# Hypothesis Testing

## BS2 Statistical Inference, Lecture 11
## Michaelmas Term 2004

Steffen Lauritzen, University of Oxford; November 15, 2004

# Hypothesis testing

We consider a family of densities $\mathcal{F} = \{f(x;\theta), \theta \in \Theta\}$ on a sample space $\mathcal{X}$ and wish to test the *null hypothesis*

$$H_0 : \theta \in \Theta_0$$

vs. the *alternative hypothesis*

$$\theta \in \Theta_A = \Theta \setminus \Theta_0,$$

where $\Theta_0 \subseteq \Theta$.

Within the Neyman–Pearson theory, a (non-randomised) test is determined by a *critical region* $C$, so that $H_0$ is *rejected* if $X \in C$ and $H_0$ is *accepted* if $X \notin C$.

# Power and level of significance

The probability of rejecting the hypothesis in a test with critical region $C$ is the *power function*:

$$\phi(\theta) = P_\theta(X \in C).$$

The *size* or *significance level* $\alpha$ of the test is the largest possible probability of making an *error of type I*, i.e. of rejecting a true null hypothesis:

$$\alpha = \sup_{\theta \in \Theta_0} P_\theta(X \in C) = \sup_{\theta \in \Theta_0} \phi(\theta).$$

Thus the size of the test is the maximal power under the null hypothesis.

If $\theta \in \Theta_A$, the probability of making an *error of type II* is

$$\beta = \beta(\theta) = 1 - \phi(\theta)$$

i.e. the probability of accepting $H_0$ when $H_0$ is false.

# Constructing critical regions

A common way of constructing critical regions is to use a *test statistic* $T = t(X)$ in such a way that large values of $T$ are good indicators of the null hypothesis being false.

We then construct a test with critical region

$$C = \{x \mid t(x) > t_0\}.$$

If we choose the *critical value* $t_0$ to satisfy

$$\alpha = \sup_{\theta \in \Theta_0} P_\theta(T \geq t_0)$$

we get a test of size $\alpha$.

# $p$-values and significance levels

The *$p$-value* when $t(X) = t$ has been observed is

$$p = \sup_{\theta \in \Theta_0} P_\theta(T \geq t)$$

i.e. *the largest possible probability of obtaining a value of $T$ which is at least as extreme as the observed, under the assumption that the null hypothesis is true.*

If the $p$-value is very small this will then be taken as strong evidence against the null hypothesis, corresponding to rejecting the hypothesis exactly when $p \leq \alpha$.

In this sense, the $p$-value of a test outcome is the *largest possible size needed to accept the null hypothesis.*

# Example

Consider a sample $X = (X_1, \ldots, X_n)$ from a normal distribution $\mathcal{N}(\mu, \sigma^2)$, with $\mu$ and $\sigma^2 > 0$ unknown sand consider the hypothesis

$$H_0 : \sigma^2 \leq \mu^2 \text{ vs. the alternative } H_0 : \sigma^2 > \mu^2.$$

One could directly choose to use the test-statistic

$$t(X) = S^2/\bar{X}^2$$

where

$$\bar{X} = \frac{1}{n}\sum X_i, \quad S^2 = \frac{1}{n-1}\sum_i (X_i - \bar{X})^2.$$

Note that $t'(X) = S^2 - k\bar{X}^2$ might as well have been chosen but this leads to quite different critical regions.

# Optimal testing

The Neyman–Pearson theory is concerned with finding optimal tests, i.e. *optimal critical regions* or, equivalently, *optimal test statistics.*

Most commonly one attempts to *maximize power under the alternative* while *keeping the size under control* or, in other words, minimize the type II error probability while controlling the probability of an error of type I.

A more symmetric approach is to *minimize a given linear combination of type I and type II error probabilites,* but this leads essentially to the same testing procedures.

# Controversy between Fisher and Neyman

One of the deepest controversies in the history of statistics is that between J. Neyman and R. A. Fisher which went on as long as both were alive.

The fierce discussions between these two giants of statistical science were held in an ever more unpleasant and relentless tone, orally as well as in public writing.

The dispute was concerned with many aspects of statistics, but in particular with that of hypothesis testing.

This divide is, in my personal opinion, much deeper than that between Bayesian and Frequentist inference.

# Cournot's principle

Briefly and inadequately, the dispute on testing was rooted in Fisher's strong opposed to the mechanical 'acceptance/rejection' which plays a dominating role in the Neyman–Pearson theory.

Fisher would rather *report the $p$-values* and then *let the scientist interpret the evidence.*

The $p$-value itself may best be interpreted according towhat we could call *Cournot's principle:*

*Events with small probability do not happen!*

Thus, if $p$ is small, the null hypothesis cannot be sustained and must therefore be considered as falsified.

# Testing in different contexts

Personally, I find point of view behind the Neyman–Pearson theory appropriate when testing has to be performed repeatedly, for example in quality control, say when a low propotion of defective items must be maintained by routine inspection procedures, and decisions about accepting or rejecting batches of items must be taken.

For scientific inference, the decision aspect is less prominent and I find concepts such as critical region, power and size to be less illuminating.

Hypothesis testing is made in many different contexts and each context may emphasize particular aspects of the mathematical theory as being more and less relevant.

# Neyman–Pearson lemma

This fundamental lemma in the theory of hypothesis testing is concerned with the case of a *simple hypothesis* having $\Theta_0 = \{\theta_0\}$ vs. a *simple alternative* $\Theta_A = \{\theta_1\}$.

In this case it holds that *any test with critical region of the form*

$$C = \left\{ x \mid \frac{L(\theta_1; x)}{L(\theta_0; x)} > K \right\}$$

*is optimal of its size, i.e. for fixed size* $\alpha = P_{\theta_0}(C)$*, it has maximal power* $\phi(\theta_1) = P_{\theta_1}(C)$ *under the alternative.*

In other words, if the observation is much more likely under the alternative hypothesis than under the null, we reject the null hypothesis. This test is the *Likelihood Ratio Test* (LRT).

# LRT in exponential families

The Likelihood Ratio Test has a simple form in a one-dimensional exponential family. We get

$$\log \frac{L(\theta_1; x)}{L(\theta_0; x)} = (\theta_1 - \theta_0)t(x) + c(\theta_0) - c(\theta_1).$$

So, *if $\theta_1 > \theta_0$ the critical region has the form*

$$C = \{x \mid t(x) > t_0\},$$

where $T = t(x)$ is the canonical sufficient statistic. If $\theta_1 < \theta_0$ the inequality in the expression for the critical region must be reversed.

# One-sided hypothesis and alternative

Two important issues about the LRT for a simple hypothesis vs. a simple alternative should be noted:

- The critical region has a simple form, i.e. a one-sided interval for the canonical statistic;
- The critical region does not depend on the specific values $(\theta_0, \theta_1)$ as long as $\theta_0 < \theta_1$ (or the converse).

It follows that *the critical region for the LRT in a one-dimensional canonical exponential family is also optimal for the hypotheses*

$$H_0 : \theta \leq \theta_0 \text{ vs. the alternative } H_A : \theta > \theta_0,$$

i.e. for a *one-sided alternative* to a *one-sided hypothesis.*

# Problems with the Neyman–Pearson theory

A major weakness of the theory is that in other than those very special cases just mentioned, there is typically no optimal (Uniformly Most Powerful) test.

This has been sought remedied by demanding in addition that a test should be *unbiased*

$$\phi(\theta) \geq \alpha \text{ for } \theta \in \Theta_A,$$

*$\alpha$-similar*

$$\phi(\theta) = \alpha \text{ for all } \theta \in \Theta_0,$$

or have certain *invariance* properties.

These additional demands do to some extent ensure the existence of optimal tests in many cases, but far from all.