Introduction to Statistical Modelling

T.J. McKinley (t.mckinley@exeter.ac.uk)

Statistical inference

Statistical inference can be thought of the ***inverse*** of simulation.

That is, we observe some data and want to know:

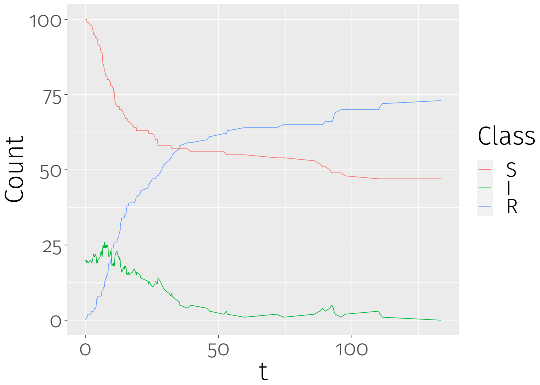
> *What **parameter values** for a model produce the 'best fit' to the data?*

We can use this to provide insights into key epidemiological processes (e.g. e.g. estimating the transmission rate, R_0 etc.). We can also use this to produce **predictions** and **forecasts**.

> *Key aspect is that we wish to quantify **uncertainty**.*

Example: SIR model

As an example, let's look at some data that we've simulated from a simple SIR model in a closed population of size $N = 120$, with the introduction of 20 initial infectives at time $t = 0$.



Example: SIR model

If we assume these data come from a **stochastic** SIR model of the form:

$$
\begin{aligned}
P(S_{t+\delta t} \to S_t - 1 \text{ and } I_{t+\delta t} \to I_t + 1) &\approx \beta S_t I_t, \\
P(I_{t+\delta t} \to I_t - 1 \text{ and } R_{t+\delta t} \to R_t + 1) &\approx \gamma I_t
\end{aligned}
$$

for small δt. We can then ask the question:

> *"What values of β and γ produce epidemic curves that are the most consistent with the* ***observations****?"*

Likelihood functions

> *"What values of β and γ produce epidemic curves that are the most consistent with the **observations**?"*

This question can be tackled by appealing to the **likelihood function**.

The *likelihood function*, $f(\mathbf{y} \mid \theta)$, gives the **likelihood**[†] of observing the data $(\mathbf{y})$ ***given*** a set of parameters (θ).

> *The exact form of the **likelihood** function depends on the **specific model** and **data**.*

[†] if the data, $\mathbf{y}$, are **discrete**, then this is a **probability**

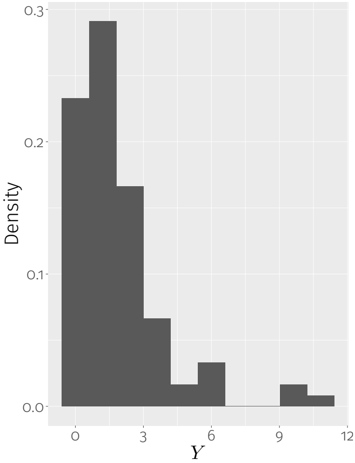
Likelihood functions

The exact form of the **likelihood** function depends on the **specific model** and **data**.

For example, imagine we have $n = 100$ **independent** samples from an **exponential** distribution:

$$Y_i \sim \text{Exp}(\lambda)$$

where λ is **unknown**.



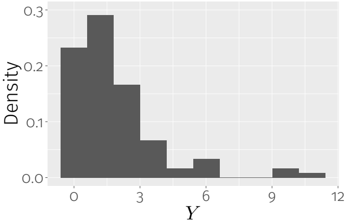
Likelihood functions

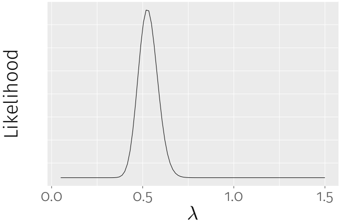
The exact form of the **likelihood** function depends on the **specific model** and **data**.

If the data are **independent**, then

$$f(\mathbf{y} \mid \lambda) = \prod_{i=1}^{n} f(y_i \mid \lambda)$$
$$= \prod_{i=1}^{n} \lambda e^{-\lambda y_i}$$

which is a function of λ and is dependent on the **probability density function** for each **observation** y_i.





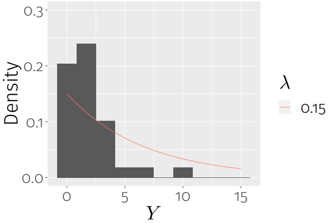
Likelihood functions

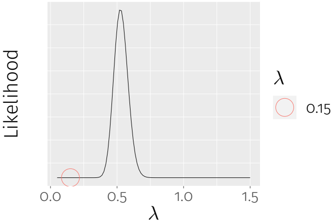
The exact form of the **likelihood** function depends on the **specific model** and **data**.

If the data are **independent**, then

$$
\begin{aligned}
f(\mathbf{y} \mid \lambda) &= \prod_{i=1}^{n} f(y_i \mid \lambda) \\
&= \prod_{i=1}^{n} \lambda e^{-\lambda y_i}
\end{aligned}
$$

which is a function of λ and is dependent on the **probability density function** for each **observation** y_i.





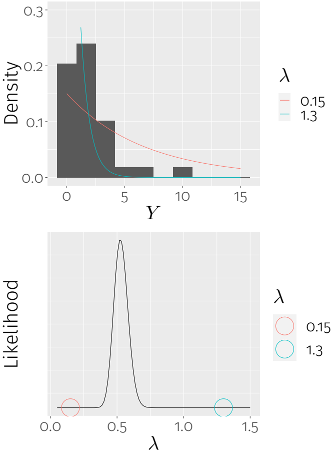
Likelihood functions

The exact form of the **likelihood** function depends on the **specific model** and **data**.

If the data are **independent**, then

$$
\begin{aligned}
f(\mathbf{y} \mid \lambda) &= \prod_{i=1}^{n} f(y_i \mid \lambda) \\
&= \prod_{i=1}^{n} \lambda e^{-\lambda y_i}
\end{aligned}
$$

which is a function of λ and is dependent on the **probability density function** for each **observation** y_i.



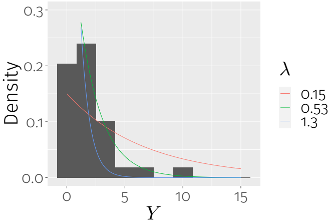
Likelihood functions

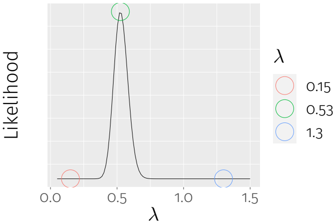
The exact form of the **likelihood** function depends on the **specific model** and **data**.

If the data are **independent**, then

$$
\begin{aligned}
f(\mathbf{y} \mid \lambda) &= \prod_{i=1}^{n} f(y_i \mid \lambda) \\
&= \prod_{i=1}^{n} \lambda e^{-\lambda y_i}
\end{aligned}
$$

which is a function of λ and is dependent on the **probability density function** for each **observation** y_i.



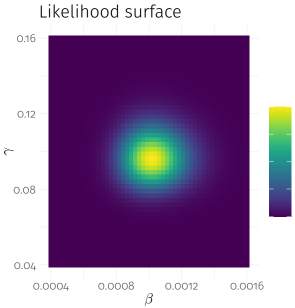


Likelihood functions

The **likelihood function** can be thought of as a **function** of the **unknown parameters** θ.

In the case of our SIR model, we have $\theta = (\beta, \gamma)$.

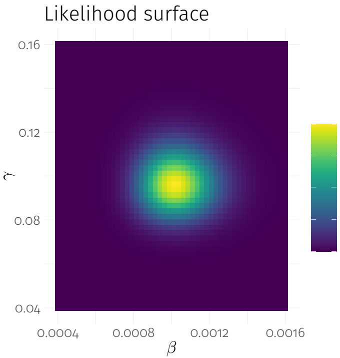
The **likelihood surface** (for different values of β and γ) looks like the plot opposite.



Note that in general **likelihoods** for compartmental models like this are **intractable**†, but in this simulated setting we can write it down directly.

†since **data** points are generally **not independent**, and typically the likelihood also depends on **unobserved variables**—we will return to this later

Likelihood functions



We can see that parameter values in the **yellow** region, produce **higher** likelihood values than parameter values in the **dark blue** regions.

*This means that parameters in the **yellow** region would produce simulations that are **more consistent** with the observed data than parameters in the **dark blue** regions.*

Maximum likelihood

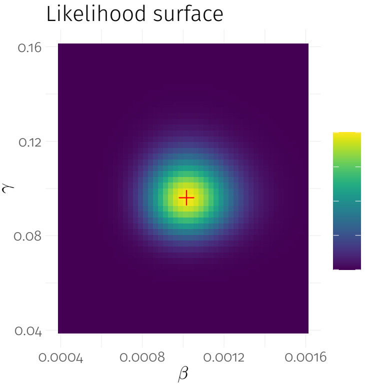
A natural way to estimate the parameters is to ask:

What parameter values ***maximise*** *the likelihood function*[†]*?*

Here the **maximum likelihood** estimates are shown with a **red cross**, and are given by:

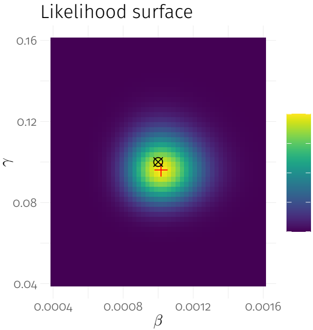
$$\hat{\beta} = 0.00102 \text{ and } \hat{\gamma} = 0.0961,$$

to 3 *significant figures*.



[†] we will see an alternative approach—using the **Bayesian** framework—later

Likelihood functions



- The **absolute value** of the likelihood is rarely interpretable, only **relative** values.
- The likelihood is based on the **data** and the choice of **model**, and thus will change for different data sets and different models.
- ML estimates do not guarantee a **good fit**.
- Similar parameter values can give similar fits (**uncertainty**).

Confidence intervals

Uncertainties in the parameter estimates can be quantified using **confidence intervals**. **Wider** confidence intervals signal **larger** uncertainties.

Here 95% confidence intervals[†] are:

- β: (0.000743, 0.00129)
- γ: (0.074, 0.118)

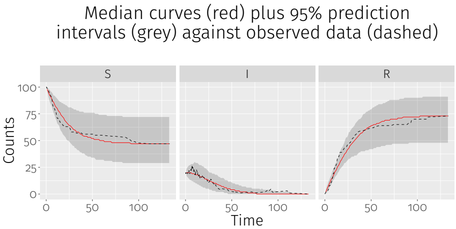
> ***Note***: *these do **not** correspond to a 95% probability that the true value is between the limits. Rather, it means that if the experiment were to be conducted an **infinite** number of times, 95% of the time the calculated CI would contain the true value*[‡].

[†] based on a **large sample** approximation

[‡] this is so-called **frequentist** inference, as opposed to **Bayesian** inference that we will cover shortly

Model checking and prediction

We can check the model fit using the ML estimates to seed a large number of simulations from the model, and plot these against the observed data.



Here the model produces simulations that are consistent with the data[†].

Note that the uncertainty bounds here **do not** account for the **parameter uncertainty**[‡]; to calculate a **true prediction interval** for these types of model is harder (see Gelman and Hill (2007) for *simulation-based* approaches).

[†] be careful, simulations from stochastic models can be tricky—see McKinley, Cook, and Deardon (2009)

[‡] the parameters are **fixed** at the MLEs

Practical

In the first practical we will explore fitting the **catalytic model** for endemic diseases to serology data for ***rubella***.

To do this, we will need to write down a **likelihood function**, and then use one of R's in-build **optimisation** functions (`optim()`) to maximise with respect to the parameters to find the **maximum likelihood estimates**.

References i

Gelman, Andrew, and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

McKinley, Trevelyan J., Alex R. Cook, and Robert Deardon. 2009. “Inference in Epidemic Models Without Likelihoods.” *The International Journal of Biostatistics* 5 (1). https://doi.org/10.2202/1557-4679.1171.