Statistical Inference: One Parameter

Bob O'Hara & Emily Simmonds

Administration Matters

- Reference Group
- Blackboard

Simple modelling

Today we will start modelling

- start to think about variability

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)

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

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

Resampling 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

Resampling The Earth On the Computer

Now we will simulate the resampling

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

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

The Simulation

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

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

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

We can interpret 1 as Land and 0 as Sea.

The Simulation Function I

We will build the function: first a function that returns 0s or 1s

```
simGlobe <- function(probability=0.5, NTrials=10) {
  sim <- rbinom(NTrials, 1, probability)
  return(sim)
}
simGlobe(probability = 0.4, NTrials = 5)
```

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

The Simulation Function II

Next, return Land or Sea

```
simGlobe <- function(probability=0.5, NTrials=10) {
  sim <- rbinom(NTrials, 1, probability)
  Res <- c("Sea", "Land")[1+sim]
  return(Res)
}
simGlobe(probability = 0.4, NTrials = 5)
```

```
## [1] "Land" "Land" "Land" "Sea"  "Land"
```

The Simulation Function III

Count the number of Land and Sea

```
simGlobe <- function(probability=0.5, NTrials=10) {
  sim <- rbinom(NTrials, 1, probability)
  Res <- c("Sea", "Land")[1+sim]
  return(table(Res))
}
# Count number of times the simulation returns "Land"
simGlobe(probability = 0.4, NTrials = 5)
```

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

Repeating Simulations

We can repeat a function several times:

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

```
##
## Res     [,1] [,2] [,3]
##   Land     9    6    8
##   Sea     11   14   12
```

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 `prob`

Compare your simulations with your guesses, and with what we actually got

(try the `hist()` function to plot the data)

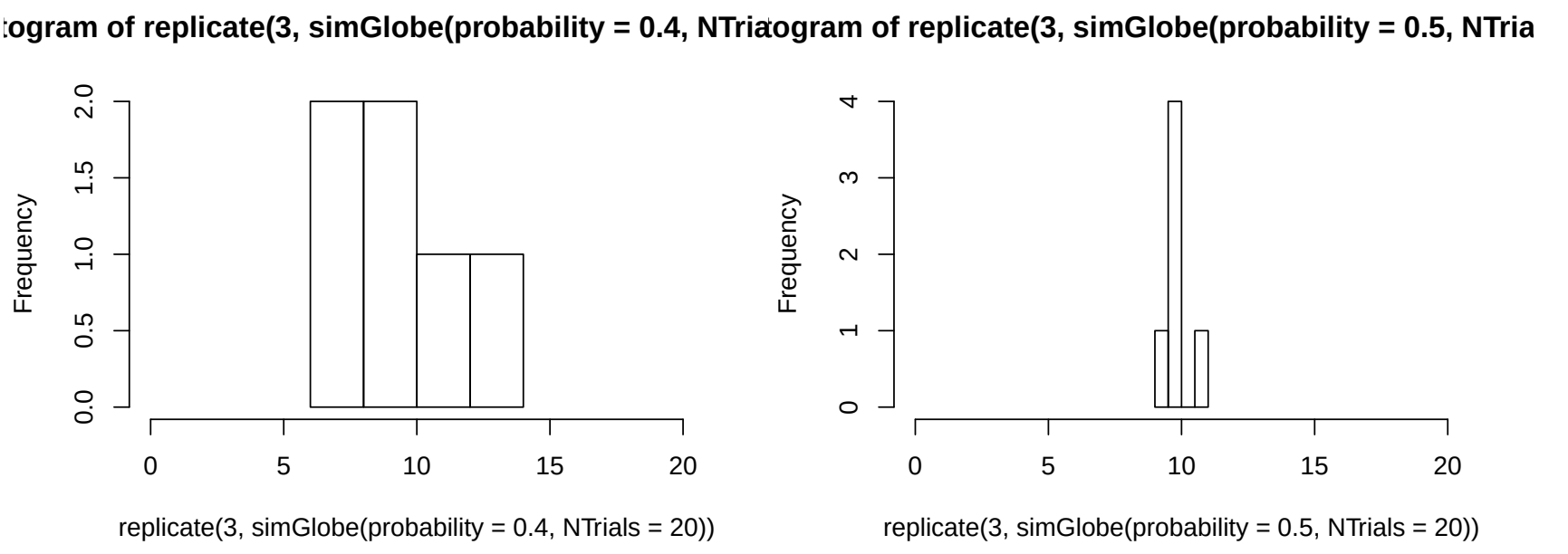
Different Probabilities

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

Different Probabilities

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

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



The Inference Problem

We have seen that even with one probability, we can get a range of observations. And we can get the same observation with a range of probabilities

So how do we find a good probability?

How do we know what are reasonable probabilities?

Likelihood

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

n follows a binomial distribution, with an unknown p (the population-level mean)

Some Maths

If we have 1 trial, the probability of Land is p

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

So

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

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

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

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

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

Using the Likelihood

We can calculate the likelihood for different values of p

In R:

```
NLand <- 4
NSea <- 6
N <- NLand + NSea

dbinom(NLand, N, 0.4)
```

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

We can also calculate several values:

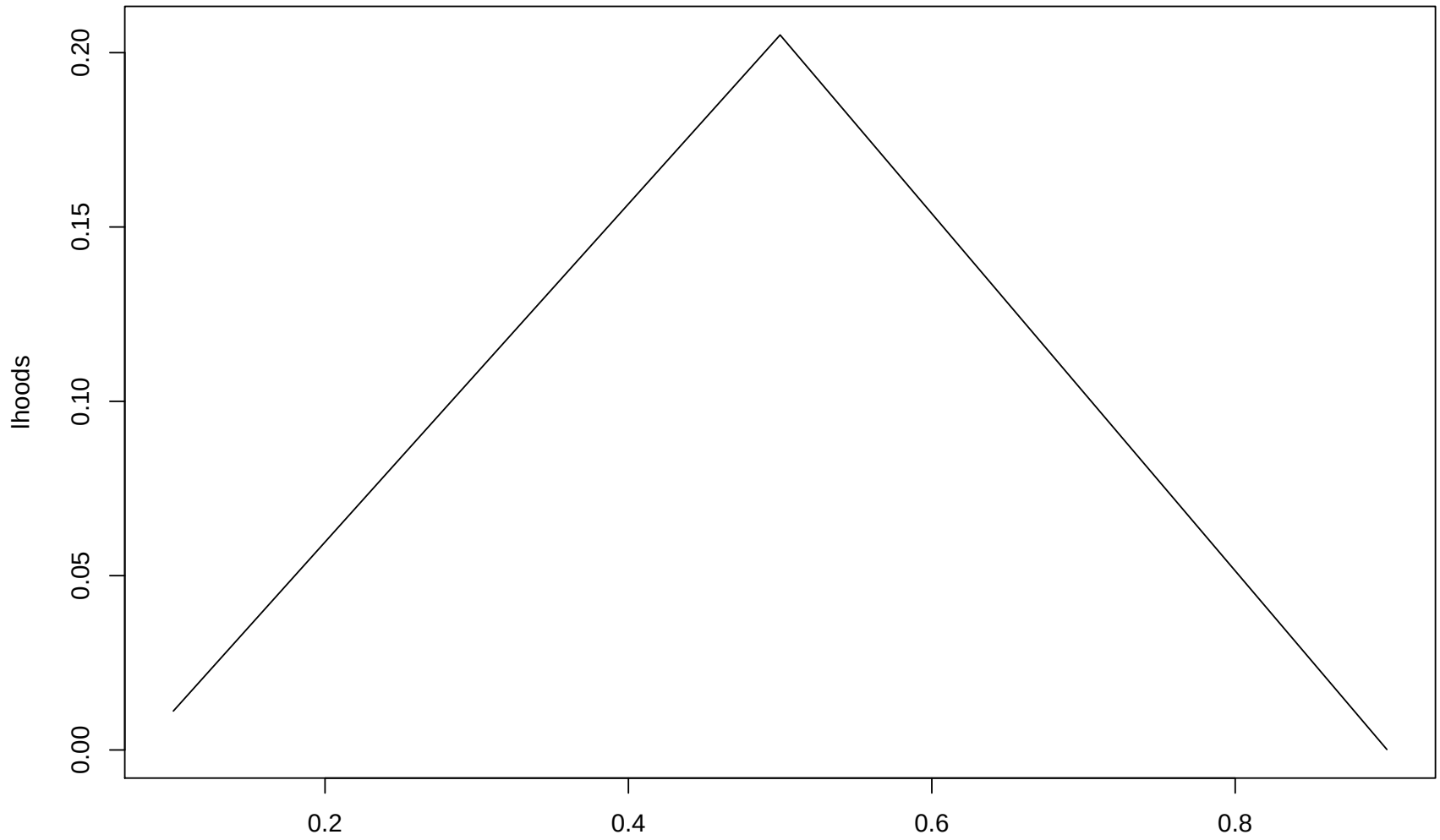
```
# seq() creates a sequence of numbers
Probs <- seq(from = 0.1, to = 0.9, length.out = 3)
(lhoods <- dbinom(NLand, size = N, prob = Probs))
```

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

Finding a Good Likelihood

Your task: calculate the likelihood for the data for different values of p

```
plot(Probs, lhoods, type="l")
```



Finding the Best Likelihood

From this, can you find the best likelihood?

Terminology

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

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