Statistical Inference: One Parameter

Bob O'Hara & Emily Simmonds

Administration Matters

- Reference Group
- Blackboard

This week

Estimating parameters and Likelihood

- Sampling the earth. What proportion is land?
- resampling the earth
  - sampling *variation* in estimates
- Binomial likelihood
- Estimation: Maximising the likelihood to find the best estimate

Our Problem

What proportion of the earth is land?

If we have a globe, how can we estimate what proportion in land and what proportion sea?

(plant cover is a real example of this problem)

The strategy

- Get Some data
- Learn about variation in the data
  - need a model for the data
- Inference: work out the distribution of estimates
  - find the "best" estimate from the distribution

Next week we can use that to decide how confident we are

Get Some Data: Sampling The Earth

Toss the globe around

When you catch it. put your finger on a point, and say whether it lands on the land or sea

Then toss it to someone else

We will record the number of times we get `Land` or `Sea`, and use this as an estimate of the proportion of the globe that is land

Data Variation: Resampling The Earth in your heads

In a moment we will do the same exercise again, but first I want you to think about what numbers you might get.

If we did this exercise in 10 classes, what values do you think we would get? Guess at some possible values

e.g. if we had 3 "earths" out of 12, we might imagine getting 3, 6, 3, 2, 1, ..., 9

Data Variation: Resampling The Earth

As before. . .

Toss the globe around

When you catch it. put your finger on a point, and say whether it lands on the land or sea

Then toss it to someone else

Data Variation: Resampling The Earth On the Computer

Now we will simulate the resampling

Data Variation: The Model I

Each observation is a sample from the real world

- "Bernoulli trial"

We observe N trials, of which n are land, and $(N - n)$ are water

Data Variation: The Model II

We can assume that each time we look at whether the sampling is "land" or "sea", there is a probability that it is "land"

- probability constant
- each trial is independent

If we know the probability we can simulate this

Data Variation: The Simulation

R has a function `rbinom()`. We can use it like this:

```
prob <- 0.4
sim <- rbinom(10, 1, prob)
sim
```

```
##  [1] 0 1 1 0 0 1 0 0 0 1
```

We can interpret 1 as Land and 0 as Sea.

Data Variation: The Simulation Function

We will build the function: first a function that returns 0 or 1 1s -> Land Count the number of 'Land*s

```
source("Week2Functions.R")

simGlobe(probability = 0.4, NTrials = 5)
```

```
##
##  Sea Land
##    5    0
```

```
simGlobe(0.4, 5)
```

```
##
##  Sea Land
##    3    2
```

Sidenote: passing arguments to R functions

```
simGlobe(probability = 0.4, NTrials = 5)
```

```
##
##  Sea Land
##    3    2
```

```
simGlobe(0.4, 5)
```

```
##
##  Sea Land
##    1    4
```

Three arguments are defined for simGlobe().

If we pass 2 arguments to the function without giving names, R will use the first as the first argument & the second as the second.

For the third the default will be used

- if there is no default, R will throw an error

Data Variation: Repeating Simulations

We can repeat a function several times:

```
simGlobe(probability = 0.4, NTrials = 20, nSims = 3)
```

```
##
##         [,1] [,2] [,3]
##    Sea    10   12   13
##    Land   10    8    7
```

```
(Lands <- simGlobe(0.4, 20, 3)["Land",])
```

```
## [1]  5 10  8
```

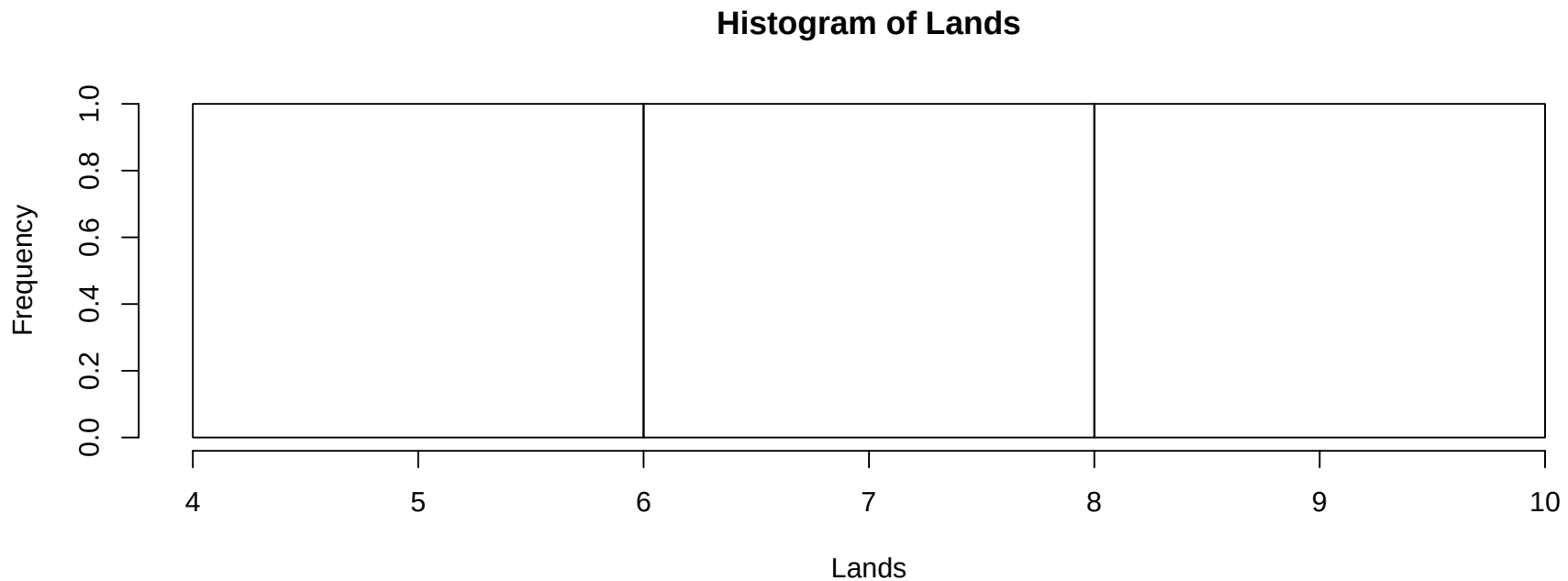
Data Variation: What to do

Use the `simGlobe()` function to simulate getting more data

- use the same `NTrials` as we used
- decide on a "good" value of the probability

Do 1000 simulations (i.e. set nSims $= 3$). Compare your simulations with your guesses, and with what we actually got. You can use hist():

```
hist(Lands)
```



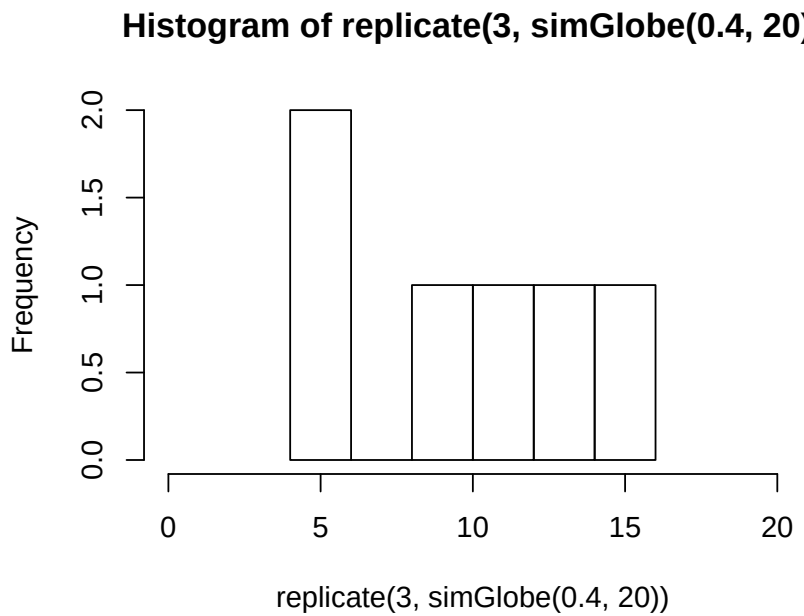
Data Variation: Different Probabilities

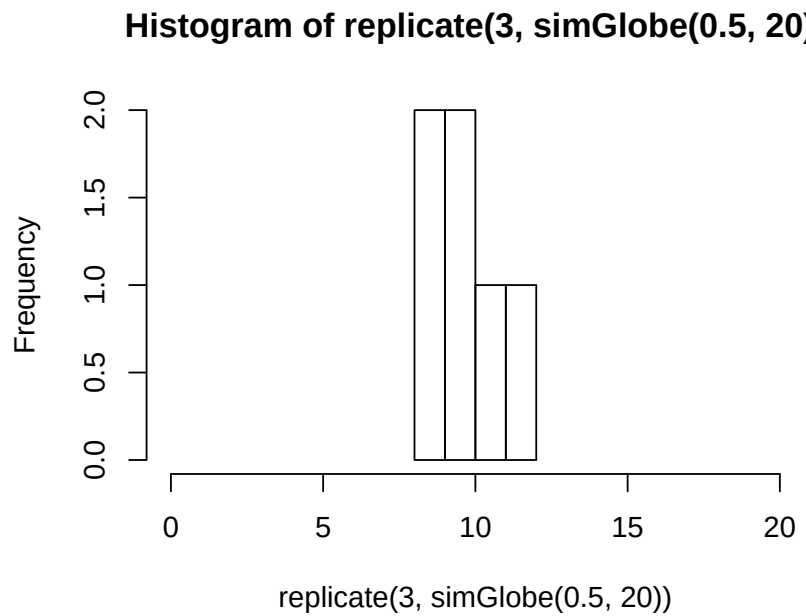
Now we have some idea about the variation in the results we could get from one parameter, what if there is another parameter?

Data Variation: Different Probabilities

Simulate the data with a value of `probability` that is 0.2 higher

```
par(mfrow=c(1,2))
hist(replicate(3, simGlobe(0.4, 20)), xlim=c(0,20))
hist(replicate(3, simGlobe(0.5, 20)), xlim=c(0,20))
```





Data Variation: Recap

We now have some idea about how much variation there could be in the data.

Even with a fixed parameter, the results could be quite varied

With one data set, a range of parameters are reasonable

Inference: Finding Good Estimates

So how do we find a good estimate of the probability?

How do we know what are reasonable probabilities?

Inference: Likelihood

One way: find the `probability` that makes the data most likely

We know n follows a binomial distribution, with an unknown p

Inference: Some Maths

If we have 1 trial, the probability of Land is $\mathtt{p}$

If we have 2 trials, we could have Land-Land, Land-Sea, Sea-Land, Sea-Sea

So

$$Pr(2Land) = p^2$$
$$Pr(1Land) = 2p(1-p)$$
$$Pr(1Land) = (1-p)^2$$

Inference: A Mathematical Shortcut

If we have N trials, and observe r "successes" then the probability of this is

$$Pr(n = r|N, p) = \frac{N!}{r!(N-r)!} p^r (1-p)^{N-r}$$

Which has 2 parts. The important part is $p^r(1-p)^{N-r}$ which is

$$p^{successs}(1-p)^{failures}$$

Inference: The other part

The other part is

$$\frac{N!}{r!(N-r)!}$$

which counts the number of combinations of r successes and $N-r$ failures

e.g. if $N = 3$ and $r = 1$ we have

- success - failure - failure
- failure - success - failure
- failure - failure - success

So $\frac{3!}{1!(3-1)!} = 3$

Inference: Likelihood

If we know p (the probability of `Land`), we can calculate the probability of obtaining the data, given the parameter

- this is called the *likelihood*

But we don't know p: this is what we want to estimate

Inference: Using the Likelihood

We can calculate the likelihood for different values of p

```
NLand <- 4; NSea <- 6; N <- NLand + NSea
dbinom(NLand, N, 0.4) # calculate the likelihood
```

```
## [1] 0.2508227
```

```
dbinom(NLand, N, 0.4, log=TRUE) # the log-likelihood
```

```
## [1] -1.383009
```

We can also calculate several values:

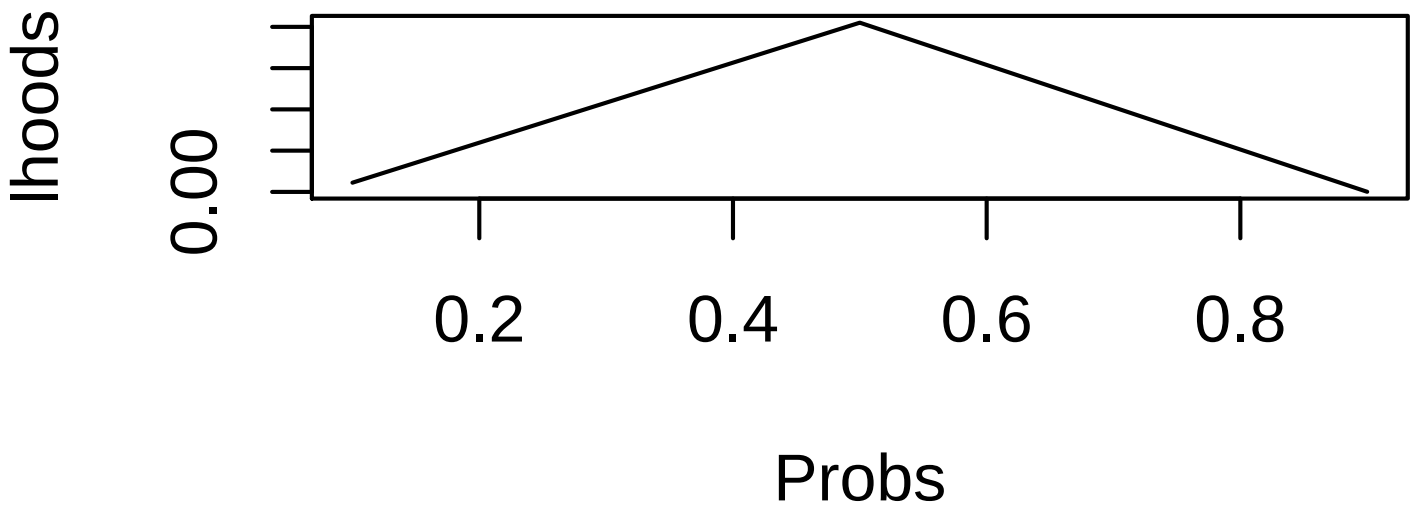
```
# seq() creates a sequence of numbers
Probs <- seq(from = 0.1, to = 0.9, length.out = 3)
(lhoods <- dbinom(NLand, size = N, prob = Probs))
```

```
## [1] 0.011160261 0.205078125 0.000137781
```

Inference: Your task, to Find a Good Likelihood

Calculate the likelihood for the data for different values (>3) of p

```
Probs <- seq(from = 0.1, to = 0.9, length.out = 3)
# NLand=data
lhoods <- dbinom(NLand, size = N, prob = Probs)
plot(Probs, lhoods, type="l")
```



From this plot, and trying a few values, can find the best likelihood?

Inference: The Philosophy

The likelihood is a data generating mechanism: it is a statistical model

We assume that the data are random, and the parameters (and model) are fixed

We want to find the parameters which are most likely to give rise to the data

- we maximise the likelihood

Inference: Maximising the likelihood

Poking around and trying values is not the best way to find the maximum.

Alternatives are:

- analytic: do the maths (works for this problem)
- numerical: use an algorithm that finds the maximum
- simulation: simulate the likelihood & find the best value

The maximum of the likelihood is the same as the maximum of the log-likelihood, so we usually work on the log scale

Inference: Maximising the log likelihood for the binomial

We can do this analytically. We want to get find an equation for the slope, then set this to zero.

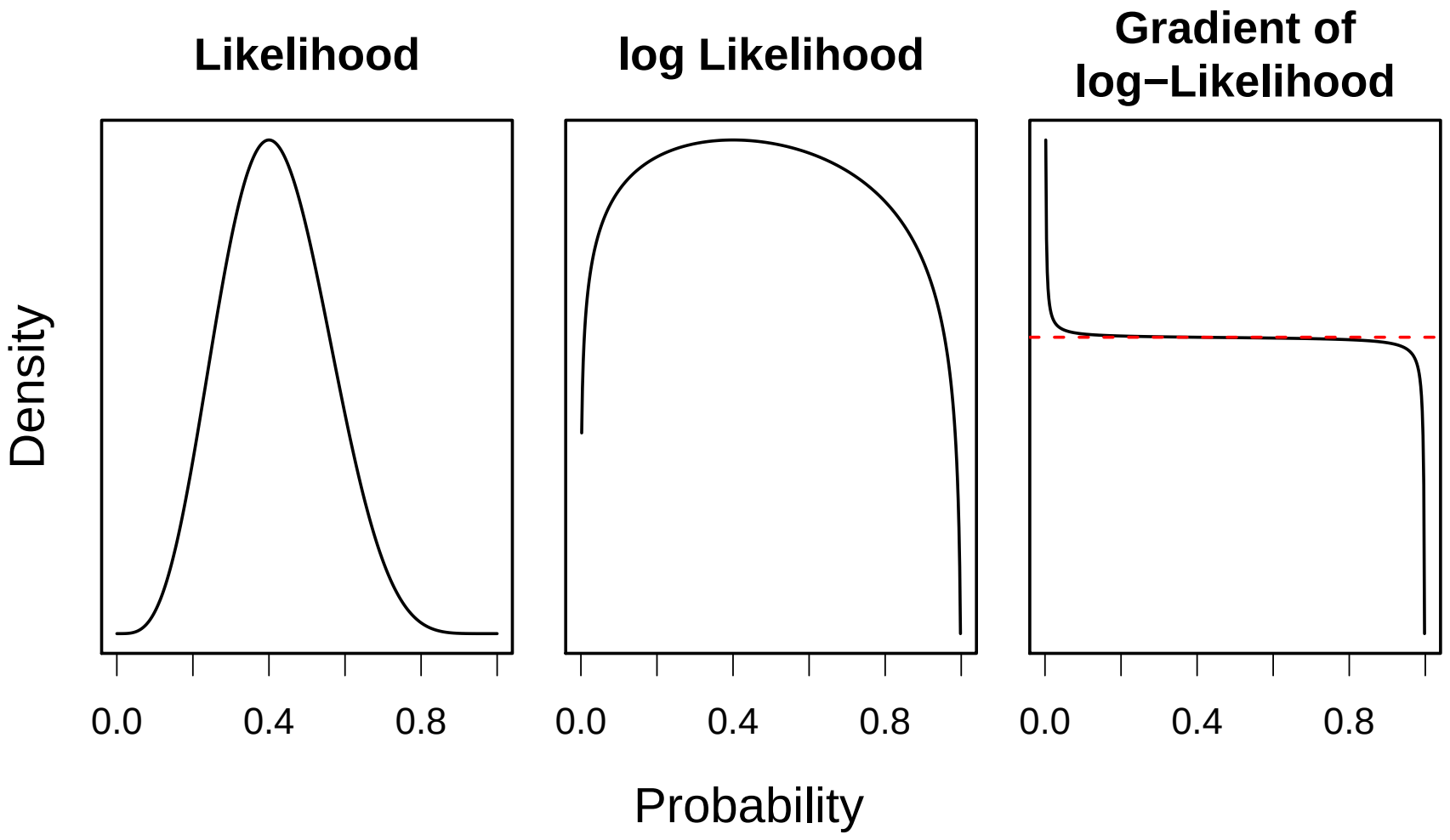
The likelihood is

$$l(p|n) = r \log(p) + (N - r) \log(1 - p) + C$$

and after we differentiate (to get the slope) we have

$$\frac{dl(p|n)}{dp} = \frac{r}{p} - \frac{N - r}{1 - p}$$

Inference: In Figures



Inference: Maximising

Set the gradient to 0:

$$0 = \frac{r}{p} - \frac{N-r}{1-p}$$

So

$$\frac{p}{1-p} = \frac{r}{N-r}$$

i.e. the *odds* of success are equal to the ratio of successes to failure.

We can re-arrange to get

$$\hat{p} = \frac{r}{N}$$

So...

We have maximised the likelihood to get an estimator of p

$$\hat{p} = \frac{r}{N}$$

In more complicated problems we do the same thing, but sometimes the maximisation is done numerically (or even through simulation)

But we always use the log-likelihood & ignore the normalising constants

Terminology

We call p the *estimand*: this is what we want an estimator of

We will call the *estimator* $\hat{p}$

Because we will get $\hat{p}$ by maximising the likelihood, we call it the *maximum likelihood estimator* (MLE).

Inference: What happens if we take another sample?

e.g. the second time we sampled the earth, we had 6 `Land` and 4 `Sea`

Will we get the same estimate?

Would this mean the estimand is different?

Inference: What happens if we take another sample?

Each sample gives us a different $\hat{p}$

p is fixed, and the data are random, so $\hat{p}$ is a property of the data

We can sample repeatedly many times, and each time get a different $\hat{p}$

The likelihood is the distribution of $\hat{p}$

More samples: your task

Look at the distribution of possible estimates of p.

- Assume the "true" value is 0.4.
- Simulate the data for 10 trials
- Calculate the maximum likelihood estimator, using the `mleGlobe()` function

```
mleGlobe(NLand=5, NTrials=10)
```

```
## [1] 0.5
```

```
mleGlobe(NLand=simGlobe(0.5, 10, 3)["Land",], NTrials=10)
```

```
## [1] 0.6 0.4 0.7
```

Repeat this many times, and plot a histogram of the estimates

Feedback

We want to know how well we are doing, and if we can improve things. So please fill in this form:

https://docs.google.com/forms/d/e/1FAIpQLSepCvw6pdNC8LMjlHUtqzNvP9uTl3KJOEeSB7l9lb6aqfLh3w/viewform?usp=sf_link

More data

So far we have only looked at one data point. But in reality we may have several.

e.g. when we sampled the globe, we had - Land 6 times, Sea 7 times, and then - Land 7 times, Sea 6 times

For “real” problems we usually have many more than 1 observation.

More than one datum: the probability

If data are independent, then

$$Pr(X_1 \& X_2) = Pr(X_1)Pr(X_2)$$

So we can multiply the probabilities

In general, then

$$Pr(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} Pr(X_i)$$

More than one datum: the likelihood

The likelihood for the parameters (θ) given the data is

$$L(\theta|X_1, X_2, ..., X_n) = \prod_{i=1}^{n} Pr(X_i|\theta)$$

So, on the log scale

$$l(\theta|X_1, X_2, ..., X_n) = \sum_{i=1}^{n} \log(Pr(X_i|\theta))$$

we just add the log-likelihoods together

- easier than multiplying!

Summary

We have seen that data vary

We can estimate the “best” parameters from the data, using the likelihood

Different data will give different “best” parameters, even if the process is the same

Next Week

Summarising the uncertainty in the estimates