making $\alpha$ small and power large.

Our main approach will be to consider tests with level of significance set at some desired value (such as 0.05), and to try and select from these tests one that maximizes power.
By taking $\alpha$ to be fairly small, this approach implicitly says that a Type I error is more serious than a Type II error.

Choosing $\alpha$ to be small makes a Type I error unlikely, but may mean that our test has low power, and hence a high probability of Type II error.

**Example 29 (continued)** The way we set up the hypotheses, a Type I error means concluding that the new process is better when it really isn’t.

Taking $\alpha$ small may reflect our reluctance to change to the new process unless there’s very convincing evidence that it’s really better.
**Type II error:** missing out on a better process.

**Type I error:** make a (perhaps costly) switch to a new process that is no better than the old one.

If the latter error is more serious, it would be advisable to choose $\alpha$ small and to live with the resulting Type II error probability.

**Testing a simple null against a simple alternative**

**Definition 15**  A test $\phi$ is said to be a most powerful test of size $\alpha$ for testing

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

if $E_{\theta_0}[\phi(X)] = \alpha$ and for any other test $\phi^*$ of size $\alpha$, $E_{\theta_1}[\phi(X)] \geq E_{\theta_1}[\phi^*(X)]$. 

137
We observe a value of $X$, whose distribution is either $f(x|\theta_0)$ or $f(x|\theta_1)$.

Suppose we observe $X$ to be $x$. Consider the likelihood ratio

$$L(x|\theta_0, \theta_1) = \frac{f(x|\theta_1)}{f(x|\theta_0)}.$$

It would be sensible to reject $H_0$ in favor of $H_1$ only when $L(x|\theta_0, \theta_1)$ is sufficiently large.

**Neyman-Pearson Lemma** Let $X$ be a random vector with pdf or pmf $f(x|\theta)$, where $\theta \in \Theta = \{\theta_0, \theta_1\}$. Define the hypotheses

$$H_0 : \theta = \theta_0 \quad H_1 : \theta = \theta_1.$$
a) Any test \( \phi \) of the form

\[
\phi(x) = \begin{cases} 
1, & \text{if } f(x|\theta_1) > kf(x|\theta_0) \\
0, & \text{if } f(x|\theta_1) < kf(x|\theta_0)
\end{cases}
\]

with \( E_{\theta_0}[\phi(X)] = \alpha \) is a most powerful level \( \alpha \) test of \( H_0 \) vs. \( H_1 \).

b) If there exists a test of the form in a) with \( k > 0 \), then any other most powerful level \( \alpha \) test, call it \( \phi^* \), has size \( \alpha \) and is such that \( \phi(x) = \phi^*(x) \) for almost all \( x \) in \( \{x : f(x|\theta_1) \neq kf(x|\theta_0)\} \).