

# Sampling and maximum likelihood estimation

Bert van der Veen

Department of Mathematical Sciences, NTNU

## Outline Today

---

- ▶ An orchid example
- ▶ Sampling variation
- ▶ Maximum likelihood estimation
- ▶ Normal distribution
- ▶ Linear models

## Sampling data



Figure 1: www.ugent.be

## What is the proportion of orchids?



We decide to walk through the field and at 10 places record when we find an orchid (1) or not (0)

1. First time: 5 orchids from 10 picks ( $5/10 = 0.5$ )

## What is the proportion of orchids?



We decide to walk through the field and at 10 places record when we find an orchid (1) or not (0)

1. First time: 5 orchids from 10 picks ( $5/10 = 0.5$ )
2. Second time: 2 orchids from 10 picks ( $2/10 = 0.2$ )
3. Third time: 8 orchids from 10 picks ( $8/10 = 0.8$ )

## What is the proportion of orchids?

We conclude, half of the flowers are orchids ( $15/30 = 0.5$ ). But encounter this guy:



- ▶ What caused our estimate of the proportion of orchids to be inaccurate?
- ▶ And why did we not get the same proportion of orchids every time?

He tells us that the true proportion of orchids is 0.4.

## The binomial distribution

---

$$f(y_i; n_{picks}, \pi) = y_i \log\{\pi\} + (n_{picks} - y_i) \log\{1 - \pi\} + \text{constant} \quad (1)$$

- ▶ Density: `dbinom`
- ▶ Number generator: ‘`rbinom`’

## Simulation: counting orchids once

---

```
set.seed(12345) # For reproducibility
p.orchid = 0.4 # The true proportion of orchids
n.picks = 10 # The number of picks in the field
n.times = 1
# Collect data
y <- rbinom(n.times, size = n.picks, prob = p.orchid)
y/n.picks # Proportion of orchids

## [1] 0.5
```

## Simulation: counting orchids once

---

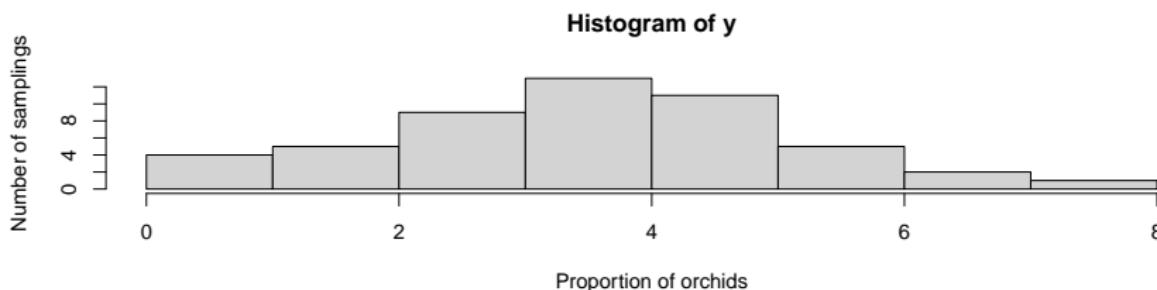
What if we sample the whole field once?

```
set.seed(12345) # For reproducibility
n.times = 1e5 # The number of picks in the field
n.picks <- 1
# Collect data
y <- rbinom(n.times, size = n.picks, prob = p.orchid)
mean(y/n.picks) # Proportion of orchids
```

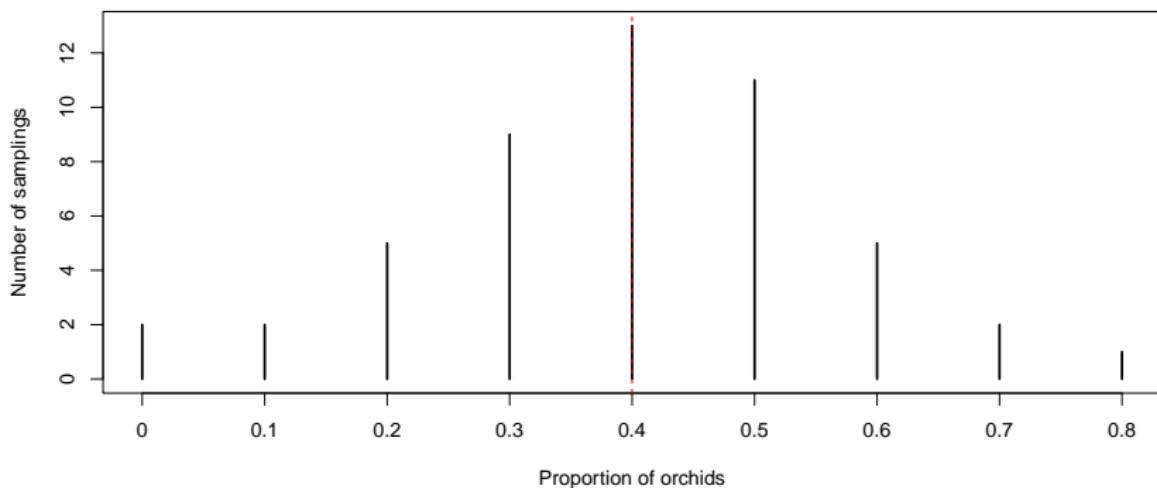
```
## [1] 0.40021
```

## Simulation: counting orchids 50x10 times

```
set.seed(12345) # For reproducibility
n.times <- 50
n.picks = 10 # The number of picks in the field
# Collect data
y <- rbinom(n.times, size = n.picks, prob = p.orchid)
hist(y, xlab = "Proportion of orchids",
      ylab = "Number of samplings")
```



## Simulation: counting orchids 50x10 times



As you see, we have variability in our estimate of the proportion of orchids.

- ▶ Can we summarize this variation?
- ▶ Preferably without collecting data many times

## The strategy

---

- ▶ Collect data
- ▶ Learn about the variation in that data
  - ▶ We need a model for that
- ▶ Work out distribution of the estimates
  - ▶ And find the “best” estimate
- ▶ Conclude if our answer is robust
  - ▶ Q: Are more than half of the flowers in this field orchids?

Q: Are more than half of the flowers in this field orchids?

---

- ▶ Exclude rare events, and decide on an acceptable margin of error (5%)
- ▶ What is the range our parameter is estimated to be 95% of the time?

```
quantile(y/n.picks,c(0.025,.975))
```

```
##    2.5%   97.5%
## 0.0225 0.7000
```

So 50 times 10 picks tell us little.

Q: Are more than half of the flowers in this field orchids?

---

```
set.seed(12345)
n.picks = 100
n.times <- 50
# Collecting data
y <- rbinom(n.times, size = n.picks, prob = p.orchid)
quantile(y/n.picks,c(0.025,.975))

##      2.5%    97.5%
## 0.31225 0.48000
```

50 times 100 picks reduces the variability in the proportion of observed orchids. Most of the time we will find fewer orchids than

On average, we will get it right.

```
set.seed(12345)
n.picks = 1;n.times <- 100
y <- rbinom(n.times, size = n.picks, prob = p.orchid)
quantile(y/n.picks,c(0.025,.975))
```

```
## 2.5% 97.5%
##      0      1
```

```
mean(y/n.picks)
```

```
## [1] 0.45
```

## Estimator and Estimand

---

- ▶ Estimand: the parameter we want to estimate, the true population level parameter
- ▶ Estimator: how we are estimating it  $\hat{p}_{orchid} = \frac{y}{n.picks}$ 
  - ▶ Why is this a good estimator?
- ▶ Estimate: the parameter value based on the data

## Estimating parameters and quantifying uncertainty

---

- ▶ We do not usually have infinite amounts of data
- ▶ So how can we quantify variability of the estimator?
- ▶ We need a model for the process that generates the data



## The likelihood: single data point

---

$$\mathcal{L}(y_i; \Theta) = f(y_i; \Theta) \quad (2)$$

The probability of obtaining our data assuming  $\Theta$  is the true parameter(s).

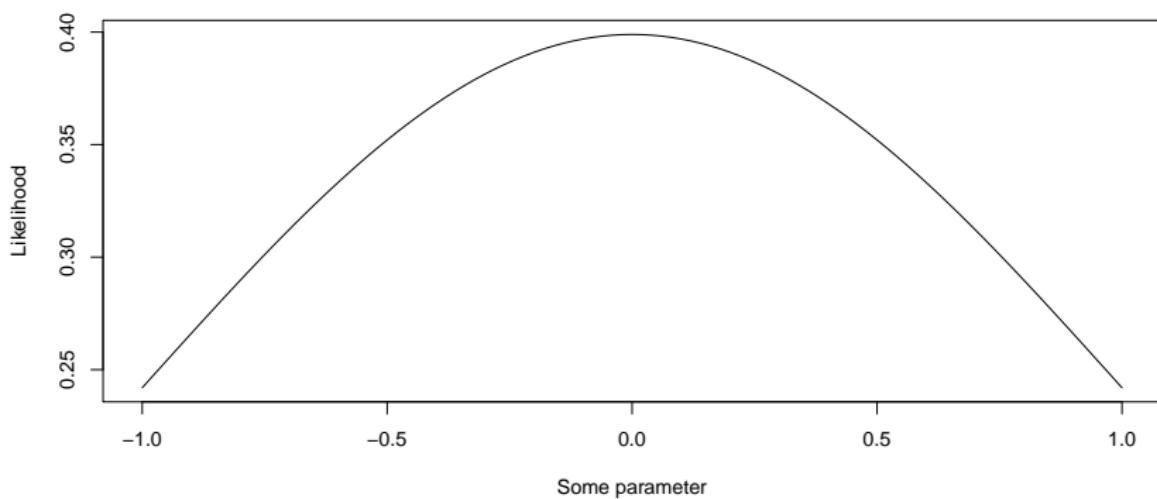
## The likelihood: multiple data points (2)

---

$$\mathcal{L}(\mathbf{y}; \Theta) = \prod_i^n f(y_i; \Theta) \quad (3)$$

We just multiply! (assumes independence)

## The likelihood (3)

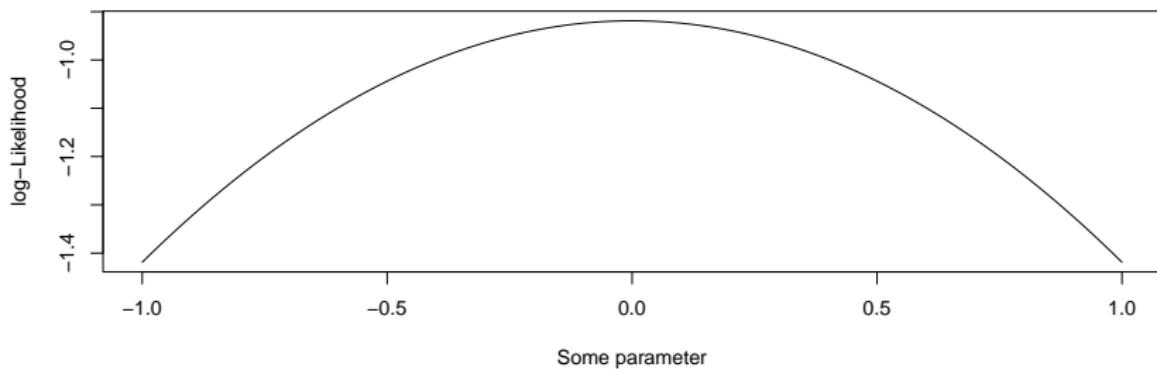


Likelihood tells us about:

- ▶ The (set of) parameter estimates that most likely generated the data
- ▶ The information contained in our data

## The log-likelihood

$$\log\{\mathcal{L}(\mathbf{y}; \Theta)\} = \sum_i^n \log\{f(y_i; \Theta)\} \quad (4)$$



Usually, we work with the log-likelihood. The maximum is the same and it is easier. So we just add things together.

## Back to the orchid example

---

- ▶ Data is binary or counts
- ▶ Which follows a binomial likelihood (only as example, for now)

$$f(y_i; n_{picks}, \pi) = y_i \log\{\pi\} + (n_{picks} - y_i) \log\{1 - \pi\} + \text{constant} \quad (5)$$

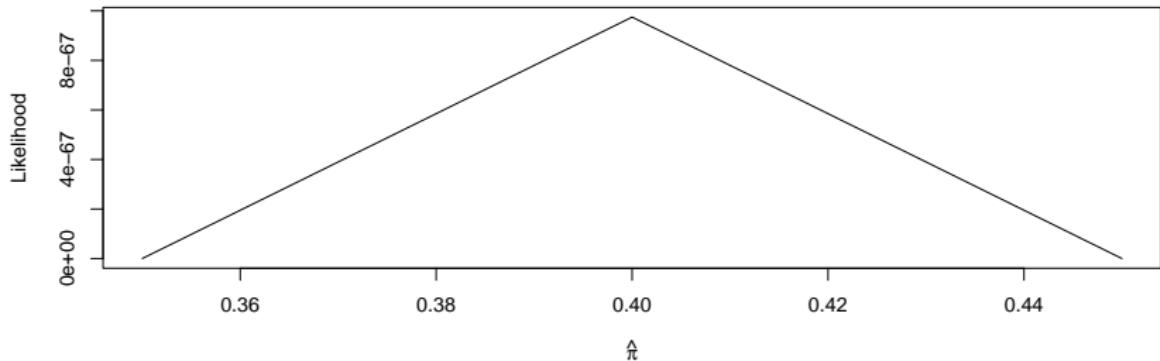
- ▶  $N = n_{picks}$  trials
- ▶  $r = y_i$  successes (orchids)
- ▶  $\pi$  the probability of success (proportion of orchids)

### Assumptions

- ▶ Each pick is independent from the next
- ▶  $\pi$  is the same for all observations

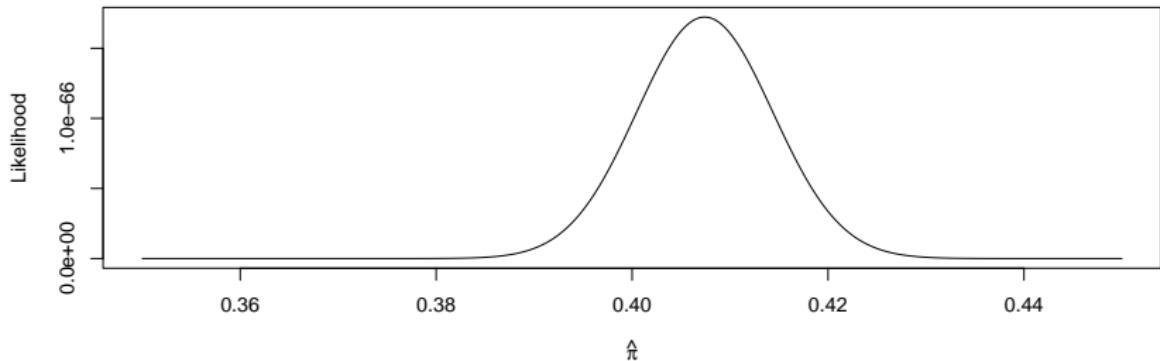
## Finding the proportion of orchids

```
ll <- function(p, n.picks, y)prod(dbinom(y, n.picks,p))
phat <- seq(0.35,0.45,length.out=3)
plot(sapply(phat, ll, n.picks = n.picks, y = y),
     x = phat, type = "l", xlab=expression(hat(pi)), ylab="Likelihood")
```



## Finding the proportion of orchids (2)

```
ll <- function(p, n.picks, y)prod(dbinom(y, n.picks,p))
phat <- seq(0.35,0.45,length.out=1000)
plot(sapply(phat, ll, n.picks = n.picks, y = y),
     x = phat, type = "l", xlab=expression(hat(pi)), ylab="Li
```



## Maximising the likelihood

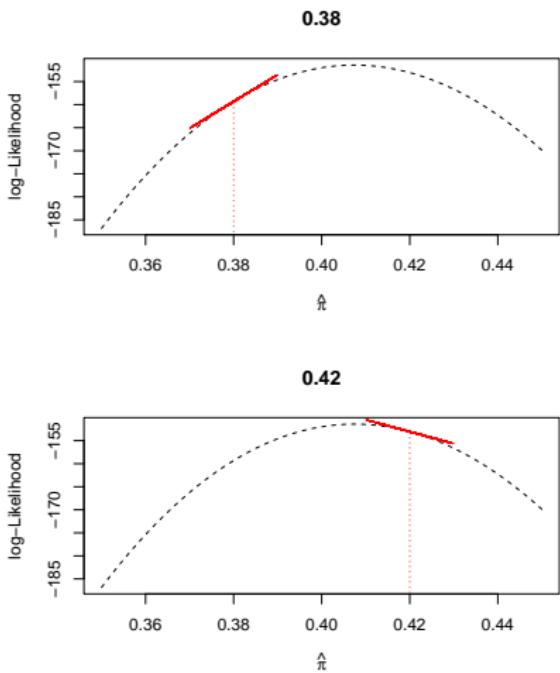
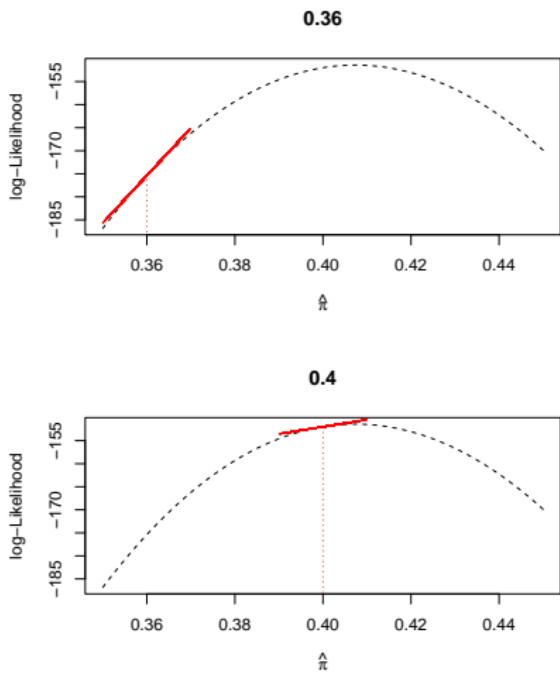
---

Trying many values (grid) is very inefficient

We can:

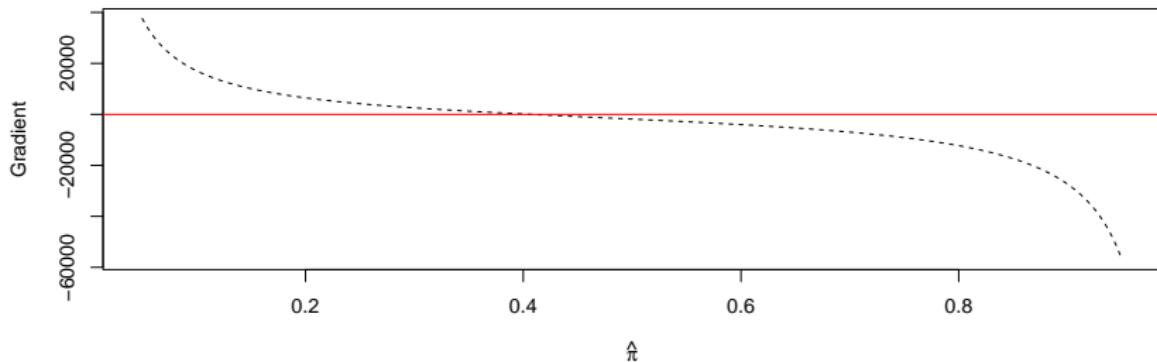
- ▶ analytically: do mathematics
- ▶ numerically: use an algorithm (as in GLMs)
- ▶ simulation: try many values

# Finding the maximum



## Finding the maximum: gradient

```
phat = seq(0.05,0.95,length.out=1000)
plot(sapply(phat, grad.dbinom, y=y, n.picks = n.picks),
      x = phat, type = "l", xlab=expression(hat(pi)),
      ylab="Gradient", lty = "dashed")
abline(h=0, col = "red")
```



## Finding the maximum: mathy bits

Our function:

$$\log\{\mathcal{L}(\mathbf{y}; n_{picks})\} = \sum_{i=1}^{n_{times}} y_i \log\{\pi\} + (n_{picks} - y_i) \log\{1 - \pi\} + \text{constant} \quad (6)$$

Slope of the likelihood:

$$\frac{\partial \log\{\mathcal{L}(\mathbf{y}; n_{picks})\}}{\partial \pi} = \sum_{i=1}^{n_{times}} \frac{y_i}{\pi} - \frac{n_{picks} - y_i}{1 - \pi} \quad (7)$$

Estimator:

$$0 = \sum_{i=1}^{n_{times}} \frac{y_i}{\pi} - \frac{n_{picks} - y_i}{1 - \pi} \quad (8)$$

(i.e., the proportion of successes!)

## Maximum likelihood estimation: summary

---

No ideal properties in finite samples. But as we get more data:

- ▶ Get things right with a lot of data (consistency)
- ▶ Is the best estimator (in terms of RMSE) in large samples of consistent estimators

Often does well in practice anyway 😊

## Letting R do the work

---

```
optimize(ll, n.picks = n.picks, y=y,  
        lower = 0, upper = 1, maximum = TRUE)
```

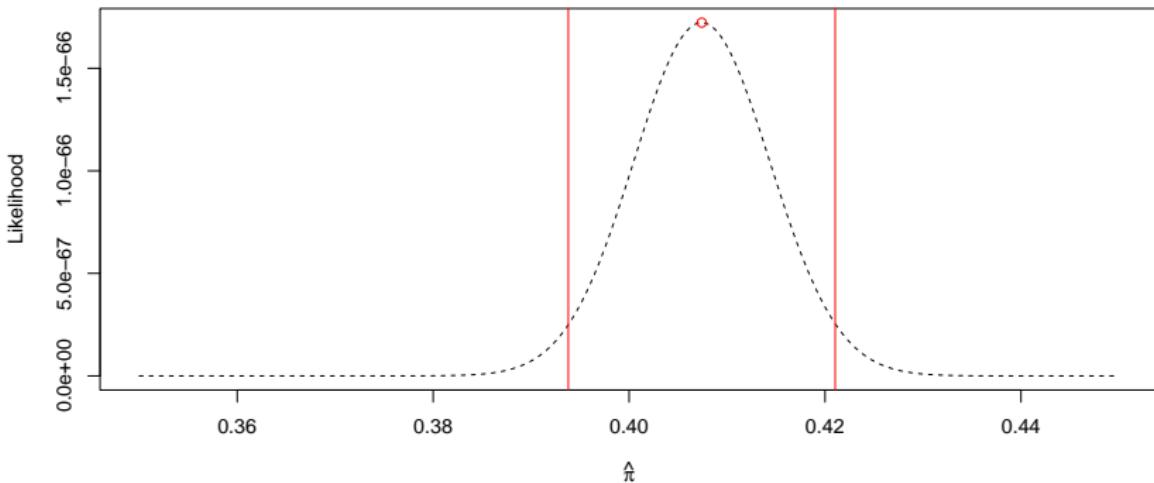
```
## $maximum  
## [1] 0.4074027  
##  
## $objective  
## [1] -151.4268
```

- ▶ Usually there is more than 1 parameter, much harder

# Uncertainty

(an estimate of) Width of the likelihood:

$$\frac{\partial \log\{\mathcal{L}(\mathbf{y}; n_{picks})\}}{\partial \pi^2} = \sum_{i=1}^{n_{times}} -\frac{y_i}{\pi^2} - \frac{n_{picks} - y_i}{(1 - \pi)^2} \quad (9)$$



## Putting it all together

- ▶ We collect data
  - ▶ We estimate a parameter of interest
  - ▶ If we collected data again, we get many different estimates
    - ▶ This forms a *sampling* distribution
  - ▶ We summarize this variability
  - ▶ The width of this sampling distribution tells us the variability
  - ▶ Instead of collecting data many times, we estimate parameters with MLE
    - ▶ This also allows us to quantify the variability



## 95% Confidence intervals

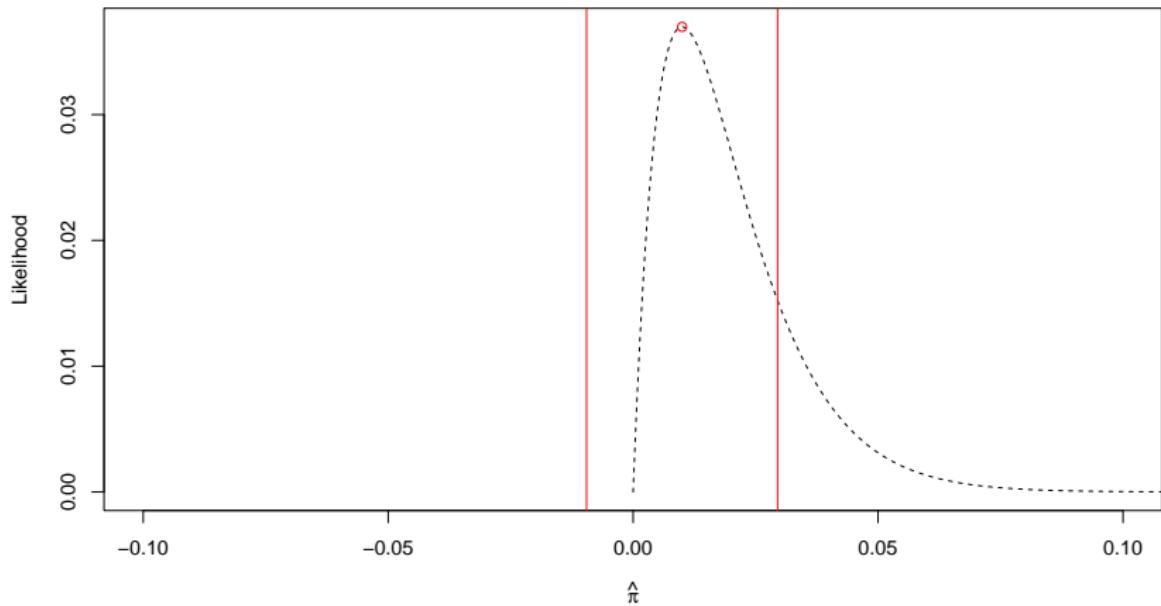
---

Estimate of the range that the true parameter will fall in 95% of the time.

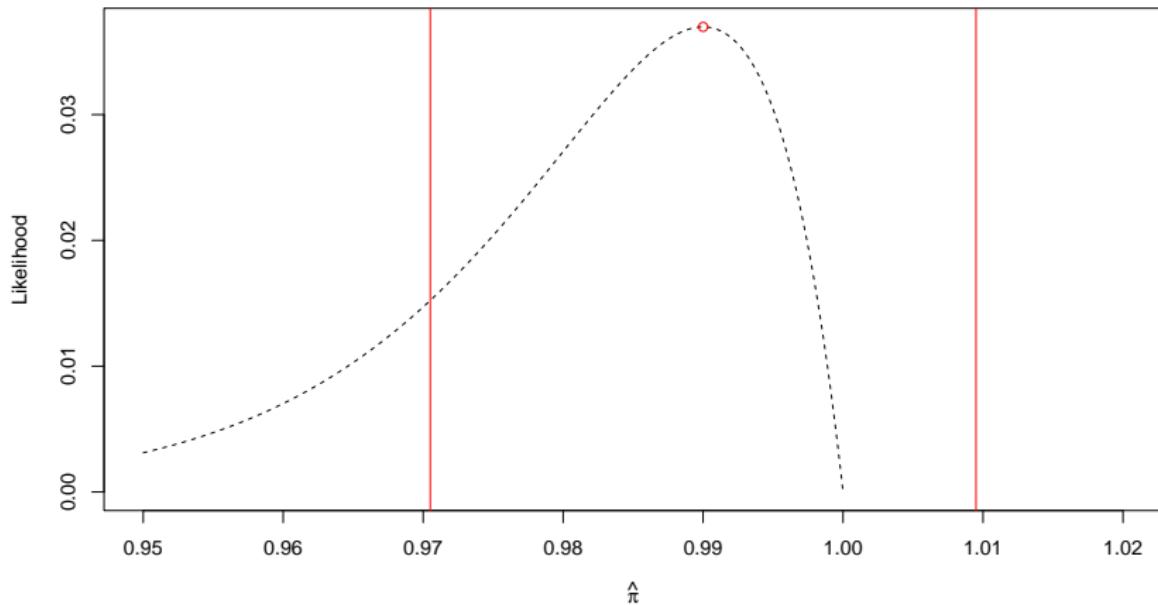
Different methods exist. Based on MLE not quite correct for the Binomial distribution due to **assumptions**.

## Why not?

---



## Why not? (2)



## Why is uncertainty so important

---

- ▶ Our sampling distribution estimates needs summarizing
- ▶ Tells us if we expect our answer to be the same if we repeated the study
- ▶ I.e., not so important for the *dataset* but important for *multiple datasets*

Afterall, we are looking for a robust recommendation.

## Confidence intervals

*An interval that contains the true value in 95% of repeated samples  
(in large samples)*

Be careful with interpretation, and with assumptions.

- ▶ Any computed interval either contains the truth, or it does not
- ▶ Not the range that the true parameter falls in with 95
- ▶ Other misinterpretations
- ▶ Can be interpreted as a kind of statistical test
- ▶ Or generally as “evidence”

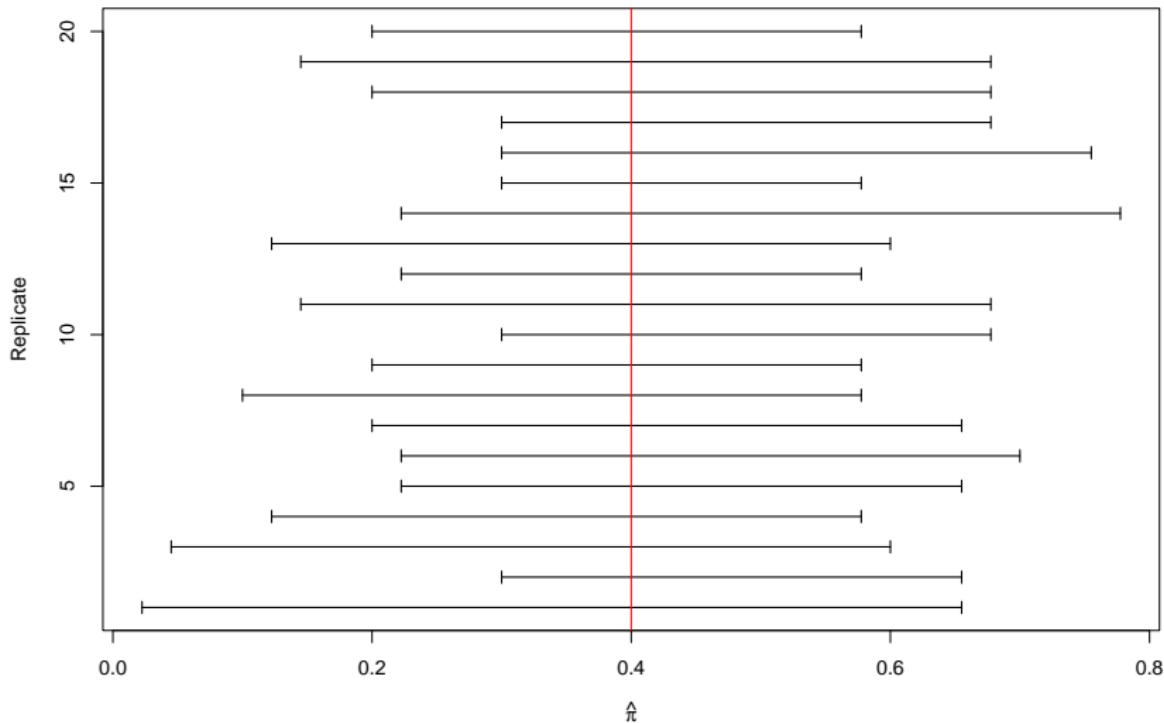
### Gets smaller with:

- ▶ More information
- ▶ Less variability
- ▶ The confidence level

### Assumes:

- ▶ Asymptotic normality
- ▶ inverse Hessian gives covariance of estimators

# Repetition



## Summary

---

- ▶ Our data comes from a population
- ▶ That population is generated by a model
- ▶ The model has parameters that we want to find
- ▶ We do that based on our data
- ▶ Our data is a sample, so it is not a perfect representation of the population
- ▶ We need to sample many times to get an idea of variability due to our sampling
- ▶ We summarize this with a sampling distribution of our estimates
- ▶ And use that to draw a conclusion

## My questions

---

Menti for “quiz”? Why not just use a single dataset? –TODO

## Questions?

---

