# Statistical Inference: Uncertainty About One Parameter

# Recap of Last Week

We tossed a beach ball around

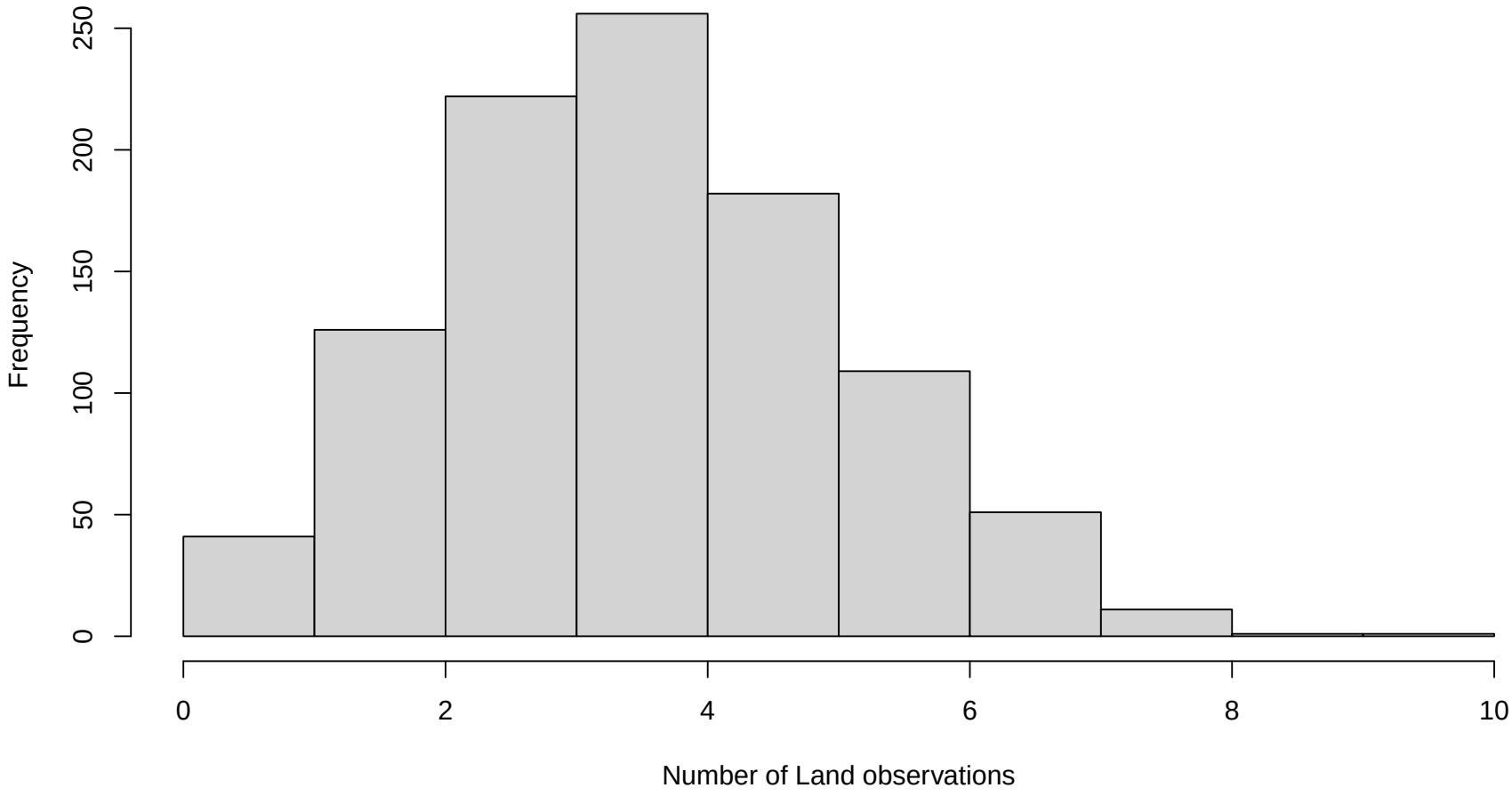
We saw land 6 times and sea 7

The second time we saw land 7 times and sea 6

We want to estimate the proportion of land

# Recap of Last Week

We saw that there would be variation when we replicate the experiment



# Recap of Last Week

We can estimate the proportion by

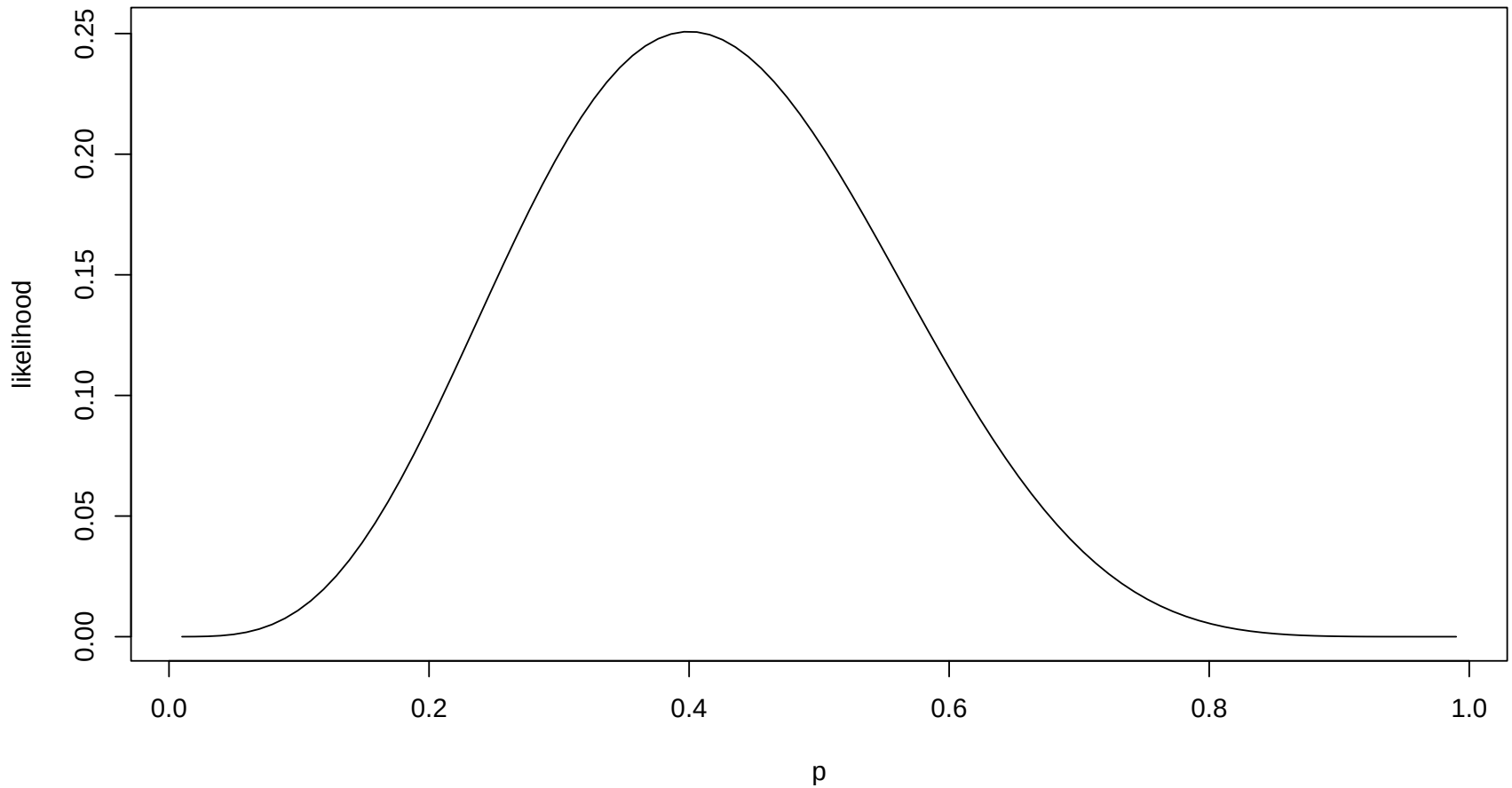
- building a model
- finding the parameters that are model likely to give the data

This is the *maximum likelihood estimate*

- maximise Pr(Data;parameters) with respect to the parameters

# Recap of Last Week

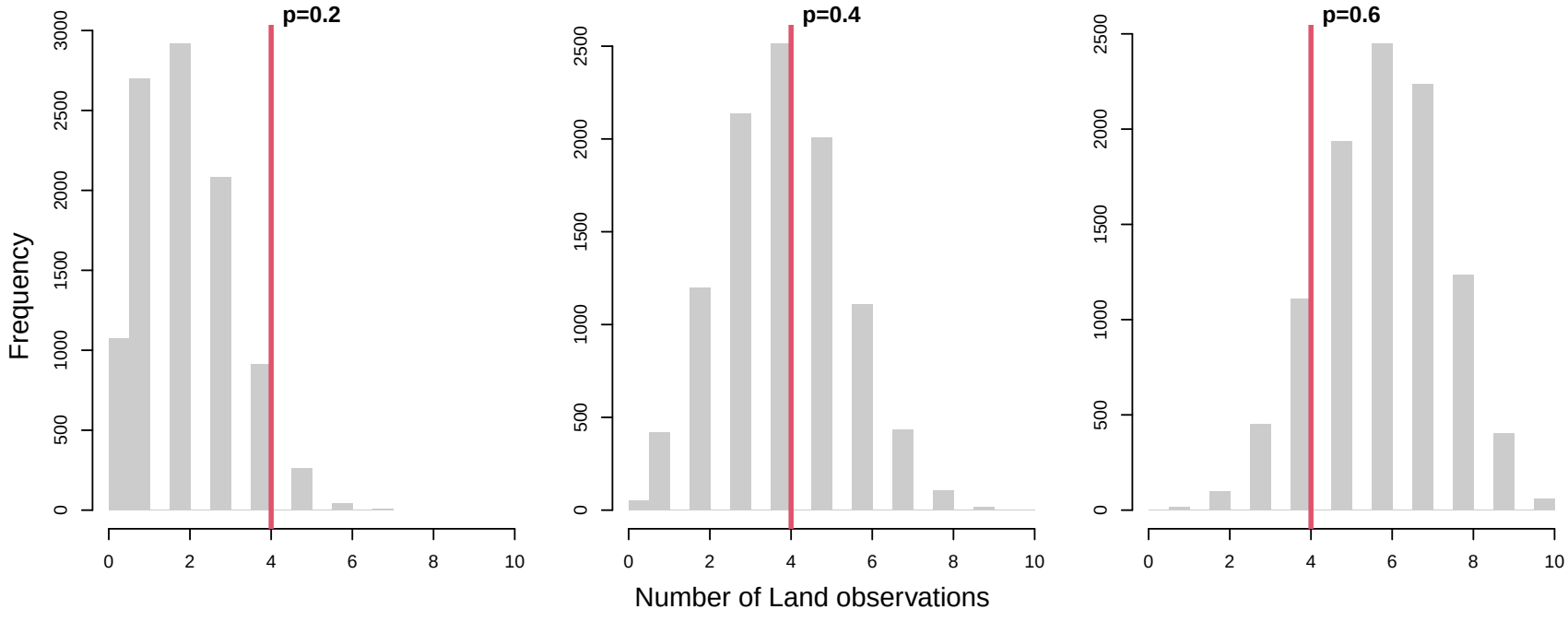
For this problem we can maximise the likelihood analytically



# This week

How good is our estimate?

We saw last week that different values of $p$ can give the same data



# Outline

Repeated sampling of data

Summarising the variation in the resamples - confidence intervals

What is a confidence interval?

Asymptotics: approximations when the numbers are big

Standard errors

# The Question

Because different samples give different estimates, we want to quantify this - suggest plausible values

What summaries could we use?

(What summaries do we use for simple statistics?)

# Simulating the Sampling Distribution

From our data, we have our estimate of $p$ (which we call $\hat{p}$)

If this is the true value, what values are we likely to estimate?

# What to do

Simulate the data. For each simulation calculate $\hat{p}$, the maxmimum likelihood estimate of $p$.

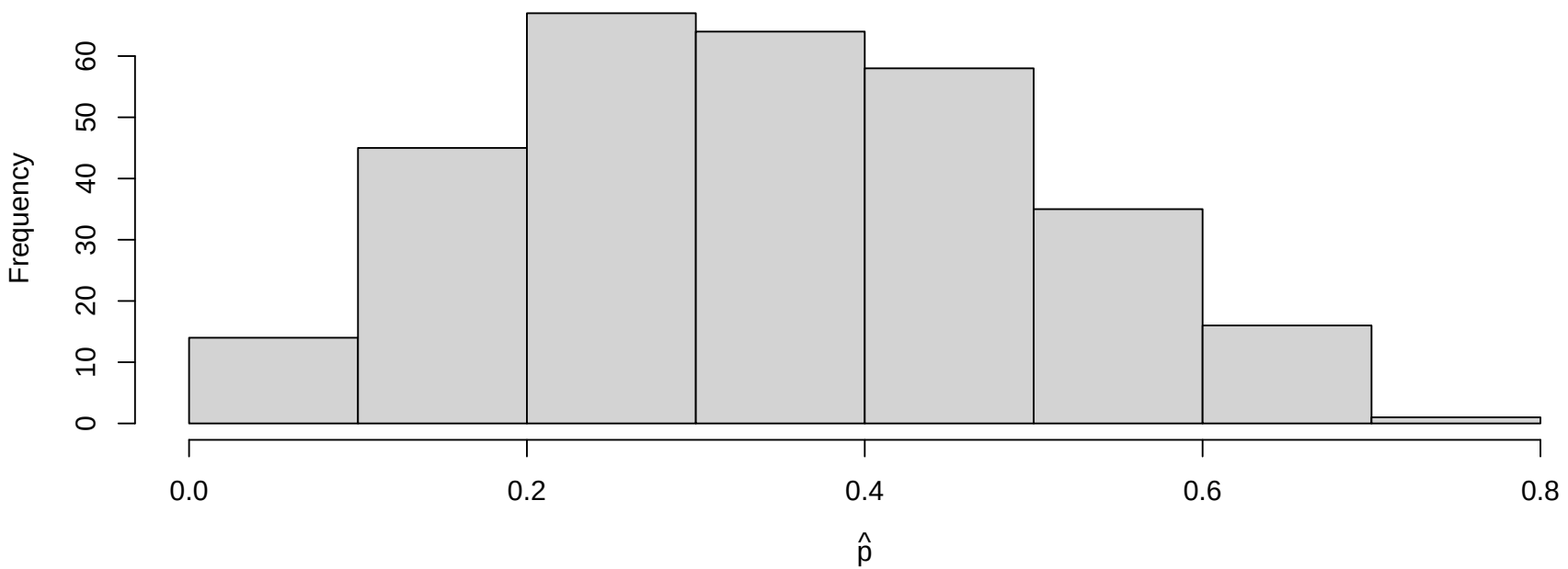
Look at the histogram of the distribution

- code on next 2 slides

# Simulations of the sampling distribution

We know that the MLE for $p$ is $r/N$, e.g. Land/(Land + Sea), so we can calculate it from the simulations: the web link is "https://www.math.ntnu.no/emner/ST2304/2024v/Module02/Module02Functions.R"

```
source("https://www.math.ntnu.no/emner/ST2304/2024v/Module0
sim <- simGlobe(probability=0.4, NTrials=10, nSims = 300)
hist(mleGlobe(sim["Land",], NTrials = 10),
     xlab=expression(hat(p)), main="")
```



# How can we summarise the distribution?

Can we give a range of probable values?

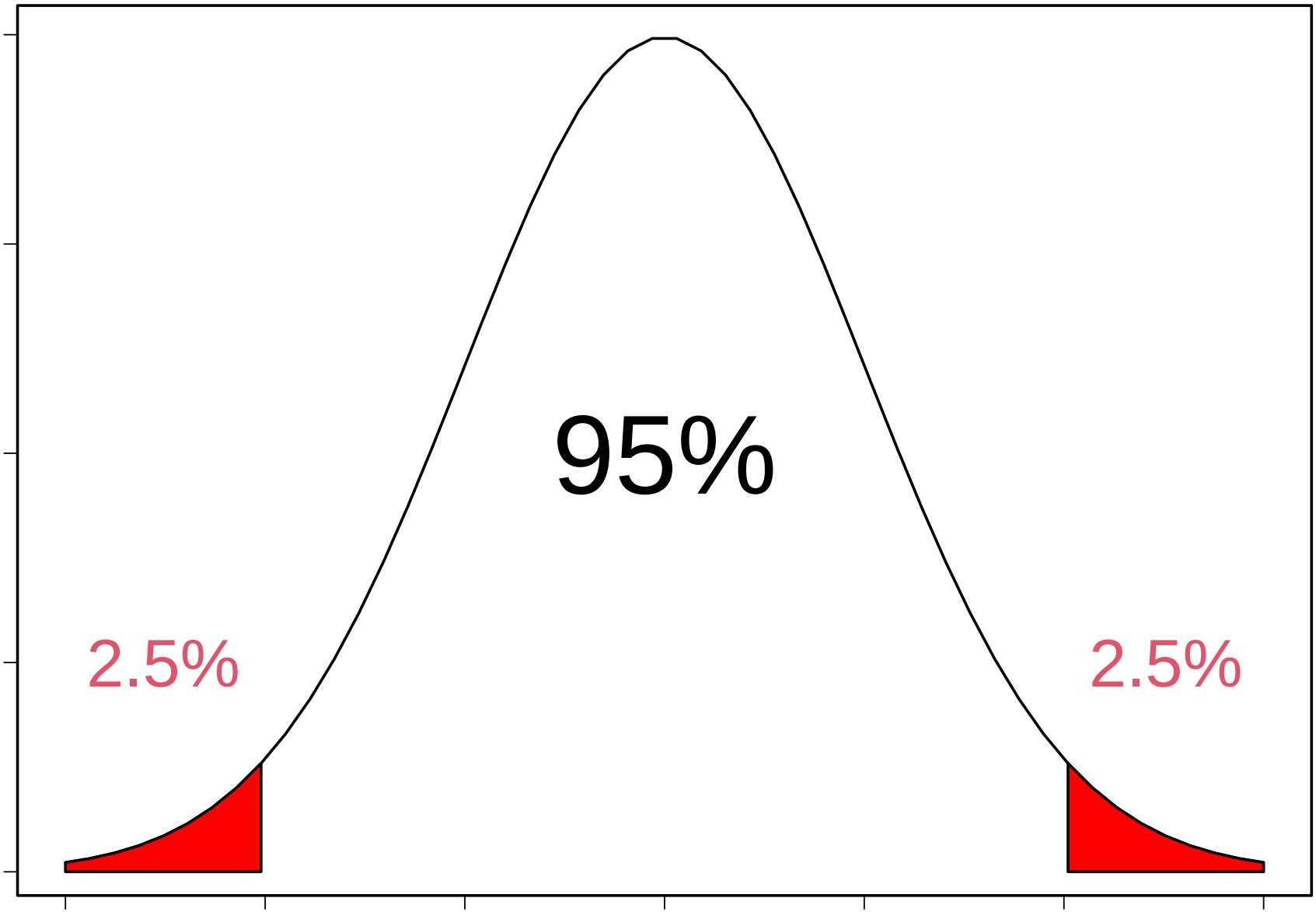
# Confidence Intervals

We can give an interval within which we think we would see the sample statistic

- the confidence interval
- usually use 95%

# Confidence Intervals

For continuous data the 95% confidence interval is constructed like this



# Confidence Intervals

Your task: try to calculate an approximate 95% confidence interval for your data

Your task: follow the "Constructing a Confidence interval" section

(for discrete data it is a bit more difficult to get an exact interval)

# Confidence Intervals and Quantiles I

There are a few ways to calculate confidence intervals. One way is to sort the numbers from lowest to highest

```
SimDist1k <- simGlobe(probability=0.4,
                      NTrials=1e3,
                      nSims = 1e3)["Land",]
sort(SimDist1k)[1:10]
```

```
##  [1] 357 360 361 362 363 363 363 364 364 365
```

and then take the values that are 2.5% of the way from the bottom, and 2.5% of the way from the top:

```
sort(SimDist1k)[c(0.025*length(SimDist1k),
                  0.975*length(SimDist1k))]
```

```
## [1] 368 431
```

# Confidence Intervals and Quantiles II

The values 2.5% of the way from the bottom, and 2.5% of the way from the top are called **quantiles**, specifically the 2.5% and 97.5% quantiles.

A x% quantile is a values of a distribution with x% of the distribution less than it

- a median is the 50% quantile
- the 25% and 75% quantiles are called quartiles (they plus the median split the data into 4 quarters)

So, we can just need the 2.5% and 97.5% quantiles. There is a function in R to do this:

```
quantile(SimDist1k, c(0.025, 0.975))
```

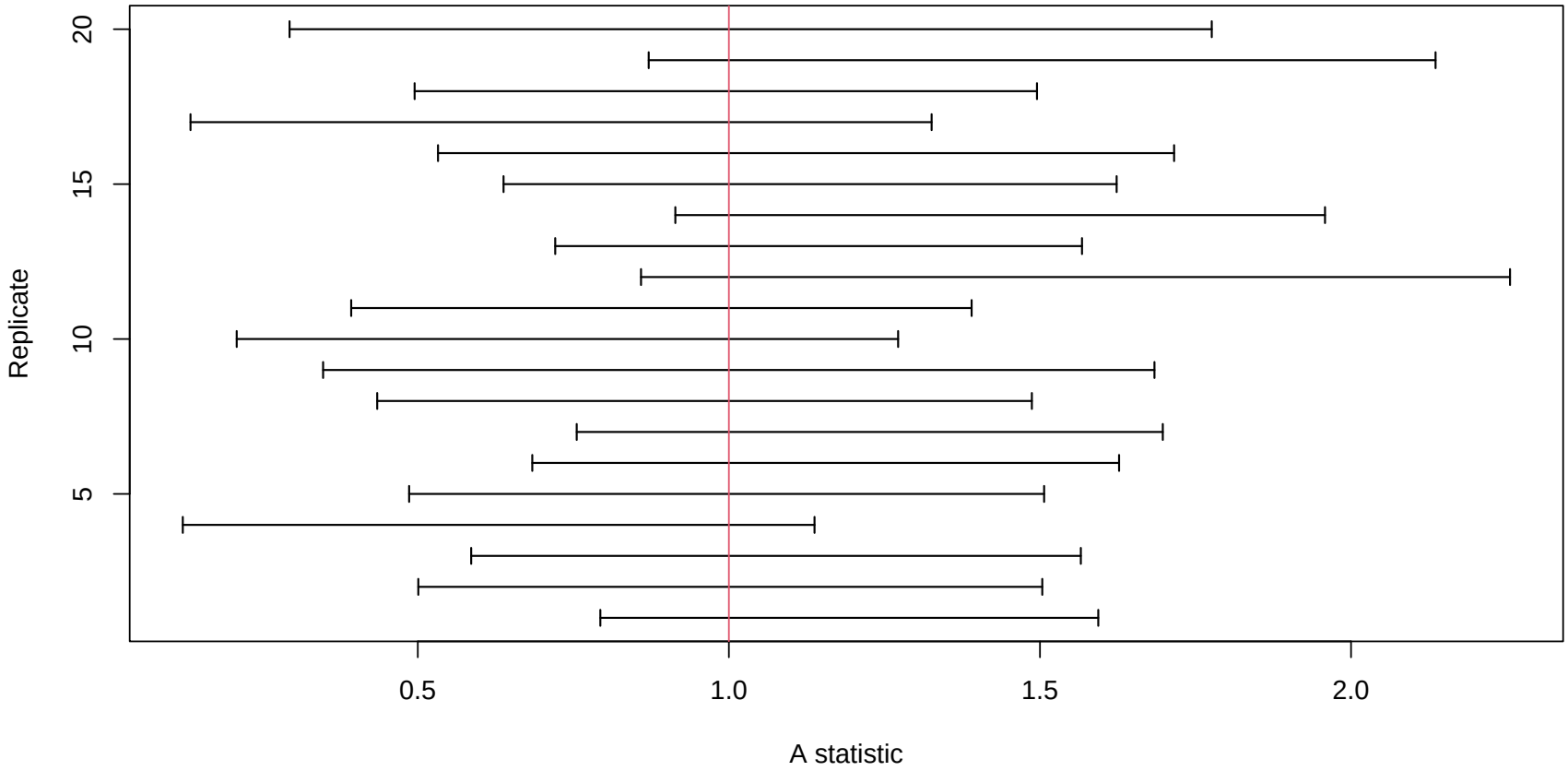
```
##  2.5% 97.5%
##   368   431
```

# Confidence Intervals and Quantiles: Your tasks

Your task: follow the "Confidence Intervals and Quantiles" section

# OK, so what, exactly, is a confidence interval?

A confidence interval is an interval that will contain a population parameter a specified proportion of the time.




i.e. if we repeatedly sample the same population, 95% of confidence intervals will include the “true” parameter Your task: follow the "What, exactly, is a confidence interval?" section

# Confidence Intervals with more data

Now imagine that rather than 10 trials, you have 1000. As before, you see 40% of the observations are land (i.e. 400 out of 1000)

```
# 1e3 = 1x10^3 = 1000
sim <- simGlobe(probability=0.4, NTrials=1e3, nSims = 3)
```

Try to find a 95% confidence interval for this

Basically, we want to remove the outer 2.5% of values, and see what is left.

# Confidence Intervals with more data

Your task: follow the "Confidence Interval for a Binomial with differnet Ns" section

What are the differences in the confidence intervals?

- in their size
- in how well they cover 95% of the sampling distribution

# Asymptotic Confidence Intervals

In statistics, large numbers usually make things much nicer: there are a lot of asymptotic results (i.e. approximations that work well when there is a lot of data).

One of these is the that most sampling distributions of statistics look like normal distributions, with enough data.

So, if we can construct a normal distribution's CI, we can make an approximation.

# Normal Confidence Intervals

We can calculate a normal confidence interval like this:

```
c(qnorm(0.025, mu, sigma), qnorm(0.975, mu, sigma))
```

The parameters are the mean and standard deviation, e.g.

```
## [1] -5.839856  9.839856
```

# Normal Approximations

If we know the mean and standard deviation of the sampling distribution, then we can use a normal approximation.

- the standard deviation of the sampling distribution is called the standard error

# Normal Approximation for the Binomial

We can use the MLE, $\hat{p}$ as the mean of the normal

The standard error for the binomial distribution is

$$\sqrt{\frac{p(1-p)}{N}}$$

# Approximatioms: Your tasks

Your task: follow the "Approximations" section

# How well are we doing?

With small N, we cannot usually get a perfect 95% confidence interval, because the possible estimates are discrete $(0/N, 1/N, ..., N/N)$, so our interval might be slightly smaller or larger

Asymptotic intervals may not be perfect either, although they should get better as $N$ increases

Your task: follow the "How well are we doing?" section

# Standard Errors

We used the **standard error** to calculate the asymptotic confidence. But we could use it to summarise the uncertainty in a single parameter

Standard Deviations of Statistics: $s$

Binomial variance of $n$: $Var(n|N,p) = Np(1-p)$

Our statistic: $n/N$

$Var(n/N) = 1/N^2\, Var(n) = p(1-p)/N$

Standard error:

$s = \sqrt{p(1-p)/N}$

# Standard Errors: Your Turn

Guess what?

Your task: follow the "Standard Errors " section

# Exercise Hand-in

- Form groups: sign up to BB
- **ONE** hand-in per group
- Deadline: 3rd of February (ish)