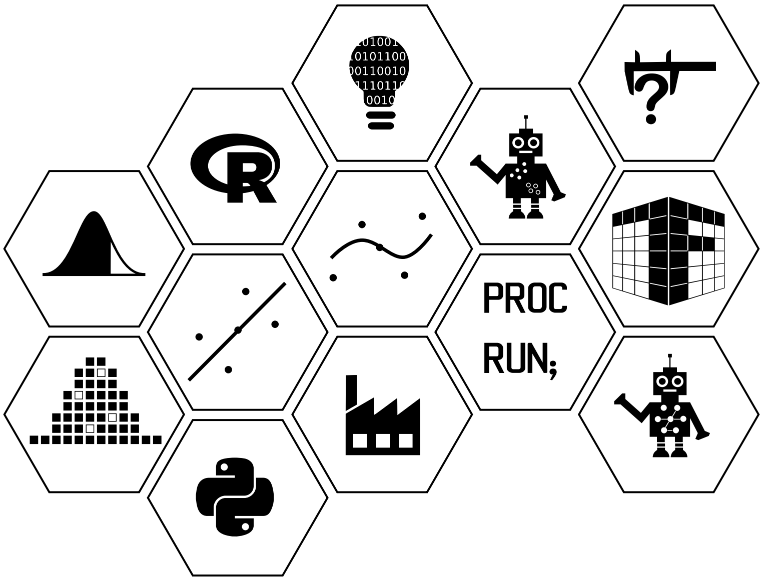
Learning from Data/Data Science Foundations

Week 9: Interval estimation II and hypotheses testing using likelihood



DATA ANALYTICS
GLASGOW

Interval estimation II and hypotheses testing using likelihood

In this week we'll extend the large sample results introduced to obtain approximate confidence intervals for one parameter situations (in week 7) to situations where we wish to compare the parameters from multiple independent populations. Specifically, we'll see how these results can be used to compute a confidence interval to compare population parameters for more than one population, e.g. constructing a confidence interval for a difference in population proportions $\theta_1 - \theta_2$, when our data have not arisen from a normal distribution.

In the second part of this week, we'll return to the idea of hypothesis testing and introduce the idea of comparing hypotheses tests using likelihood, and definitions for Type I/Type II errors and power for a statistical test.

Week 9 learning material aims

The material in week 9 covers:

- approximate confidence intervals to compare parameters from independent populations;
- interpreting the results of these intervals;
- comparing hypotheses using likelihood;
- Type I and Type II errors and statistical power.

Confidence intervals for linear functions of parameters

In week 8, we introduced the idea of combining likelihoods for multiple independent populations. For example, we might be interested in comparing a measured variable between two independent groups (assuming the same form of distribution), and therefore we might wish to compare the parameters for the assumed distributions. In week 8 we have seen how to derive point estimators, and hence calculate point estimates for population parameters. However, in order to establish conclusions around comparisons of population parameters we also need to perform inference and one way to do this is to construct confidence intervals to compare the population parameters. An alternative way to phrase this is that we would like a confidence interval for a particular linear combination of parameters, say $\theta_1 - \theta_2$, which we'll write here as $\mathbf{b}^T\boldsymbol{\theta}$.

The large sample distributional result using the properties of MLEs that we established in week 8, give us (for k parameters) that:

$$\hat{\boldsymbol{\theta}}_{MLE} \sim N_k(\boldsymbol{\theta}, \mathbf{K}(\mathbf{x})^{-1}).$$

It then follows from the results introduced in the *Probability and Stochastic models* course or *Probability and Sampling Fundamentals/Sampling Fundamentals* courses that[1],

$$\mathbf{b}^T\hat{\boldsymbol{\theta}}_{MLE} \sim N(\mathbf{b}^T\boldsymbol{\theta}, \mathbf{b}^T\mathbf{K}(\mathbf{x})^{-1}\mathbf{b}),$$

where $\mathbf{b}$ is a non-zero vector.

An approximate pivotal function for $\mathbf{b}^T\boldsymbol{\theta}$ is then given by

$$\frac{\mathbf{b}^T\hat{\boldsymbol{\theta}}_{MLE} - \mathbf{b}^T\boldsymbol{\theta}}{\sqrt{\mathbf{b}^T\mathbf{K}(\mathbf{x})^{-1}\mathbf{b}}}.$$

An **approximate 95% confidence interval for $\mathbf{b}^T\boldsymbol{\theta}$** is then given by

$$\mathbf{b}^T\hat{\boldsymbol{\theta}}_{MLE} \pm z\sqrt{\mathbf{b}^T\mathbf{K}(\mathbf{x})^{-1}\mathbf{b}},$$

where, $z = \Phi^{-1}\left(1 - \frac{(1-c)}{2}\right) = \Phi^{-1}(0.975) = 1.96$, for $c = 0.95$.

This result enables us to compute approximate confidence intervals for the difference in population parameters for any known distribution.

The first video for this week explores this result and provides an example of using $\mathbf{b}$ to provide a linear function of our parameters[2]:

Video

Linear functions of parameters

Duration 5:05

University of Glasgow

Linear functions of parameters

Claire Miller, Eilidh Jack and Suzy Whoriskey

Let's explore how we use these results in practice through the next example.

Example 1

Comparing two population proportions

We considered the following example, as example 2 in week 8.

In an experiment to investigate the bacteria present in the home, 15 households had the number of bacteria (in 100,000's) recorded from a sample taken from the kitchen sink (control surface) and 15 different households had the number of bacteria recorded from a sample taken from the kitchen sink immediately after cleaning (treated surface). The data are shown below.

```
Control surface: 37 35 31 36 25 43 41 33 25 37 27 30 32 35 38

Treated surface: 4 4 3 3 3 5 3 4 5 4 4 3 4 4 3
```

We will assume that these data follow independent Poisson probability models with parameters λ_1 and λ_2 respectively.

Let's use the maximum likelihood estimates and sample information that we found in week 8 to now compute an interval to investigate if there is a difference in bacteria present in the home for the two surfaces.

In week 8, we found that: $\hat{\lambda}_1 = 33.67$ and $\hat{\lambda}_2 = 3.73$, and

$$\mathbf{K}(\mathbf{x})^{-1} = \begin{pmatrix} 2.245 & 0 \\ 0 & 0.248 \end{pmatrix}.$$

If we now wish to estimate $\lambda_1 - \lambda_2$ then we can use the result on linear combinations, with $\mathbf{b = (1, -1)^T}$, to form a **95% confidence interval for $\boldsymbol{\lambda_1 - \lambda_2}$**. This gives

$$\mathbf{b}^T \hat{\boldsymbol{\lambda}}_{MLE} \pm 1.96\sqrt{\mathbf{b}^T \mathbf{K}(\mathbf{x})^{-1} \mathbf{b}}$$

$$\mathbf{b} = (1, -1)^T$$

$$\hat{\lambda}_1 - \hat{\lambda}_2 \pm 1.96\sqrt{\begin{pmatrix} 1 & -1 \end{pmatrix} \mathbf{K}(\mathbf{x})^{-1} \begin{pmatrix} 1 \\ -1 \end{pmatrix}}$$

$$(33.67 - 3.73) \pm 1.96\sqrt{2.245 + 0.248}$$

$$(26.85, 33.03)$$

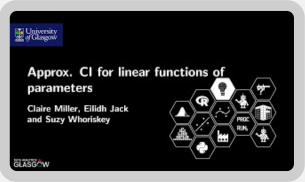
Since the CI does not contain zero there is a statistically significant difference between the two population parameters λ_1 and λ_2. The population mean number of bacteria on a control surface is highly likely to be larger than the population mean number of bacteria on a treated surface by between 2,700,000 and 3,300,000 bacteria.

The second video for this week, provides an overview of example 2, focussing on the interpretation of the interval and matrix multiplication:

Video

CI for linear functions of parameters

Duration 9:18


Approx. CI for linear functions of parameters

Claire Miller, Eilidh Jack and Suzy Whoriskey

Task 1

The following example was introduced in a task in week 8.

Patients with high blood pressure were randomly allocated to receive one of two treatments. The patients had been treated with cognitive behaviour therapy (CBT) and were then given no further treatment (NFT) or CBT and beta-blockers (BB). Six weeks later it was noted whether or not each of the patients showed a decrease in their blood pressure. For the NFT group 15 out of 50 patients showed a decrease after 6 weeks, whereas this happened for 26 out of 50 of the BB patients.

We found in week 8, after assuming a binomial distribution for each group, with individual sample sizes n_i and individuals parameters θ_i, that the maximum likelihood estimates are:

$$\hat{\theta}_1 = 15/50 = 0.30,$$

$$\hat{\theta}_2 = 26/50 = 0.52,$$

and the Hessian matrix is:

$$\begin{pmatrix} \frac{\partial^2 \ell}{\partial \theta_1^2} & \frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_2} \\ \frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_2} & \frac{\partial^2 \ell}{\partial \theta_2^2} \end{pmatrix} = \begin{pmatrix} -\frac{15}{\theta_1^2} - \frac{35}{(1-\theta_1)^2} & 0 \\ 0 & -\frac{26}{\theta_2^2} - \frac{24}{(1-\theta_2)^2} \end{pmatrix}.$$

1. Evaluate the sample information matrix $\mathbf{K}(\mathbf{x})$;
2. State the estimated variance for each of $\hat{\theta}_1$, $\hat{\theta}_2$;
3. Produce an approximate 95% confidence interval for $\theta_1 - \theta_2$;
4. Interpret this interval to form a conclusion on the differences between the two population proportions.

Note for the example in task 1 here:

- this is a confidence interval for a difference in population proportions;
- compare the theoretical result for the confidence interval to the result that we used to compare binomial population proportions in week 4 (for which we used the function `prop.test` in `R`), you'll see that we have derived the same approximate form for the interval here;
- for this type of problem $\log_e\left(\frac{\theta_1}{\theta_2}\right)$ would often be considered to ensure that the bounds of the interval do not go beyond 0 or 1. This will be considered, and these results will be extended, in your course on *Advanced Predictive Models*.

Confidence Regions

Our results for Wilks and Wald intervals in the one parameter case, and the result that we have used above can be extended to the multiparameter case in general. In this course, we will only focus on the situation of comparing independent populations, as above. However, the supplementary material below provides a more general framework for extending the results here.

Supplement 1

Wilks Confidence Regions

A $100p\%$ likelihood region can be defined in terms of the relative log-likelihood function for a vector of parameters $\boldsymbol{\theta}$,

$$r(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) - \ell(\hat{\boldsymbol{\theta}}_{\boldsymbol{MLE}}).$$

An approximate $\mathbf{100c\%}$ Wilks confidence region for $\boldsymbol{\theta}$ is then defined by

$$\{\boldsymbol{\theta} : -2r(\boldsymbol{\theta}) \leq \chi^2_k(c)\},$$

(where k is the number of parameters in the model and c is between 0 and 1), or, equivalently,

$$\{\boldsymbol{\theta} : r(\boldsymbol{\theta}) \geq -\frac{1}{2}\chi^2_k(c)\}.$$

For example, in week 8 example 1 we looked at the annual mean temperature data from New Haven. We assumed that our data were independent observations from a normal distribution with two parameters, μ and σ, therefore $k = 2$. Let's visualise an **approximate 95% Wilks confidence region for** μ and σ for this example.

Since we have 2 parameters of interest here μ and σ, $\chi^2_2(0.95) = 5.99$.

So the threshold is $-5.99/2 = -3.00$ to 2 decimal places.

Visualising this on the contour plot of the relative log-likelihood for the normal distribution we have the figure below where the dashed line contour is at -3:

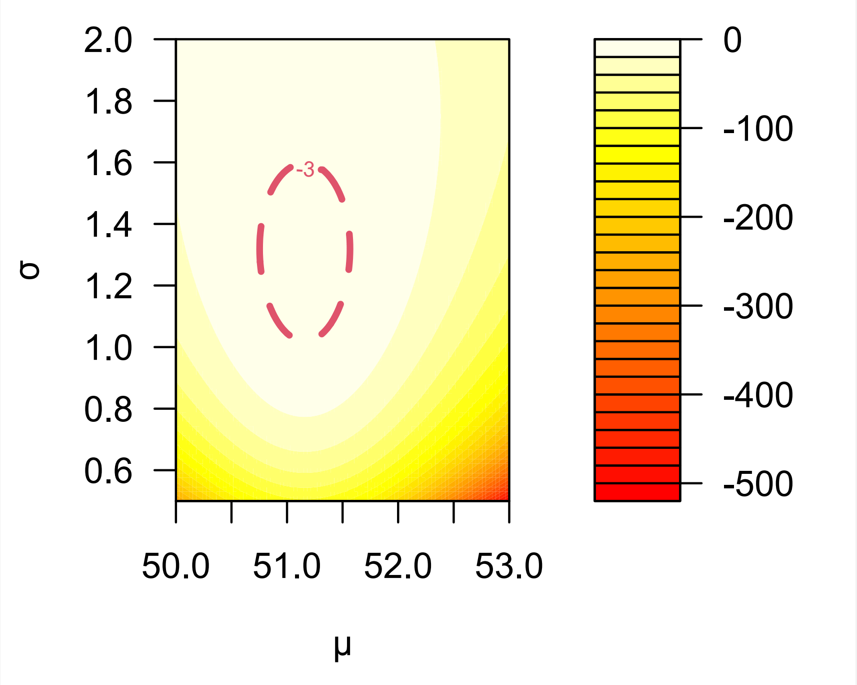


Figure 1

The values of μ and σ that are within this contour line at -3 form the confidence region.

Wald confidence regions

In order to extend our Wald confidence interval for more than one parameter, we can use a quadratic approximation (this is derived from a Taylor Series expansion) to define a **Wald** confidence region.

We create a quadratic approximation around the MLE itself, $\hat{\boldsymbol{\theta}}_{MLE}$.

The quadratic approximation is

$$\ell(\boldsymbol{\theta}) \simeq \ell(\hat{\boldsymbol{\theta}}_{MLE}) + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{MLE})^T \mathbf{g} + \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{MLE})^T \mathbf{H}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{MLE}),$$

where $\mathbf{g}$ denotes the vector of first derivatives of ℓ evaluated at $\hat{\boldsymbol{\theta}}_{MLE}$ and $\mathbf{H}$ is the usual Hessian matrix of second derivatives, evaluated again at $\hat{\boldsymbol{\theta}}_{MLE}$.

However, since $\hat{\boldsymbol{\theta}}_{MLE}$ maximises the log-likelihood, the first derivatives at $\hat{\boldsymbol{\theta}}_{MLE}$ are $\mathbf{0}$ i.e. $\mathbf{g} = \mathbf{0}$.

So,

$$\ell(\boldsymbol{\theta}) \simeq \ell(\hat{\boldsymbol{\theta}}_{MLE}) + \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{MLE})^T \mathbf{H}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{MLE}).$$

By subtracting $\ell(\hat{\boldsymbol{\theta}}_{MLE})$, from both sides, we can write this as

$$r(\boldsymbol{\theta}) \simeq \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{MLE})^T \mathbf{H}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{MLE}).$$

The Wilks confidence region

$$\{\boldsymbol{\theta} : r(\boldsymbol{\theta}) \geq -\frac{1}{2}\chi^2_k(c)\},$$

can therefore be approximated by the Wald region

$$\{\boldsymbol{\theta} : \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{MLE})^T \mathbf{H}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{MLE}) \geq -\frac{1}{2}\chi^2_k(c)\},$$

or, equivalently,

$$\{\boldsymbol{\theta} : (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{MLE})^T \mathbf{K}(\mathbf{x})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{MLE}) \leq \chi^2_k(c)\},$$

where $\mathbf{K}(\mathbf{x})$ denotes the sample information matrix, namely $-\mathbf{H}$.

Visualising likelihood

The following `R` code can be used to visualise a 3D surface of the log-likelihood for the normal distribution and the temperature in New Haven example over a range of values for both μ and σ, and to visualise the Wilks and Wald confidence regions.

```r
airtemp <- c(49.9, 52.3, 49.4, 51.1, 49.4, 47.9, 49.8, 50.9,
             49.3, 51.9, 50.8, 49.6, 49.3, 50.6, 48.4, 50.7,
             50.9, 50.6, 51.5, 52.8, 51.8, 51.1, 49.8, 50.2,
             50.4, 51.6, 51.8, 50.9, 48.8, 51.7, 51.0, 50.6,
             51.7, 51.5, 52.1, 51.3, 51.0, 54.0, 51.4, 52.7,
             53.1, 54.6, 52.0, 52.0, 50.9, 52.6, 50.2, 52.6,
             51.6, 51.9, 50.5, 50.9, 51.7, 51.4, 51.7, 50.8,
             51.9, 51.8, 51.9, 53.0)

library(rpanel)
rp.likelihood("sum(log(dnorm(data, theta[1], theta[2])))",
airtemp, c(50, 0.5), c(53, 2))
```

After running these commands, selecting the radio button 'ci' provides a plane for the Wilks confidence region i.e. at -3.00, selecting 'quadratic' adds the log-likelihood surface corresponding to that assumed by the symmetric Wald region, and selecting 'transparent' makes for easier visualisation of the two surfaces.

Hypothesis Testing using Likelihood

For our final section on likelihood, we now want to return to hypothesis testing and extend the ideas that we presented in week 4 to enable us to investigate the relative plausibility of other values for our parameters $\boldsymbol{\theta}$ through hypothesis testing using likelihood.

Some general principles

Suppose we have data $x_1, x_2, \ldots, x_n$ and a probability model with unknown parameters $\boldsymbol{\theta}$ describing how these data were generated. We want to test a **null hypothesis** (H_0), which specifies values for the parameters (or restrictions on the values of the parameters) against some **alternative hypothesis** (H_1) about the values that the parameters can take. The likelihood under each hypothesis can be used to compare the relative plausibility of the hypotheses.

Generally hypotheses fall into two main classes.

Simple hypotheses completely specify the values taken by the parameters.

For example,

$$H_0 : \theta = 2.5,$$

is an example of a simple hypothesis.

Composite hypotheses are hypotheses that allow the parameter(s) to take a range of possible values.

One parameter problem

$$H_1 : \theta > 2.5,$$

is an example of a composite (one-sided) hypothesis.

Two parameter problem

$$H_0 : \theta_1 = \theta_2.$$

There is a range of values of θ_1 and θ_2 satisfying the (two-sided) hypothesis.

Task 2

For the following hypotheses, state whether or not they are simple or composite:

1. $H_0 : \theta_1 = \theta_2 = \theta_3$
2. $H_0 : \lambda_1 = 5$
3. $H_1 : \gamma_1 = 3$
4. $H_1 : \theta_1 > \theta_2$

Comparing hypotheses using likelihood

Here are three examples of situations involving different types of null and alternative hypotheses.

Example 2

Scenario 1

Suppose there is a single parameter, θ, and we wish to test:

$H_0 : \theta = \theta_{H_0}$,

against the alternative,

$H_1 : \theta = \theta_{H_1}$,

where θ_{H_0} and θ_{H_1} are specific (known) values.

In this case both hypotheses are simple and it is straightforward to compare their likelihoods.

The ratio of the likelihoods:

$\Lambda = L(\theta_{H_1})/L(\theta_{H_0})$,

would be a sensible statistic to use for the comparison.

This is referred to as a **Likelihood Ratio statistic**.

In fact this situation, while of considerable theoretical interest, is not often encountered in practice.

Example 3

Scenario 2

We may be interested in comparing a simple null hypothesis:

$$H_0 : \theta = \theta_{H_0},$$

(where θ_{H_0} is a specific (known) value) and a composite alternative,

$$H_1 : \theta \neq \theta_{H_0}.$$

Here the likelihood for the null is well defined, but under the alternative θ can take almost any value. In this case the maximum of the likelihood under the alternative $\hat{\theta}_{H_1}$, i.e. the likelihood at the MLE, is used for the alternative hypothesis.

Then the **(Generalized) Likelihood Ratio**:

$$\Lambda = L(\hat{\boldsymbol{\theta}}_{H_1})/L(\theta_{H_0}),$$

is a sensible measure of the relative plausibility of the hypotheses.

Example 4

Scenario 3

It is common to wish to compare two composite hypotheses. For example, in a two parameter situation:

$$H_0 : \theta_1 = \theta_2$$

vs.

H_1 : no restriction ($\theta_1 \neq \theta_2$),

involves two composite hypotheses.

Under H_0 the parameters would have to lie on the line $\theta_1 = \theta_2$. The obvious way to compare the hypotheses now is to compare the maximum likelihood achievable under $H_0(\hat{\boldsymbol{\theta}}_{H_0})$ with the maximum likelihood achievable under $H_1(\hat{\boldsymbol{\theta}}_{H_1})$.

So in this case the **Generalized Likelihood Ratio statistic** would be:

$$\Lambda = L(\hat{\boldsymbol{\theta}}_{H_1})/L(\hat{\boldsymbol{\theta}}_{H_0}).$$

In each testing situation some variation on the generalized likelihood ratio statistic (our **test statistic** here) provides a reasonable measure of the relative plausibility of the hypotheses, with **high** values supporting H_1 and **low** values supporting H_0. However, in many situations there is some prior reason to favour H_0. Often it is favoured because it is the simpler hypothesis, and we would like to have the simplest model consistent with the data. In other circumstances there may be more pressing considerations: for example in a drug trial it makes sense not to reject the null hypothesis that a new drug is no more effective than the current standard one, unless the data provide strong evidence to the contrary.

Also, the null hypothesis usually states that the data were generated by a more restricted version of the model assumed under the alternative hypothesis. In this circumstance it is always the case that $L(\hat{\boldsymbol{\theta}}_{H_1}) \geq L(\hat{\boldsymbol{\theta}}_{H_0}) \Rightarrow \Lambda \geq 1.$

For these reasons it is usual not to reject H_0 unless Λ is so large that, in effect, the evidence against H_0 has become too strong to ignore. We do not reject the null hypothesis until there is enough evidence against it. The strength of the evidence against H_0 (and for H_1) is judged using a $\boldsymbol{p}$**-value**, as we saw in weeks 3 & 4, or by using a **rejection region**.

The **rejection region** is the set of values of the test statistic for which the null hypothesis is rejected.

Reminders from weeks 3 & 4:

The $\boldsymbol{p}$**-value** is the probability of obtaining a value for the test statistic (assuming H0 is true) that is at least as extreme as the observed value of the test statistic.

The p-value is an attempt to measure the consistency of the data with the null hypothesis.

Generally, a value for $\boldsymbol{\alpha}$, the **significance level** is chosen such that H_0 will be rejected if the p-value is $\leq \alpha$. As we have seen, 0.05 is the most common choice for a two-sided test.

What to look for in a good testing procedure?

In testing procedures in which a significance level is chosen and the null hypothesis is either not rejected or rejected, two errors are possible.

1. A **Type I error** is rejection of the null hypothesis when it is true.

2. A **Type II error** is not rejecting the null hypothesis when it is false.

This is illustrated in the table below:

	do not reject $\mathbf{H_0}$	**reject** $\mathbf{H_0}$
H_0 true		type I error
H_1 true	type II error	

Example 5

A manufacturer is trying to market a new drug for the common cold. However, the truth is that the drug is a placebo (i.e. it has 'no effect' - there is no reason that it should improve cold symptoms).

A medical centre is interested in testing whether or not the new drug improves a person's condition, and considers the following hypotheses.

H_0 : There is no effect of the drug on population mean condition;

H_1 : There is an effect of the drug on population mean condition.

The medical centre notices that the condition of most of their patients who received this drug has improved and therefore rejects the null hypothesis and concludes (wrongly) that the new drug improves cold symptoms.

This would be an example of a **Type I error**.

Example 6

A manufacturer is trying to market a new drug, which is effective in improving the condition of a person suffering from the common cold.

A medical centre is interested in testing whether or not the new drug improves a person's condition, and considers the following hypotheses.

H_0 : There is no effect of the drug on population mean condition;

H_1 : There is an effect of the drug on population mean condition.

The medical centre notices that the condition of most of their patients who received this drug has NOT improved. Therefore, the surgery does not reject the null hypothesis, and concludes that there is no evidence of an effect of the new drug.

This would be an example of a **Type II error**.

The following scenarios provide some interesting contexts to consider to help motivate the ideas of Type I and Type II errors a bit more. For each of them, state if it is an example of a Type I or Type II error.

Task 3

You are investigating whether or not people believe the popular myth that "Newton was hit by an apple" (he wasn't). Assume that the truth is that most people do NOT believe this and consider the following hypotheses:

H_0: population proportion of people that do believe this statement ≥ 0.5

H_1: population proportion of people that do believe this statement < 0.5

You conduct your research by polling local residents at a retirement community and to your surprise find out that most people do, in fact, believe the statement. Therefore, you do not reject H_0.

- Is this a Type I or Type II error?
- What appears to have caused this error?

Task 4

In a company's computer security system, the default position is that the person accessing part of the system is the authorised user. If there are a number of discrepancies in the person's access data then the person is flagged as an imposter.

A new user (an authorised user) tries to access the system. Suppose that we test the following hypotheses for a new user:

H_0 : population mean discrepancies = 0;

H_1 : population mean discrepancies $\neq$ 0.

The new user supplies all of the correct login data, except that their fingerprint is not a match (because of dirt on the sensor). After considering the data supplied by the new user the computer decides to reject the null hypothesis, and the new user is flagged as an imposter and hence not given access.

- Is this a Type I or a Type II error?
- What appears to have caused this error?

Probabilities for Type I and Type II error

A good testing procedure should keep the probabilities of both types of error low, but there is a trade-off between keeping the Type I error probability low and keeping the Type II error probability low. Generally a test procedure which tries to avoid falsely rejecting H_0 will require a great deal of evidence in order to reject H_0, but this is bound to lead to a quite high probability of not rejecting H_0 when it is false.

To move beyond generalities error probabilities must be calculated. The **significance level** of a test sets the probability of a **Type I error**. For example, when testing at the 5% level we reject H_0 for 5% of datasets for which it is true.

So the probability of a Type I error is set by the chosen significance level, which suggests that given a choice of testing procedures we should select the one that gives the lowest probability of a **Type II error** given our chosen significance level. Since 1 - Pr[Type II error] is the probability of correctly rejecting the null when it is false, it is known as the **power** of a test. So at a given significance level, we would generally choose the **most powerful test** available: this principle is often referred to as the Neyman-Pearson approach to hypothesis testing.

Supplement 2

A visualisation for power, Type I and Type II errors can be seen here: Power, for different significance levels and different sample sizes.

You migth also be interested in this NetFlix article on the use of Type I and Type II errors NetFlix

Likelihood ratio test

If both hypotheses are simple it is then possible to prove that **the most powerful test** will always be based on the likelihood ratio test statistic. This result is known as the **Neyman-Pearson lemma**.

To construct a **likelihood ratio test** for simple hypotheses, consider the simple hypotheses, where θ_0 and θ_1 are specific values:

$\mathrm{H}_0 : \theta = \theta_0$

$\mathrm{H}_1 : \theta = \theta_1$

We would reject H_0 when the ratio of the likelihoods for H_0 and H_1 is large enough, i.e. when,

$$\frac{L(\theta_1)}{L(\theta_0)} > k$$

where k is some constant (greater than 1) to be chosen such that:

Pr(Type I error) $\leq \alpha$

where α is the significance level.

The rejection region (RR) is therefore:

$$RR = \left\{ \mathbf{x}; \frac{L(\theta_1)}{L(\theta_0)} > k \right\}$$

with k defined as above. The test based on this Rejection Region is termed a **Likelihood Ratio Test**. Therefore, for simple hypotheses it is possible to find the exact distribution of the likelihood ratio statistic and hence calculate a p-value exactly. However, in most circumstances we are interested in a composite hypothesis and so this is not possible, and we have to use approximate p-values (or a rejection region) through a **Generalised Likelihood Ratio Test** (GLRT) using the results we have already seen in week 7 based on large sample properties (along with theoretical support from the **Neyman-Pearson lemma**). The GLRT will be the focus of week 10.

Learning outcomes for week 9

- use the general formula for Wald confidence intervals based on linear combinations of the parameters to compute and interpret confidence intervals for linear functions;
- define the terms *simple* and *composite* hypotheses;
- state the likelihood ratio statistic and the generalised likelihood ratio statistic and explain when each should be used;
- define the terms, *type I* and *type II error* and *power*.

Review exercises, selected video solutions and written answers to all tasks/review exercises are provided overleaf.

Review exercises

Task 5

A clinical trial was conducted to compare four drugs used to treat hypertension. Each patient was randomly allocated to receive one of the four drugs during the course of the trial. The numbers of side effects which were recorded for the patients during the course of the trial are given below.

Drug	**1**	**2**	**3**	**4**
Number of patients	29	27	29	26
Number of side effects	10	85	55	39

It may be assumed that the data are described well by independent Poisson probability models $X_i \sim \mathrm{Po}(n_i\theta_i)$, where,

X_i denotes the number of side effects for the ith drug,

n_i denotes the number of patients on the ith drug,

θ_i denotes the mean number of side effects per patient on the ith drug.

The general model for these data is:

$$X_1 \sim \mathrm{Po}(29\theta_1),\ X_2 \sim \mathrm{Po}(27\theta_2),\ X_3 \sim \mathrm{Po}(29\theta_3),\ X_4 \sim \mathrm{Po}(26\theta_4)$$

and hence, since the models are independent, the likelihood is:

$$L(\theta, x) = K\theta_1^{10}e^{-29\theta_1}\theta_2^{85}e^{-27\theta_2}\theta_3^{55}e^{-29\theta_3}\theta_4^{39}e^{-26\theta_4}$$

You found the Hessian matrix (in week 8) to be,

$$\boldsymbol{H} = \begin{pmatrix} -10/\theta_1^2 & 0 & 0 & 0 \\ 0 & -85/\theta_2^2 & 0 & 0 \\ 0 & 0 & -55/\theta_3^2 & 0 \\ 0 & 0 & 0 & -39/\theta_4^2 \end{pmatrix}.$$

Use this to find an approximate 95% confidence interval for $\theta_1 - \theta_3$ by the Wald method. (Reminder: $\hat{\theta}_{i\,MLE} = \frac{X_i}{n_i}$).

Task 6

Looking back to week 8 review exercise task 5, we had observations (x_{1i}, x_{2i}) for $i = 1, \ldots, n$. An appropriate model for these was that each pair was an independent observation of the random variables with joint p.d.f.

$$f(x_1, x_2) = \lambda_1 \lambda_2 e^{-\lambda_1 x_1 - \lambda_2 x_2} \text{ where } x_1 > 0,\ x_2 > 0,\ \lambda_1 > 0,\ \lambda_2 > 0.$$

You found that the likelihood of $\lambda_1,\ \lambda_2$ was,

$$L(\boldsymbol{\lambda}) \propto \prod_{i=1}^{n} \lambda_1 \lambda_2 e^{-\lambda_1 x_{1i} - \lambda_2 x_{2i}} = \lambda_1^n \lambda_2^n e^{-\lambda_1 \sum x_{1i} - \lambda_2 \sum x_{2i}}$$

and the corresponding log-likelihood,

$$\ell(\boldsymbol{\lambda}) = n \log_e \lambda_1 + n \log_e \lambda_2 - \lambda_1 \sum x_{1i} - \lambda_2 \sum x_{2i}$$

This gave the maximum likelihood estimates to be $\hat{\lambda}_j = \frac{n}{\sum_i x_{ji}}$ for $j = 1, 2$.

Lastly, the Hessian matrix was calculated to be,

$$\mathbf{H} = \begin{pmatrix} -n/\lambda_1^2 & 0 \\ 0 & -n/\lambda_2^2 \end{pmatrix}.$$

- Use the Hessian matrix to find expressions for approximate 95% confidence intervals for λ_1 and λ_2 by the Wald method.

- Use the Hessian matrix again, and the following data, to contruct an approximate 95% confidence interval for $\lambda_1 - \lambda_2$ by the Wald method.

x_1	1	3	2	1
x_2	5	2	4	6

Task 7

This question relates to Type I and Type II errors. A man goes on trial and is tried for the murder of his ex-wife. He is acquitted in the criminal trial by the jury, but convicted in a subsequent civil lawsuit based on the same evidence. The default position is that the null hypothesis of any test here suggests innocence.

- What type of error do we have when the man is not guilty but found guilty?
- What type of error do we have when the man is guilty but found not guilty?

Explain your answers.

Answer 1

The sample information matrix evaluates the Hessian matrix at the MLEs.

Notice that since $15 = 50\hat{\theta}_1$ and $26 = 50\hat{\theta}_2$:

$$\begin{aligned} -\frac{15}{\hat{\theta}_1^2} - \frac{35}{(1-\hat{\theta}_1)^2} &= -\frac{50\hat{\theta}_1}{\hat{\theta}_1^2} - \frac{50(1-\hat{\theta}_1)}{(1-\hat{\theta}_1)^2} \\ &= -\frac{50}{\hat{\theta}_1} - \frac{50}{(1-\hat{\theta}_1)} \\ &= -\frac{50}{\hat{\theta}_1(1-\hat{\theta}_1)} \end{aligned}$$

and similarly for the term in $\hat{\theta}_2$.

The sample information matrix is therefore

$$\mathbf{K}(\mathbf{x}) = \begin{pmatrix} \frac{50}{\hat{\theta}_1(1-\hat{\theta}_1)} & 0 \\ 0 & \frac{50}{\hat{\theta}_2(1-\hat{\theta}_2)} \end{pmatrix}$$

and its inverse is

$$\mathbf{K}(\mathbf{x})^{-1} = \begin{pmatrix} \frac{\hat{\theta}_1(1-\hat{\theta}_1)}{50} & 0 \\ 0 & \frac{\hat{\theta}_2(1-\hat{\theta}_2)}{50} \end{pmatrix}$$

$\mathbf{K(x)^{-1} = (-H)^{-1}}$ is the variance covariance matrix of $\hat{\theta}_1$ and $\hat{\theta}_2$ and so the diagonal entries are the variances of $\hat{\theta}_1$ and $\hat{\theta}_2$.

Notice that the variance of the sample proportions has been automatically generated by the likelihood procedure.

Numerically,

$$\mathbf{K(x)^{-1}} = \begin{pmatrix} 0.00420 & 0 \\ 0 & 0.004992 \end{pmatrix}$$

Diagonal terms:

$$\mathrm{Var}(\hat{\theta}_1) = 0.004$$

$$\mathrm{Var}(\hat{\theta}_2) = 0.005$$

Off-diagonal terms:

$$\mathrm{Cov}(\hat{\theta}_1, \hat{\theta}_2) = 0$$

$$\mathrm{Cov}(\hat{\theta}_2, \hat{\theta}_1) = 0$$

If we now wish to estimate $\theta_1 - \theta_2$ then we can use the result on linear combinations, with $\mathbf{b = (1, -1)^T}$, to form **a 95% confidence interval for $\boldsymbol{\theta_1 - \theta_2}$**. This gives

$$\mathbf{b}^T \hat{\boldsymbol{\theta}}_{MLE} \pm 1.96\sqrt{\mathbf{b^T K(x)^{-1} b}}$$

$$\mathbf{b} = (1, -1)^T$$

$$\hat{\theta}_1 - \hat{\theta}_2 \pm 1.96\sqrt{\begin{pmatrix} 1 & -1 \end{pmatrix} \mathbf{K(x)^{-1}} \begin{pmatrix} 1 \\ -1 \end{pmatrix}}$$

$$(0.30 - 0.52) \pm 1.96\sqrt{0.00420 + 0.004992}$$

$$(-0.408, -0.032)$$

Since the CI does not contain zero there is a statistically significant difference between the two population proportions. The population proportion of BB patients experiencing a decrease is highly likely to be larger than the population proportion of NFT patients by between 3% and 41%.

Note: that the form of the confidence interval here is the same as the one that we derived in week 4:

$$\hat{\theta}_1 - \hat{\theta}_2 \pm 1.96\sqrt{\frac{\hat{\theta}_1(1-\hat{\theta}_1)}{n_1} + \frac{\hat{\theta}_2(1-\hat{\theta}_2)}{n_2}}$$

Answer 2

Hypotheses:

1. Composite
2. Simple
3. Simple
4. Composite

Answer 3

Scenario 1:

- Type II error.
- Your sample is not representative of the whole population (older generation might be more inclined to respond based on their belief where as younger generation may check information using e.g. smart phone).

Answer 4

Scenario 2

- Type I error.
- The fingerprint match software sensitivity needs adjusted - it's currently providing too much data against the null hypothesis.

Answer 5

Based on the large sample distribution of $\hat{\boldsymbol{\theta}}$ an approximate 95% CI for $\theta_1 - \theta_3$ is

$$\boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \end{pmatrix}$$

$$\mathbf{b}^T\hat{\boldsymbol{\theta}}_{MLE} \pm 1.96\sqrt{\mathbf{b^T K^{-1}(x) b}}$$

Take $\mathbf{b^T = c(1, 0, -1, 0)}$

$$\hat{\theta}_1 - \hat{\theta}_3 \pm 1.96\sqrt{\hat{\theta}_1^2/10 + \hat{\theta}_3^2/55}$$

$$10/29 - 55/29 \pm 1.96\sqrt{(10/29)^2/10 + (55/29)^2/55}$$

$$0.34 - 1.897 \pm 1.96\sqrt{0.012 + 0.065}$$

$$-1.557 \pm 0.544$$

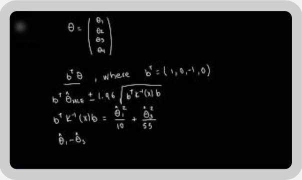
$$(-2.10, -1.01)$$

Since the confidence interval does not include 0, there is evidence of a statistically significant difference between θ_1 and θ_3. θ_3 is highly likely to be larger than θ_1 and so we interpret this as the mean number of side effects per patient is highly likely to be larger for Drug 3 than for 1 by between 1.01 and 2.10 side effects per patient.

Video

Video model answers for task 5

Duration 3:45



$$\theta = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \end{pmatrix}$$

$\underline{b^T \theta}$, where $b^T = (1, 0, -1, 0)$

$$b^T \hat{\theta}_{MLE} \pm 1.96 \sqrt{b^T K^{-1}(x) b}$$

$$b^T K^{-1}(x) b = \frac{\hat{\theta}_1^2}{10} + \frac{\hat{\theta}_3^2}{55}$$

$\hat{\theta}_1 - \hat{\theta}_3$

Answer 6

Based on the large sample distribution of $\hat{\boldsymbol{\lambda}}$ an approximate 95% CI for λ_j is

$$\boldsymbol{\lambda} = \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix}$$

$$\mathbf{b}^T \hat{\boldsymbol{\lambda}}_{MLE} \pm 1.96 \sqrt{\mathbf{b}^T \mathbf{K^{-1}(x)b}}$$

Take $\mathbf{b^T = c(1, 0)}$ for λ_1 and $c(0, 1)$ for λ_2

$$\hat{\lambda}_j \pm 1.96 \sqrt{\hat{\lambda}_j^2 / n}$$

where $\hat{\lambda}_j$ is as defined as $\hat{\lambda}_j = \frac{n}{\sum_i x_{ji}}$ for $j = 1, 2$.

From the question we know that $\lambda_1 = \frac{n}{\sum x_{1i}}, \lambda_2 = \frac{n}{\sum x_{2i}}$, and so $\hat{\lambda}_1 = 4/7$ and $\hat{\lambda}_2 = 4/17$.

Hence, based on the large sample distribution of $\hat{\boldsymbol{\lambda}}$ an approximate 95% CI for $\lambda_1 - \lambda_2$ is

$$\boldsymbol{\lambda} = \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix}$$

$$\mathbf{b}^T \hat{\boldsymbol{\lambda}}_{MLE} \pm 1.96 \sqrt{\mathbf{b}^T \mathbf{K^{-1}(x)b}}$$

Take $\mathbf{b^T = c(1, -1)}$

$$\hat{\lambda}_1 - \hat{\lambda}_2 \pm 1.96 \sqrt{\lambda_1^2 / n + \lambda_2^2 / n}$$

$$4/7 - 4/17 \pm 1.96\sqrt{\lambda_1^2/n + \lambda_2^2/n}$$

$$0.57 - 0.24 \pm 1.96\sqrt{0.082 + 0.014}$$

$$0.33 \pm 0.61$$

Since the confidence interval includes 0, we have insufficient evidence of a difference between λ_1 and λ_2. However, the sample size is very small, a larger sample size may have a different result here. This indicates that we cannot reject that $\lambda_1 = \lambda_2$.

Answer 7

Type I/II errors:

- Type I error (rejecting H_0 when it is true).
- Type II error (not rejecting H_0 when it is false).

Footnotes

1. Note: that a similar result is used in the Predictive Modelling course but for an exact interval for the normal distribution ↩

2. Many thanks to Suzy Whoriskey for all her contributions to the development of the course material ↩