Statistics - 01 Introduction

Eric Stemmler

Khovd University

20.01.2021

Section 1

Personal Introduction

Personal Introduction

- Eric Stemmler
- M.Sc. Computational Science (Technical University of Chemnitz)
- M.Sc. Human Factors (Technical University of Berlin)
- Statistics, Data Science

Contact

- email (en): rcst@posteo.de
- email (mn): byambaa3007@yahoo.com
- Room: 415 (please send an email before visiting)
- phone: +976 8868 3742

Please provide your name and email so I can send you my presentations and other material

Personal Introduction

What is statistics?

*Statistics is the discipline that concerns the **collection**, **organization**, **analysis**, **interpretation**, and **presentation** of data.*

Cambridge Dictionary

Personal Introduction

- What is your experience with statistics?
- What kind of data do you analyse and how did you do it?
- About what topics do you want to learn about?

Section 2

Learning Goals

Learning Goals

- Formulate statistical modelling problems
- Exploratory data analysis
- Basic computations in R

Section 3

Why is statistics important?

Why is statistics important?

- Learning from data about the world
- Randomness is omnipresent
- Estimation of uncertainty vs. establishing facts
- Making decisions

Why is statistics important?

- Learning from data about the world
- Randomness is omnipresent
- Estimation of uncertainty vs. establishing facts
- Making decisions

Why is statistics important?

- Learning from data about the world
- Randomness is omnipresent
- Estimation of uncertainty vs. establishing facts
- Making decisions

Why is statistics important?

- Learning from data about the world
- Randomness is omnipresent
- Estimation of uncertainty vs. establishing facts
- Making decisions

Section 4

Vocabulary

Subsection 1

Randomness

Randomness

Figure 1: Uncertainty: Flipping a coin

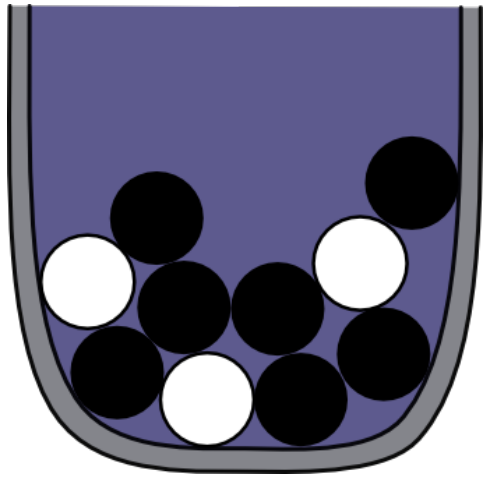


Figure 2: Variation: blindly drawing balls from an urn

Randomness

Figure 1: Uncertainty: Flipping a coin

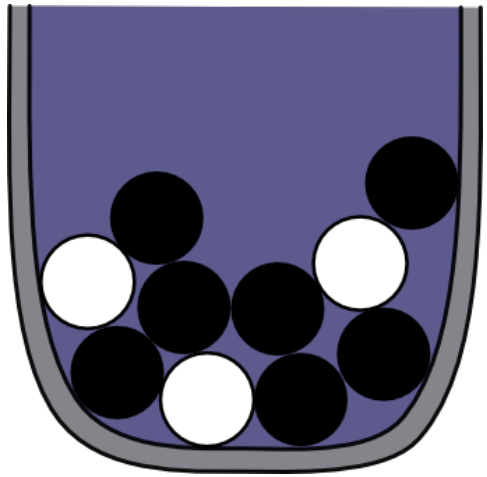


Figure 2: Variation: blindly drawing balls from an urn

- randomness can never be removed completely
- Law of large numbers $\rightarrow$ Estimation of parameters

Randomness

Figure 1: Uncertainty: Flipping a coin

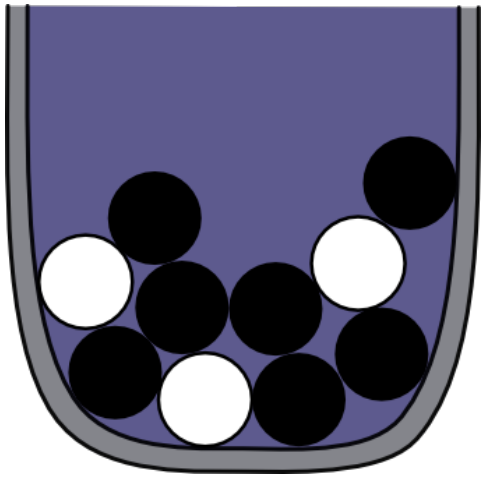


Figure 2: Variation: blindly drawing balls from an urn

- randomness can never be removed completely
- Law of large numbers $\rightarrow$ Estimation of parameters

Randomness

What are examples for infinite populations and why?

- probability for a coin to land on head

Randomness

What are examples for infinite populations and why?

- probability for a coin to land on head
- (Human) gender ratio

Randomness

What are examples for infinite populations and why?

- probability for a coin to land on head
- (Human) gender ratio
- Measurement errors in physics

Randomness

What are examples for infinite populations and why?

- probability for a coin to land on head
- (Human) gender ratio
- Measurement errors in physics
- Effectiveness of a vaccine

Randomness

What are examples for infinite populations and why?

- probability for a coin to land on head
- (Human) gender ratio
- Measurement errors in physics
- Effectiveness of a vaccine
- The probability of getting cancer from smoking

Randomness

What are examples for infinite populations and why?

- probability for a coin to land on head
- (Human) gender ratio
- Measurement errors in physics
- Effectiveness of a vaccine
- The probability of getting cancer from smoking
- Temperature-dependent sex determination of *Crocodylus niloticus*
- ...

Randomness

Important terms:

- variation:
- uncertainty:
- trial:
- population:
- population parameter:

Randomness

Important terms:

- variation: the outcome of a sample varies randomly
- uncertainty:
- trial:
- population:
- population parameter:

Randomness

Important terms:

- variation: the outcome of a sample varies randomly
- uncertainty: lack of knowledge of about a true value
- trial:
- population:
- population parameter:

Randomness

Important terms:

- variation: the outcome of a sample varies randomly
- uncertainty: lack of knowledge of about a true value
- trial: the realization of an experiment
- population:
- population parameter:

Randomness

Important terms:

- variation: the outcome of a sample varies randomly
- uncertainty: lack of knowledge of about a true value
- trial: the realization of an experiment
- population: all possible events or items
- population parameter:

Randomness

Important terms:

- variation: the outcome of a sample varies randomly
- uncertainty: lack of knowledge of about a true value
- trial: the realization of an experiment
- population: all possible events or items
- population parameter: the true value

Randomness

Demonstration: Real vs. fake coin flips

- 2 judges
- 1 recorder
- 2 groups
  - group 1: note down the result of 100 real coin flips
  - group 2: note down 100 invented/ fake coin flips that *look* random

Randomness

Demonstration: Real vs. fake coin flips

each group:

1. count the length of the longest run
2. count the number of runs
3. mark the location on the plot

Randomness

Demonstration: Real vs. fake coin flips

each group:

1. count the length of the longest run
2. count the number of runs
3. mark the location on the plot

example: 0, 0, 1, 1, 1, 1, 0, 0, 0, 1, 1

- length of longest run: 4

Randomness

Demonstration: Real vs. fake coin flips

each group:

1. count the length of the longest run
2. count the number of runs
3. mark the location on the plot

example: 0, 0, 1, 1, 1, 1, 0, 0, 0, 1, 1

- length of longest run: 4
- no. runs: 4

Simulation Results

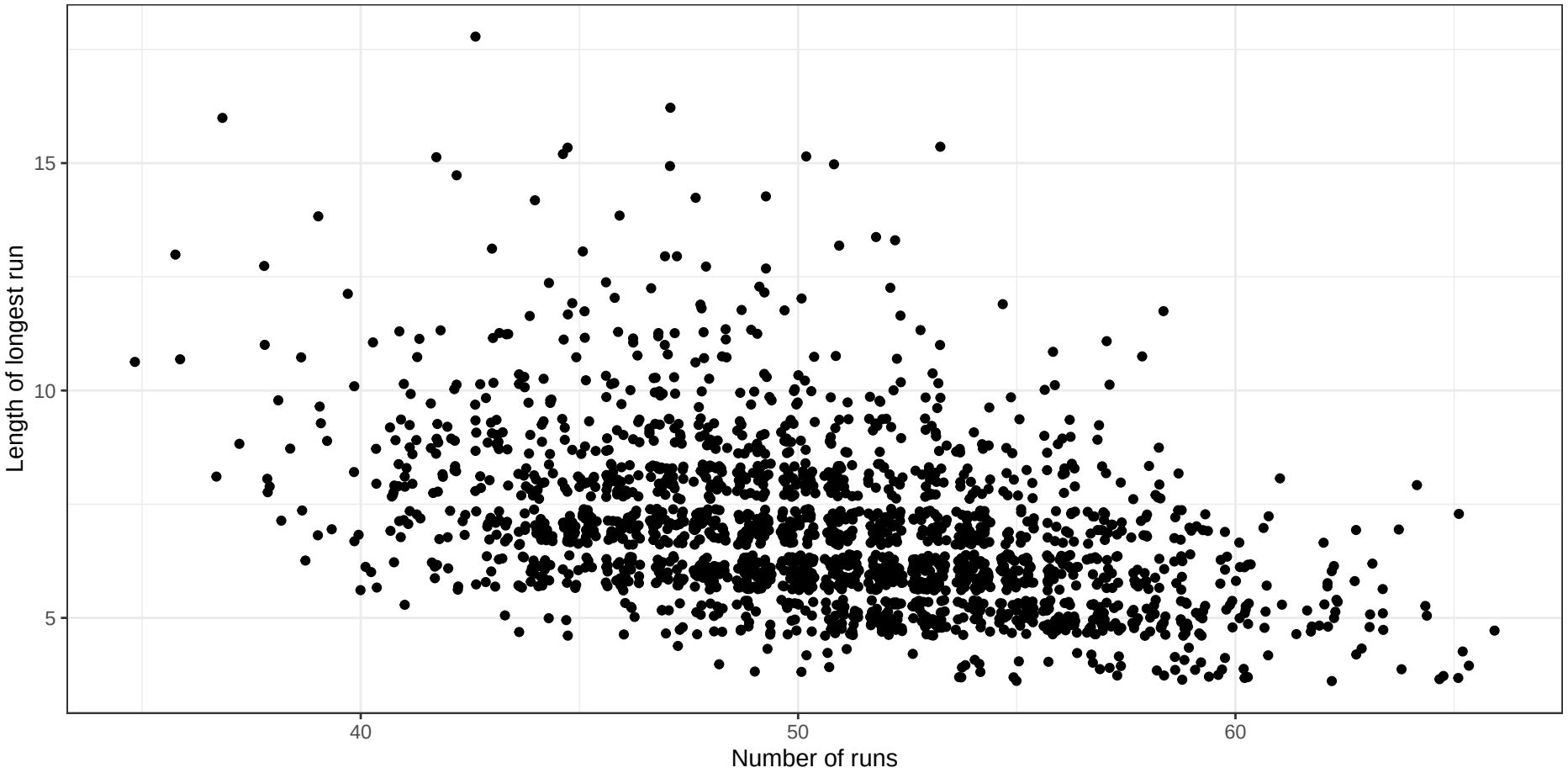


Figure 3: Length of longest run vs. number of runs from 2000 simulated experiments of 100 coin flips.

Subsection 2

Coin flipping

Coin flipping

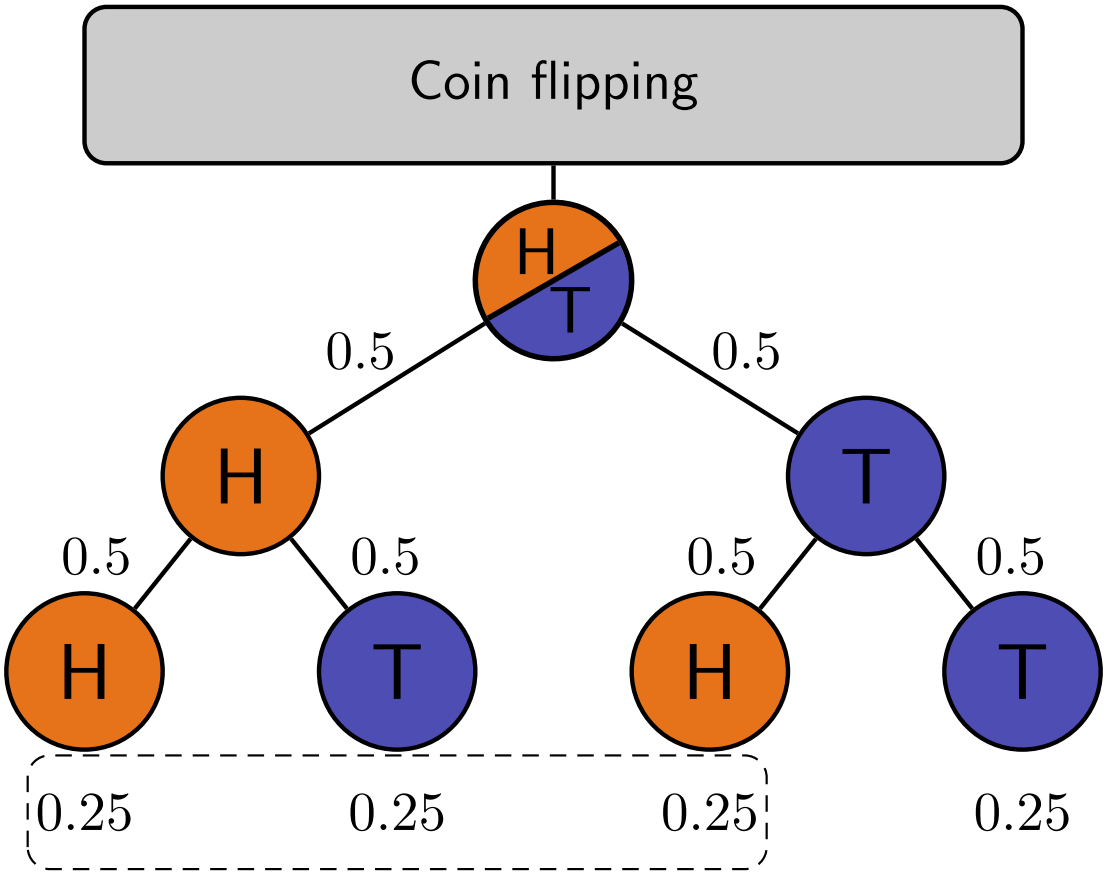


Figure 4: Probability tree for the outcomes of a coin flipping experiment

Coin flipping

Exercise: What is the probability of getting 5?

Coin flipping

Exercise: What is the probability of getting 5?

Exercise: What is the probability of getting 5 heads in a row during 100 coin flips?

Coin flipping

Exercise: What is the probability of getting 5?

Exercise: What is the probability of getting 5 heads in a row during 100 coin flips?

Hint: How many possibilities for 3 heads in a row exist in 10 coin flips?
$10 - (3 - 1) = 8$

Coin flipping

Exercise: What is the probability of getting 5?

Exercise: What is the probability of getting 5 heads in a row during 100 coin flips?

Hint: How many possibilities for 3 heads in a row exist in 10 coin flips?
$10 - (3 - 1) = 8$

Solution: $96 \times (1/32)^1 \times (31/32)^{95} \approx 0.15$

Note: This is only the probability of getting 5 heads **exactly once**!

Coin flipping

Exercise: What is the probability of getting 5?

Exercise: What is the probability of getting 5 heads in a row during 100 coin flips?

Hint: How many possibilities for 3 heads in a row exist in 10 coin flips?
$10 - (3 - 1) = 8$

Solution: $96 \times (1/32)^1 \times (31/32)^{95} \approx 0.15$

Note: This is only the probability of getting 5 heads **exactly once**!

Binomial Distribution: $p = \binom{n}{k} \theta^k (1 - \theta)^{n-k}$

Subsection 3

Binomial Distribution

Binomial Distribution

$$p(k \mid n, \theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k} \tag{1}$$

- k - number of "successes"
- n - number of trials
- θ - probability of "success"

Binomial Distribution

$$p(k \mid n, \theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k} \tag{1}$$

- k - number of "successes"
- n - number of trials
- θ - probability of "success"

Subsection 4

Estimating fish population

Estimating fish population

Figure 5: Fishes in a lake

Estimating fish population

Definition

A **random sample** is a subset of a population such that each individual random sample is chosen with equal probability.

Subsection 5

Modelling Fish Population - Binomial distribution

Modelling Fish Population - Binomial distribution

- (finite) population of fish in a lake of size N.
- One possible choice of a model is the Binomial distribution

$$p(y \mid N, \theta) = \binom{N}{y} \theta^y (1-\theta)^{N-y}$$

- sampling/ fishing: y out of N in total
- θ is *capture probability*
- N and θ are generally called *parameters*
- y is called *data*

Modelling Fish Population - Binomial distribution

- (finite) population of fish in a lake of size N.
- One possible choice of a model is the Binomial distribution

$$p(y \mid N, \theta) = \binom{N}{y} \theta^y (1-\theta)^{N-y}$$

- sampling/ fishing: y out of N in total
- θ is *capture probability*
- N and θ are generally called *parameters*
- y is called *data*

Modelling Fish Population - Binomial distribution

- (finite) population of fish in a lake of size N.
- One possible choice of a model is the Binomial distribution

$$p(y \mid N, \theta) = \binom{N}{y} \theta^y (1-\theta)^{N-y}$$

- sampling/ fishing: y out of N in total
- θ is *capture probability*
- N and θ are generally called *parameters*
- y is called *data*

Modelling Fish Population - Binomial distribution

- (finite) population of fish in a lake of size N.
- One possible choice of a model is the Binomial distribution

$$p(y \mid N, \theta) = \binom{N}{y} \theta^y (1-\theta)^{N-y}$$

- sampling/ fishing: y out of N in total
- θ is *capture probability*
- N and θ are generally called *parameters*
- y is called *data*

Modelling Fish Population - Binomial distribution

- (finite) population of fish in a lake of size N.
- One possible choice of a model is the Binomial distribution

$$p(y \mid N, \theta) = \binom{N}{y} \theta^y (1-\theta)^{N-y}$$

- sampling/ fishing: y out of N in total
- θ is *capture probability*
- N and θ are generally called *parameters*
- y is called *data*

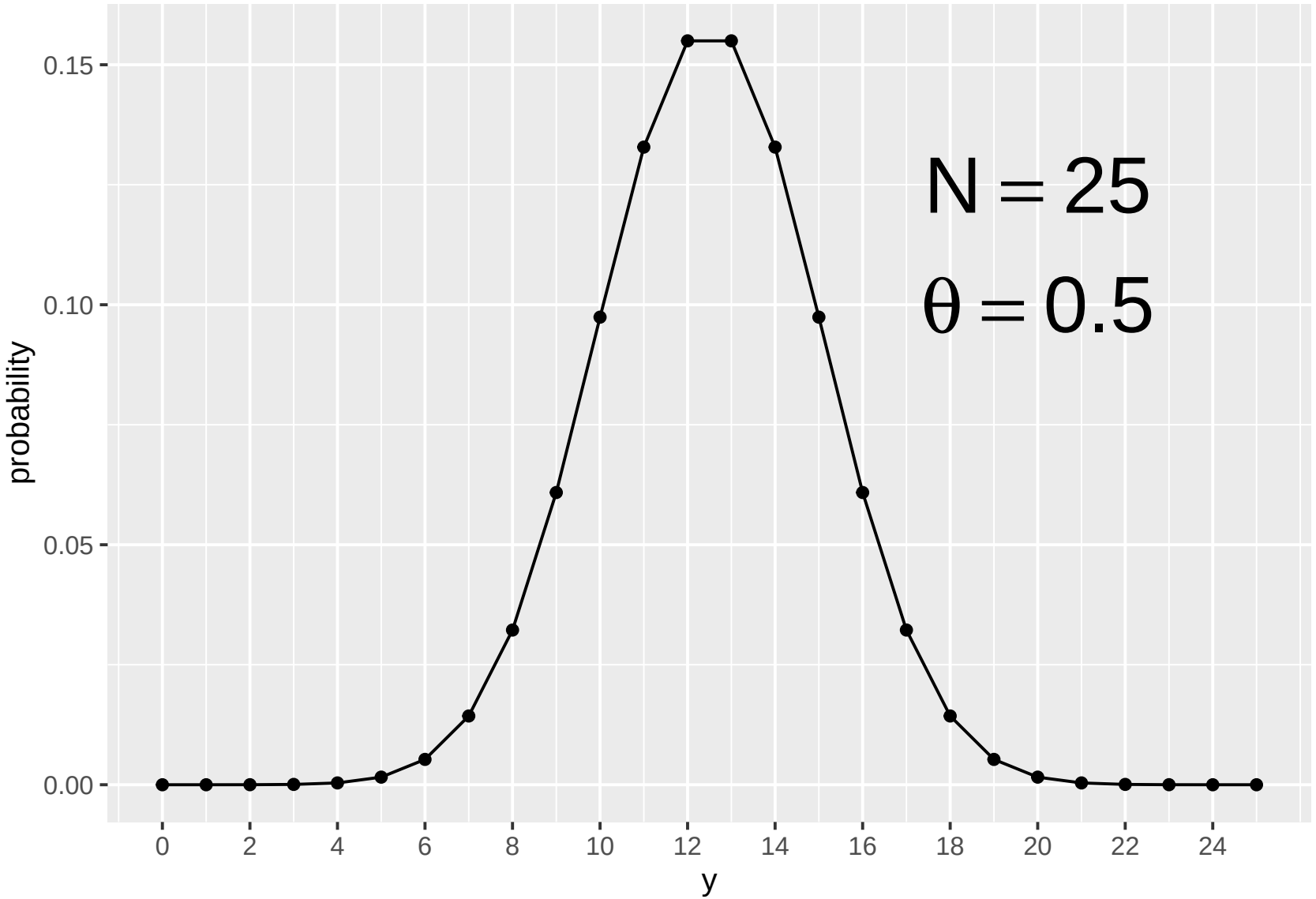
Modelling Fish Population - Binomial distribution

- (finite) population of fish in a lake of size N.
- One possible choice of a model is the Binomial distribution

$$p(y \mid N, \theta) = \binom{N}{y} \theta^y (1-\theta)^{N-y}$$

- sampling/ fishing: y out of N in total
- θ is *capture probability*
- N and θ are generally called *parameters*
- y is called *data*

Modelling Fish Population - Binomial distribution



Subsection 6

Data Set

Data Set

Table 1: Collected fish data: number of caught fish in 5 locations at 3 different time points.

	sampling occasions		
site	t1	t2	t3
1	2	1	2
2	3	5	5
3	0	1	1
4	2	2	1
5	3	3	3

Data Set

Table 1: Collected fish data: number of caught fish in 5 locations at 3 different time points.

	sampling occasions		
site	t1	t2	t3
1	2	1	2
2	3	5	5
3	0	1	1
4	2	2	1
5	3	3	3

- The total number of fish over all locations varies between 10 to 12.

Data Set

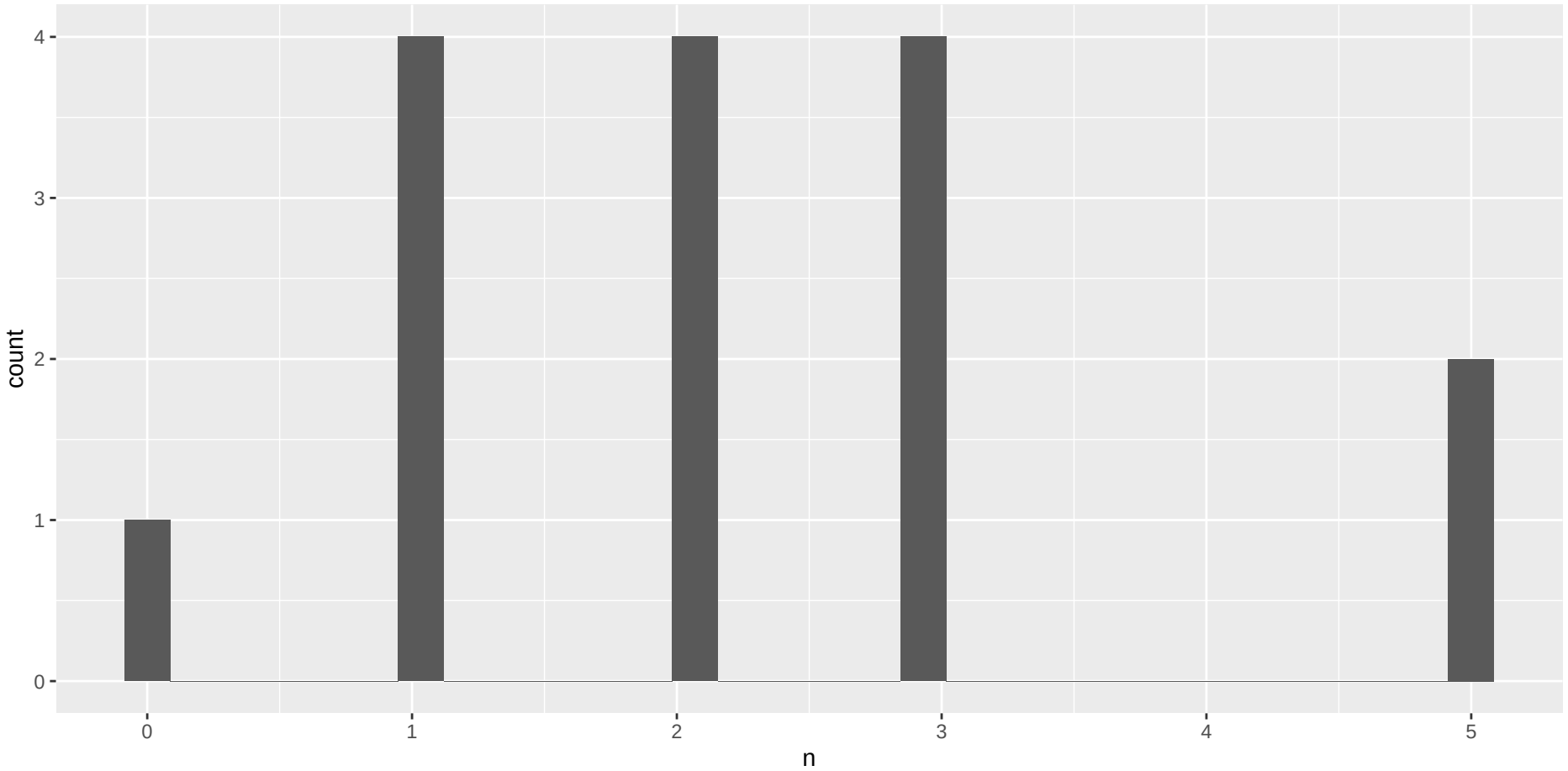


Figure 7: Histogram of the collected fish capture data.

Subsection 7

Fitting the model

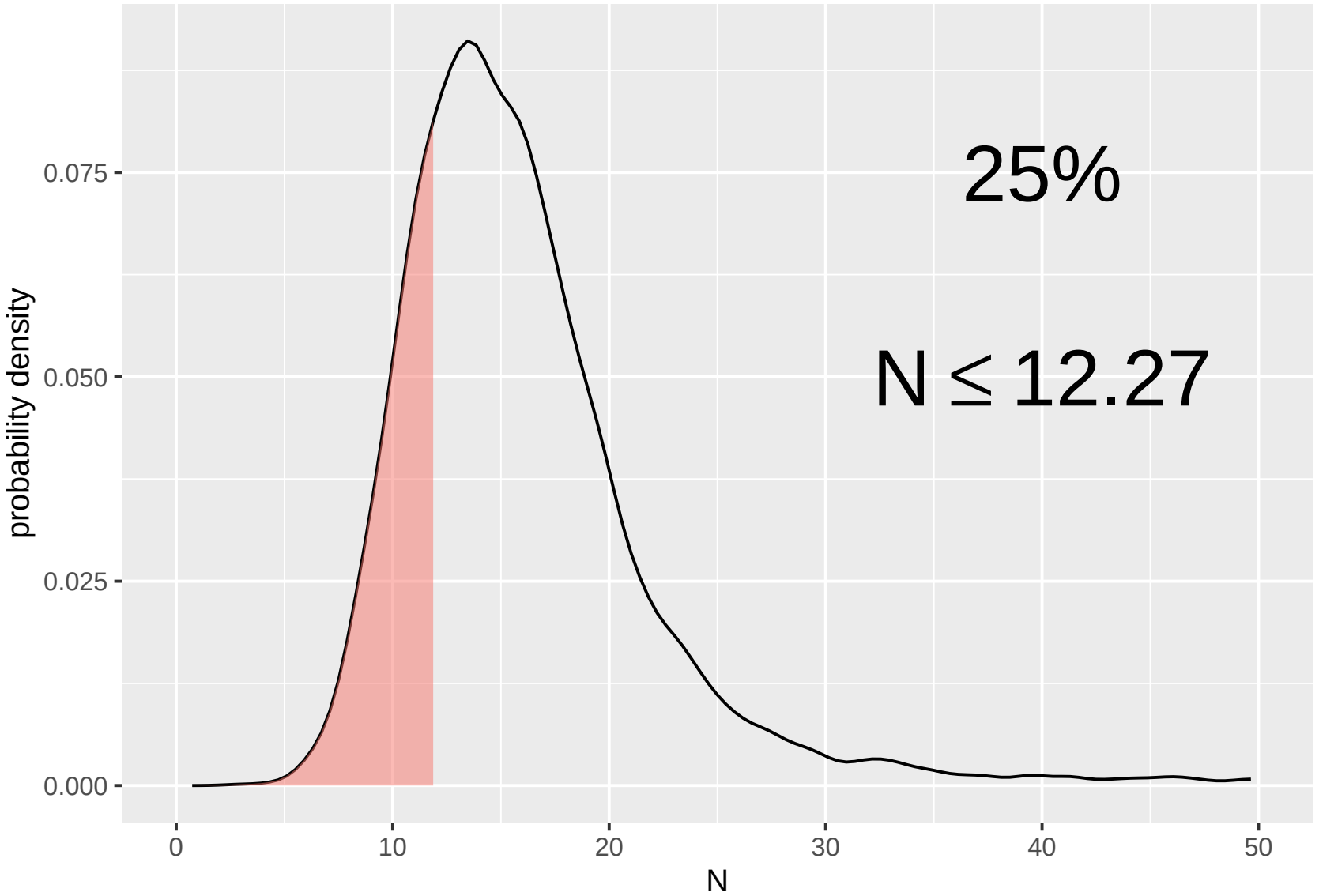
Fitting the model

```
## Inference for Stan model: fish.
## 4 chains, each with iter=4000; warmup=1000; thin=1;
## post-warmup draws per chain=3000, total post-warmup draws=12000.
##
##          mean se_mean    sd 2.5%   25%   50%   75% 97.5% n_eff Rhat
## p        0.75    0.00  0.15 0.30  0.70  0.80  0.86  0.93  2115    1
## lambda   3.39    0.06  2.04 1.66  2.45  3.01  3.71  7.79  1350    1
## Ntotal  16.93    0.28 10.22 8.30 12.27 15.07 18.55 38.96  1350    1
##
## Samples were drawn using NUTS(diag_e) at Mon Jan 18 18:19:08 2021.
## For each parameter, n_eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor on split chains (at
## convergence, Rhat=1).
```

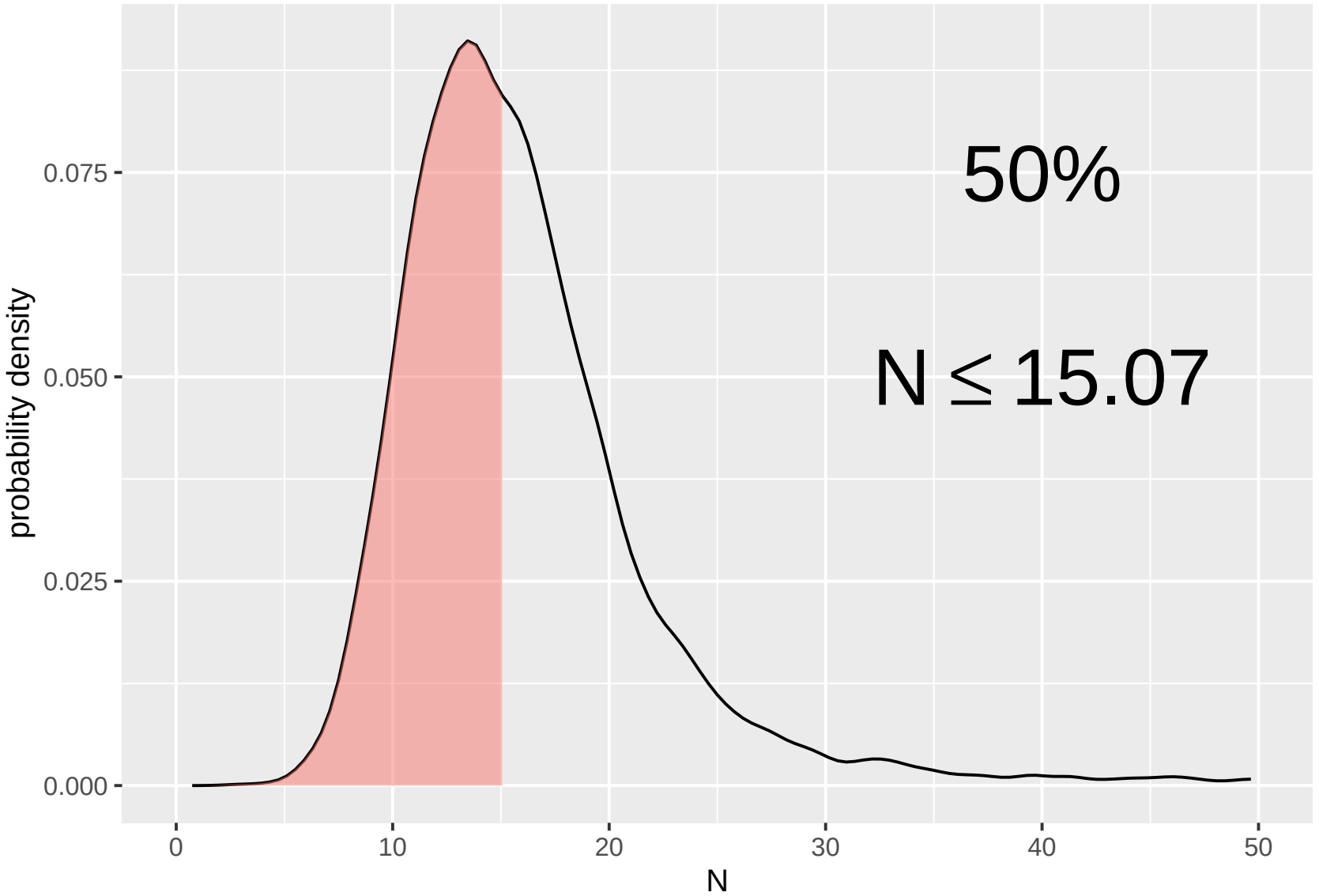
Subsection 8

Inference - Parameters as estimates

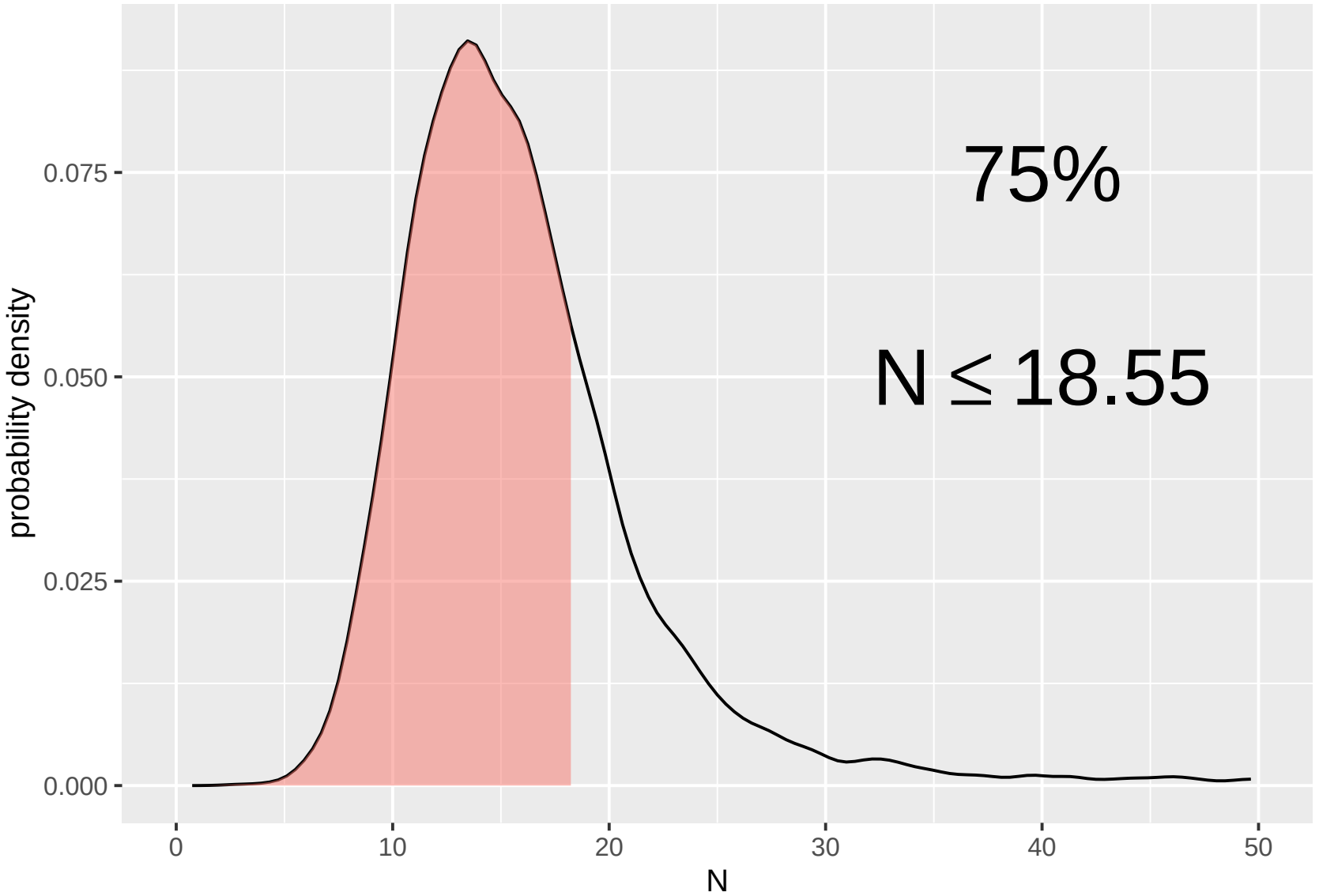
Inference - Parameters as estimates



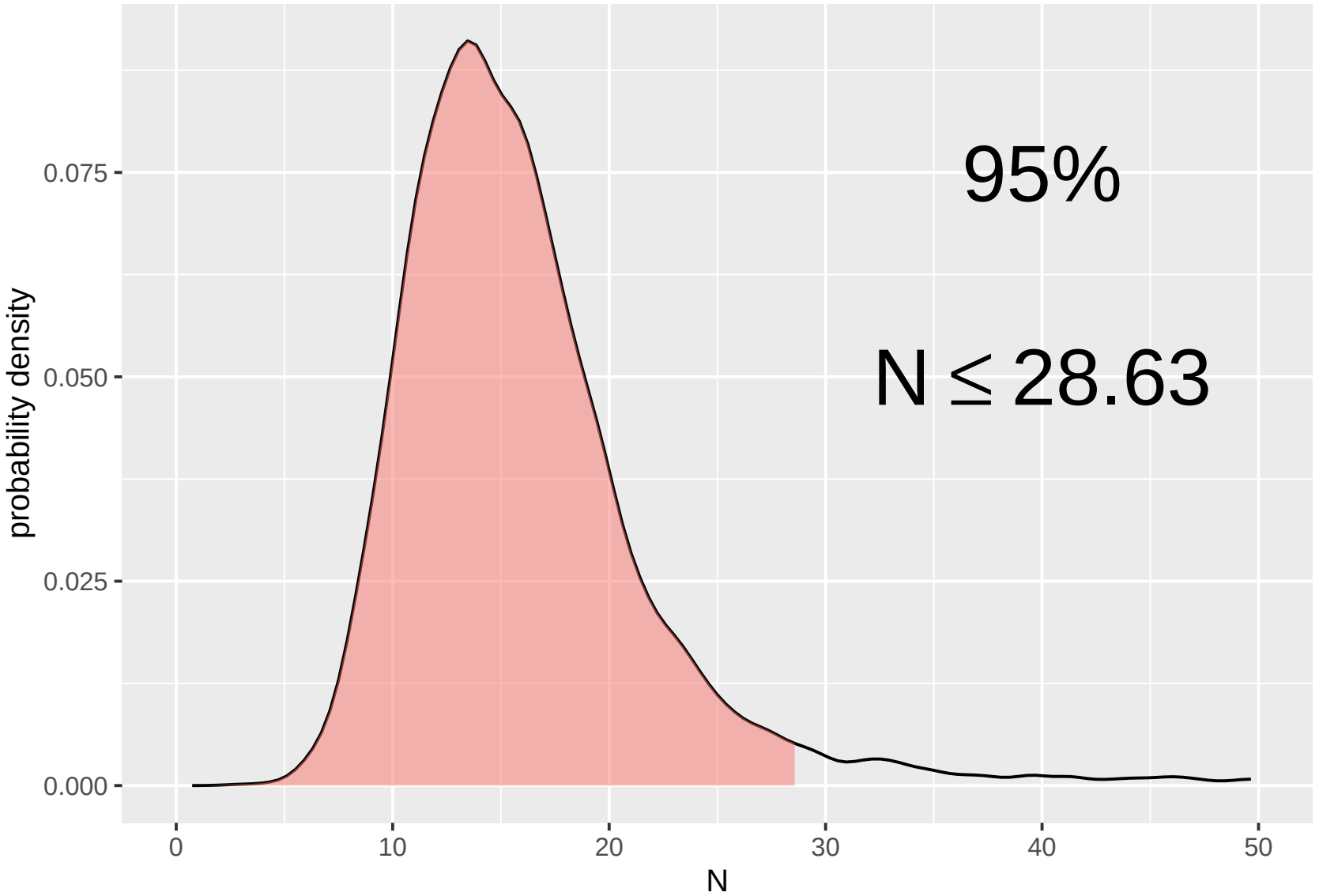
Inference - Parameters as estimates



Inference - Parameters as estimates



Inference - Parameters as estimates



Section 5

Summary

Summary

- the role of statistics
- vocabulary: uncertainty, variation, population, parameters, data
- probability distributions
- parameter estimation

Summary

- the role of statistics
- vocabulary: uncertainty, variation, population, parameters, data
- probability distributions
- parameter estimation

Summary

- the role of statistics
- vocabulary: uncertainty, variation, population, parameters, data
- probability distributions
- parameter estimation

Summary

- the role of statistics
- vocabulary: uncertainty, variation, population, parameters, data
- probability distributions
- parameter estimation

Ingram Olkin, A John Petkau, and James V Zidek. A comparison of n estimators for the binomial distribution. *Journal of the American Statistical Association*, 76(375):637–642, 1981.