Course Arrangements

Lectures:
Wednesday 9am: Herschel LT3
Wednesday 11am: Herschel LT1

Computer Classes (Weeks 3, 7)
Group 1: Monday 10am: Herschel Brig/Moss/Pond PC
Group 2: Monday 1pm: Herschel Brig/Moss/Pond PC

Problems Classes (Weeks 3, 5, 7, 9)
Thursday 4pm: Herschel LT3

Drop—in Classes (Weeks 4, 6, 8, 10)
Thursday 4pm: Herschel LT3

Assessment:
10% Coursework; 10% Test Exercises; 80% Exam
Course Structure

1. Revision of likelihood and Bayesian inference
2. Maximum–likelihood estimation
3. Hypothesis testing
4. Small sample methods
1. Revision of likelihood and Bayesian inference

- In MAS1301, Semester 2, you considered the concept of likelihood.

  You saw how the likelihood function could be maximized for a set of data, in order to estimate some parameters of interest.

- You were also introduced to the concept of Bayesian inference, where the likelihood function is used in combination with prior knowledge, in order to estimate the parameters in a completely different way.

  In Section 1 we will revise these ideas.
1.1 The likelihood function

The likelihood function is used in both Classical (i.e. non-Bayesian) and Bayesian inference.

If we have some data, and a statistical model, then the likelihood is the way in which the data gives us information about the model.

The model in question usually takes the form of a probability distribution which is specified by parameters.
Example 1.1.1 The likelihood function for the Poisson distribution

Suppose we model the number of goals scored in any Premier League football match as a Poisson random variable with parameter $\mu$.

Consider the following data: in the last full weekend of Premier League matches, the numbers of goals scored were:

1, 1, 3, 11, 4, 2, 0, 1, 2, 8.

We wish to evaluate the likelihood function.
Solution

If each recorded number of goals is $x_i, i = 1, \ldots, 10$, then the likelihood for each value of $x_i$ is given by

$$L(\mu|x_i) = \Pr(X = x_i|\mu) = \frac{\mu^{x_i} e^{-\mu}}{x_i!}.$$ 

The likelihood function for $\mu$ given all 10 values of $x_i$ is the product of the individual likelihood functions for each $x_i$.

We write this as

$$L(\mu|\mathbf{x}) = \prod_{i=1}^{10} L(\mu|x_i) = \prod_{i=1}^{10} \frac{\mu^{x_i} e^{-\mu}}{x_i!},$$

where $\mathbf{x} = (x_1, \ldots, x_{10})$. 

Dr. David Walshaw

MAS2302/MAS3302: Introduction to Statistical Inference: Week 1
The likelihood function $L(\theta|x)$ for a parameter $\theta$ given observations $x = (x_1, x_2, \ldots, x_n)$ on a random sample $x_1, x_2, \ldots, x_n$ is

$$L(\theta|x) = Pr(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n|\theta) = p(x_1|\theta) \times p(x_2|\theta) \times \cdots \times p(x_n|\theta)$$

when the $X$s are discrete random variables, each with probability mass function $p(x|\theta)$, and

$$L(\theta|x) = f(x_1|\theta) \times f(x_2|\theta) \times \cdots \times f(x_n|\theta)$$

when the $X$s are continuous random variables, each with probability density function $f(x|\theta)$. 
Now suppose we want to estimate the parameter $\mu$ from these data.

We do this by finding the value of the parameter $\mu$ which maximizes the likelihood.

This is usually most easily achieved using the logarithm of the likelihood function, known as the *log–likelihood*, and denoted $l(x|\mu)$.

Let’s see how this is achieved for our football data.
Solution

- The log–likelihood is given by

\[
\ell(\mu|x) = \log[L(\mu|x)] = \log \left[ \prod_{i=1}^{10} \frac{\mu^{x_i} e^{-\mu}}{x_i!} \right]
\]

\[
= (\log \mu) \sum_{i=1}^{10} x_i - 10\mu - \sum_{i=1}^{10} \log(x_i!).
\]

- If we differentiate this with respect to \( \mu \), and set equal to zero, we obtain the maximum–likelihood estimate:

\[
\hat{\mu} = \bar{x} = 3.3.
\]

[See exercises!]
Example 1.1.2 The likelihood function for the Normal distribution

An exploratory data analysis of a sample of cholesterol measurements from a group of patients suggests that the Normal distribution might be a good model for cholesterol levels for people with this condition. I.e. the distribution of the cholesterol level $X$ for this group is $X \sim N(\mu, \sigma^2)$.

Now suppose the data consist of the following 25 measurements (units are mmol/L):

4.57  5.20  6.16  6.20  5.33  5.57  4.43  5.31  5.72  6.36
4.28  4.71  4.23  5.48  6.09  6.11  5.80  5.50  3.99  5.41
5.90  5.81  6.17  5.59  5.65

We wish to evaluate the likelihood function for the parameters $\mu$ and $\sigma^2$. 
Here, for each individual’s cholesterol measurement $x_i, i = 1, \ldots, 25$, the likelihood is given by

$$L(\mu, \sigma^2 | x_i) = f_X(x_i | \mu, \sigma^2) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{x_i - \mu}{\sigma}\right)^2}.$$ 

Hence the likelihood given all the data $x = (x_1, \ldots, x_{25})$ is

$$L(\mu, \sigma^2 | x) = \prod_{i=1}^{25} L(\mu, \sigma^2 | x_i) = \prod_{i=1}^{25} \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{x_i - \mu}{\sigma}\right)^2}.$$
As before, we can estimate the parameters $\mu$ and $\sigma$ by maximizing the likelihood function with respect to $\mu$ and $\sigma$.

Once again, this is most easily achieved using the log–likelihood $l(\mu, \sigma^2 | x) = \log[L(\mu, \sigma^2 | x)]$. 
Solution

The log–likelihood is given by

\[
I(\mu, \sigma^2 | x) = \log[\mathcal{L}(\mu, \sigma^2 | x)] = \log \left[ \prod_{i=1}^{25} \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left( \frac{x_i - \mu}{\sigma} \right)^2} \right]
\]

\[
= -25 \log(\sigma \sqrt{2\pi}) - \frac{1}{2} \sum_{i=1}^{25} \left( \frac{x_i - \mu}{\sigma} \right)^2.
\]

Differentiating this with respect to each of \( \mu \) and \( \sigma^2 \) and setting equal to zero gives the maximum likelihood estimates:

\[
\hat{\mu} = \bar{x} = 5.4228; \quad \hat{\sigma}^2 = \frac{1}{25} \sum_{1}^{25} (x_i - \bar{x})^2 = 0.4517.
\]
1.2 Bayesian inference

- Bayesian inference is not just a different statistical technique to classical inference. It is a completely separate approach to inference, with a different underlying philosophy.

  People who believe that the Bayesian approach is the most sensible one are often referred to as *Bayesian statisticians*, or just *Bayesians*.

- One of the key differences between the Bayesian approach and the classical approach is that Bayesians believe that the best analysis is *not* necessarily objective.

  Bayesian inference works by combining objective information, obtained through the data, with *subjective* information, expressed in the form of prior beliefs.
Suppose that we have a single parameter $\theta$. Suppose also that we have a prior distribution $\pi(\theta)$ which expresses our beliefs about $\theta$.

To get information about $\theta$ we observe $x_1, x_2, \ldots, x_n$ from a distribution with probability (mass or density) function $f(x|\theta)$. The likelihood function for $\theta$ is therefore given by:

$$L(\theta|\underline{x}) = f(x_1|\theta) \times f(x_2|\theta) \times \cdots \times f(x_n|\theta).$$
Bayes Theorem

- The posterior (density) function for \( \theta \) is given by

\[
\pi(\theta|\mathbf{x}) \propto \pi(\theta)L(\theta|\mathbf{x})
\]

i.e.

\[
\text{posterior} \propto \text{prior} \times \text{likelihood}.
\]

- The posterior distribution represents our beliefs about \( \theta \) after taking into account both our prior beliefs and the data.
We now take a Bayesian approach to inference on the parameter $\mu$ which specifies the Poisson model for the number of goals per Premier League match.

It is useful to remember here that if the Poisson model were correct, then $\mu$ would be the mean number of goals per game.

Now $\mu$ is considered to be a random variable, and we must specify our prior beliefs about $\mu$ as a probability distribution.

As a prior distribution for $\mu$, I express my personal beliefs as a Gamma distribution with mean 2.25, and standard deviation 0.25. We wish to obtain the posterior distribution.
Solution

From MAS1301 we know that if \( Y \) is a Gamma random variable with parameters \( \alpha \) and \( \lambda \), i.e. \( Y \sim \Gamma(\alpha, \lambda) \), then \( Y \) has mean \( \alpha/\lambda \), and variance \( \alpha/\lambda^2 \), giving standard deviation \( \sqrt{\alpha}/\lambda \).

So here we determine the values of \( \alpha \) and \( \lambda \) by setting

\[
\frac{\alpha}{\lambda} = 2.25; \\
\frac{\sqrt{\alpha}}{\lambda} = 0.25.
\]

Solving these gives us \( \alpha = 81 \) and \( \lambda = 36 \), so our prior for \( \mu \) is

\( \mu \sim \Gamma(81, 36) \).

We must now calculate the posterior distribution.
Solution

- Recall the numbers of goals scored are 1, 1, 3, 11, 4, 2, 0, 1, 2, 8.

  Hence the sample size is \( n = 10 \), and the sample mean is \( \bar{x} = 3.3 \).

- From MAS1301, we know that the posterior distribution for \( \mu \) is obtained directly as a \( \Gamma(\alpha + n\bar{x}, \lambda + n) \) distribution.

Here, this gives us the posterior distribution for \( \mu \) as

\[
\mu | \bar{x} \sim \Gamma(114, 46).
\]
Revision of likelihood and Bayesian inference

Figure 1

Dr. David Walshaw
MAS2302/MAS3302: Introduction to Statistical Inference: Week 1
The prior distribution I used to specify my beliefs about $\mu$ was rather strongly peaked, because I felt confident that the parameter lies in quite a narrow range of possible values. For this reason, the prior and posterior densities are rather similar to each other — the data had little impact (see Figure 1).

The Gamma distribution has a very special property when used as a prior for the Poisson parameter: it causes the posterior distribution to have the same form, i.e. to be another Gamma distribution. This property is called conjugacy.

We could use a different form for our prior distribution if it represented our beliefs, but it would not have the property of conjugacy, and obtaining the posterior distribution would be a hard task!
Example 1.1.2 revisited

- Remember the cholesterol level for a patient is given by $X \sim N(\mu, \sigma^2)$. Suppose we are interested in making inference on the mean $\mu$, and we assume that the variance $\sigma^2$ is equal to the population variance for healthy individuals, known to be $0.6^2$.

- Following an example in MAS1301, we express the distribution of $X$ as $X \sim N(\mu, \frac{1}{w})$. So here we are assuming we know $w$ to be $\frac{1}{0.6^2} = 2.778$.

- I need to specify my prior for $\mu$. I will use a Normal distribution for the prior, so it will be of the form $\mu \sim N(\mu_0, \frac{1}{kw})$. 
Since I don’t have much idea of the effect that the condition in question has on cholesterol levels, I will specify a fairly vague (flat) prior distribution:

\[ \mu \sim N(6, 2^2). \]

I.e. \( \mu_0 = 6 \) and \( k = \frac{0.6^2}{2^2} = 0.3^2 = 0.09. \)

We now need to calculate the posterior distribution.
Using the result from MAS1301, the posterior is given by

$$\mu \mid \bar{x} \sim N \left( \frac{k\mu_0 + n\bar{x}}{k + n}, \frac{1}{(k + n)\omega} \right).$$

I.e. $\mu \mid \bar{x} \sim N(5.425, 0.1198^2)$. 
Revision of likelihood and Bayesian inference

Figure 2

Dr. David Walshaw
MAS2302/MAS3302: Introduction to Statistical Inference: Week 1
Notes

- Note the contrast with the previous example!

- This time the data has provided us with relatively much more information compared to the prior.

- The prior is rather flat, whereas the posterior is quite sharply peaked.

- Because the data has provided much new information this time, the prior and posterior distributions are very different from each other.
Review of Section 1

We have done the following:

1. Revised the idea of the likelihood function for discrete and continuous distributions, and given an example of each.
2. Obtained the maximum likelihood estimates for the parameters in each of the examples above.
3. In addition, adopted a Bayesian approach to inference for each of the examples.
4. Contrasted the philosophies underlying the classical and Bayesian approaches.