STAT0008 Lecture 8
Hypothesis Testing II

Dr. Aidan O'Keeffe

Department of Statistical Science
University College London

26th November 2018

Outline

- Structure of a hypothesis test
- Power function
- Neyman Pearson lemma – most powerful test for simple hypotheses
- Uniformly most powerful tests
- Three important parametric hypothesis tests:
  - Generalised likelihood ratio test
  - Wald test
  - Score test

Recap

Recall from Lecture 7 that we use the following steps when performing a hypothesis test

1. Decide on a distribution for the data (i.e. choose $X_i \sim \mathcal{D}(\theta)$).
2. State hypotheses (H_0 and H_1 – in terms of θ).
3. Assume that H_0 is true.
4. Decide whether or not the observed data, $\mathbf{x}$, are compatible with H_0, by assessing whether or not $\mathbf{x}$ belongs to the critical region $\mathcal{C}$.
5. Based on 4., either reject H_0 (and conclude that H_1 holds) or retain H_0.

Decision Function

When conducting a parametric hypothesis test, we state hypothesis in terms of the parameter of interest, θ, as

$$\mathsf{H}_0 : \theta \in \Theta_0 \quad \text{versus} \quad \mathsf{H}_1 : \theta \in \Theta_1.$$

If $\mathcal{C}$ is the critical region of the test, then we can define a **decision function** $\phi(\mathbf{x})$ such that

$$\phi(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x} \in \mathcal{C}; \\ 0 & \text{if } \mathbf{x} \notin \mathcal{C}. \end{cases}$$

In words, $\phi(\mathbf{x})$ is simply an indicator variable that takes the value 1 when we reject H_0 and takes the value 0 when we do not reject H_0.

Testing Framework

The following table summarises the possible decisions that we might take when performing a hypothesis test.

	True state of nature	
	$\theta \in \Theta_0$ (H_0 true)	$\theta \in \Theta_1$ (H_1 true)
Decision	$\phi(\mathbf{x}) = 0$ (Retain H_0)	$\phi(\mathbf{x}) = 0$ (Type II Error)
	$\phi(\mathbf{x}) = 1$ (Type I Error)	$\phi(\mathbf{x}) = 1$ (Reject H_0)

Recall that the **size** (or significance level) of the test is defined as the probability of a Type I error. Here, size α implies that

$$\alpha = \sup_{\theta \in \Theta_0} \mathbb{P}(\mathbf{X} \in \mathcal{C} \mid \theta).$$

The **power** of the test is defined as

$$1 - \beta = 1 - \mathbb{P}(\text{Type II Error}).$$

In order to calculate the power of a test, we consider the **power function**.

Power Function

For a **parametric hypothesis** test (i.e. a hypothesis test with hypotheses that involve some parameter $\theta \in \Theta$) with a critical region $\mathcal{C}$, the **power function** is defined

$$W(\theta) = \mathbb{P}(\mathbf{X} \in \mathcal{C} \mid \theta).$$

Note that the power function is function of θ and gives the probability of rejecting H_0 (i.e. the event $\mathbf{X} \in \mathcal{C}$) for a given value of θ.

Note also that both the size and the power of the test can be specified in terms of the power function since

$$\alpha = \sup_{\theta \in \Theta_0} W(\theta);$$

$$1 - \beta = \sup_{\theta \in \Theta_1} W(\theta).$$

As mentioned in Lecture 7, we usually set a significance level α and then choose a test that has a high power. But how do we choose such a test...?

The Neyman Pearson Lemma

Suppose that we have data $\mathbf{X}$ with pdf/pmf f and we wish to test the **simple hypotheses**

$$\mathsf{H}_0\colon f = f_0 \;\text{ versus }\; \mathsf{H}_1\colon f = f_1.$$

Then if $\mathcal{L}_0(\mathbf{X})$ and $\mathcal{L}_1(\mathbf{X})$ are the corresponding likelihood functions under H_0 and H_1, respectively, the **most powerful test** of size $\leq \alpha$ has the critical region

$$\mathcal{C} = \left\{ \mathbf{x}\colon \frac{\mathcal{L}_1(\mathbf{x})}{\mathcal{L}_0(\mathbf{x})} > k \right\}$$

where k is such that

$$\mathbb{P}(\mathbf{X} \in \mathcal{C} \mid \mathsf{H}_0) = \alpha.$$

We note that a test of this type is a **likelihood ratio test**.

Proof of the Neyman Pearson Lemma

Proof of the Neyman Pearson Lemma

Uniformly Most Powerful Tests

A uniformly most powerful test is a hypothesis test of a given size, α, that is optimal (i.e. most powerful) for any value within the parameter space Θ_1 (the parameter space under H_1).

More formally, a uniformly most powerful (UMP) test of size α is such that

(i) The test has size α.

(ii) The power function, $W(\theta)$, is as large as possible for all $\theta \in \Theta_1$.

Sometimes, a uniformly most powerful test may not exist. In many situations, likelihood ratio tests are uniformly most powerful.

Example 1: Normal Distribution LRT

Suppose that $X_1, \ldots, X_n$ are independent and identically distributed random variables with $X_i \sim \mathcal{N}(\mu, \sigma^2)$ with σ known. We wish to test

$$\mathsf{H}_0 \colon \mu = \mu_0 \quad \text{versus} \quad \mathsf{H}_1 \colon \mu = \mu_1 \quad (\text{with } \mu_1 > \mu_0).$$

(a) Construct the most powerful test of size α.

(b) Is the test that you constructed in (a) uniformly most powerful? Justify your answer.

Example 1: Normal Distribution LRT

Example 1: Normal Distribution LRT

Generalised Likelihood Ratio Test

When forming the critical region

$$\mathcal{C} = \left\{ \mathbf{x} \colon \frac{\mathcal{L}_1(\mathbf{x})}{\mathcal{L}_0(\mathbf{x})} > k \right\}$$

we reject H_0 when the likelihood function under the assumption of H_1 (the numerator) is significantly larger than the likelihood function under the assumption of H_0.

This makes sense - a model that results in large valued likelihood function would imply that the data are more consistent with this model than a model with a smaller valued likelihood function.

But what about a situation where we do not have simple hypotheses?

Generalised Likelihood Ratio Test

Suppose that we have $X_1, \ldots, X_n$ with a probability distribution parameterised by θ and corresponding likelihood function $\mathcal{L}(\theta \mid \mathbf{X})$.

Suppose we wish to test the null hypothesis $\theta = \theta_0$ against a **general alternative**. That is

$$\mathsf{H}_0 \colon \theta = \theta_0 \quad \text{versus} \quad \mathsf{H}_1 \colon \theta \neq \theta_0 \quad (\text{i.e. } \theta \in \Theta, \theta \neq \theta_0).$$

Then the test statistic for the **generalised likelihood ratio test** is

$$2 \log \left[\frac{\sup_{\theta \in \Theta} \mathcal{L}(\theta \mid \mathbf{X})}{\mathcal{L}(\theta_0 \mid \mathbf{X})} \right] = 2 \left[\ell(\hat{\theta} \mid \mathbf{X}) - \ell(\theta_0 \mid \mathbf{X}) \right].$$

Under regularity conditions and for large n

$$2 \left[\ell(\hat{\theta} \mid \mathbf{X}) - \ell(\theta_0 \mid \mathbf{X}) \right] \sim \chi^2_1$$

where $\hat{\theta}$ is the maximum likelihood estimator of θ.

Generalised Likelihood Ratio Test Statistic - Proof (non-examinable)

For notational simplicity, we write the log-likelihood function as

$$\ell(\theta \mid \mathbf{X}) = \ell(\theta).$$

We Taylor expand the log-likelihood about the mle $\hat{\theta}$ to obtain

$$\ell(\theta) \approx \ell(\hat{\theta}) + (\theta - \hat{\theta}) \left(\frac{\partial \ell(\theta)}{\partial \theta} \Big|_{\theta=\hat{\theta}} \right) + \frac{1}{2}(\theta - \hat{\theta})^2 \left(\frac{\partial^2 \ell(\theta)}{\partial \theta^2} \Big|_{\theta=\hat{\theta}} \right).$$

By definition

$$\frac{\partial \ell(\theta)}{\partial \theta} \Big|_{\theta=\hat{\theta}} = 0$$

because this is the first derivative of the log-likelihood function evaluated at the mle (i.e. a solution to the score equation). Hence

$$\ell(\theta) \approx \ell(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^2 \left(\frac{\partial^2 \ell(\theta)}{\partial \theta^2} \Big|_{\theta=\hat{\theta}} \right). \tag{1}$$

Generalised Likelihood Ratio Test Statistic - Proof (non-examinable)

Re-arranging (1), we obtain

$$2\left[\ell(\hat{\theta}) - \ell(\theta)\right] \approx (\hat{\theta} - \theta)^2 \left(-\frac{\partial^2 \ell(\theta)}{\partial \theta^2}\bigg|_{\theta=\hat{\theta}}\right). \qquad (2)$$

Since the mle, $\hat{\theta}$, is consistent for θ (Lecture 6) it follows that, for large n, using the law of large numbers

$$-\frac{\partial^2 \ell(\theta)}{\partial \theta^2}\bigg|_{\theta=\hat{\theta}} \approx \mathcal{I}(\theta).$$

Then (2) may be re-written as

$$\begin{aligned} 2\left[\ell(\hat{\theta}) - \ell(\theta)\right] &\approx (\hat{\theta} - \theta)^2 \mathcal{I}(\theta) \\ &= \left(\frac{\hat{\theta} - \theta}{1/\sqrt{\mathcal{I}(\theta)}}\right)^2 \qquad (3) \end{aligned}$$

Generalised Likelihood Ratio Test Statistic - Proof (non-examinable)

Recall, from Lecture 6, the asymptotic distribution of the maximum likelihood estimator, $\hat{\theta}$, is

$$\hat{\theta} \sim \mathcal{N}\left(\theta, \frac{1}{\mathcal{I}(\theta)}\right).$$

Therefore, in (3)

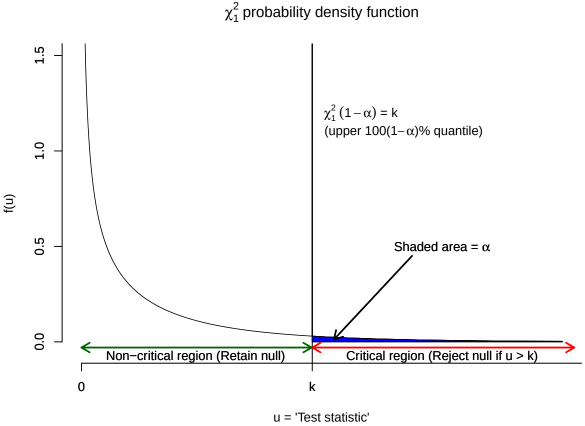
$$\frac{\hat{\theta} - \theta}{1/\sqrt{\mathcal{I}(\theta)}} \sim \mathcal{N}(0, 1)$$

and it follows that

$$2\left[\ell(\hat{\theta}) - \ell(\theta)\right] \sim \chi_1^2.$$

Generalised Likelihood Ratio Test

The generalised likelihood ratio test statistic has a chi-squared distribution with one degree of freedom under H_0. For a test of size α, we reject or retain H_0 by comparing the value of the test statistic to the upper $100(1-\alpha)\%$ quantile of a χ^2_1 distribution.



Generalised Likelihood Ratio Test: Example

Suppose that $X_1, \ldots, X_n$ are independent and identically distributed Bernoulli(θ) random variables.

(a) Construct a size α generalised likelihood ratio test of the hypotheses

$$\mathsf{H}_0\colon \theta = \theta_0 \quad \text{versus} \quad \mathsf{H}_1\colon \theta \neq \theta_0.$$

(b) In the run-up to a city's mayoral election, a polling company hypothesises that 30% of eligible voters intend to vote for the incumbent mayor. When a simple random sample of 80 eligible voters was conducted (with no non-response), 18 voters said that that they would vote for the incumbent mayor. Test the polling company's null hypothesis using a generalised likelihood ratio test at the 5% level.

Generalised Likelihood Ratio Test: Example

Generalised Likelihood Ratio Test: Example

Generalised Likelihood Ratio Test: Example

The Wald Test

When conducting likelihood ratio tests, the critical region is formed using likelihood (or log-likelihood) functions.

However, given a null hypothesis $\mathsf{H}_0 : \theta = \theta_0$ we could form a test statistic using the distribution of the maximum likelihood estimator, $\hat{\theta}$, directly. In words, we might think about forming a test where we reject H_0 if $\hat{\theta}$ 'lies far away from' θ_0. In other words, where $\hat{\theta}$ is 'extreme' when compared with θ_0.

From Lecture 6, we know that, under regularity conditions, the maximum likelihood estimator, $\hat{\theta}$, is such that

$$\hat{\theta} \sim \mathcal{N}\left(\theta, \frac{1}{\mathcal{I}(\theta)}\right).$$

The Wald Test

Then, we wish test the hypotheses

$$\mathsf{H}_0 \colon \theta = \theta_0 \;\; \text{versus} \;\; \mathsf{H}_1 \colon \theta \neq \theta_0.$$

We assume that H_0 is true (as usual) and it follows that

$$\sqrt{\mathcal{I}(\theta_0)}(\hat{\theta} - \theta_0) \sim \mathcal{N}(0, 1).$$

Then, squaring the above, the test statistic for a Wald test of these hypotheses is

$$\mathcal{I}(\theta_0)(\hat{\theta} - \theta_0)^2.$$

For large n, this test statistic has a χ_1^2 distribution under the null hypothesis. In other words,

$$\mathcal{I}(\theta_0)(\hat{\theta} - \theta_0)^2 \sim \chi_1^2.$$

The Wald Test

Thus, when using the Wald test, we reject H_0 when values of

$$\mathcal{I}(\theta_0)(\hat{\theta} - \theta_0)^2$$

are extreme.

For a test of size α, we reject or retain H_0 by comparing the value of the test statistic to the upper $100(1-\alpha)\%$ quantile of a χ^2_1 distribution.

Note, in some situations, $\mathcal{I}(\theta_0)$ may be difficult to evaluate. Then, we might replace $\mathcal{I}(\theta_0)$ in the above test statistic with $\mathcal{J}(\hat{\theta})$ where $\mathcal{J}(\hat{\theta})$ is the **observed information**, evaluated at $\theta = \hat{\theta}$, defined

$$\mathcal{J}(\hat{\theta}) = -\frac{\partial^2 \ell(\theta \mid \mathbf{x})}{\partial \theta^2}\bigg|_{\theta=\hat{\theta}}.$$

The Wald Test: Example

(a) Construct a size α Wald test of the hypotheses

$$\mathsf{H}_0\colon \theta = \theta_0 \quad \text{versus} \quad \mathsf{H}_1\colon \theta \neq \theta_0.$$

(b) In the run-up to a city's mayoral election, a polling company hypothesises that 30% of eligible voters intend to vote for the incumbent mayor. When a simple random sample of 80 eligible voters was conducted (with no non-response), 18 voters said that that they would vote for the incumbent mayor. Test the polling company's null hypothesis using a Wald test at the 5% level.

Wald Test: Example

Wald Test: Example

The Score Test

A third important (and common) hypothesis test is based on the distribution of the score function. Recall, the score function $U(\theta; \mathbf{X})$ is

$$U(\theta; \mathbf{X}) = \frac{\partial}{\partial \theta} \ell(\theta \mid \mathbf{X}) = \frac{\partial}{\partial \theta} \log(\mathcal{L}(\theta \mid \mathbf{X})).$$

We know that

$$\mathbb{E}\left(U(\theta; \mathbf{X})\right) = 0 \quad \text{and} \quad \mathsf{Var}\left(U(\theta; \mathbf{X})\right) = \mathcal{I}(\theta).$$

For large n and, under regularity conditions, we can show that

$$\frac{\left[U(\theta; \mathbf{X})\right]^2}{\mathcal{I}(\theta)} \sim \chi_1^2.$$

We can use the above result for a test of

$$\mathsf{H}_0 \colon \theta = \theta_0 \quad \text{versus} \quad \mathsf{H}_1 \colon \theta \neq \theta_0.$$

Score Test Statistic: Proof (non-examinable)

The Taylor series approximation of $U(\hat{\theta}; \mathbf{X})$ about θ is

$$U(\hat{\theta}; \mathbf{X}) \approx U(\theta; \mathbf{X}) + (\hat{\theta} - \theta)U'(\theta; \mathbf{X}).$$

Since $U(\hat{\theta}; \mathbf{X}) = 0$, we can re-write the above as

$$\begin{aligned} U(\theta; \mathbf{X}) + (\hat{\theta} - \theta)U'(\theta; \mathbf{X}) &\approx 0 \\ \implies U(\theta; \mathbf{X}) &\approx (\hat{\theta} - \theta)\left(-U'(\theta; \mathbf{X})\right) \\ &= (\hat{\theta} - \theta)\left(-\frac{\partial^2 \ell(\theta \mid \mathbf{X})}{\partial \theta^2}\right). \qquad (4) \end{aligned}$$

We know that $\hat{\theta} - \theta$ has an asymptotic normal distribution. Also (4) implies that the score function is a linear transformation of $\hat{\theta} - \theta$. It follows that the score function $U(\theta; \mathbf{X})$ has a normal distribution too (for large n).

In addition, in (4), the term

$$-\frac{\partial^2 \ell(\theta \mid \mathbf{X})}{\partial \theta^2} \approx \mathcal{I}(\theta) \text{ for large } n.$$

So we can re-write (4) as

$$\begin{aligned} U(\theta; \mathbf{X}) &\approx (\hat{\theta} - \theta)\mathcal{I}(\theta) \\ \implies \frac{U(\theta; \mathbf{X})}{\sqrt{\mathcal{I}(\theta)}} &\approx \sqrt{\mathcal{I}(\theta)}(\hat{\theta} - \theta) \quad \text{(dividing both sides by } \sqrt{\mathcal{I}(\theta)}\text{).} \end{aligned}$$

Since $\sqrt{\mathcal{I}(\theta)}(\hat{\theta} - \theta) \sim \mathcal{N}(0, 1)$, squaring both sides of the above it follows that

$$\frac{[U(\theta; \mathbf{X})]^2}{\mathcal{I}(\theta)} \sim \chi_1^2 \text{ as required.}$$

The Score Test

Thus, when using the Score test, for a test of the hypotheses

$$\mathsf{H}_0\colon \theta = \theta_0 \quad \text{versus} \quad \mathsf{H}_1\colon \theta \neq \theta_0$$

we reject H_0 when the value of

$$\frac{\left[U(\theta_0; \mathbf{X})\right]^2}{\mathcal{I}(\theta_0)} \quad \text{is extreme.}$$

For a test of size α, we reject or retain H_0 by comparing the value of the test statistic to the upper $100(1-\alpha)\%$ quantile of a χ^2_1 distribution.

The Score Test: Example

(a) Construct a size α Score test of the hypotheses

$$\mathsf{H}_0\colon \theta = \theta_0 \;\; \text{versus} \;\; \mathsf{H}_1\colon \theta \neq \theta_0.$$

(b) In the run-up to a city's mayoral election, a polling company hypothesises that 30% of eligible voters intend to vote for the incumbent mayor. When a simple random sample of 80 eligible voters was conducted (with no non-response), 18 voters said that that they would vote for the incumbent mayor. Test the polling company's null hypothesis using a Score test of size $\alpha = 0.05$.

Score Test: Example

Score Test: Example

Three Important Likelihood-Based Tests

We have derived three important tests for a hypothesis test of the form

$$\mathsf{H}_0 : \theta = \theta_0 \quad \text{versus} \quad \mathsf{H}_1 : \theta \neq \theta_0$$

where n is large. The following table summarises these tests.

Test	**Test Statistic**	**Distribution under $\mathbf{H}_0$**
GLRT	$2[\ell(\hat{\theta}; \mathbf{X}) - \ell(\theta_0; \mathbf{X})]$	χ^2_1
Wald Test	$\mathcal{I}(\theta_0)(\hat{\theta} - \theta_0)^2$	χ^2_1
Score Test	$\frac{[U(\theta_0; \mathbf{X})]^2}{\mathcal{I}(\theta_0)}$	χ^2_1

Here, GLRT $=$ 'Generalised likelihood ratio test'

We note that asymptotically (i.e. as $n \to \infty$) these three tests are equivalent. In words we say that they are **asymptotically equivalent**.

Three Important Likelihood-Based Tests

We might be tempted think - why not just have one test. . . ?

In answer, there are two things we might like to consider before deciding which test to use.

1. How easily can we evaluate the test statistic? For example, if the Fisher information is difficult to evaluate, then we might prefer to use the GLRT. Conversely, if the mle is difficult to evaluate, we might prefer to use the Score test.

2. Which of the tests is the most powerful for a given set of hypotheses and data distribution? For example, the score test is particularly powerful compared to the others when the the alternative hypothesis has a small departure from the null hypothesis.

Multi-parameter Case

We have focused on and derived our three tests for situations where we have one parameter θ. However, we can use these tests where $\boldsymbol{\theta}$ has dimension k.

If $\boldsymbol{\theta}$ is a $k \times 1$ vector and our hypotheses are

$$\mathsf{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \quad \text{versus} \quad \mathsf{H}_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$$

then the GLRT, Wald and Score test statistics and distributions are as follows

Test	**Test Statistic**	**Distribution under $\mathbf{H}_0$**
GLRT	$2[\ell(\hat{\boldsymbol{\theta}}; \mathbf{X}) - \ell(\boldsymbol{\theta}_0; \mathbf{X})]$	χ^2_k
Wald Test	$(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top \mathcal{I}(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$	χ^2_k
Score Test	$[U(\theta_0; \mathbf{X})]^\top \mathcal{I}^{-1}(\boldsymbol{\theta}_0) U(\theta_0; \mathbf{X})$	χ^2_k

Multi-parameter Case

These tests are used commonly when model fitting and model checking. For example, consider the standard normal linear model of the form

$$\mathbf{Y} = \boldsymbol{\alpha} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where X is an $n \times p$ design matrix of explanatory variables, $\mathbf{Y}$ a set of outcome variables, $\boldsymbol{\alpha}$ an intercept term, and $\boldsymbol{\epsilon}$ zero-mean error term with

$$\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 I).$$

When testing the hypotheses

$$\mathsf{H}_0 \colon \boldsymbol{\beta} = 0 \quad \text{versus} \quad \mathsf{H}_0 \colon \boldsymbol{\beta} \neq 0$$

we could use any of these tests (subject to n being large enough).
Note that n should be fairly large anyway if we're using linear regression.

Learning Outcomes

- Understand the steps to use when testing hypotheses
- Know the definition of the power function and be able to calculate the power function for hypothesis tests.
- Understand and be able to prove the Neyman Pearson lemma.
- Be able to construct most powerful tests of simple hypotheses for a given size.
- Understand and be able to carry out the three important tests in both single parameter and multi-parameter cases.
  - Generalised likelihood ratio test (sometimes called a 'Wilks test')
  - Wald test
  - Score test