MATH 203

Principles of Statistics I

Fall 2024

Dr David A. Stephens
Section 001
david.stephens@mcgill.ca

Dr José Correa
Section 002
jose.a.correa@mcgill.ca

Department of Mathematics & Statistics
McGill University

Overview

This course presents an introduction to probability and statistics.

Prerequisite knowledge:

- Elementary mathematical concepts
  - logarithms, exponents
  - functions

If we use language or notation you do not recognize, please let us know.

Overview

The two Sections will

- follow the same lecture slides
- have the same in-term assessments
- have the final examination
- have a shared myCourses page

Overview

Available via myCourses

- Course materials
- Lecture recordings
- Discussion board

In addition, there is a Teams channel.

Overview

If you have questions about course management, issues etc., please contact the instructors.

- TAs will be available to address questions on the course content.

Course content

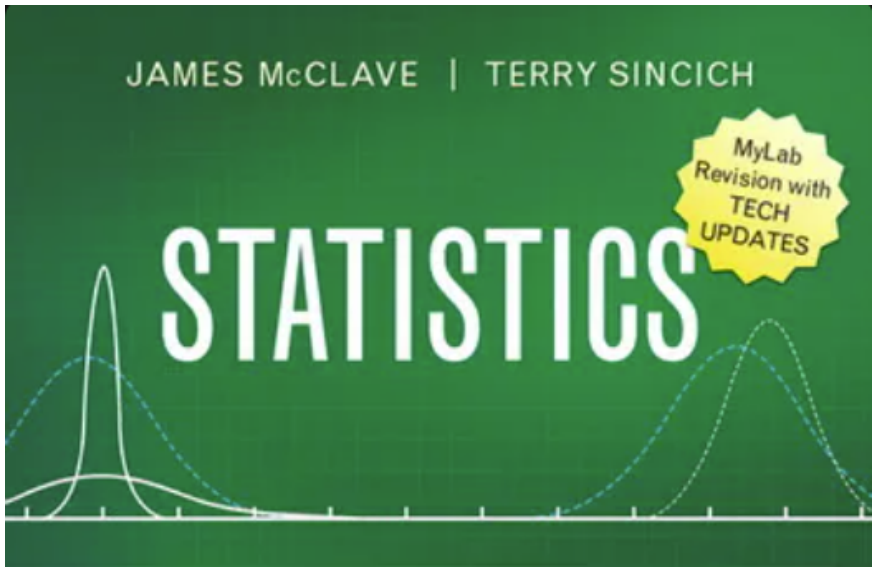
The course is broken down into four parts

1. Data
2. Introduction to Probability
3. Probability Distributions
4. Statistical methods

Textbook

A useful[1] book for the course is:

Statistics (13th edition, 2021), J. T. McClave, T. T Sincich



We will be covering Chapters 1 to 9.

[1] useful but not mandatory

Statistical computing

In this course, we will be illustrating some of the key calculations using the statistical software R.

- basic knowledge of statistical computing is a necessary skill for modern statistics and data science;
- computing is intended to be a labour-saving device;
- we will not be assessing your computing skills,
- however, anyone who wishes to learn the basics will have the opportunity;
- the myCourses page will have some relevant content

Chapter 1
Data

Motivation

Statistical methods are particularly useful for studying, analyzing, and learning about populations of experimental units.

- Experimental (or observational) unit
  - An “object” about which we collect data
  - e.g. person, thing, transaction, tree, event
- Population
  - the set of all units (people, transactions, trees, animals etc.) that we want to study

Basic components

- Variable
  - A characteristic or property of an individual experimental (or observational) unit in the population
  - e.g. age, weight, counts etc.
- Sample
  - a subset of units in a population

Basic components

- Statistical inference
  - An estimate, prediction, or some other generalization about a population based on information contained in a sample
- Measure of reliability
  - A statement (usually quantitative) about a degree of uncertainty associated with a statistical inference

Elements of an analysis

In any study we need to determine:

1. Objective(s) of a study: what question(s) do we want to answer in our study?
2. Experimental units: people, machines, transactions, ..., that we want to study
3. Population under study: a set of experimental units
    - All first-year McGill undergrad students
    - All transactions made at Walmart in Montreal between 2008 and 2009

Elements of an analysis

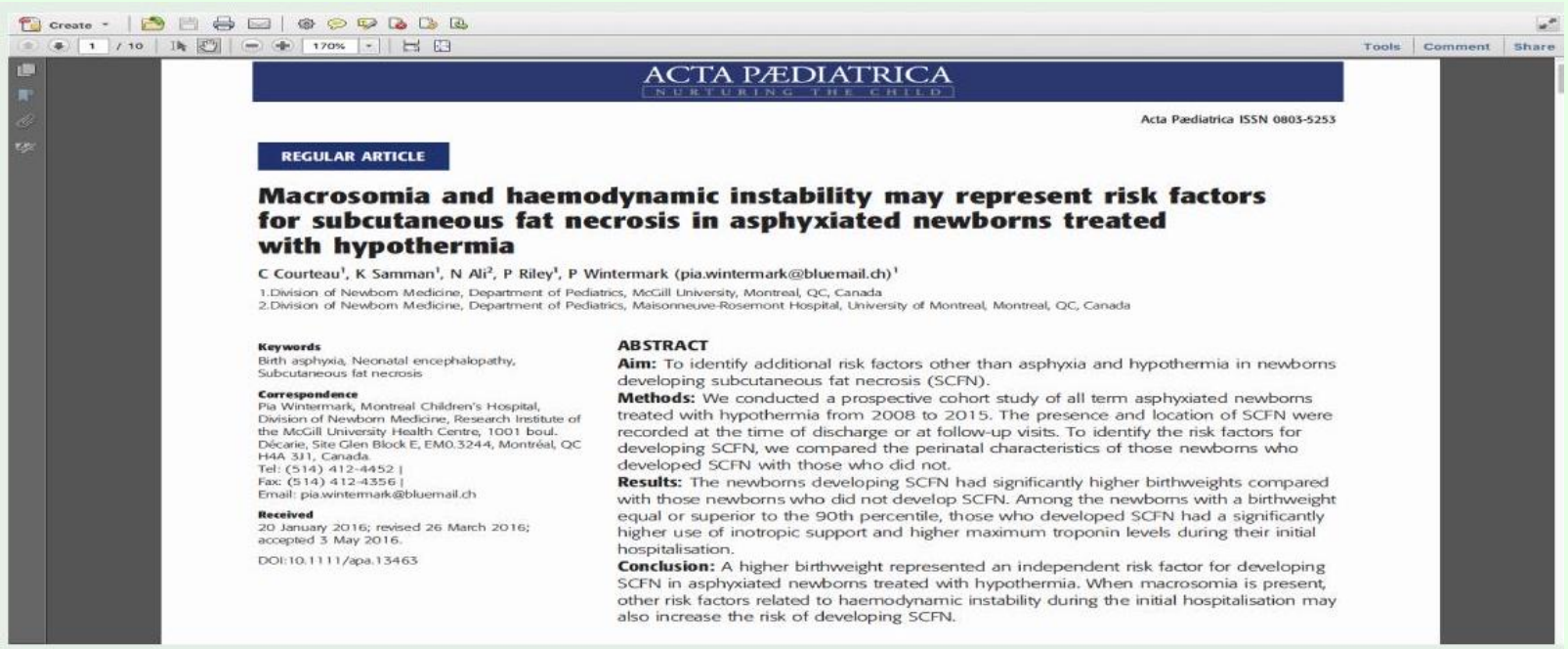
4. Characteristic(s) of interest: variable(s) measured on the experimental units
5. Sample: a subset of a population under study
    - If we could, we would look at the whole population (census); otherwise we need to take a sample

Elements of an analysis

6. Inference(s) of interest about the population based on the sample
7. Measure(s) of reliability of the inference(s)

Example

Example: Sub-Cutaneous Fat Necrosis in newborns



ACTA PÆDIATRICA
NURTURING THE CHILD

Acta Pædiatrica ISSN 0803-5253

REGULAR ARTICLE

Macrosomia and haemodynamic instability may represent risk factors for subcutaneous fat necrosis in asphyxiated newborns treated with hypothermia

C Courteau[1], K Samman[1], N Ali[2], P Riley[1], P Wintermark (pia.wintermark@bluemail.ch)[1]

1.Division of Newborn Medicine, Department of Pediatrics, McGill University, Montreal, QC, Canada
2.Division of Newborn Medicine, Department of Pediatrics, Maisonneuve-Rosemont Hospital, University of Montreal, Montreal, QC, Canada

Keywords
Birth asphyxia, Neonatal encephalopathy, Subcutaneous fat necrosis

Correspondence
Pia Wintermark, Montreal Children's Hospital, Division of Newborn Medicine, Research Institute of the McGill University Health Centre, 1001 boul. Décarie, Site Glen Block E, EM0.3244, Montréal, QC H4A 3J1, Canada.
Tel: (514) 412-4452 |
Fax: (514) 412-4356 |
Email: pia.wintermark@bluemail.ch

Received
20 January 2016; revised 26 March 2016; accepted 3 May 2016.

DOI:10.1111/apa.13463

ABSTRACT

Aim: To identify additional risk factors other than asphyxia and hypothermia in newborns developing subcutaneous fat necrosis (SCFN).

Methods: We conducted a prospective cohort study of all term asphyxiated newborns treated with hypothermia from 2008 to 2015. The presence and location of SCFN were recorded at the time of discharge or at follow-up visits. To identify the risk factors for developing SCFN, we compared the perinatal characteristics of those newborns who developed SCFN with those who did not.

Results: The newborns developing SCFN had significantly higher birthweights compared with those newborns who did not develop SCFN. Among the newborns with a birthweight equal or superior to the 90th percentile, those who developed SCFN had a significantly higher use of inotropic support and higher maximum troponin levels during their initial hospitalisation.

Conclusion: A higher birthweight represented an independent risk factor for developing SCFN in asphyxiated newborns treated with hypothermia. When macrosomia is present, other risk factors related to haemodynamic instability during the initial hospitalisation may also increase the risk of developing SCFN.

Acta Paediatrica 2016 Sep;105(9):e396-405.
https://doi.org/10.1111/apa. 13463

Example

Example: Sub-Cutaneous Fat Necrosis in newborns

- Some newborns suffer asphyxia during birth
- A treatment for perinatal asphyxia is called therapeutic hypothermia
  - involves lowering the child's body temperature for 72 hours,
  - then the child's body temperature is slowly raised to normal.

Example

Example: Sub-Cutaneous Fat Necrosis in newborns

- Some children develop the rare condition sub-Cutaneous fat necrosis (SCFN)
  - this usually disappears within a few weeks, however
  - some children develop more serious complications

Example

Example: Sub-Cutaneous Fat Necrosis in newborns

- Population: All term-newborns (i.e. not premature) who may suffer from perinatal asphyxia and will receive hypothermia as a treatment (these are the units of interest);
- Variable(s) of interest: Whether or not the child developed SCFN, birthweight and other risk factors;
- Sample: Term asphyxiated newborns admitted to the Montreal Children's Hospital Neonatal Intensive Care Unit from August 2008 to March 2015;
- Inference: One inference of interest was the estimation of birthweight for those who developed SCFN.

Basic types of Data

- Quantitative data
  - Measurements are recorded on a naturally occurring numerical scale
  - Height, blood pressure, speed, temperature, counts, ...

Basic types of Data

- Qualitative data
  - Measurements cannot be made in a natural numerical scale
  - Can only be ascertained and classified into one of several groups or categories
  - Gender, presence or not of a disease, religion, ethnicity, ...

Data collection

Once we have decided on the units, population, sample and the quantities of interest, we need to collect data

There are in general three methods of obtaining data

1. Published sources
    - books, journals, government publications etc..
2. Designed experiments.
3. Observational studies.

Designed experiments

In a designed experiment

- The researcher designs the experiment and has control over the experimental units sampled
- Typically involves two groups
  - A group of experimental units assigned to a treatment,
  - A group of experimental units that are untreated;
- May involve more than two groups.

Designed experiments

Example: Fertilizer study

A company claims that its new organic fertilizer increases yield and size of tomatoes, as compared to the leading brand.

Experiment: Tomato plants are randomly assigned to two groups:

- Leading fertilizer
- New organic product

The size and weight of each tomato is recorded, along with the total yield of the plants for comparison

- Key element: Researcher has control over what plant gets which fertilizer

Observational studies

- Sometimes it is either unethical or not possible to assign a treatment or group to a subject
- The researcher can only observe (directly or indirectly) the experimental units and record the variables of interest

Observational studies

Example: Smoking and Lung cancer

- We cannot assign people to smoking/non smoking groups and then look at incidence of lung cancer
- Possible approach: Select cancer and non-cancer patients, and then determine whether they were smokers or not from hospital records
- Key element: Researcher can only observe (directly or indirectly) the experimental units and records the variables of interest (e.g. smoking status)

Example: Iron Man competition

Oral Hypertonic Saline Is Effective in Reversing Acute Mild-to-Moderate Symptomatic Exercise-Associated Hyponatremia

Eileen Bridges, MD, MSc,* Tawfeeq Altherwi, MD,† José A. Correa, PhD,‡ and Tamara Hew-Butler, DPM, PhD§

Abstract

Objectives: To determine whether oral administration of 3% hypertonic saline (HTS) is as efficacious as intravenous (IV) 3% saline in reversing symptoms of mild-to-moderate symptomatic exercise-associated hyponatremia (EAH) in athletes during and after a long-distance triathlon. **Design:** Noninferiority, open-label, parallel-group, randomized control trial to IV or oral HTS. We used permuted block randomization with sealed envelopes, containing the word either "oral" or "IV." **Setting:** Annual long-distance triathlon (3.8-km swim, 180-km bike, and 42-km run) at Mont-Tremblant, Quebec, Canada. **Participants:** Twenty race finishers with mild to moderately symptomatic EAH. **Independent Variables:** Age, sex, race finish time, and 9 clinical symptoms. **Main Outcome Measures:** Time from treatment to discharge. **Methods:** We successfully randomized 20 participants to receive either an oral (n = 11) or IV (n = 9) bolus of HTS. We performed venipuncture to measure serum sodium (Na) at presentation to the medical clinic and at time of symptom resolution after the intervention. **Results:** The average time from treatment to discharge was 75.8 minutes (SD 29.7) for the IV treatment group and 50.3 minutes (SD 26.8) for the oral treatment group (*t* test, $P = 0.02$). Serum Na before and after treatment was not significantly different in both groups. There was no difference on presentation between groups in age, sex, or race finish time, both groups presented with an average of 6 symptoms. **Conclusions:** Oral HTS is effective in reversing symptoms of mild-to-moderate hyponatremia in EAH.

Key Words: acute hyponatremia, triathlete, hypertonic saline, exercise-associated hyponatremia, endurance athlete, exercise

(*Clin J Sport Med* 2020;30:8–13)

Clinical Journal of Sport Medicine. 30(1):8-13, January 2020.

Observational studies

Example: Iron Man competition

- Iron Man competition: Swim 3.8 km, Bicycle ride 180.25 km, Marathon 42.2 km, all without a break
- Some athletes develop hyponatremia during the race: levels of sodium in the blood stream get abnormally low
- Can lead to serious health complications, including death
- Mild hyponatremia can be treated very effectively with an intravenous (IV) saline solution

Observational studies

Example: Iron Man competition

Investigators at McGill are interested in

1. Determining possible causes of hyponatremia
2. Testing whether an oral saline solution is as effective as the IV saline solution for treating mild hyponatremia

Observational studies

Example: Iron Man competition

Objective 1. can only be determined with an observational study:

- We cannot induce or provoke hyponatremia in subjects in order to have a treatment group
- We can only wait for hyponatremia to happen in some athletes during the race
- Only then we can make comparisons between the group of those with hyponatremia and those without

Observational studies

Example: Iron Man competition

Objective 2. can be determined with a designed experiment:

- Consider only those who suffered hyponatremia during the race
- Make two groups by assigning (at random) some to get the oral solution and some to get the IV solution
- We have control over who gets which treatment
- Compare effectiveness of the treatments between the groups

Methods for describing qualitative data

Class: One of the categories into which qualitative data can be classified

- frequency: Number of observations in the data that fall into a particular class
- relative frequency: Class frequency divided by the total number of observations in the data set
- percentage: Class relative frequency multiplied by 100

Methods for describing qualitative data

Graphical methods:

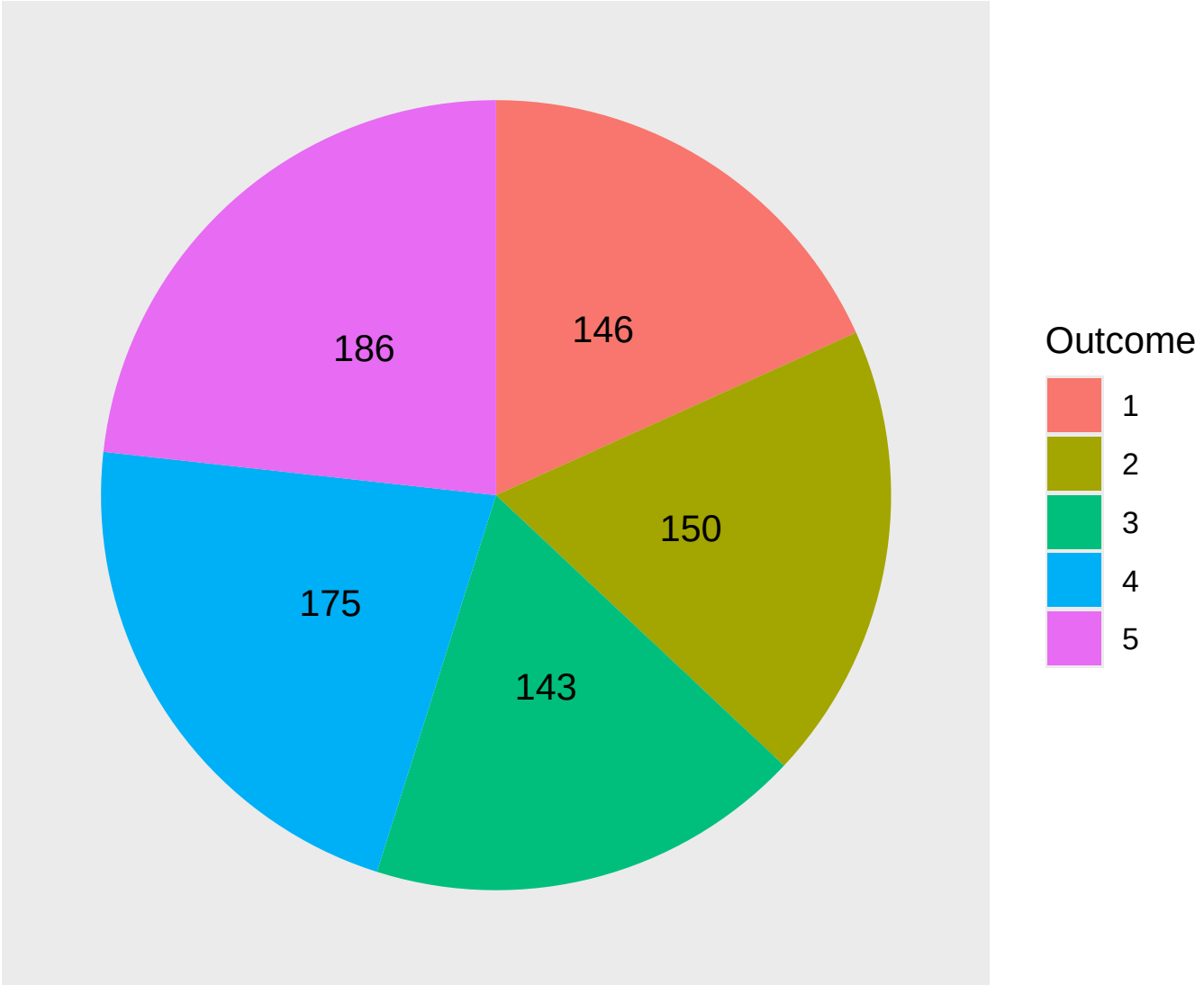
- There many types of graphical displays
- We will discuss the two that are the most commonly used

Methods for describing qualitative data

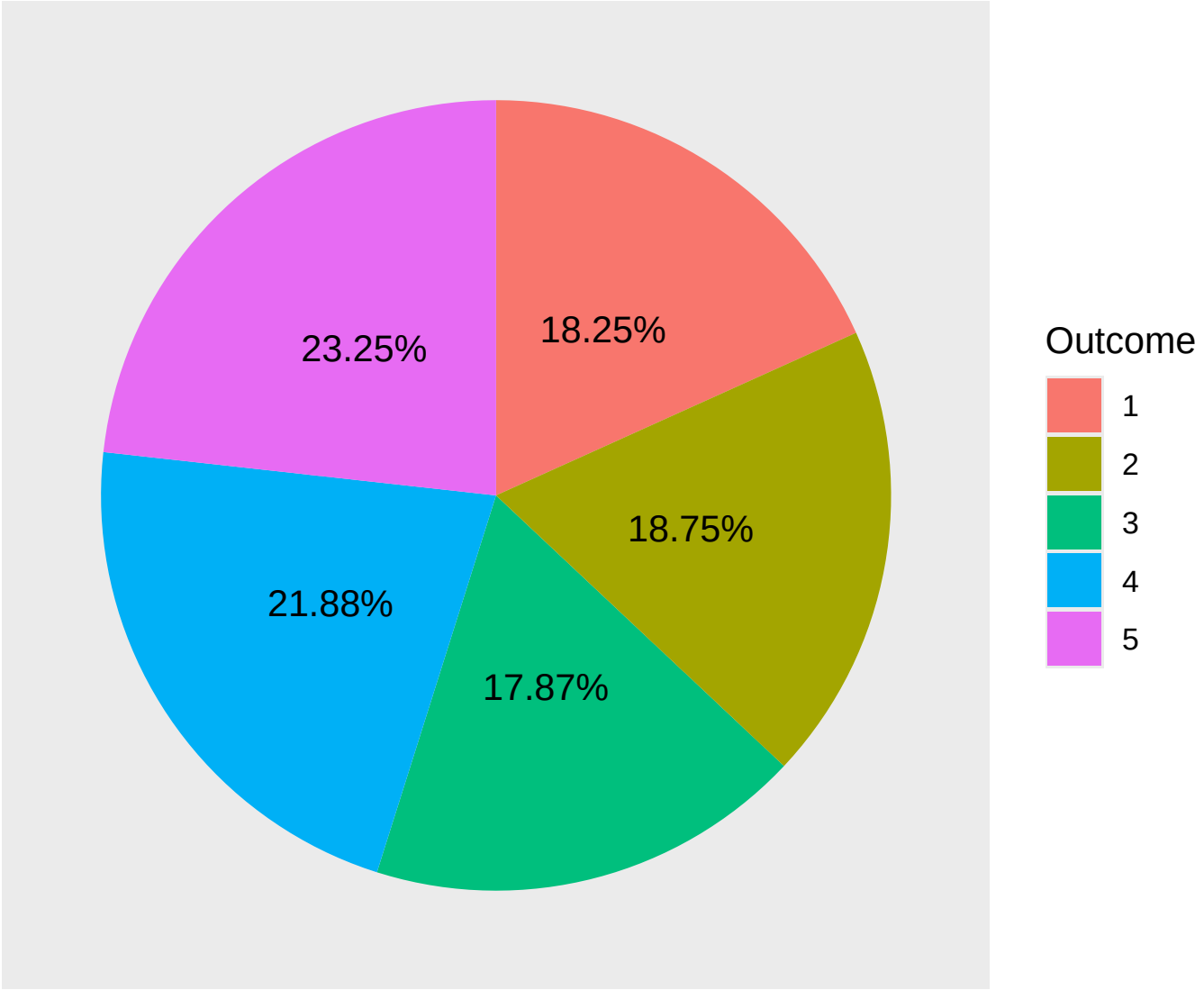
Pie chart

- Classes represented by slices of a pie (circle)
- The size of each slice is proportional to the class relative frequency

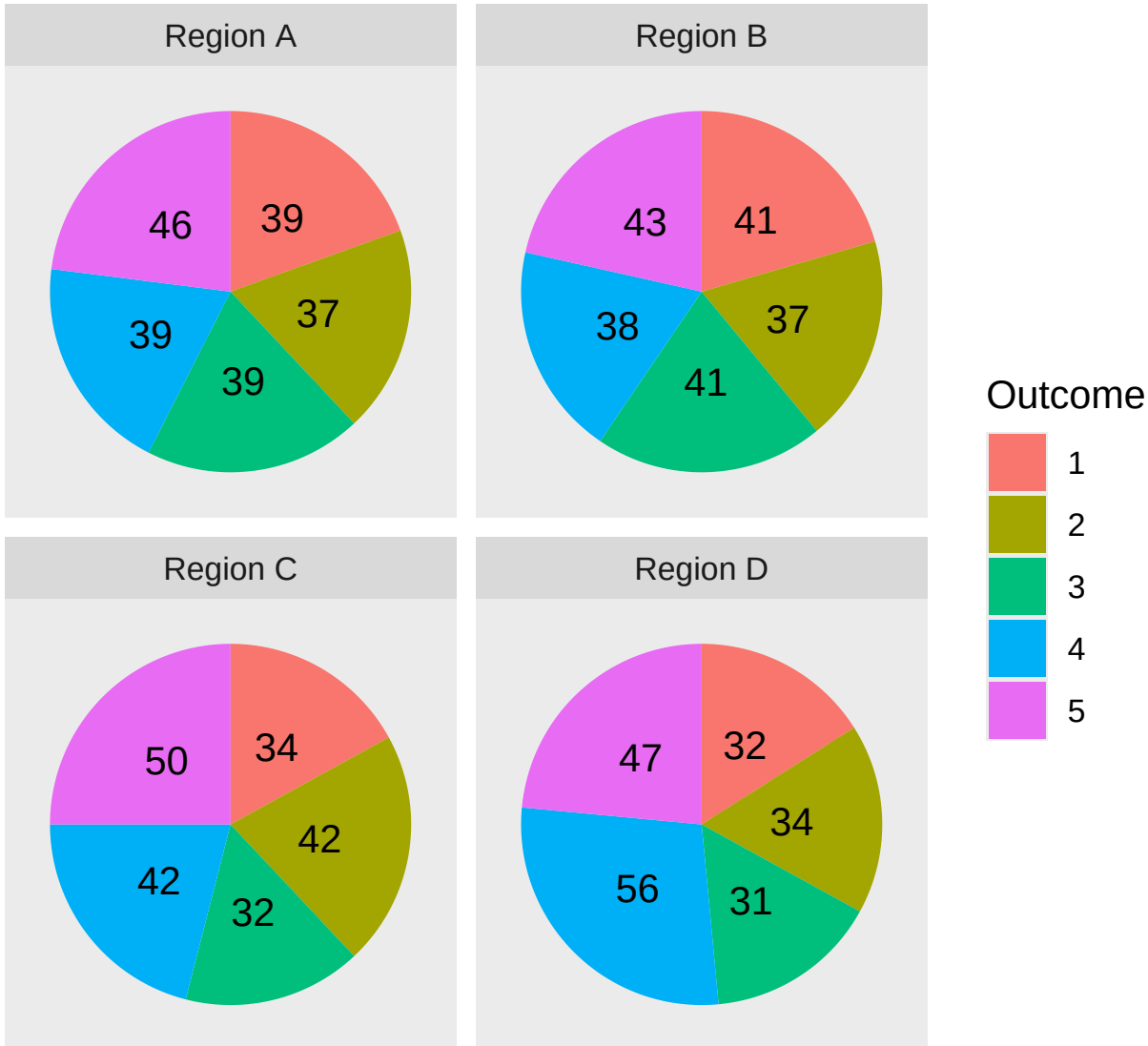
Methods for describing qualitative data



Methods for describing qualitative data



Methods for describing qualitative data

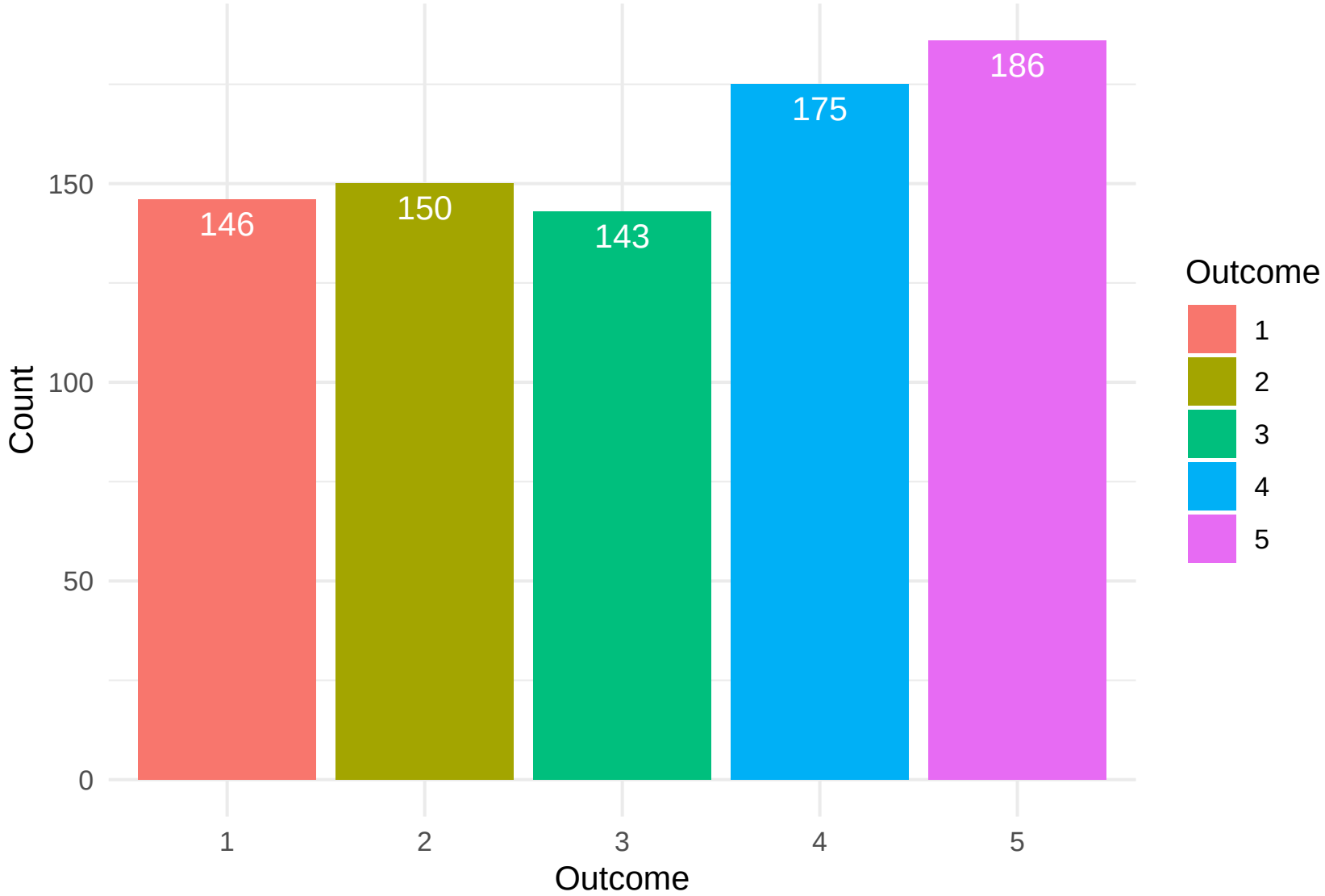


Methods for describing qualitative data

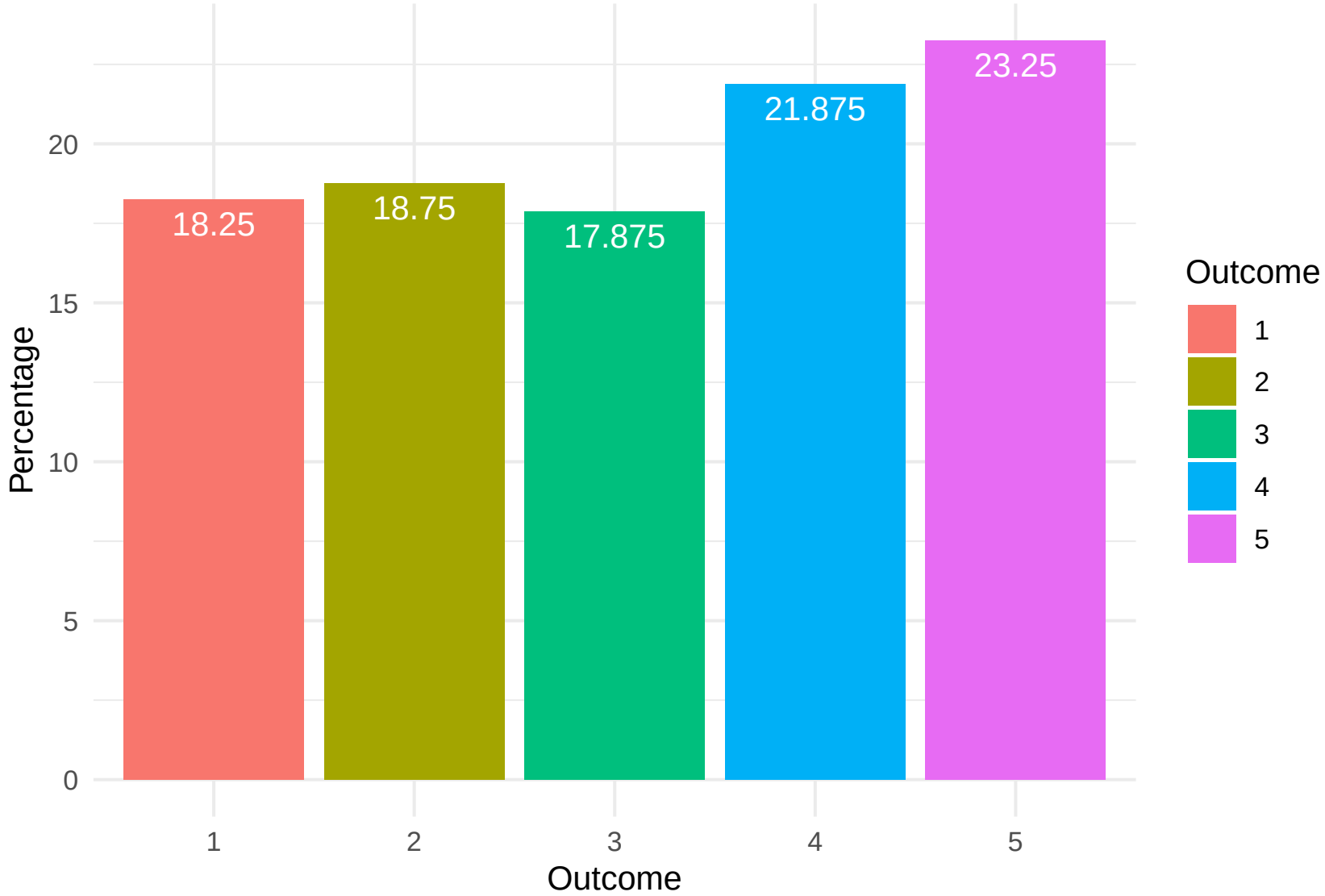
Bar plot

- Classes represented by bars
- The height of each bar is either the class frequency, class relative frequency, or class percentage
- The width of the bars is not important

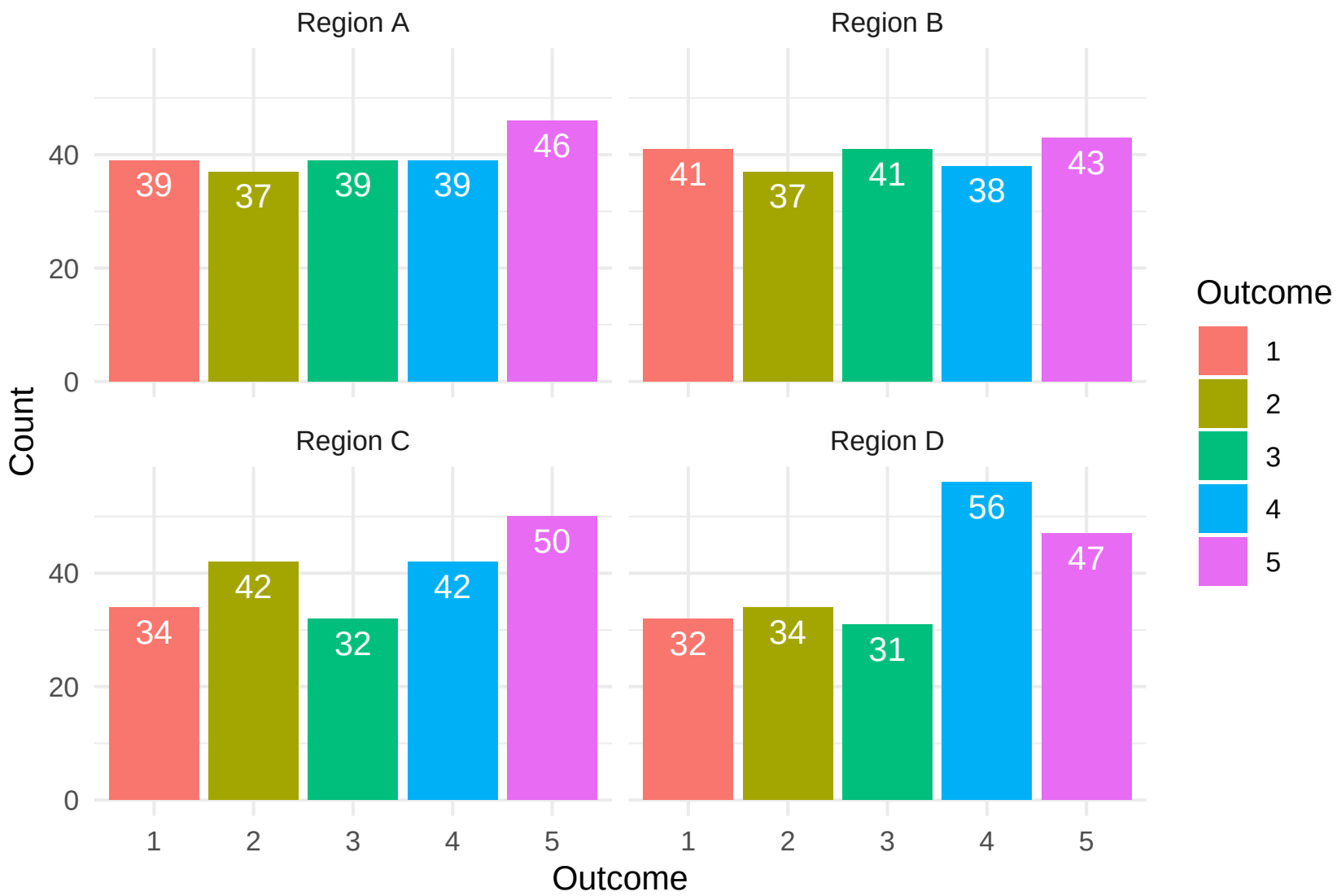
Methods for describing qualitative data



Methods for describing qualitative data



Methods for describing qualitative data



Methods for describing qualitative data

Numerical methods:

- Class frequency, relative frequency, percentage
- Usually displayed in a Table

Methods for describing qualitative data

Table: Outcome data: counts

	Outcome				
	1	2	3	4	5
Region A	39	37	39	39	46
Region B	41	37	41	38	43
Region C	34	42	32	42	50
Region D	32	34	31	56	47

Methods for describing qualitative data

Example: Kidney stone treatment

- Designed experiment. The experimental units are patients with kidney stones
- Two treatment groups: A and B
- Each patient was assigned to one of the two treatments

Methods for describing qualitative data

Example: Kidney stone treatment

The data represent three variables that are ascertained on each patient:

1. Treatment: Treatment group: (A or B)
2. Outcome: Whether or not the treatment was successful in removing the stones: (Success/Failure)
3. Size: The size of the stone.

The data produced are qualitative

Methods for describing qualitative data

Example: Kidney stone treatment

The Table shows:

- The number of patients treated per treatment.
- The number (frequency) of successes per treatment.
- The success rate (percentage) per treatment.

	Treat. A		Treat. B	
Success	273	(78 %)	289	(83 %)
Failure	77	(22 %)	61	(17 %)
Total	350		350	

Methods for describing qualitative data

Example: Kidney stone treatment

Pie charts of Success/Failure by treatment

```
Treatment<-factor(rep(c('A','B'),each=350))
Outcome<-c(rep(c('Success','Failure'),c(273,350-273)),
           rep(c('Success','Failure'),c(289,350-289)))
Outcome<-factor(Outcome,levels=c('Success','Failure'))
Size<-rep(rep(c('Small','Large'),4),c(81,192,6,71,234,55,36,25))
Size<-factor(Size,levels=c('Small','Large'))
kidney.df = data.frame(Treatment,Outcome,Size)

kidney.df1<-kidney.df %>%
    group_by(Treatment,Outcome) %>%
    summarize(Count=n())

+ `summarise()` has grouped output by 'Treatment'.  You can override using the
+ `.groups` argument.
```

Methods for describing qualitative data

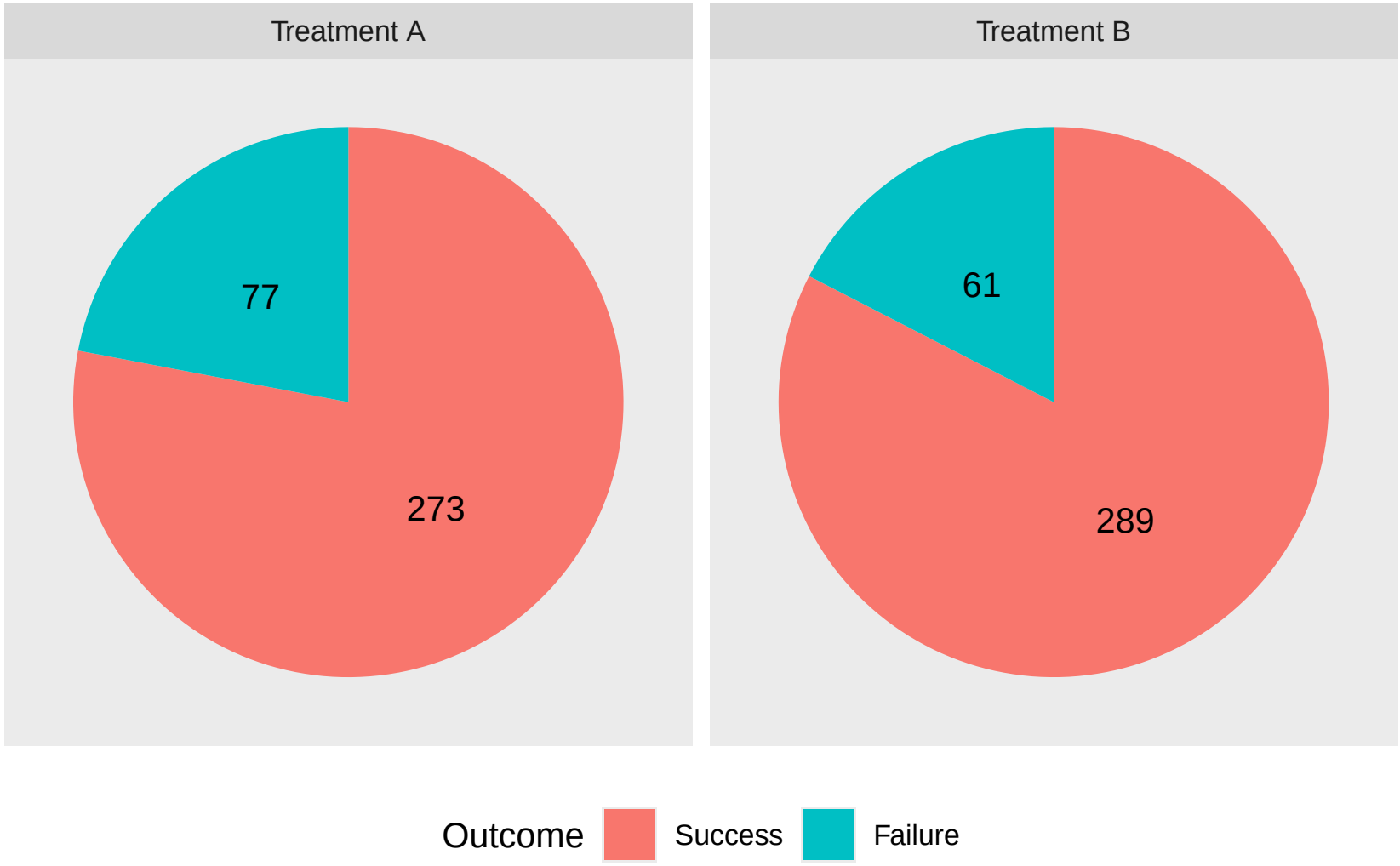
Example: Kidney stone treatment

```
kidney.df1

+ # A tibble: 4 x 3
+ # Groups:   Treatment [2]
+   Treatment Outcome Count
+   <fct>     <fct>   <int>
+ 1 A         Success   273
+ 2 A         Failure    77
+ 3 B         Success   289
+ 4 B         Failure    61
```

```
ggplot(data = kidney.df1, aes(x = "", y = Count, group=Treatment, fill = Outcome )) +
    geom_bar(stat = "identity", position = position_fill()) +
    geom_text(aes(label = Count), position = position_fill(vjust = 0.5)) +
    coord_polar(theta = "y") +
    facet_wrap(~ Treatment, labeller=labeller(Treatment =~ c('Treatment A','Treatment B')))  +
    theme(axis.title.x = element_blank(),
          axis.title.y = element_blank(),
          axis.text=element_blank(),
          axis.ticks=element_blank(),
          panel.grid = element_blank()) +
    theme(legend.position='bottom') +
    guides(fill=guide_legend(nrow=1, byrow=TRUE))
```

Methods for describing qualitative data



Methods for describing qualitative data

Example: Kidney stone treatment

```
kidney.pctA <- kidney.df %>%
  filter(Treatment=='A') %>%
  group_by(Treatment,Outcome) %>% # Variable to be transformed
  count() %>%
  ungroup() %>%
  mutate(perc = `n` / sum(`n`)) %>%
  arrange(perc) %>%
  mutate(labels = scales::percent(perc),Count=n)

kidney.pctB <- kidney.df %>%
  filter(Treatment=='B') %>%
  group_by(Treatment,Outcome) %>% # Variable to be transformed
  count() %>%
  ungroup() %>%
  mutate(perc = `n` / sum(`n`)) %>%
  arrange(perc) %>%
  mutate(labels = scales::percent(perc), Count=n)

kidney.pct<-bind_rows(kidney.pctA,kidney.pctB)
```

Methods for describing qualitative data

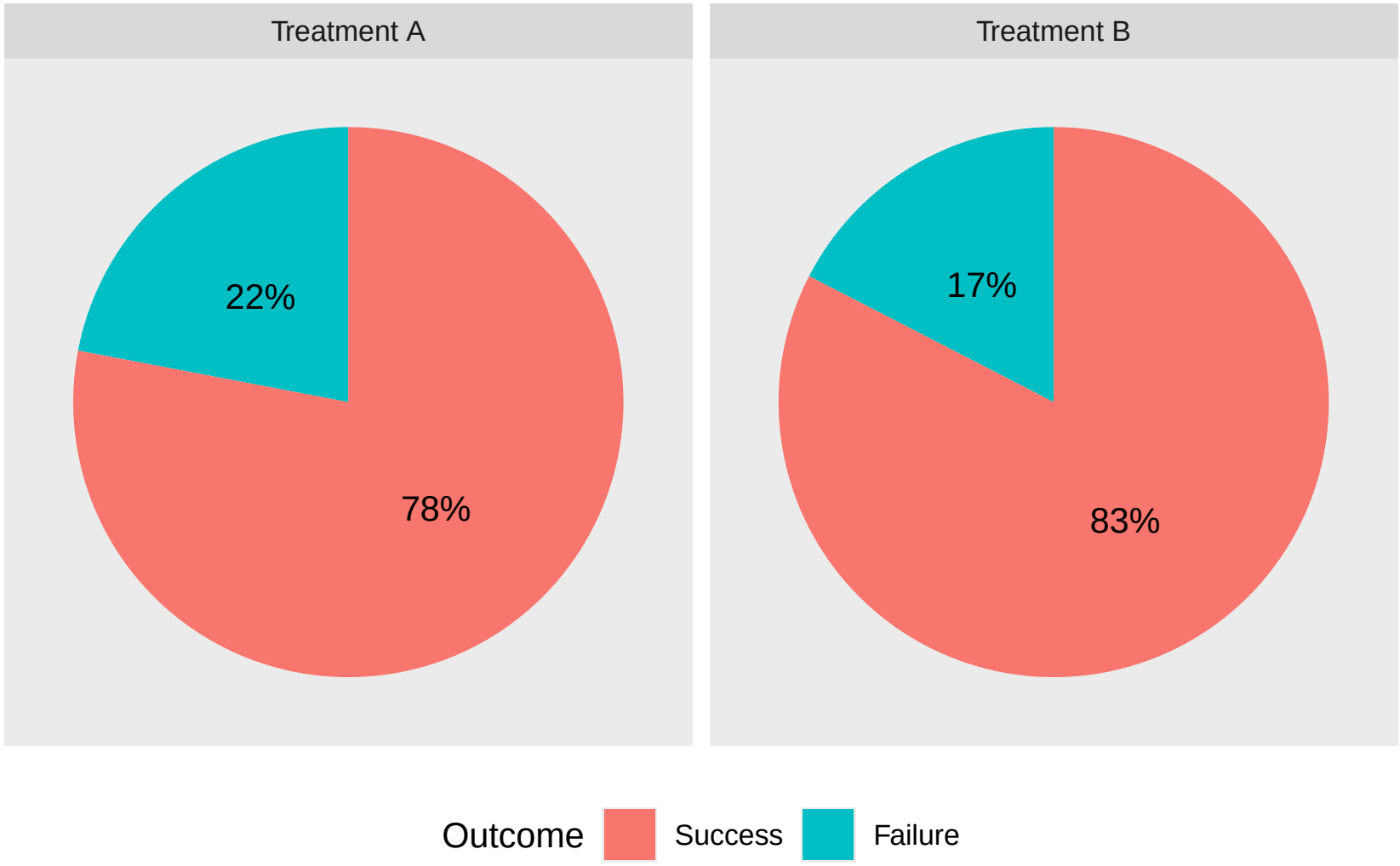
Example: Kidney stone treatment

```
kidney.pct

+ # A tibble: 4 x 6
+   Treatment Outcome     n  perc labels Count
+   <fct>     <fct>   <int> <dbl> <chr>  <int>
+ 1 A         Failure    77 0.22  22%       77
+ 2 A         Success   273 0.78  78%      273
+ 3 B         Failure    61 0.174 17%       61
+ 4 B         Success   289 0.826 83%      289
```

```
ggplot(data = kidney.pct, aes(x = "", y = Count, group=Treatment, fill = Outcome )) +
    geom_bar(stat = "identity", position = position_fill()) +
    geom_text(aes(label = labels), position = position_fill(vjust = 0.5)) +
    coord_polar(theta = "y", start=0, direction=-1) +
    facet_wrap(~ Treatment, labeller=labeller(Treatment =~ c('Treatment A','Treatment B')))  +
    theme(axis.title.x = element_blank(),
          axis.title.y = element_blank(),
          axis.text=element_blank(),
          axis.ticks=element_blank(),
          panel.grid = element_blank()) +
    theme(legend.position='bottom') +
    guides(fill=guide_legend(nrow=1, byrow=TRUE))
```

Methods for describing qualitative data



Methods for describing qualitative data

Example: Kidney stone treatment

Interpreting the results:

- It seems that treatment B is more effective; it shows a higher percentage of success

Methods for describing qualitative data

Example: Kidney stone treatment

However, breaking the results down by size of kidney stone:

		Treat. A		Treat. B	
Small	Success	81	(93 %)	234	(87 %)
	Failure	6	(7 %)	36	(13 %)
	Total	87		270	
Large	Success	192	(73 %)	55	(69 %)
	Failure	71	(27 %)	25	(31 %)
	Total	263		80	

Treatment A seems to be more effective than Treatment B, for both sizes.

Methods for describing qualitative data

Example: Kidney stone treatment

- The number of patients, when considering the stone size, are very different
- Looks like doctors tend to give treatment A to large stone cases and treatment B to small stone cases
- Patients with large stones using treatment A do worse than those with small stones, even if the latter used the inferior treatment B

Methods for describing qualitative data

Example: Kidney stone treatment

The interpretation of the relationship between

- Treatment type and
- Whether the treatment was a Success or a Failure

changed when the variable size of the stone was considered

We must be careful when interpreting results from a study

Simpson's Paradox

- The kidney stone data illustrate what is known as Simpson's Paradox:
- A third variable changes the interpretation of the relationship between two other variables;
- The third variable is termed a confounding factor.

We would say that the relationship between treatment type and success status was confounded by stone size.

Simpson's Paradox

Example: Berkeley college admission

Data from UC-Berkeley in 1973 on three variables from students who applied:

- Sex: Female/Male
- Major: A, B, C, D, E, F (for simplicity)
- Admission Status: Admitted/Not admitted.

Each of these variables provides qualitative data.

Simpson's Paradox

Example: Berkeley college admission

Frequencies of males and females admitted:

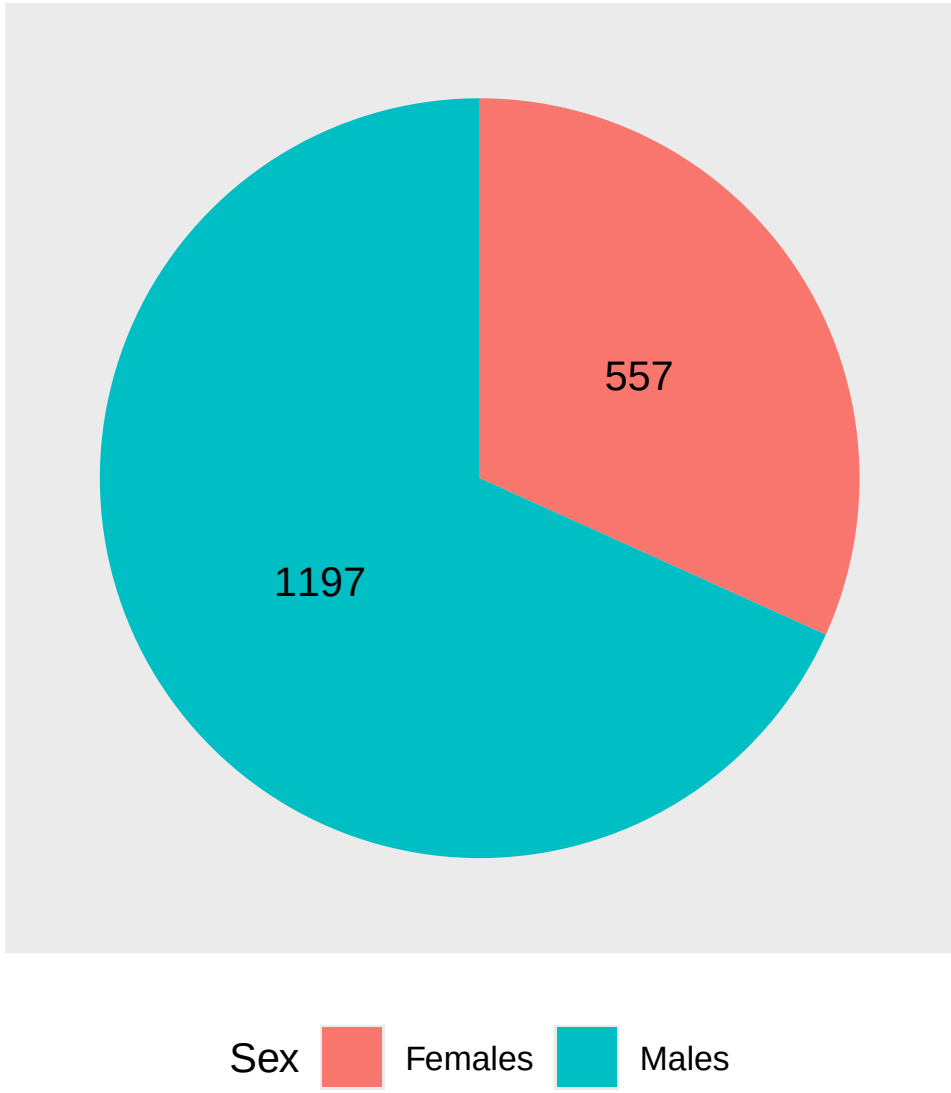
	Admitted
Males	1197
Females	557

Simpson's Paradox

Example: Berkeley college admission

```
UCB0<-data.frame(Sex=c('Females', 'Males'), Admitted=c('Yes','Yes'),Count=c(557,1197))
ggplot(data = UCB0, aes(x = "", y = Count, group=Sex, fill = Sex )) +
    geom_bar(stat = "identity", position = position_fill()) +
    geom_text(aes(label = Count), position = position_fill(vjust = 0.5)) +
    coord_polar(theta = "y", start=0, direction=-1) +
    theme(axis.title.x = element_blank(),
          axis.title.y = element_blank(),
          axis.text=element_blank(),
          axis.ticks=element_blank(),
          panel.grid = element_blank()) +
    theme(legend.position='bottom') +
    guides(fill=guide_legend(nrow=1, byrow=TRUE))
```

Simpson’s Paradox



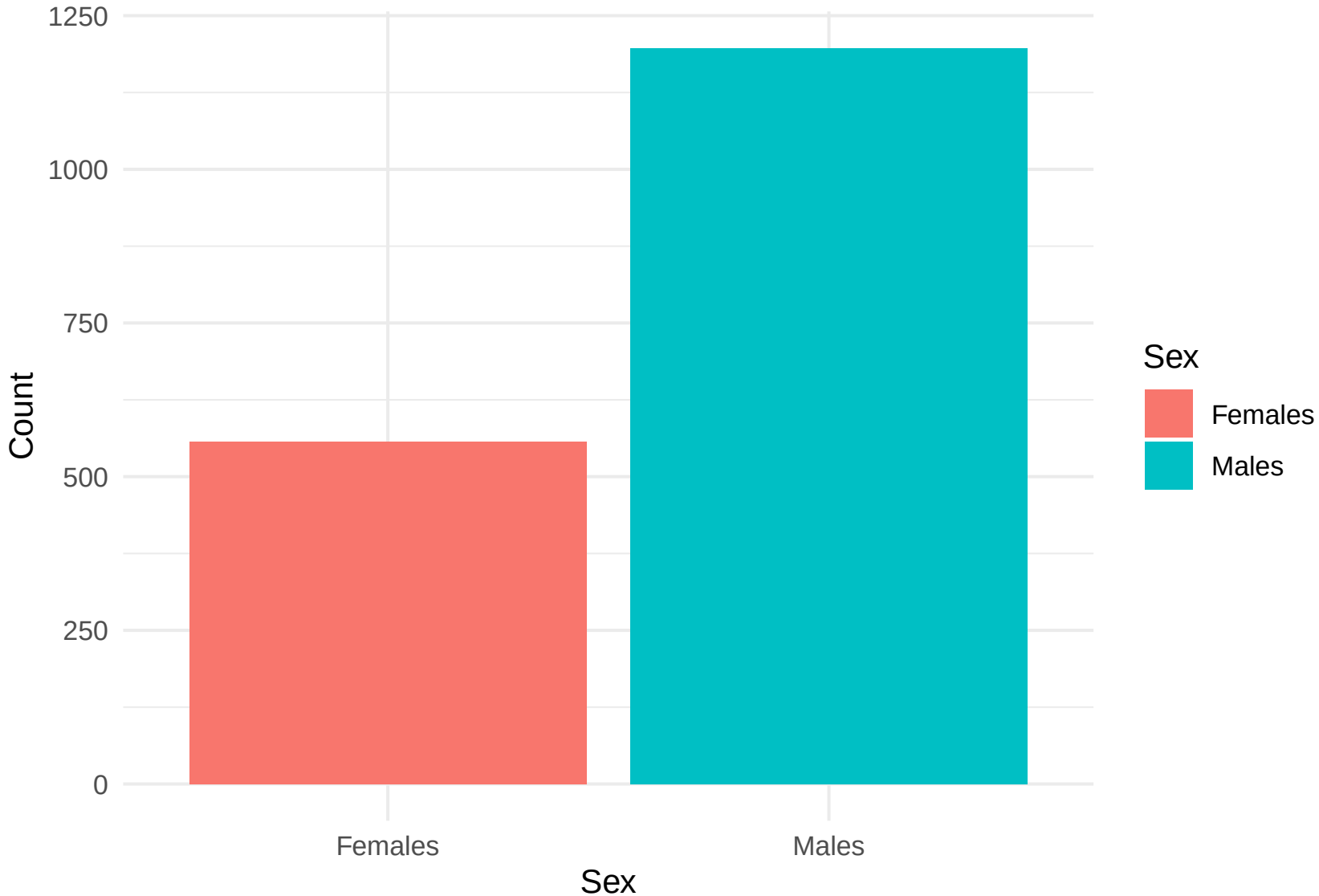
Simpson’s Paradox

Example: Berkeley college admission

Bar plot:

```
UCB0<-data.frame(Sex=c('Females', 'Males'), Admitted=c('Yes','Yes'),Count=c(557,1197))
ggplot(UCB0, aes(x=Sex, y=Count, fill=Sex)) +
  geom_bar(stat="identity")+theme_minimal()
```

Simpson’s Paradox



Simpson's Paradox

Example: Berkeley college admission

- If we only look at the frequencies of admitted, it looks like there is a bias against females;
- The table does not provide enough information;
- It is missing the number who have applied.

Simpson's Paradox

Example: Berkeley college admission

More information:

	Applied	Admitted	% Admitted
Males	2691	1197	44.5%
Females	1835	557	30.4%

Bias against females ?

Simpson's Paradox

Example: Berkeley college admission

We also have information about the Major:

	Males			Females		
Major	App.	Adm.	Adm. %	App.	Adm.	Adm. %
A	825	512	62	108	89	82
B	560	352	63	25	17	68
C	325	120	37	593	202	34
D	417	138	33	375	131	35
E	191	53	28	393	94	24
F	373	22	6	341	24	7
Total	2691	1197	45	1835	557	30

Simpson's Paradox

Example: Berkeley college admission

- We can see that when considering Major (confounding factor) no department was significantly biased against women
- We say that the relationship between Sex and Admission Status was confounded by Major

Graphical methods

- Recall that quantitative data are recorded on a meaningful numerical scale
- We can use graphical methods to summarize, describe and detect patterns in quantitative data
- We say that these methods describe the “shape" of the data

Graphical methods

We will discuss two graphical methods

- Histograms
- Boxplots

Histograms

- The values of the quantitative variable are partitioned into class intervals, each of which has the same width
- These intervals form the scale of the horizontal axis
- The frequency or relative frequency of observations in each class interval is determined
- A vertical bar is placed over each class interval, with the height of the bar equal to either the class frequency or class relative frequency

Histograms

Example: Body temperatures

Data: Body temperatures of 130 people, in degrees Celsius.

35.7 36.4 36.7 36.9 37.1 35.8 36.6 37.1 37.2 36.1 36.4 36.7 36.9 37.1 36.7 36.8 37.1
37.2 36.2 36.4 36.7 37.2 36.3 36.7 36.9 37.1 37.3 36.2 36.8 37.0 37.2 36.5 36.7 36.9
37.1 36.3 36.6 36.8 37.0 37.3 36.6 36.8 37.1 37.7 36.3 36.6 36.8 37.0 37.4 36.8 37.0
37.1 38.2 36.8 37.0 37.1 35.9 36.4 36.7 36.9 37.1 35.9 36.6 36.8 36.0 36.7 36.8 37.1
37.2 36.1 36.4 36.7 36.9 37.1 36.2 36.9 37.2 36.2 36.7 36.9 37.1 37.3 36.2 36.5 36.7
36.9 36.6 36.7 37.0 37.2 36.4 36.7 36.9 37.1 37.3 36.2 36.6 37.3 36.3 36.6 36.8 37.0
37.3 36.5 36.8 36.9 37.1 37.4 36.9 37.1 37.4 36.3 36.6 36.8 37.0 37.4 36.6 36.8 37.0
36.6 36.8 37.0 37.1 37.8 36.3 36.7 36.8 37.1 37.5 36.6

Histograms

Example: Body temperatures

```
BodyTemp<-scan('https://www.math.mcgill.ca/dstephens/203/Data/BodyTemp.txt')
BodyTemp

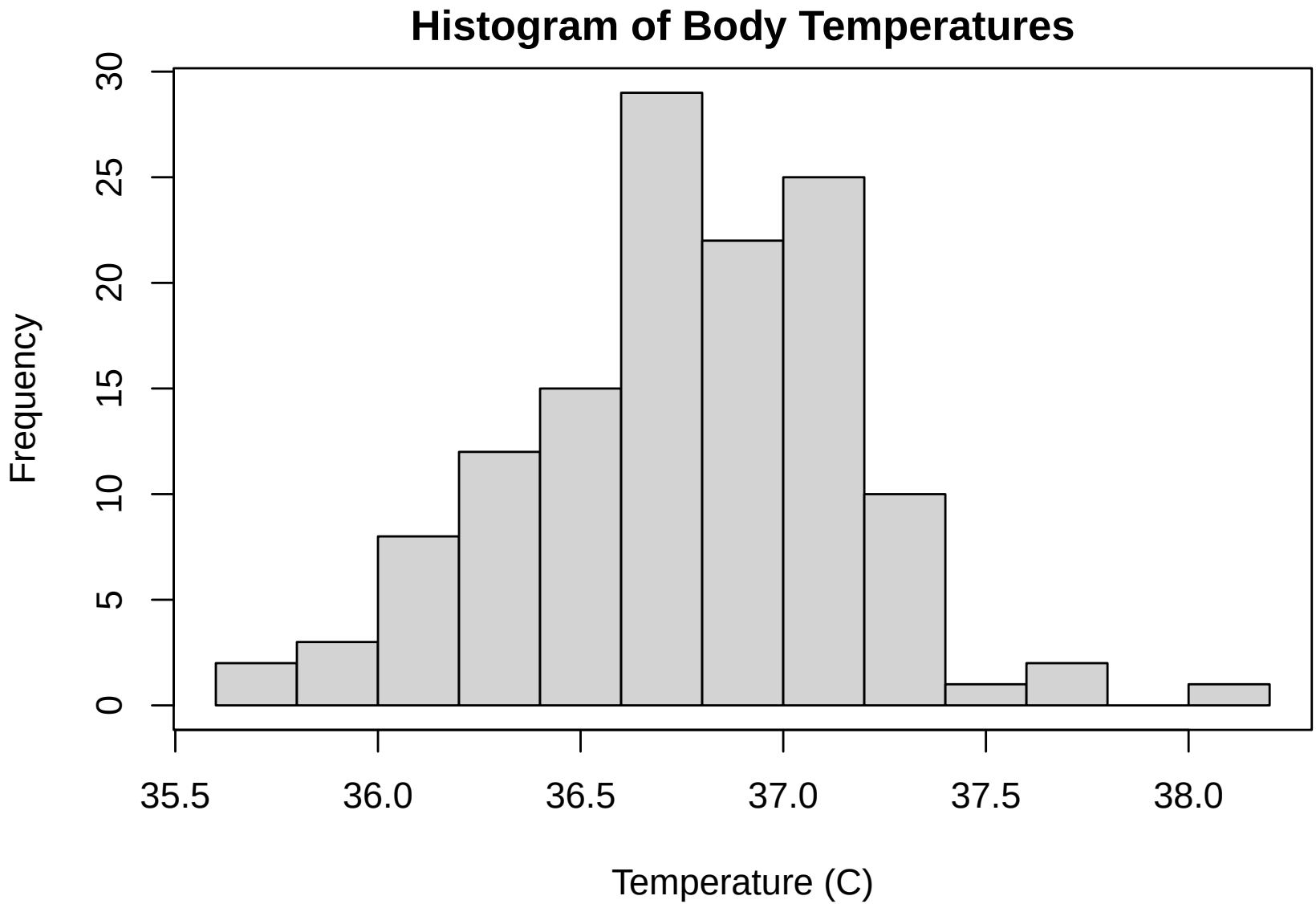
+   [1] 35.7 36.4 36.7 36.9 37.1 35.8 36.6 37.1 37.2 36.1 36.4 36.7 36.9 37.1 36.7
+  [16] 36.8 37.1 37.2 36.2 36.4 36.7 37.2 36.3 36.7 36.9 37.1 37.3 36.2 36.8 37.0
+  [31] 37.2 36.5 36.7 36.9 37.1 36.3 36.6 36.8 37.0 37.3 36.6 36.8 37.1 37.7 36.3
+  [46] 36.6 36.8 37.0 37.4 36.8 37.0 37.1 38.2 36.8 37.0 37.1 35.9 36.4 36.7 36.9
+  [61] 37.1 35.9 36.6 36.8 36.0 36.7 36.8 37.1 37.2 36.1 36.4 36.7 36.9 37.1 36.2
+  [76] 36.9 37.2 36.2 36.7 36.9 37.1 37.3 36.2 36.5 36.7 36.9 36.6 36.7 37.0 37.2
+  [91] 36.4 36.7 36.9 37.1 37.3 36.2 36.6 37.3 36.3 36.6 36.8 37.0 37.3 36.5 36.8
+ [106] 36.9 37.1 37.4 36.9 37.1 37.4 36.3 36.6 36.8 37.0 37.4 36.6 36.8 37.0 36.6
+ [121] 36.8 37.0 37.1 37.8 36.3 36.7 36.8 37.1 37.5 36.6
```

Histograms

Example: Body temperatures

```
par(mar=c(4,4,2,1))
hist(BodyTemp,xlab='Temperature (C)',main='Histogram of Body Temperatures')
box()
```

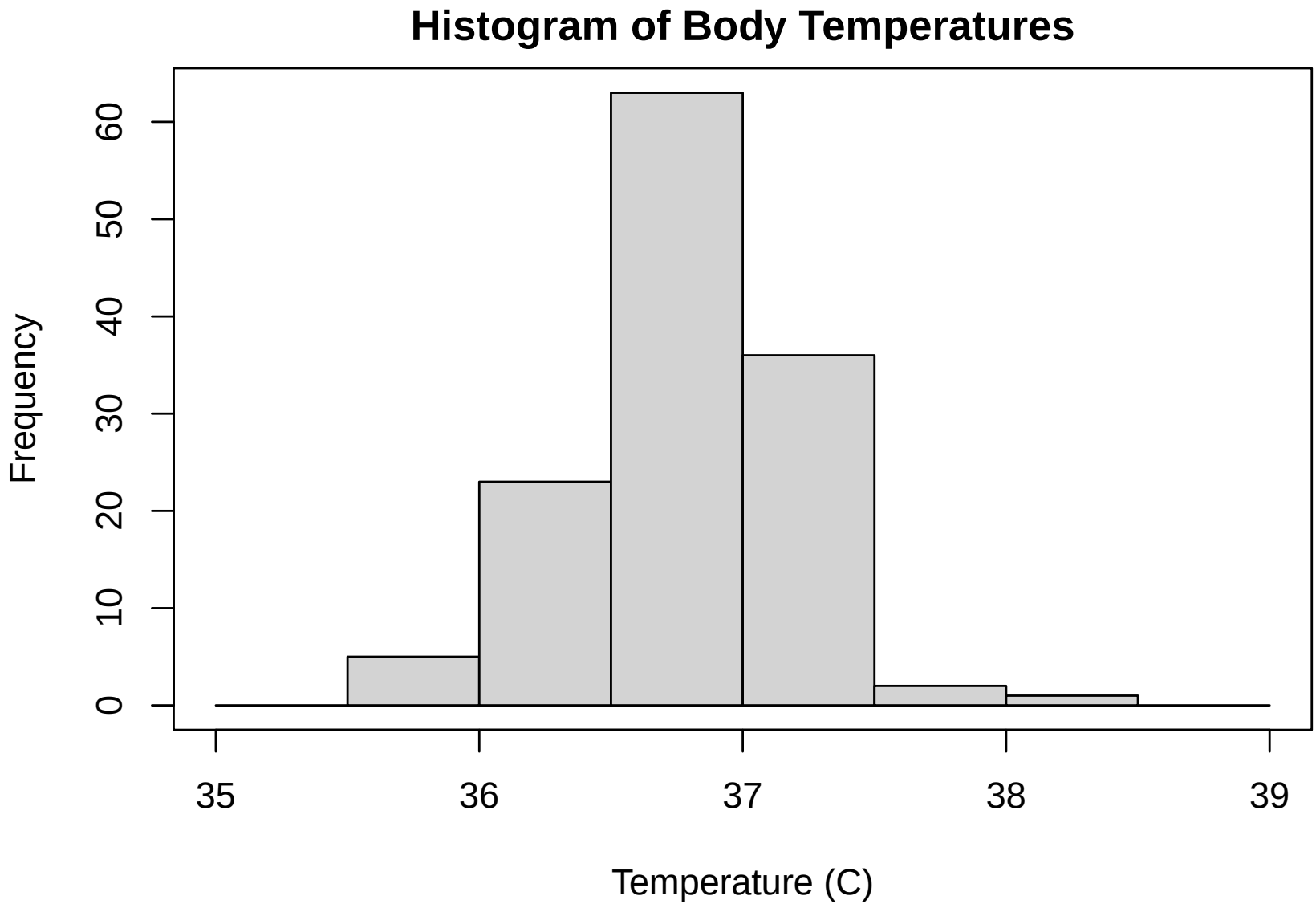
Histograms


Histogram of Body Temperatures
Frequency
0
5
10
15
20
25
30
35.5
36.0
36.5
37.0
37.5
38.0
Temperature (C)

Histograms

```
par(mar=c(4,4,2,1));
hist(BodyTemp,breaks=seq(35,39,by=0.5),
        xlab='Temperature (C)',main='Histogram of Body Temperatures');box()
```

Histograms


Histogram of Body Temperatures
Frequency
0
10
20
30
40
50
60
35
36
37
38
39
Temperature (C)

Histograms

1. Order the observations from lowest to highest
2. Let x_S be the smallest value and x_L be the largest value
3. Compute the difference $x_L - x_S$
4. Decide on a number of “classes” (intervals) k into which you want to divide your observations
   - k should not be “too small" or “too large”,
   - This is "vague" for the moment. We will come back to this point later
5. Compute $l = (x_{\mathrm{L}} - x_{\mathrm{S}})/k$, which will be the length of each interval.
6. Starting with x_S, form k subintervals of length l, one after another

Histograms

7. The intervals might be specified to be

$$[x_S, x_S + l), [x_S + l, x_S + 2l), \ldots, [x_L - l, x_L]$$

where the notation for intervals is that

$$[a, b) \text{ includes } a \text{ but not } b$$

and

$$[a, b] \text{ includes both } a \text{ and } b.$$

The first $(k - 1)$ intervals are closed on the left and open on the right; the last interval is closed on the left and on the right

Histograms

8. For each interval plot a rectangle with height equal to either
   (i) the number of observations in the data that fall in the interval (the frequency distribution), or
   (ii) the proportion (out of n, the total number of observations) of observations in the data that fall in the interval (the relative frequency distribution).

Histograms

Construction: $n = 40$:

0.4	2.8	0.8	1.4	6.4	1.3	1.8	0.1
0.0	1.6	1.2	1.4	0.4	1.5	3.9	2.0
0.3	3.8	0.1	0.6	0.4	0.3	0.7	0.5
0.1	1.2	0.1	0.7	1.3	1.3	0.6	0.3
0.2	1.6	1.6	1.1	0.6	2.1	0.4	0.0

Histograms

Sorted Data:

0.0	0.0	0.1	0.1	0.1	0.1	0.2	0.3
0.3	0.3	0.4	0.4	0.4	0.4	0.5	0.6
0.6	0.6	0.7	0.7	0.8	1.1	1.2	1.2
1.3	1.3	1.3	1.4	1.4	1.5	1.6	1.6
1.6	1.8	2.0	2.1	2.8	3.8	3.9	6.4

so $x_S = 0$ and $x_L = 6.4$

Histograms

Let us use $k = 8$: the width of each interval is

$$l = (6.4 - 0)/8 = 0.8,$$

and the $k = 8$ intervals are:

$$[0, 0.8), [0.8, 1.6), [1.6, 2.4), [2.4, 3.2), [3.2, 4), [4, 4.8), [4.8, 5.6), [5.6, 6.4]$$

Histograms

We now count how many observations fall in each interval:

Interval	Frequency	Rel. frequency
$[0, 0.8)$	20	0.5
$[0.8, 1.6)$	10	0.25
$[1.6, 2.4)$	6	0.15
$[2.4, 3.2)$	1	0.025
$[3.2, 4)$	2	0.05
$[4, 4.8)$	0	0
$[4.8, 5.6)$	0	0
$[5.6, 6.4]$	1	0.025

Histograms

```
x<-c(0.4,2.8,0.8,1.4,6.4,1.3,1.8,0.1,0.0,1.6,1.2,1.4,0.4,1.5,3.9,2.0,
0.3,3.8,0.1,0.6,0.4,0.3,0.7,0.5,0.1,1.2,0.1,0.7,1.3,1.3,0.6,0.3,
0.2,1.6,1.6,1.1,0.6,2.1,0.4,0.0)
y<-sort(x)
y

+  [1] 0.0 0.0 0.1 0.1 0.1 0.1 0.2 0.3 0.3 0.3 0.4 0.4 0.4 0.4 0.5 0.6 0.6 0.6 0.7
+ [20] 0.7 0.8 1.1 1.2 1.2 1.3 1.3 1.3 1.4 1.4 1.5 1.6 1.6 1.6 1.8 2.0 2.1 2.8 3.8
+ [39] 3.9 6.4

x.S<-min(x)
x.L<-max(x)

k<-8
l<-(x.L-x.S)/k
l

+ [1] 0.8
```

Histograms

```
xv<-seq(x.S,x.L,by=l)
table(cut(x,xv,right=FALSE,include.lowest=TRUE))
```

```
+
+   [0,0.8) [0.8,1.6) [1.6,2.4) [2.4,3.2)   [3.2,4)   [4,4.8) [4.8,5.6) [5.6,6.4]
+        20        10         6         1         2         0         0         1
```

```
hist(x,breaks=xv,right=FALSE,include.lowest=TRUE,plot=FALSE)$counts
```

```
+ [1] 20 10  6  1  2  0  0  1
```

Other “binning” options may also be used:

```
hist(x,breaks=xv,plot=FALSE)$counts
```

```
+ [1] 21 12  3  1  2  0  0  1
```

Histograms

From the R `help(hist)`:

If `right = TRUE` (default), the histogram cells are intervals of the form $(a, b]$, i.e., they include their right-hand endpoint, but not their left one, with the exception of the first cell when `include.lowest` is `TRUE`.

For `right = FALSE`, the intervals are of the form $[a, b)$, and `include.lowest` means 'include highest'.

Histograms

The number of intervals k is relatively important

- If k is too small then the histogram will convey little information.
- If k is too large then the information will be too detailed to convey the shape of the distribution.

However, there is no 'incorrect' choice.

Histograms

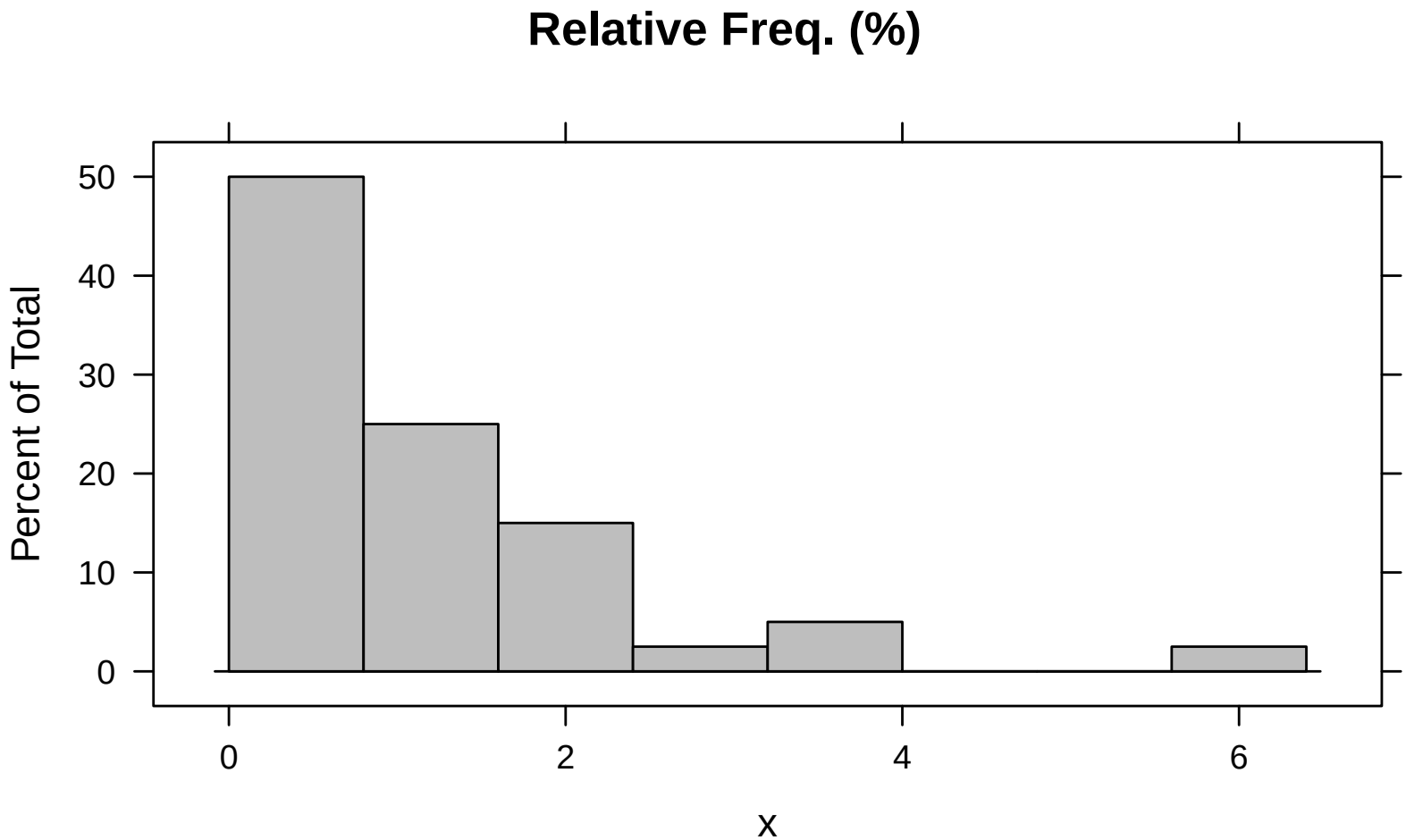
When inspecting or constructing a histogram:

- Take note of vertical and horizontal scales: is the vertical scale measuring
  - ‘Frequency’ ?
  - ‘Relative Frequency’ ?
  - ’Density’ ?
- Label the horizontal axis in an informative way.

Histograms

```
k<-8;l<-(x.L-x.S)/k
xv<-seq(x.S,x.L,by=l);par(mar=c(4,4,2,0))
lattice::histogram(x,breaks=xv,
            right=FALSE,include.lowest=TRUE,
            main='Relative Freq. (%)', col='gray')
```

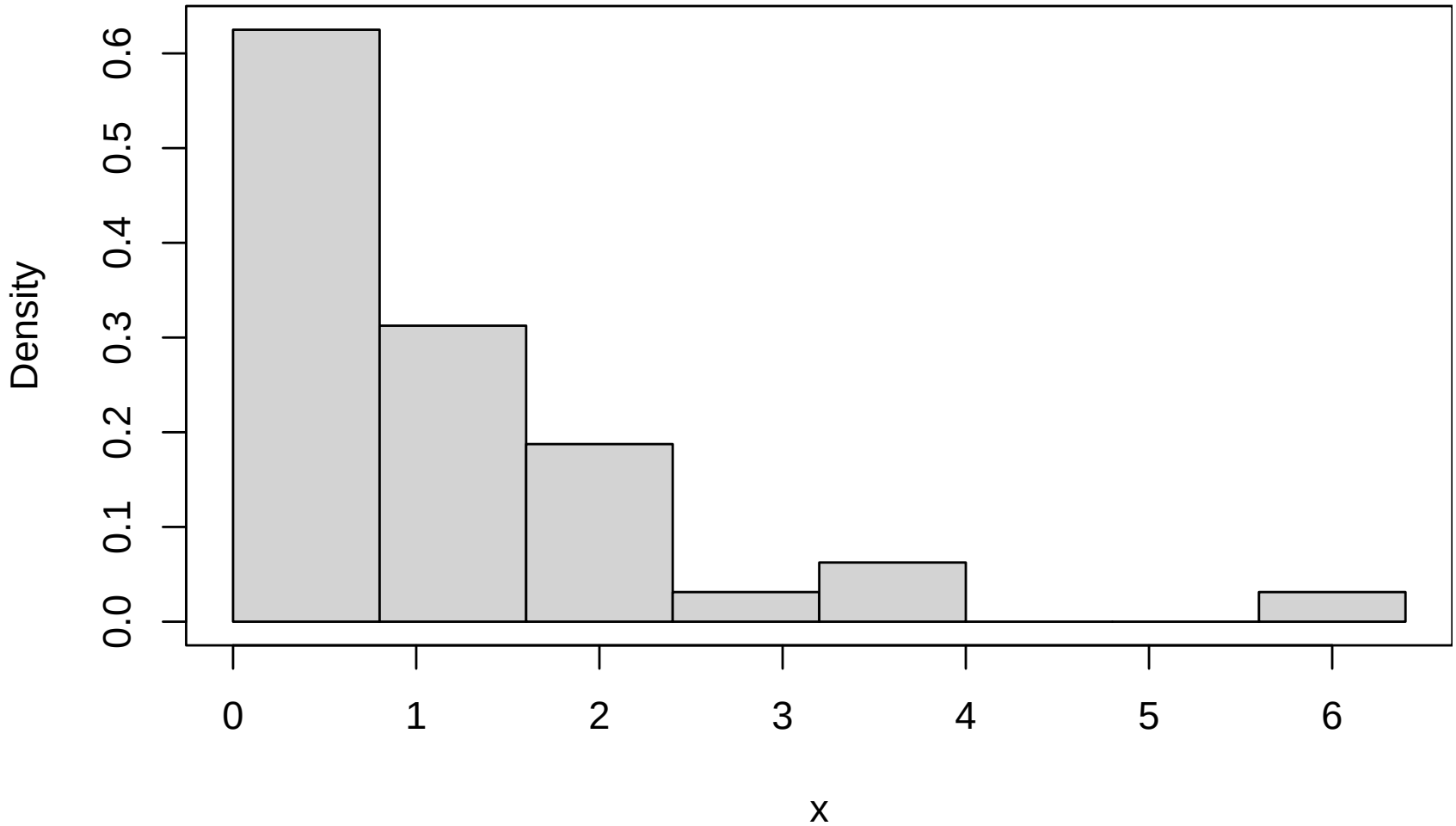
Histograms



Histograms

```
k<-8;l<-(x.L-x.S)/k
xv<-seq(x.S,x.L,by=l);par(mar=c(4,4,1,0))
hist(x,breaks=xv,right=FALSE,include.lowest=TRUE,main='',freq=FALSE);box()
```

Histograms



Histograms

```
h1<-hist(x,breaks=xv,right=FALSE,include.lowest=TRUE,plot=FALSE)

h1$density

+ [1] 0.62500 0.31250 0.18750 0.03125 0.06250 0.00000 0.00000 0.03125

h1$counts/(l*sum(h1$counts))

+ [1] 0.62500 0.31250 0.18750 0.03125 0.06250 0.00000 0.00000 0.03125
```

$$\text{Density in interval } k \ = \frac{1}{l} \times \frac{\text{Count in interval } k}{n}$$

Histograms

Choosing k:

- There are several guidelines for deciding k, e. g.

$$k = \left\lceil \frac{10}{3} \log_{10}(n) \right\rceil + 1$$

  - $\log_{10}$ is the logarithm to base 10
  - n is the total number of observations
  - k is often rounded up to the next integer

For $n = 40$

$$k = \left\lceil \frac{10}{3} \log_{10}(40) \right\rceil + 1 \approx 6.34$$

indicating $k = 6$ or $k = 7$ would be reasonable choices.

Histograms

Example: Body temperatures

```
BodyTemp<-scan('https://www.math.mcgill.ca/dstephens/203/Data/BodyTemp.txt')
n<-length(BodyTemp)
k<-(10/3)*log10(n)+1
c(n,k)

+ [1] 130.000000   8.046478

k0<-ceiling(k)
k0

+ [1] 9

k1<-round(k)
k1

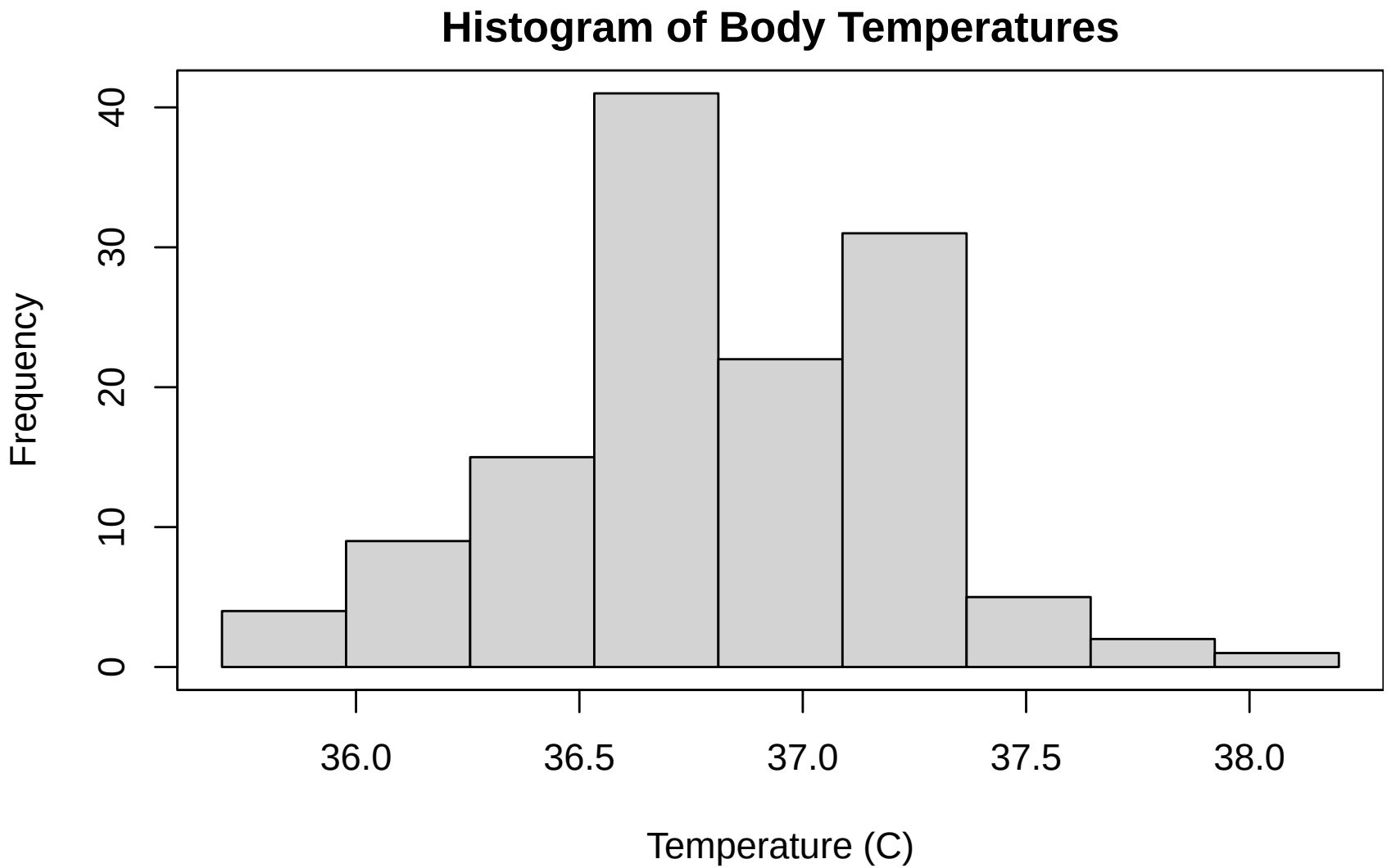
+ [1] 8
```

Histograms

Example: Body temperatures

```
x.S<-min(BodyTemp)
x.L<-max(BodyTemp)
k<-9
l<-(x.L-x.S)/k
xv<-seq(x.S,x.L,by=l);par(mar=c(4,4,2,0))
hist(BodyTemp,breaks=xv,right=FALSE,include.lowest=TRUE,
        xlab='Temperature (C)', main='Histogram of Body Temperatures')
box()
```

Histograms


Histogram of Body Temperatures
Frequency
0
10
20
30
40
36.0
36.5
37.0
37.5
38.0
Temperature (C)

Histograms

Here the histogram is adequate, but the intervals are unintuitive:

```
xv

+  [1] 35.70000 35.97778 36.25556 36.53333 36.81111 37.08889 37.36667 37.64444
+  [9] 37.92222 38.20000
```

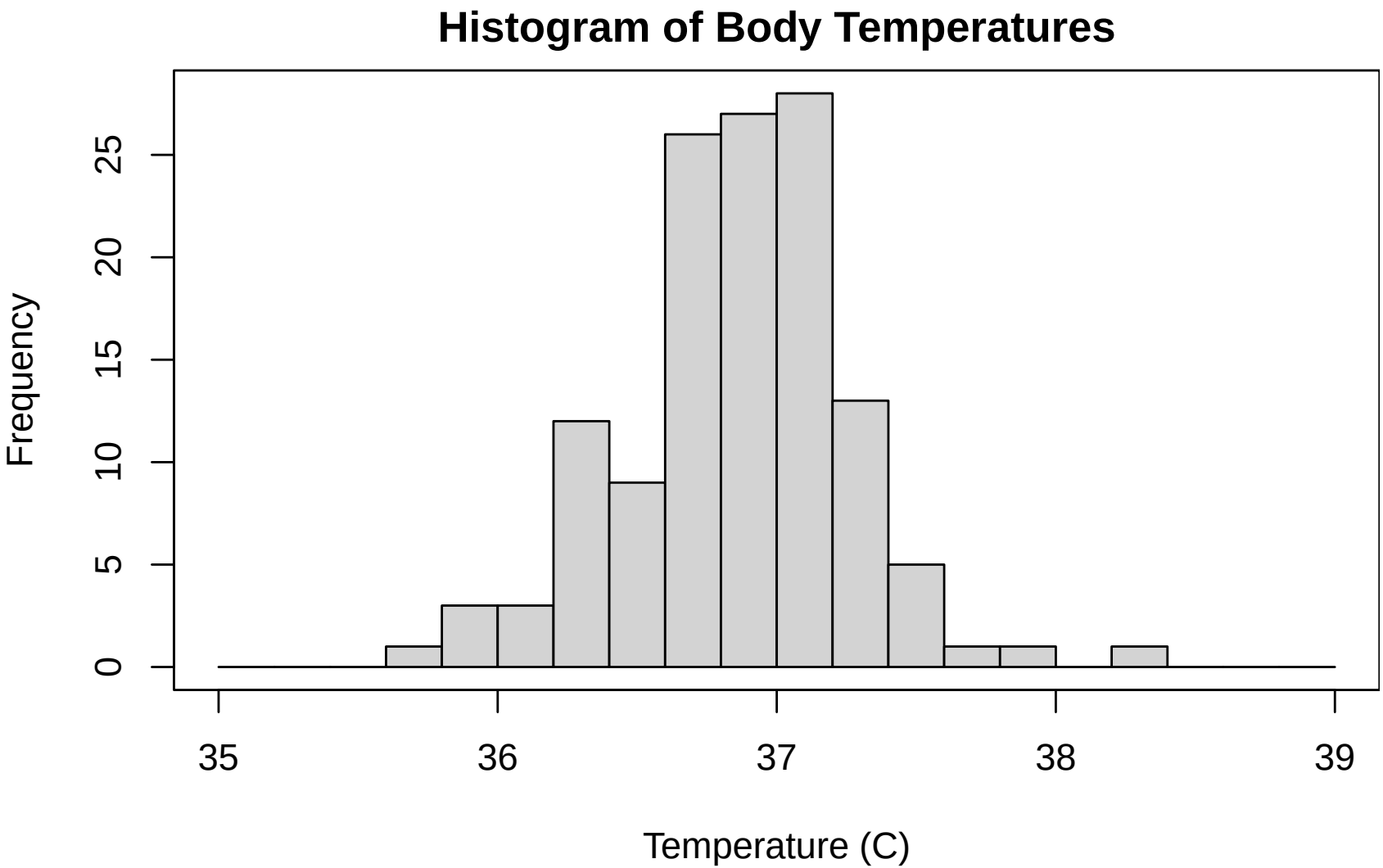
It is sometimes preferable to use more interpretable intervals.

Histograms

Example: Body temperatures

```
xv<-seq(35,39,by=0.2);par(mar=c(4,4,2,0))
hist(BodyTemp,breaks=xv,right=FALSE,include.lowest=TRUE,
        xlab='Temperature (C)', main='Histogram of Body Temperatures')
box()
```

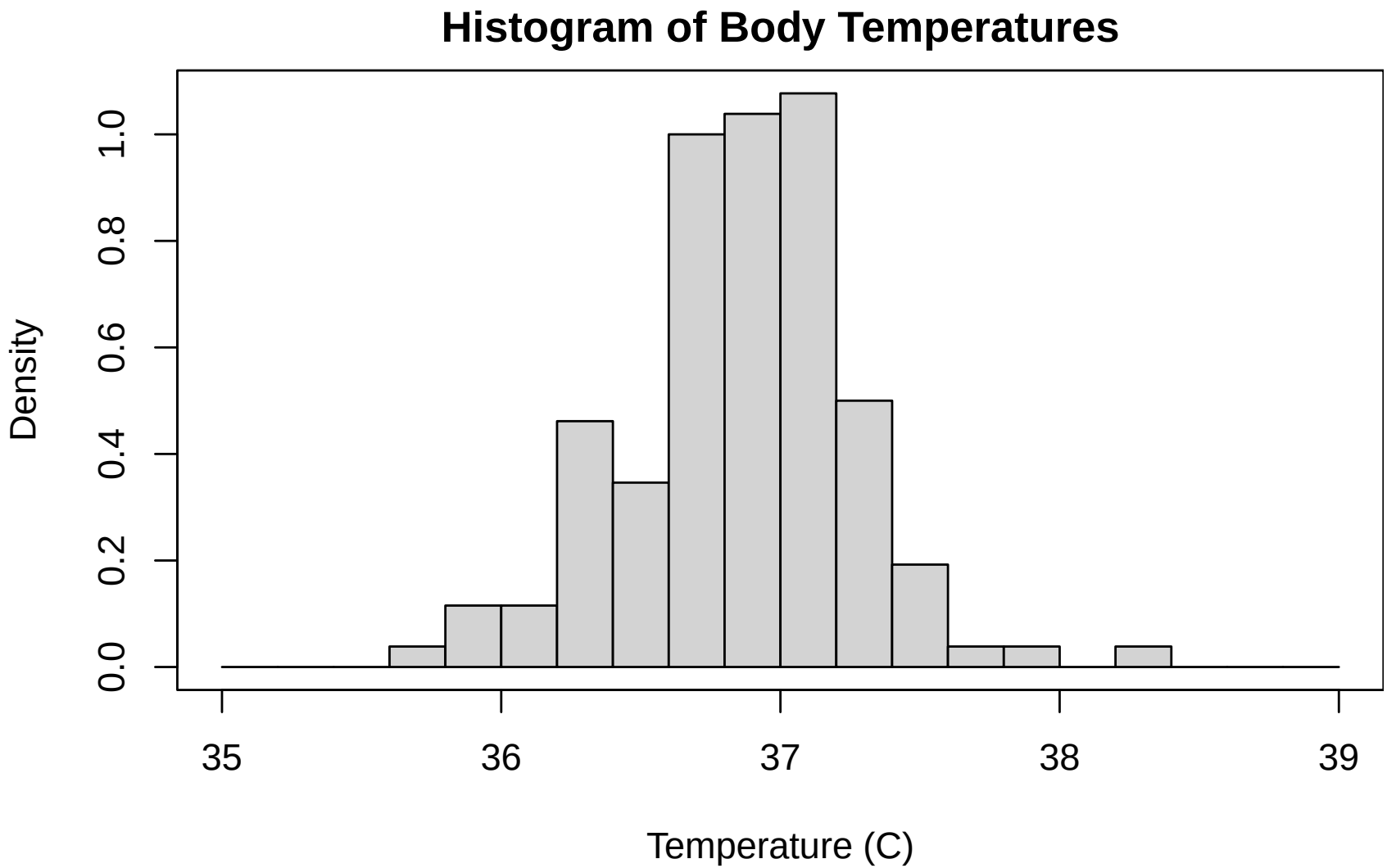
Histograms


Histogram of Body Temperatures
Frequency
0
5
10
15
20
25
35
36
37
38
39
Temperature (C)

Histograms

```
xv<-seq(35,39,by=0.2);par(mar=c(4,4,2,0))
hist(BodyTemp,breaks=xv,right=FALSE,include.lowest=TRUE,freq=FALSE,
        xlab='Temperature (C)', main='Histogram of Body Temperatures')
box()
```

Histograms


Histogram of Body Temperatures
Density
0.0
0.2
0.4
0.6
0.8
1.0
35
36
37
38
39
Temperature (C)

Histograms

Example: Simulated example

```
set.seed(101)
n<-1000
x<-rnorm(n,10,2)
x.S<-min(x)
x.L<-max(x)
c(x.S,x.L)

+ [1]  3.645581 16.356979

l<-(x.L-x.S)/k
xv<-seq(x.S,x.L,by=l)
k<-ceiling((10/3)*log10(n)+1)
c(n,k)

+ [1] 1000   11

xv

+  [1]  3.645581  5.057958  6.470336  7.882713  9.295091 10.707468 12.119846
+  [8] 13.532223 14.944601 16.356979
```

Histograms

Example: Simulated example

```
x.S<-floor(min(x))
x.L<-ceiling(max(x))
c(x.S,x.L,x.L-x.S)

+ [1]  3 17 14

xv<-seq(x.S,x.L,by=1)
xv

+  [1]  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17
```

Histograms

Example: Simulated example

```
xv<-seq(x.S,x.L,by=0.5)
xv

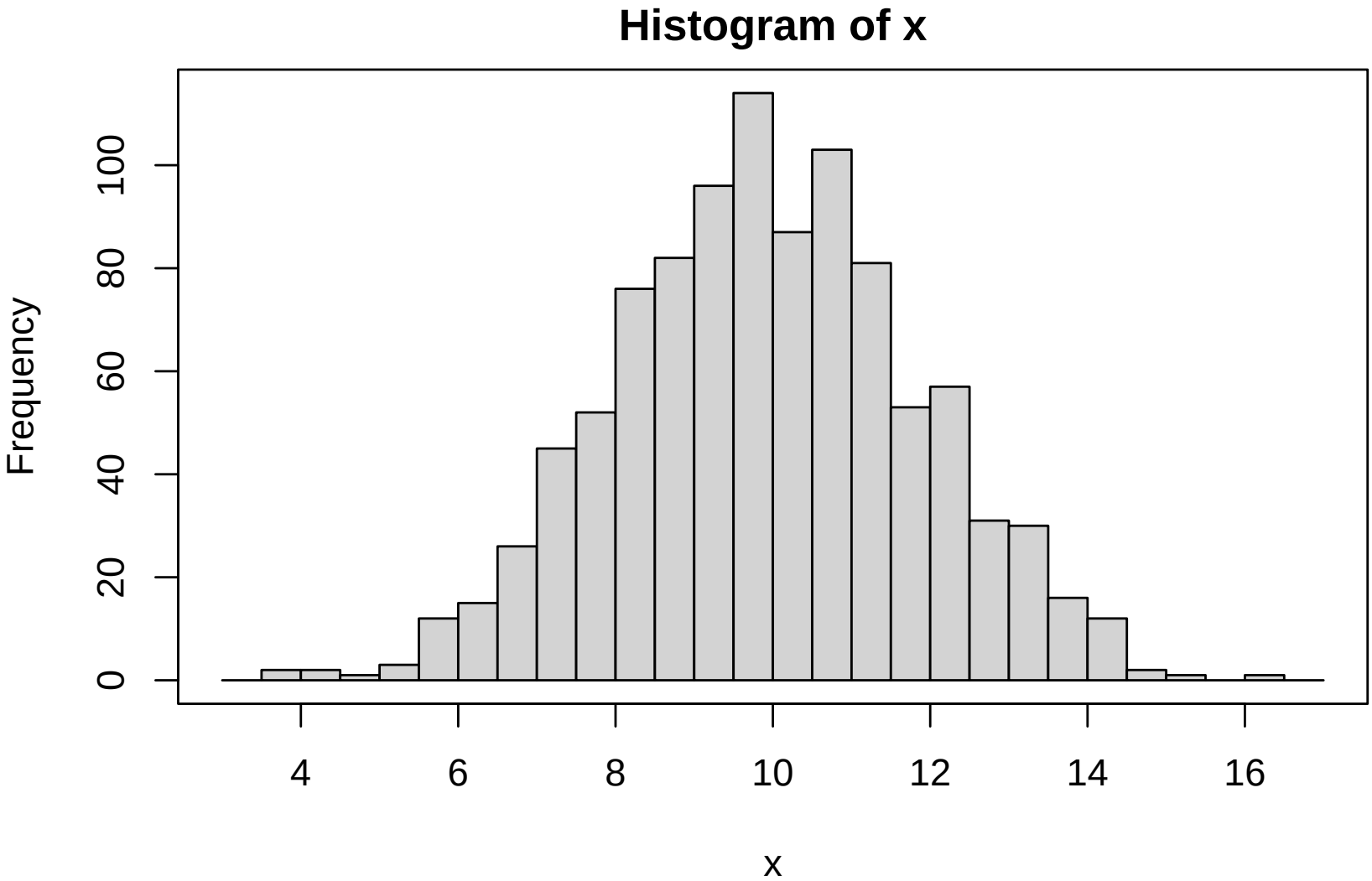

+  [1]  3.0  3.5  4.0  4.5  5.0  5.5  6.0  6.5  7.0  7.5  8.0  8.5  9.0  9.5 10.0
+ [16] 10.5 11.0 11.5 12.0 12.5 13.0 13.5 14.0 14.5 15.0 15.5 16.0 16.5 17.0
```

Histograms

Example: Simulated example

```
par(mar=c(4,4,2,1))
hist(x,breaks=xv,right=FALSE,include.lowest=TRUE,main='Histogram of x');box()
```

Histograms



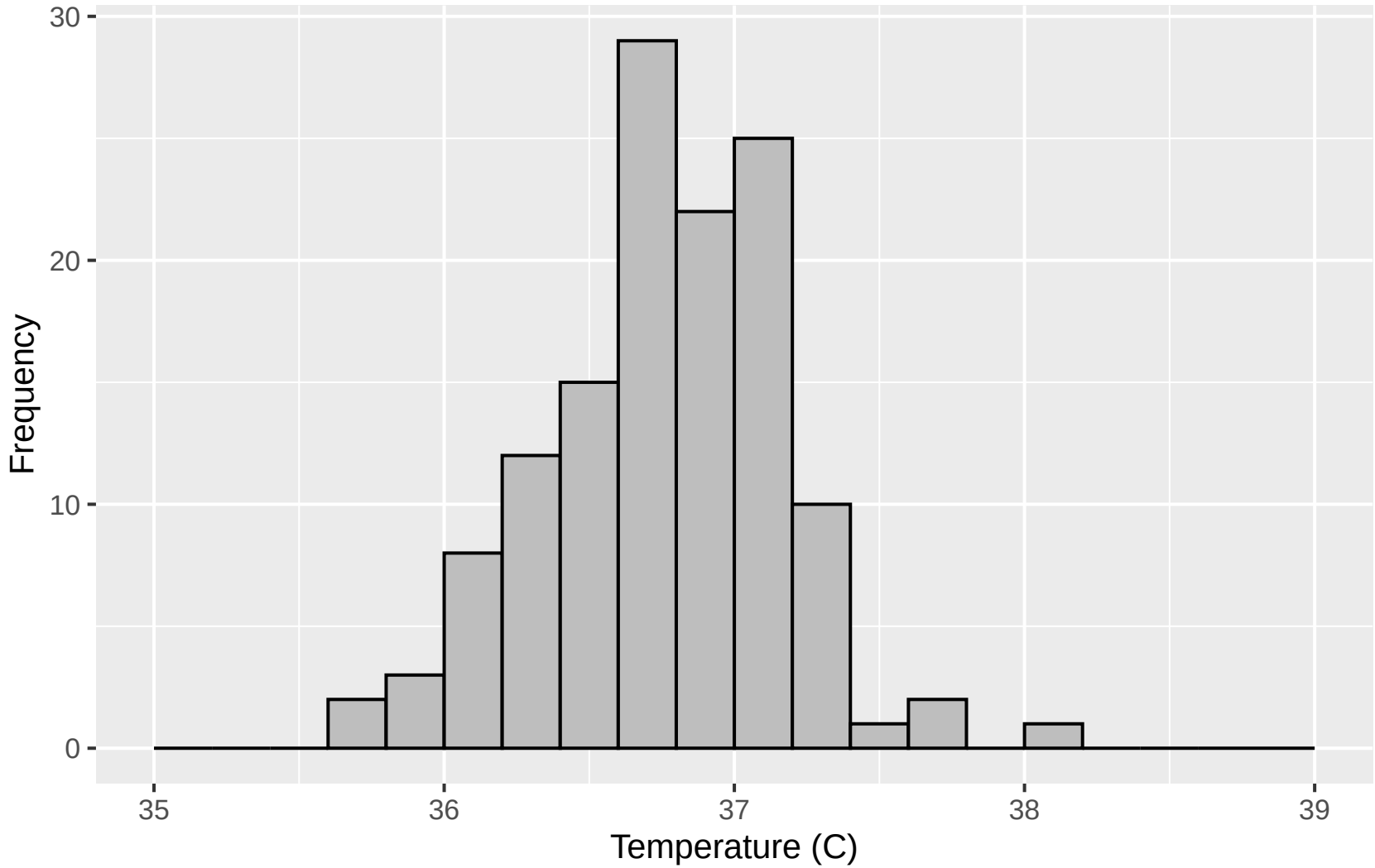
Histograms

Example: Body temperatures

Using `ggplot2`

```
xv<-seq(35,39,by=0.2);par(mar=c(4,4,2,0))
df<-data.frame(BodyTemp)
ggplot(df, aes(x=BodyTemp)) +
  geom_histogram(fill='gray',color='black',breaks=xv) +
  ylab('Frequency') + xlab('Temperature (C)')
```

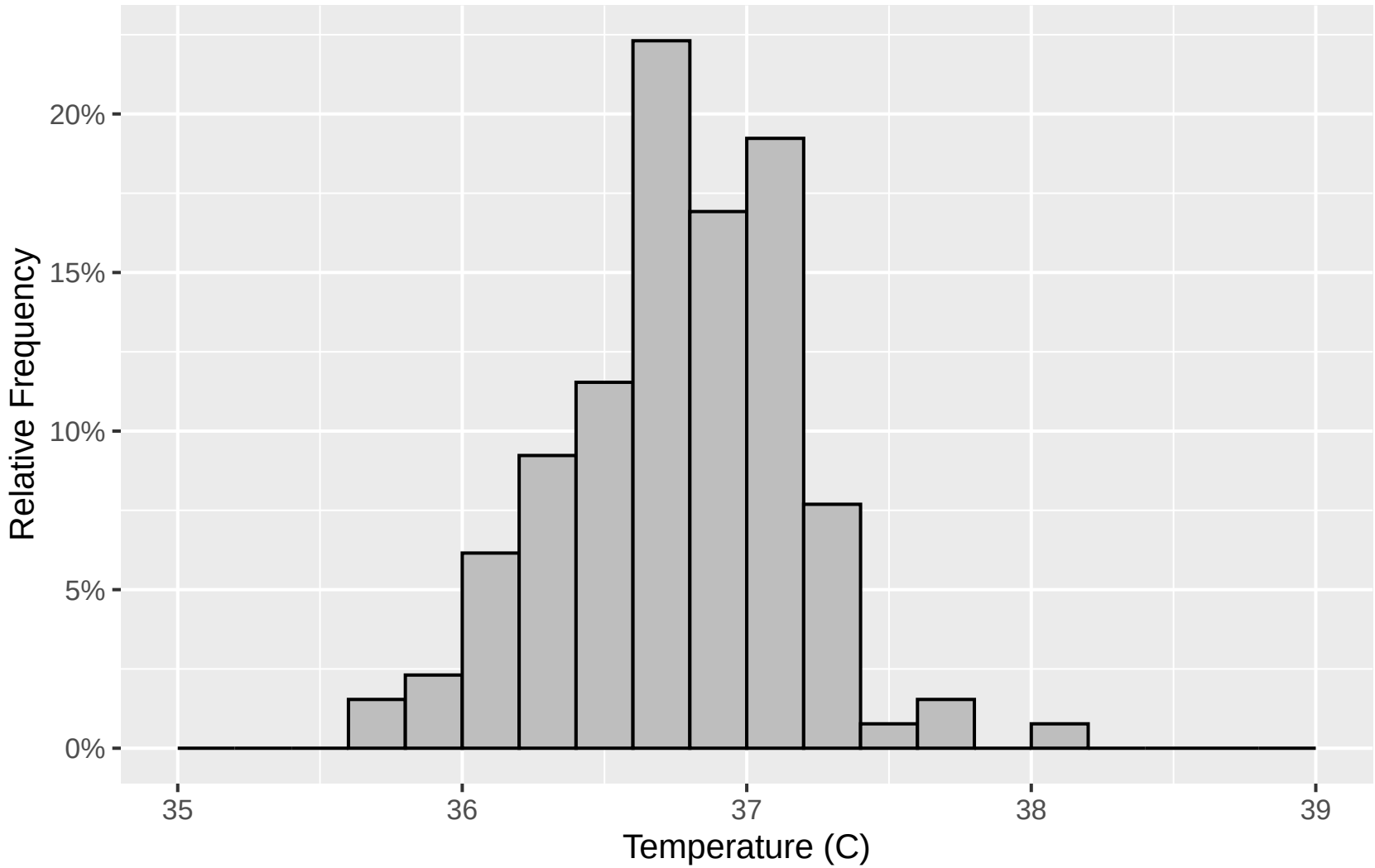
Histograms



Histograms

```
ggplot(df, aes(x=BodyTemp)) +
  geom_histogram(aes(y = after_stat(count/sum(count))), fill='gray',color='black',breaks=xv) +
  scale_y_continuous(labels = scales::percent) +
  ylab('Relative Frequency') + xlab('Temperature (C)')
```

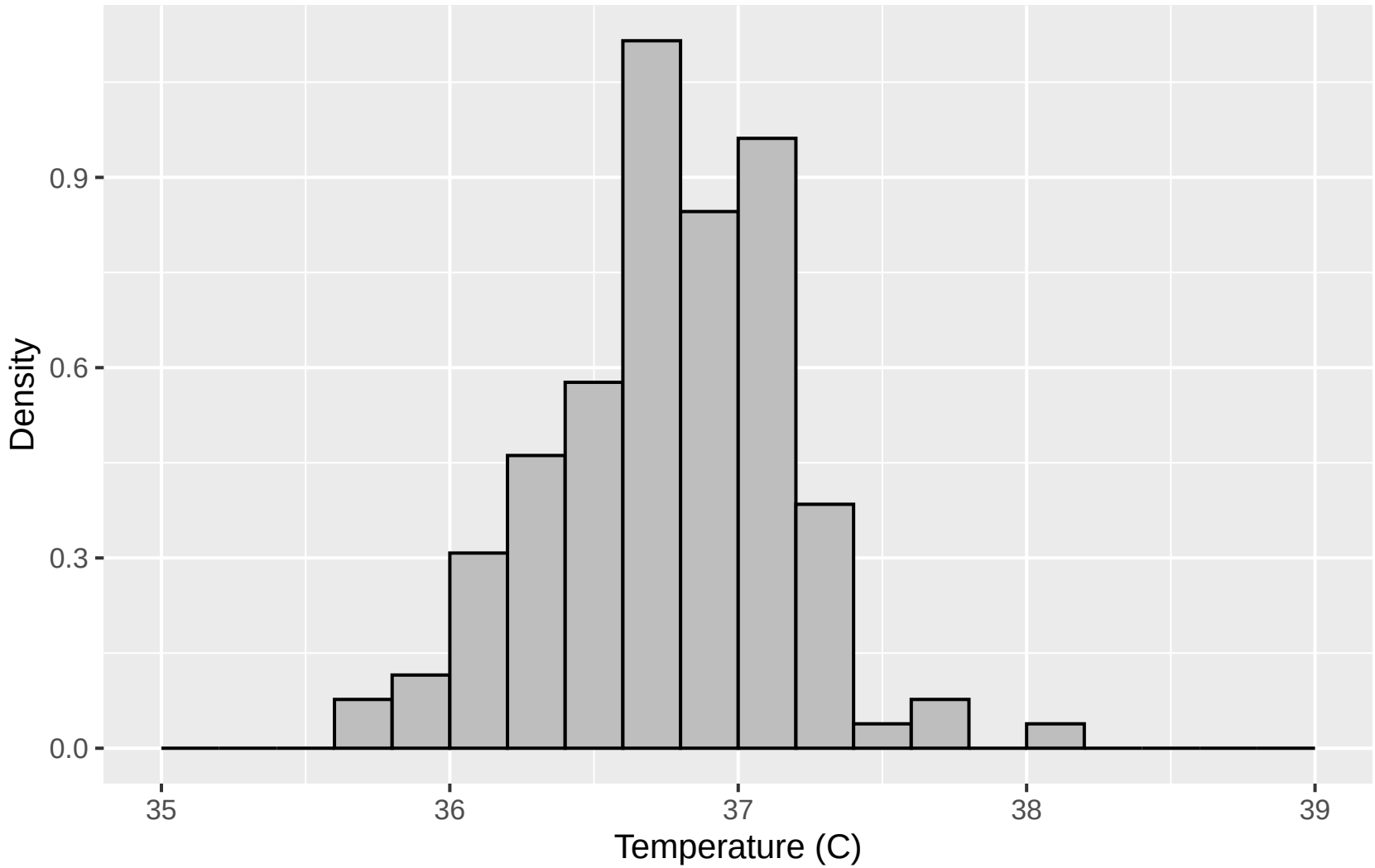
Histograms



Histograms

```
ggplot(df, aes(x=BodyTemp)) +
  geom_histogram(aes(y=after_stat(density)), fill='gray',color='black',breaks=xv) +
  ylab('Density') + xlab('Temperature (C)')
```

Histograms



Interpreting Histograms

We can use a histogram to deduce the number (or the proportion) of observations in a given interval

- This is done by adding up the heights of the rectangles within an interval
- It becomes trickier if only part of a rectangle is included in the interval
- We know nothing about the distribution of the observations within the rectangle itself

Interpreting Histograms

We have to make one of these assumptions:

- All of the observations in the rectangle are included in the interval
  - add the entire rectangle height
- None of the observations in the rectangle are included in the interval
  - do not add the rectangle at all
- Some fraction of the observations in the rectangle are included in the interval
  - add the same fraction of the rectangle height.

Numerical methods for quantitative data

Numerical measures of central tendency try to capture where the data are situated

- Mean (or average)
- Median
- Mode

Sample mean

Suppose $x_1, x_2, \ldots, x_n$ be n data values.

The sample mean is denoted by $\bar{x}$ and is defined as:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

- $\bar{x}$ is also referred to as the sample average
- we may use the summation notation

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Sample mean

Example:

Suppose that $n = 5$, and our observations are

$$x_1 = 3, x_2 = 5, x_3 = 5, x_4 = 6, x_5 = 8$$

Then

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5}$$

$$= \frac{3 + 5 + 5 + 6 + 8}{5} = 5.4$$

Sample mean

Example:

Suppose that $n = 4$, and our observations are $x_1 = 3, x_2 = 1, x_3 = 4$, and $x_4 = 100$. Then:

$$\bar{x} = \frac{3 + 1 + 4 + 100}{4} = \frac{108}{4} = 27$$

Note that $\bar{x}$ is pulled away from the smaller values towards the larger value

Sample mean

Example:

Suppose that $n = 8$, and our observations are

$$132.3, 130.7, 134.1, 5.3, 133.9, 131.3, 132.0, 133.3.$$

Then

$$\bar{x} = \frac{132.3 + 130.7 + 134.1 + 5.3 + 133.9 + 131.3 + 132.0 + 133.3}{8}$$

$$= 116.6125$$

Note that $\bar{x}$ is pulled away from the larger values towards the smaller value

Sample mean

Extreme observations in a sample are often called outliers

- outliers distort the sample mean as a measure of central tendency
- take care when outliers are present when interpreting such averages
- the sample mean may not be a good measure of central tendency in the presence of outliers

Sample mean

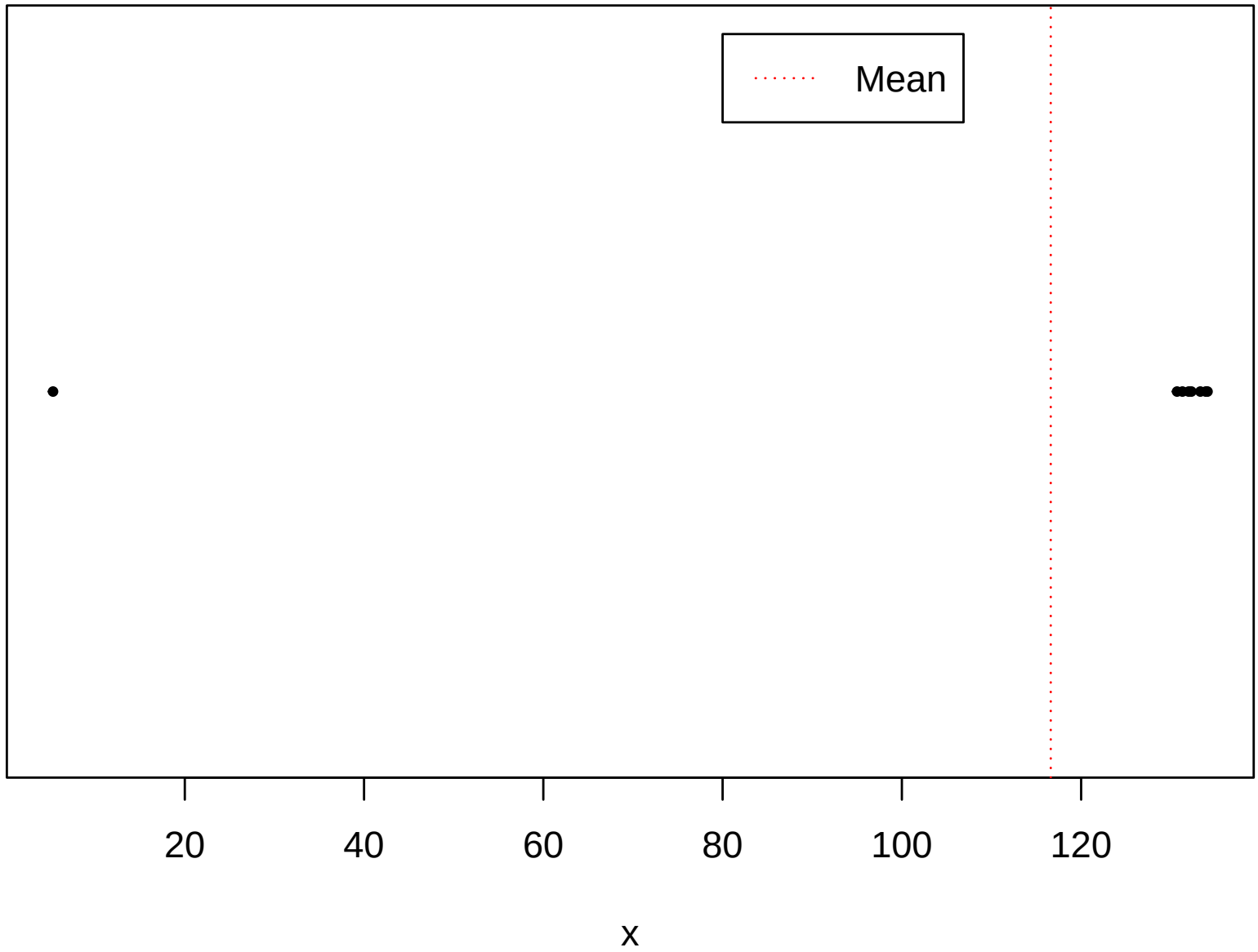
Computing in R:

```
x<-c(132.3, 130.7, 134.1, 5.3, 133.9,131.3,132.0,133.3)
mean(x)

+ [1] 116.6125

par(mar=c(4,2,1,1))
plot(x,rep(0,8),pch=19,cex=0.5,yaxt='n',ylab='')
abline(v=mean(x),col='red',lty=3)
legend(80,1,c('Mean'),lty=3,col='red')
```

Sample mean



Sample median

Let $x_1, x_2, \ldots, x_n$ be a set of n observed values.

The sample median, denoted by m, is the number such that, if we ordered these numbers from smallest to largest then m would be in the "middle"

- 50% of the data points in a sample are below the median and 50% are above the median

Sample median

Suppose that $n = 5$, and the observations are

$$2, 1, 3, 5, 4$$

If we order these observations, we get:

$$1 \quad 2 \quad 3 \quad 4 \quad 5$$

So the median is $m = 3$

Sample median

Suppose that $n = 6$ and we have

$$7, 6, 7, 5, 3, 4.$$

In order:

$$3 \quad 4 \quad 5 \quad 6 \quad 7 \quad 7$$

We may define the median as $m = (5 + 6)/2 = 5.5$

Sample median

In general, in the ordered list of n values in a sample:

- If n is odd, the sample median m is the number in the $(n+1)/2$th position
- If n is even, m is the average of the number in the $(n/2)$th position and the number in the $(n/2+1)$th position

Sample median

The sample median is not as influenced by outliers as the sample mean.

- Suppose that $n = 5$, and our observations are

$$1, 5, 2, 9, 8.$$

After ordering:

$$1 \quad 2 \quad 5 \quad 8 \quad 9$$

The median is $m = 5$

Sample median

- Now suppose that the largest observation was changed from 9 to 900. After ordering

$$1 \quad 2 \quad 5 \quad 8 \quad 900$$

the median is still $m = 5$.

Sample median

Computing in R:

```
x<-c(132.3, 130.7, 134.1, 5.3, 133.9,131.3,132.0,133.3)
sort(x)

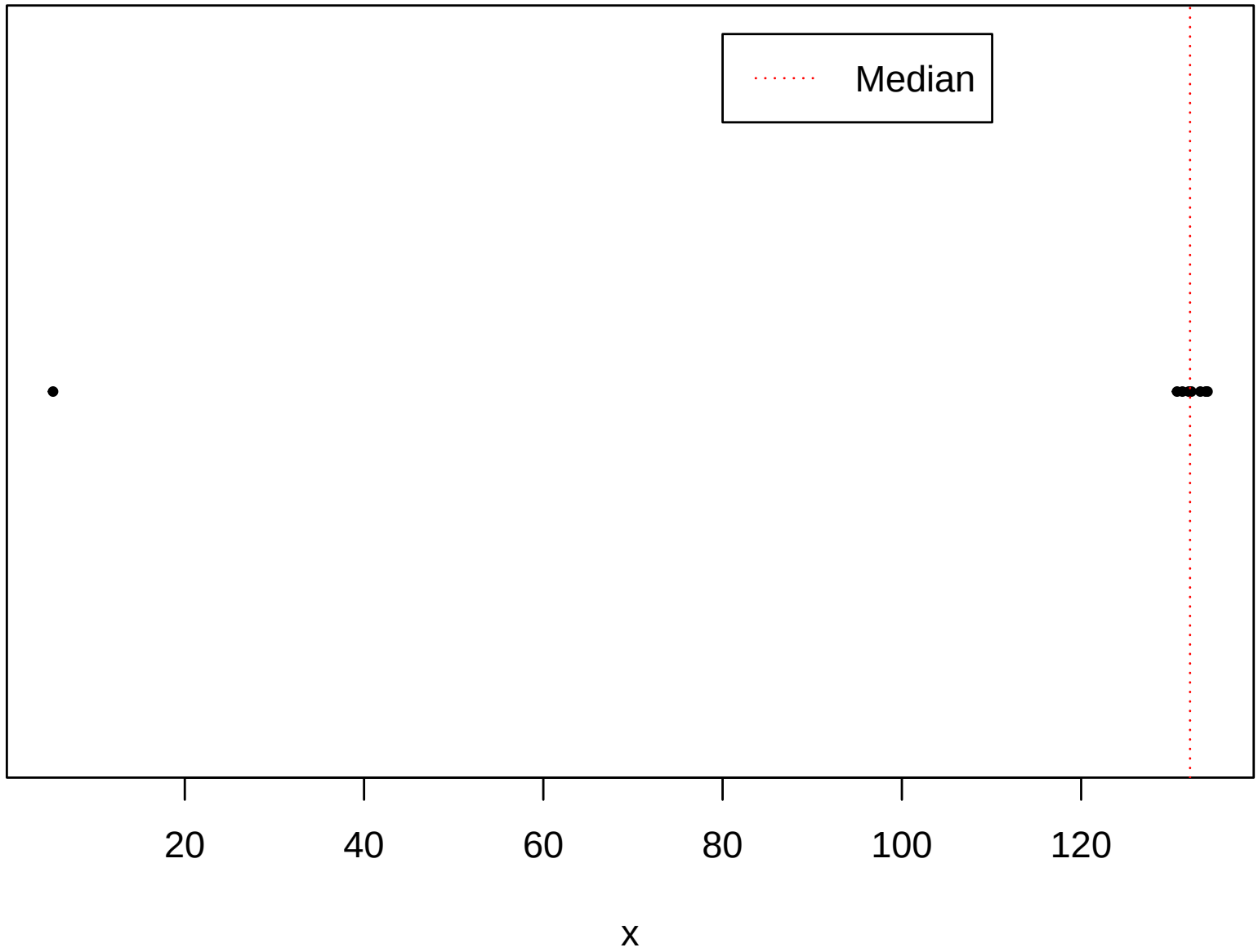
+ [1]   5.3 130.7 131.3 132.0 132.3 133.3 133.9 134.1

median(x)

+ [1] 132.15

par(mar=c(4,2,1,1))
plot(x,rep(0,8),pch=19,cex=0.5,yaxt='n',ylab='')
abline(v=median(x),col='red',lty=3)
legend(80,1,c('Median'),lty=3,col='red')
```

Sample median



Sample mode

The sample mode is the most frequently observed value.

If all observations occur exactly once, then

- every value is a mode, or
- none of the values is the mode.

A set of observations can be unimodal or multimodal.

Sample mode

- If the data points are

$$1, 5, 2, 6, 9, 8.$$

  This data set does not have any repeated observations; every data point is a mode (or there is no mode).

- If the data points are

$$1, 5, 2, 6, 2, 9, 8.$$

  The number 2 occurs twice, and is the mode.

- If the data points are

$$8, 1, 5, 2, 6, 8, 9, 8, 2.$$

  The number 8 occurs three times, and is the mode.

Sample mode

- If the data points are

$$8, 1, 5, 2, 6, 2, 9, 8.$$

The numbers 8 and 2 are each repeated twice, and are both modes; the data are bimodal.

Sample mode

Example: Body temperatures ($n = 130$)

```
BodyTemp<-scan('https://www.math.mcgill.ca/dstephens/203/Data/BodyTemp.txt')
BodyTemp

+   [1] 35.7 36.4 36.7 36.9 37.1 35.8 36.6 37.1 37.2 36.1 36.4 36.7 36.9 37.1 36.7
+  [16] 36.8 37.1 37.2 36.2 36.4 36.7 37.2 36.3 36.7 36.9 37.1 37.3 36.2 36.8 37.0
+  [31] 37.2 36.5 36.7 36.9 37.1 36.3 36.6 36.8 37.0 37.3 36.6 36.8 37.1 37.7 36.3
+  [46] 36.6 36.8 37.0 37.4 36.8 37.0 37.1 38.2 36.8 37.0 37.1 35.9 36.4 36.7 36.9
+  [61] 37.1 35.9 36.6 36.8 36.0 36.7 36.8 37.1 37.2 36.1 36.4 36.7 36.9 37.1 36.2
+  [76] 36.9 37.2 36.2 36.7 36.9 37.1 37.3 36.2 36.5 36.7 36.9 36.6 36.7 37.0 37.2
+  [91] 36.4 36.7 36.9 37.1 37.3 36.2 36.6 37.3 36.3 36.6 36.8 37.0 37.3 36.5 36.8
+ [106] 36.9 37.1 37.4 36.9 37.1 37.4 36.3 36.6 36.8 37.0 37.4 36.6 36.8 37.0 36.6
+ [121] 36.8 37.0 37.1 37.8 36.3 36.7 36.8 37.1 37.5 36.6

mean(BodyTemp)

+ [1] 36.80769

median(BodyTemp)

+ [1] 36.8
```

Sample mode

Example: Body temperatures ($n = 130$)

```
Modes <- function(x) {   #Need to write a small function to compute the mode
  ux <- unique(x)
  tab <- tabulate(match(x, ux))
  ux[tab == max(tab)]
}
Modes(BodyTemp)

+ [1] 37.1
```

Sample mode

- Sample mean: 36.8076923
- Sample median: 36.8
- Sample mode(s): 37.1

Sample mode

For the sample median: $n = 130$, i.e. n is even

- We compute $(n/2) = 65$ and $(n/2) + 1 = 65 + 1 = 66$
- In the ordered data
  - the observation in the 65th position is 36.8
  - the observation in the 66th position is 36.8
- m is the average of these two observations, i.e. $m = 36.8$

Sample mode

Example: Body temperatures ($n = 130$)

```
sort(BodyTemp)[65:66]
```

```
+ [1] 36.8 36.8
```

```
mean(sort(BodyTemp)[65:66])
```

```
+ [1] 36.8
```

Sample mode

Comparison of measures:

- Sample mean
  - Uses all observations directly
  - Sensitive to extreme observations
- Sample median
  - Uses all observations indirectly
  - Not sensitive to extreme observations
- Sample mode
  - Rarely used in practice

Measures of variability (spread)

Measures of variability (or spread) of the data include

- Range
- Variance and Standard Deviation
- Inter-quartile Range

The “centre” of the observations is not a sufficient description of the data

- We also would like to know how spread out the observations are around their centre
- e.g. two classes may each have a sample average of 75%, but different spreads of grades

Sample range

Let $x_1, x_2, \ldots, x_n$ be a set of n numbers. Let

- x_{L} be the largest of these numbers,
- x_{S} be the smallest.

Then the sample range R is

$$R = x_{\mathrm{L}} - x_{\mathrm{S}},$$

the distance between the smallest and largest observations.

- This gives an extremely crude measure of spread
- The range is affected by extreme observations

Sample variance

We define the sample variance, denoted by s^2, as

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

The sample standard deviation is defined as $s = \sqrt{s^2}$

- s is often quoted rather than s^2,
- s is measured in the same units as the observations.

Sample variance

Understanding the formula: we want to find some measure of average distance between the observations $x_1, x_2, \ldots x_n$ and their sample mean $\bar{x}$

- Distances from the sample mean

$$x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x}$$

Note that the sum of these distances is zero, i.e.

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0.$$

Sample variance

- We can make all the differences positive by squaring:

$$(x_1 - \bar{x})^2, (x_2 - \bar{x})^2, \ldots, (x_n - \bar{x})^2$$

and then take the mean of the squares

$$\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Sample variance

- However, we have "used up" one observation's worth of data by calculating $\bar{x}$, and once we know the value of

$$x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_{n-1} - \bar{x}$$

we automatically know the value of $x_n - \bar{x}$.

Thus final formula is divided by $n - 1$ and not by n :

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2$$

Sample variance

Note the following: squaring out the bracket

$$(a - b)^2 = a^2 - 2ab + b^2$$

Sample variance

we have

$$\begin{aligned}
\sum_{i=1}^{n} (x_i - \bar{x})^2 &= \sum_{i=1}^{n} \left(x_i^2 - 2\bar{x}x_i + \bar{x}^2\right) \\
&= \sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n} \bar{x}x_i + \sum_{i=1}^{n} \bar{x}^2 \\
&= \sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} \bar{x}^2 \\
&= \sum_{i=1}^{n} x_i^2 - 2\bar{x}(n\bar{x}) + n\bar{x}^2 \\
&= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2
\end{aligned}$$

Sample variance

Therefore

$$\begin{aligned}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2 &= \sum_{i=1}^{n}x_i^2-n\bar{x}^2\\ &= \sum_{i=1}^{n}x_i^2-n\frac{\left(\sum_{i=1}^{n}x_i\right)^2}{n^2}\\ &= \sum_{i=1}^{n}x_i^2-\frac{\left(\sum_{i=1}^{n}x_i\right)^2}{n}\end{aligned}$$

Sample variance

Therefore

$$
\begin{aligned}
s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2 \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2-\frac{\left(\sum_{i=1}^{n}x_i\right)^2}{n}\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2-n\left(\bar{x}\right)^2\right)
\end{aligned}
$$

We need the sum of the squared values, and the sum of the values squared.

Sample variance

Example:

Suppose we have a set of $n = 5$ observations

$$1, 6, 3, 7, 5$$

We wish to calculate the sample standard deviation s

First, we calculate the sample mean $\bar{x}$:

$$\bar{x} = \frac{1 + 6 + 3 + 7 + 5}{5} = 4.4$$

Sample variance

Example:

Next, we calculate the sample variance s^2, using the formula

$$
\begin{aligned}
s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 \\
&= \frac{(1-4.4)^2+(6-4.4)^2+(3-4.4)^2+(7-4.4)^2+(5-4.4)^2}{4} \\
&= 5.8
\end{aligned}
$$

We can now calculate the sample standard deviation s

$$s = \sqrt{5.8} = 2.4$$

Sample variance

Example:

We could also choose to use the second formula:

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n(\bar{x})^2\right)$$

We need:

$$\sum_{i=1}^{n} x_i^2 = 1^2 + 6^2 + 3^2 + 7^2 + 5^2 = 120$$

and

$$\sum_{i=1}^{n} x_i = 1 + 6 + 3 + 7 + 5 = 22$$

Sample variance

Example:

Thus

$$s^2 = \frac{1}{4}\left(120 - \frac{(22)^2}{5}\right) = 5.8 \quad \text{and} \quad s = \sqrt{5.8} = 2.4$$

Sample variance

Example: Computation in R

```
n<-5
x<-c(1,6,3,7,5)
mean(x)

+ [1] 4.4

sum(x)

+ [1] 22

sum(x^2)

+ [1] 120
```

Sample variance

Example: Computation in R

```
var(x)

+ [1] 5.8

sum((x-mean(x))^2)/(n-1)

+ [1] 5.8

(sum(x^2)-n*mean(x)^2)/(n-1)

+ [1] 5.8

sqrt(var(x))

+ [1] 2.408319

sd(x)

+ [1] 2.408319
```

Sample variance

Example: Body temperatures

```
n<-length(BodyTemp)
c(mean(BodyTemp),var(BodyTemp))

+ [1] 36.8076923  0.1659869

sum((BodyTemp-mean(BodyTemp))^2)/(n-1)

+ [1] 0.1659869

(sum(BodyTemp^2)-n*mean(BodyTemp)^2 )/(n-1)

+ [1] 0.1659869

sqrt(var(BodyTemp))

+ [1] 0.4074149

sd(BodyTemp)

+ [1] 0.4074149
```

Sample variance

If we round $\bar{x}$ to, say 36.81 before using the formula we get

```
xbar<-round(mean(BodyTemp),2)
xbar

+ [1] 36.81

var(BodyTemp)

+ [1] 0.1659869

sum((BodyTemp-xbar)^2)/(n-1)

+ [1] 0.1659922

(sum(BodyTemp^2)-n*xbar^2 )/(n-1)

+ [1] -0.005217054
```

However, the variance cannot be negative.

Data set description

Describing a data set involves providing an appropriate

- Numerical measure of centre
- Numerical measure of spread
- Graphical description (shape) (e.g. histogram)

In statistical terminology we say that we describe the distribution of values (or observations) in a sample

Symmetry and Skew

Symmetric distributions:

- We say that a sample has a (perfectly) symmetric distribution if its histogram has a symmetric shape around some x value;
- In this case, we will also have that mean = median.

Very rarely a sample has a perfectly symmetric distribution, but in some cases the approximation is very good

- In this case, we will also have that mean $\approx$ median

Symmetry and Skew

Positively (right) skewed data: histogram shows a “long tail” on the right

- median $<$ mean

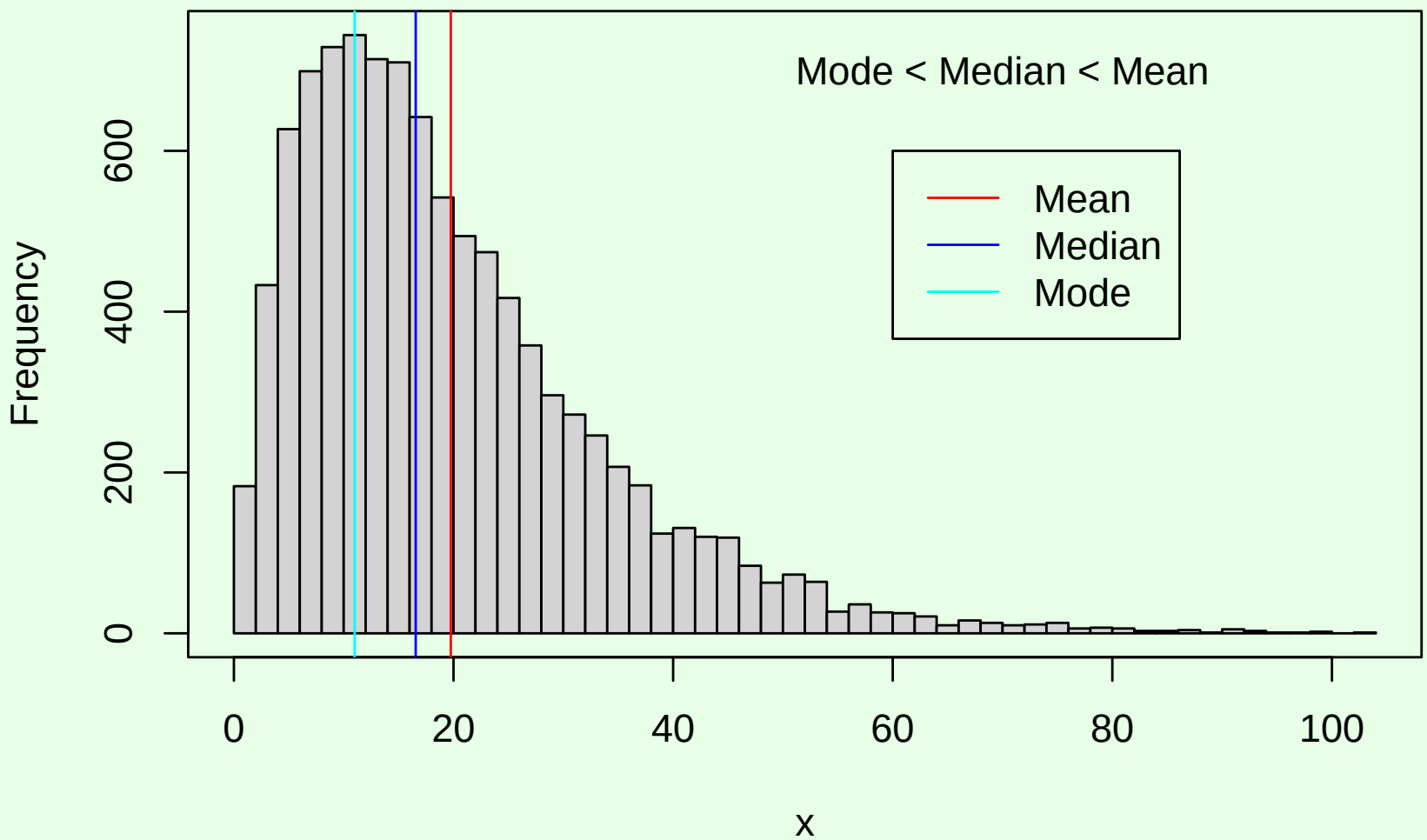
Negatively (left) skewed data: histogram shows a “long tail” on the left.

- median $>$ mean

Recall that the mean is influenced by large values but the median is not

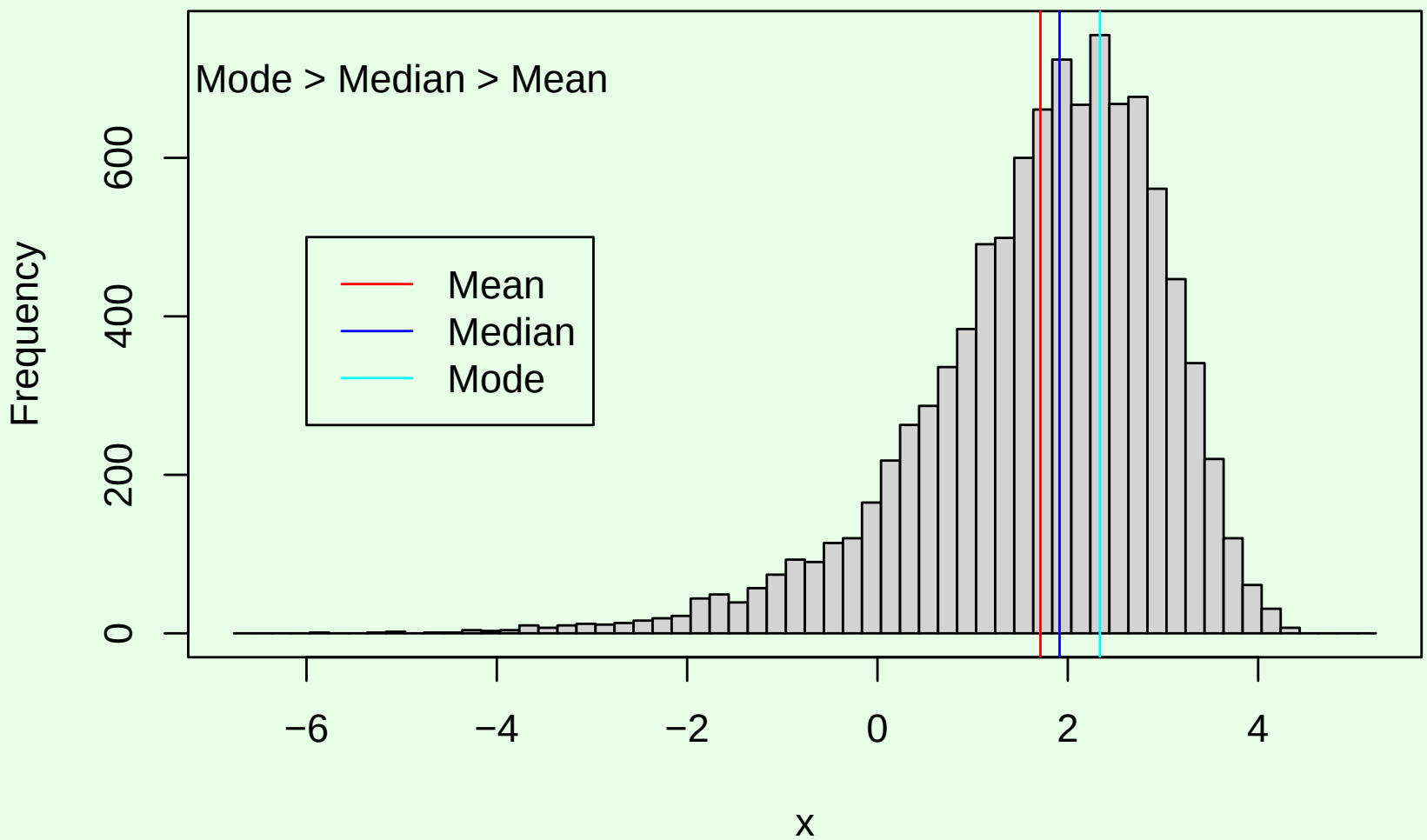
Symmetry and Skew

Example: Positive (right) skew



Symmetry and Skew

Example: Negative (left) skew



Histograms and numerical summaries

We cannot compute the sample mean, sample median and sample variance if we only have access to the histogram plot.

- In a histogram data are grouped in intervals
- By looking at the height of the rectangle you know how many data observations are in the interval, but you don’t know their exact location within the interval
- In the previous examples, the mean and the median were computed from the full data set.

From samples to populations

We recognize that our data are representative of some underlying population, and our real goal is to understand the characteristics of the population, based on the sample.

From samples to populations

Rather than

- sample mean $\bar{x}$
- sample variance s^2
- sample standard deviation (s.d.) s

we want to consider the corresponding population quantities

- population mean μ
- population variance σ^2
- population standard deviation (s.d.) σ

From samples to populations

A simple interpretation of the “population” is to think of it as a sample with an extremely large sample size:

- we keep collecting data until we have learned everything about the way the data are generated.

Empirical rule

For a perfectly symmetric, "mound-shaped" distribution:

- Roughly 68% of the observations will lie in the range

$$\text{mean} \pm \text{s.d.}$$

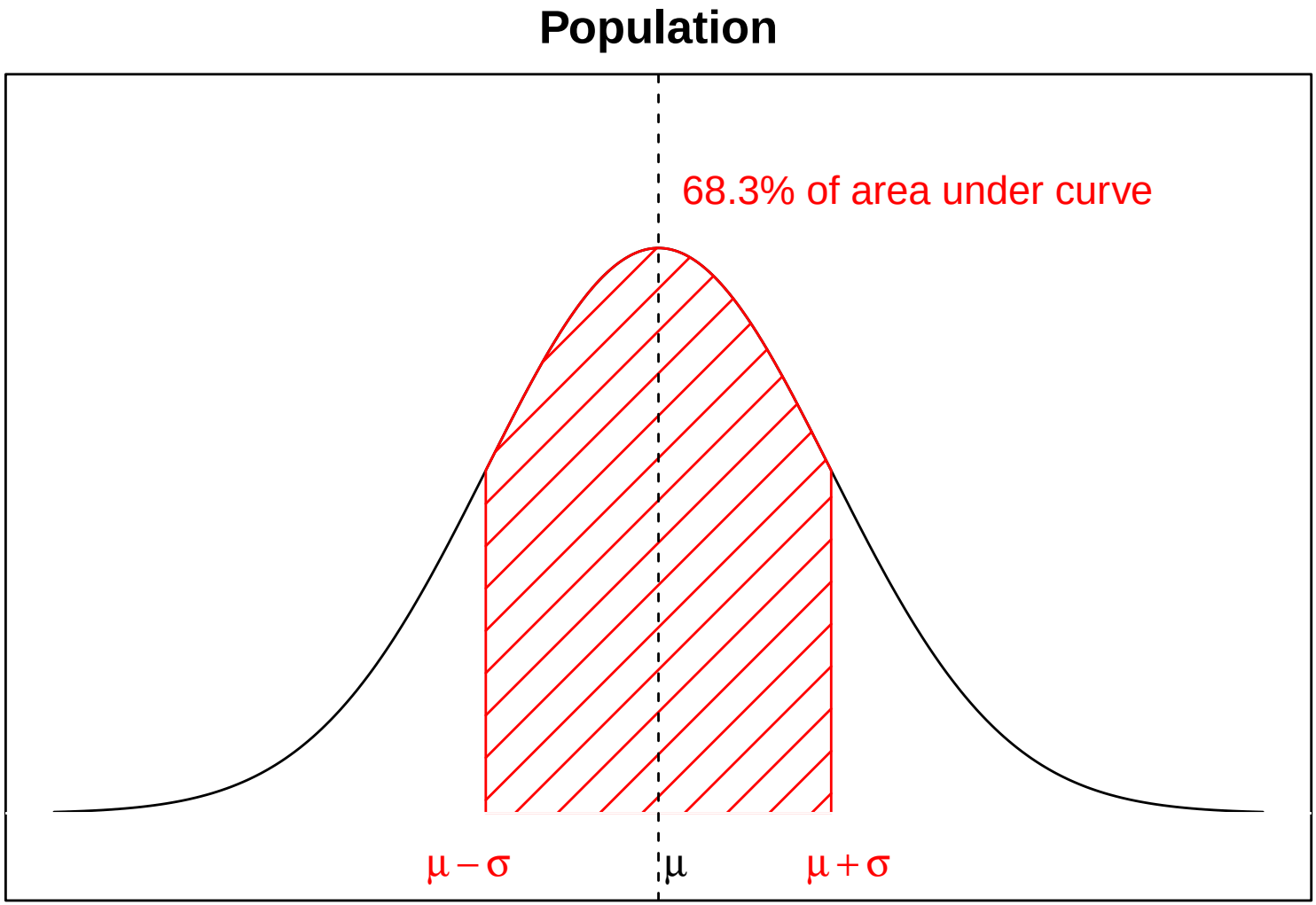
- Roughly 95% of the observations will lie in the range

$$\text{mean} \pm (2 \times \text{s.d.})$$

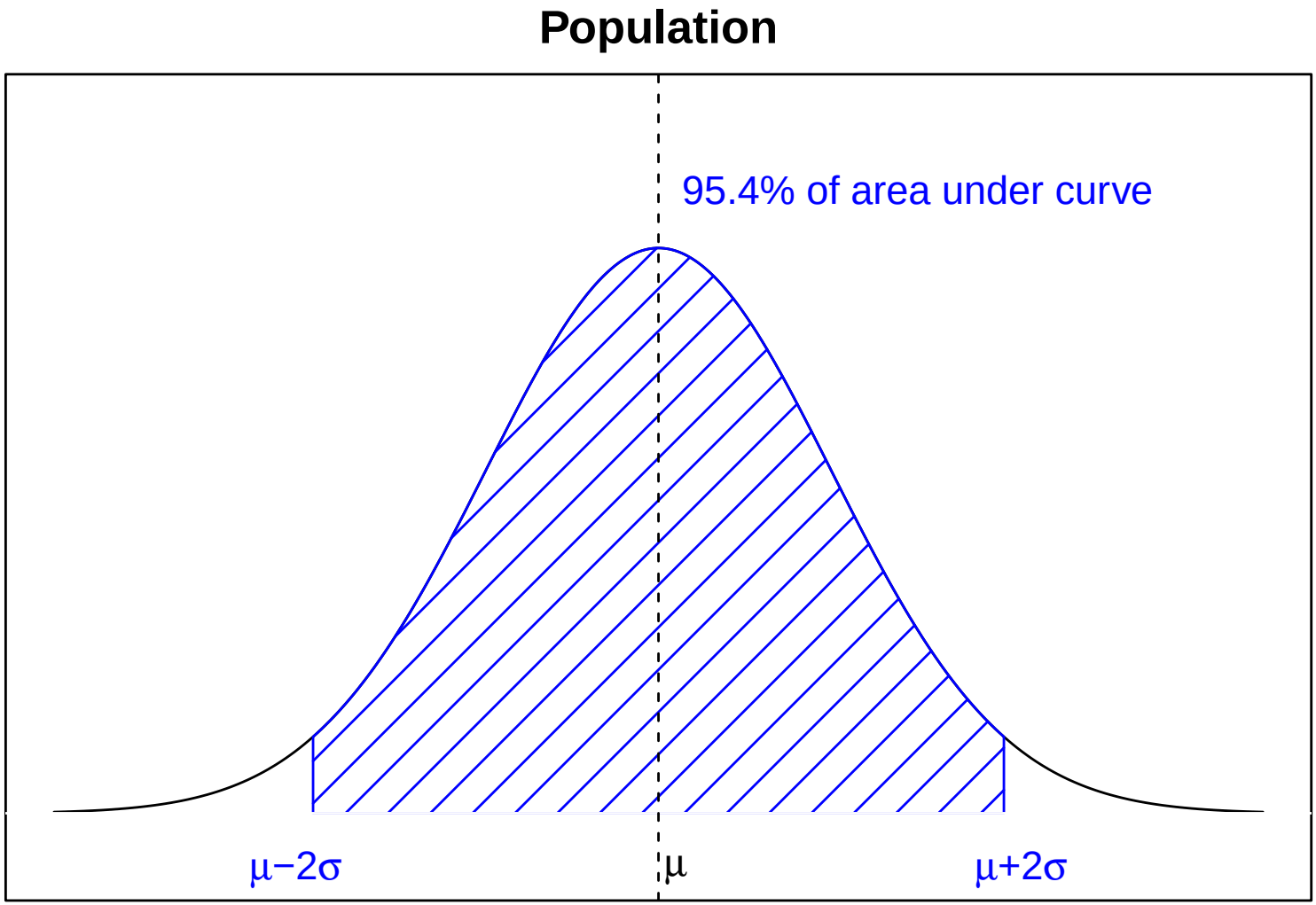
- Roughly 99.7% of the observations will lie in the range

$$\text{mean} \pm (3 \times \text{s.d.})$$

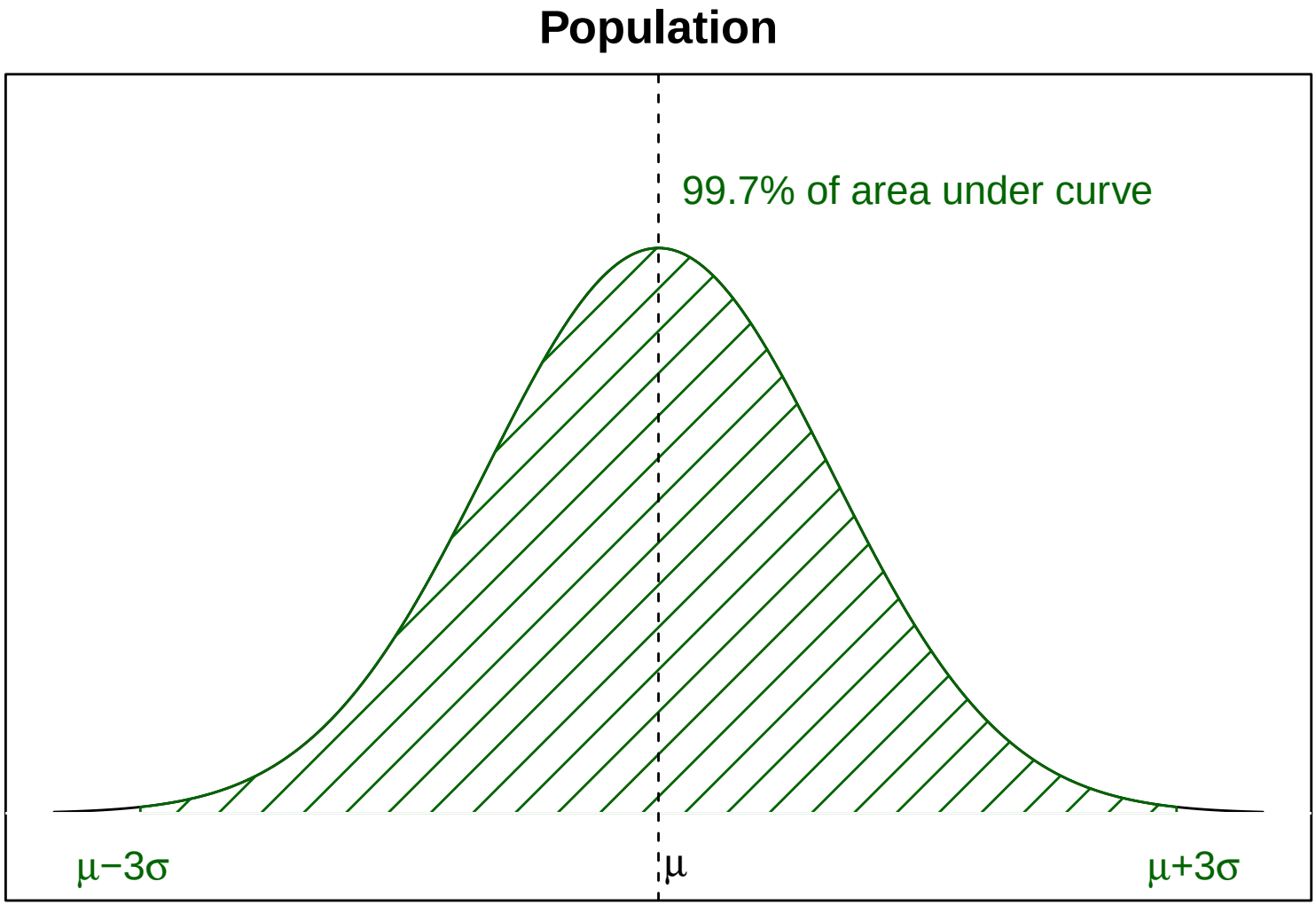
Empirical rule



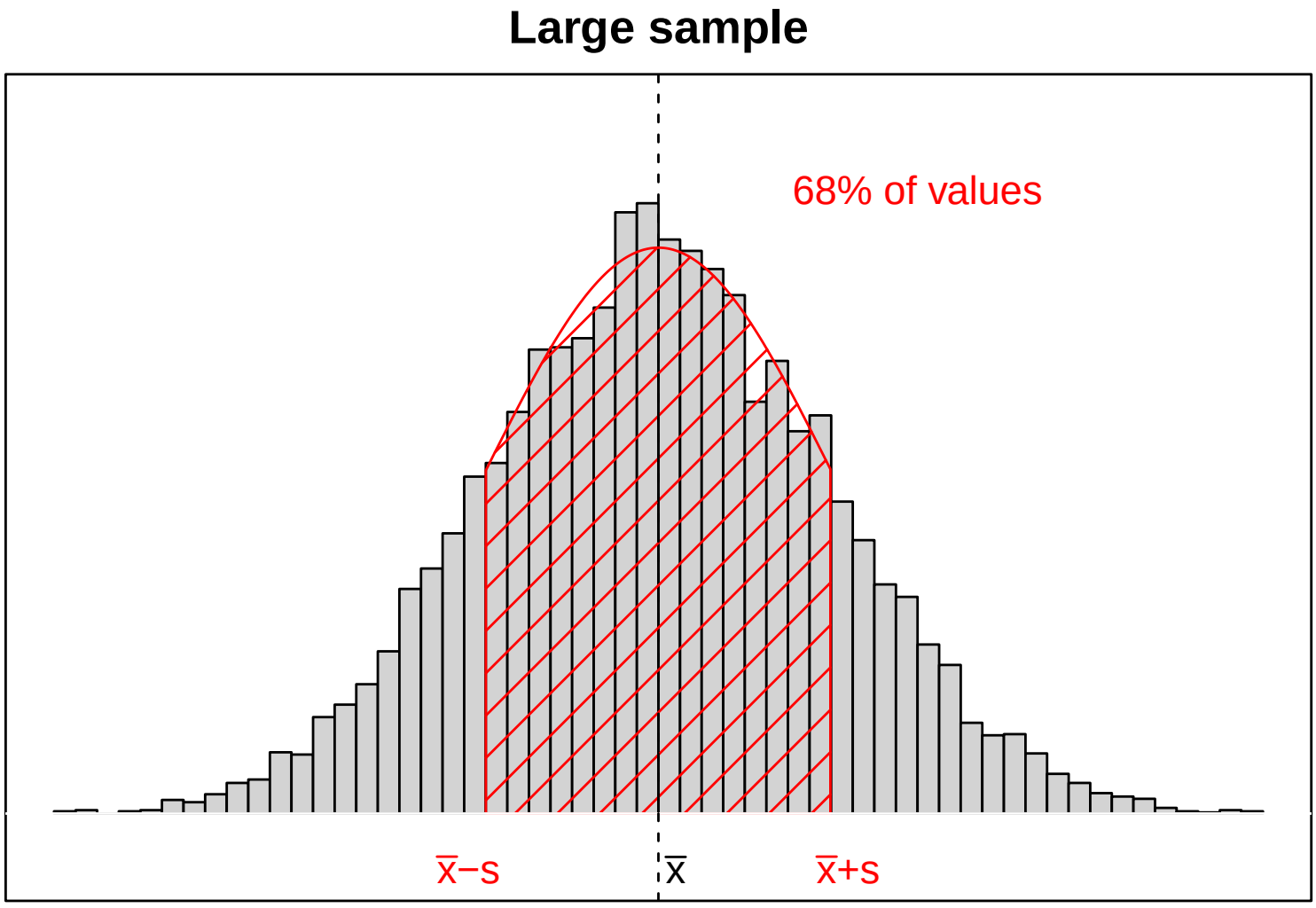
Empirical rule



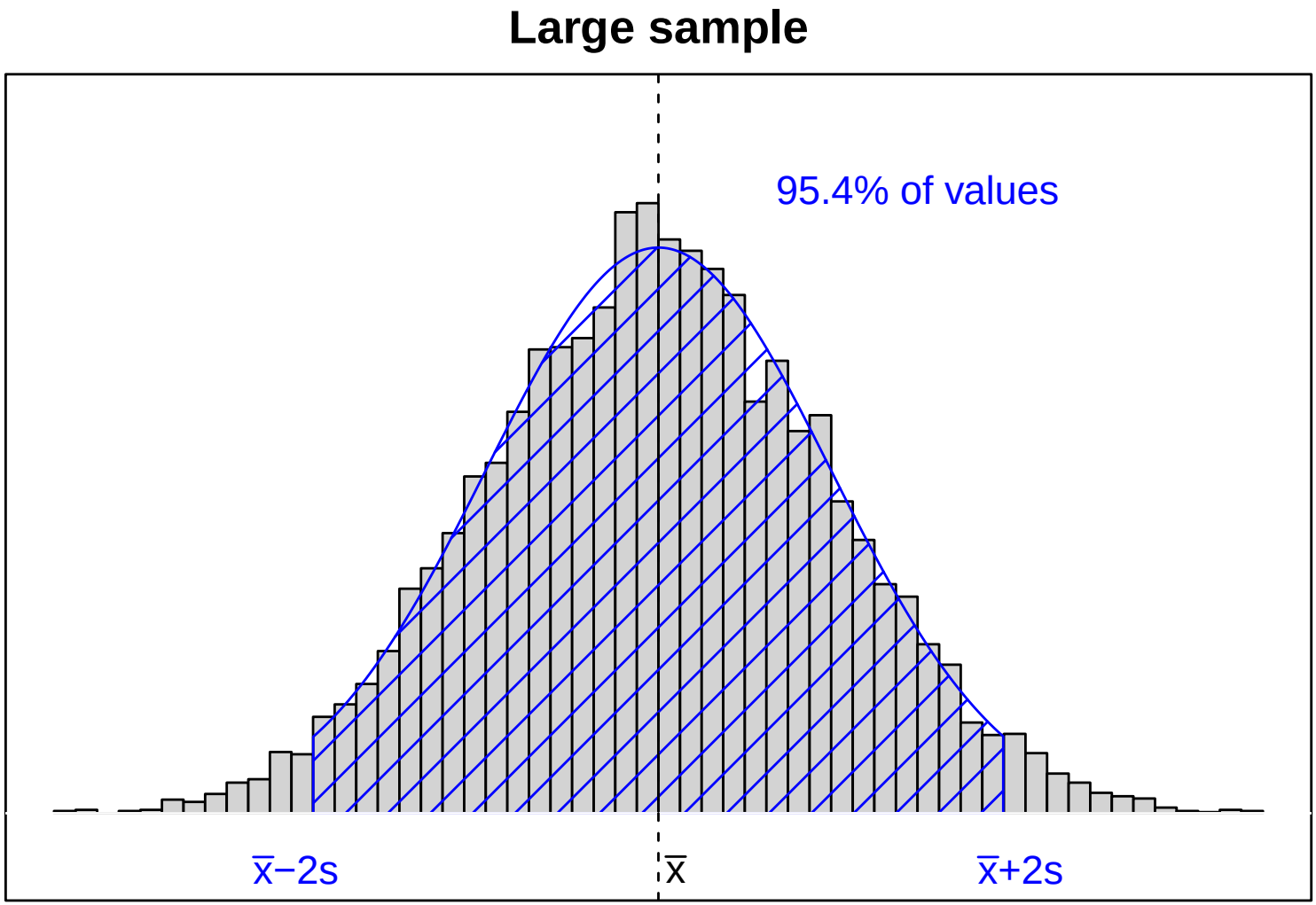
Empirical rule



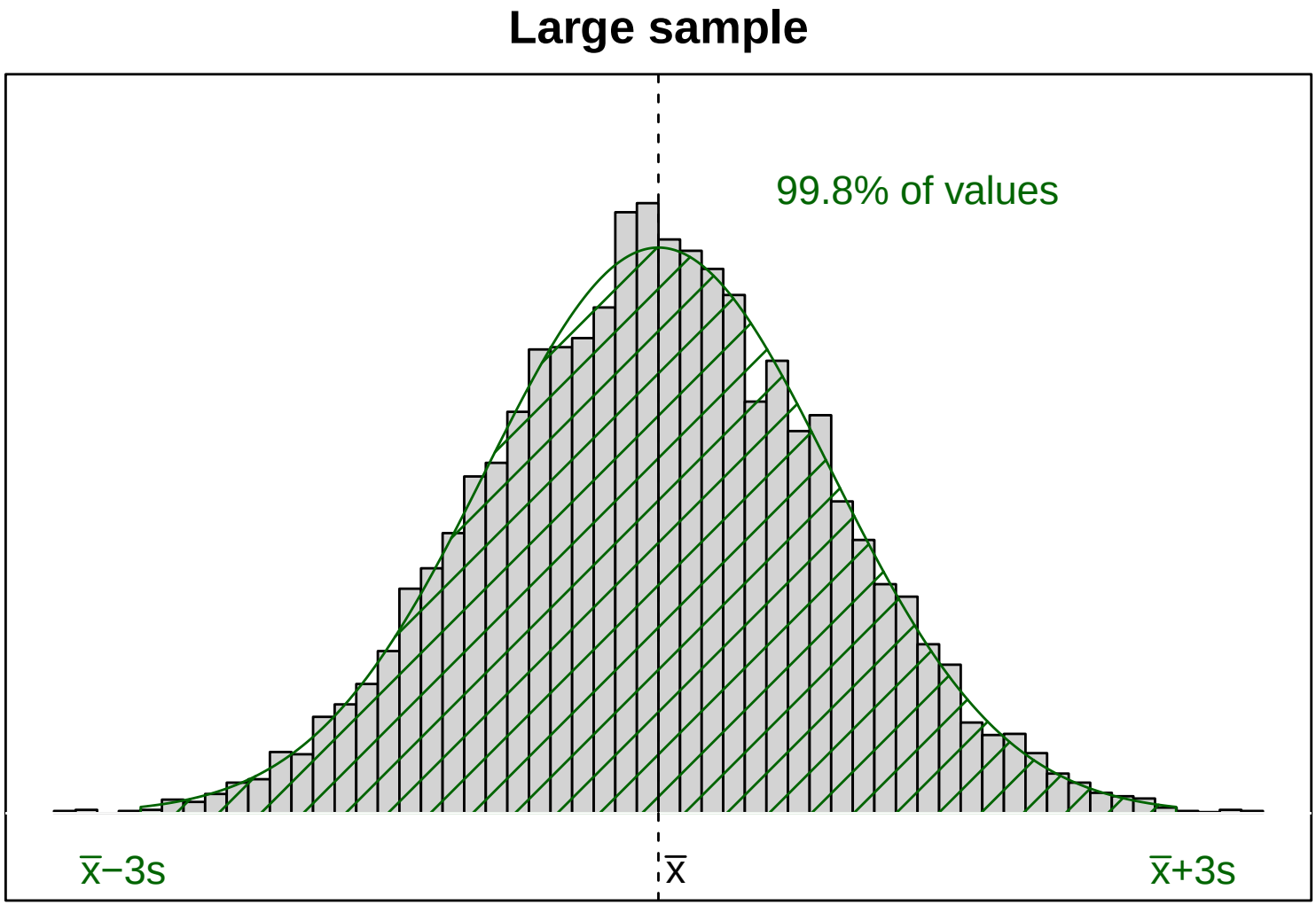
Empirical rule



Empirical rule



Empirical rule



Empirical rule

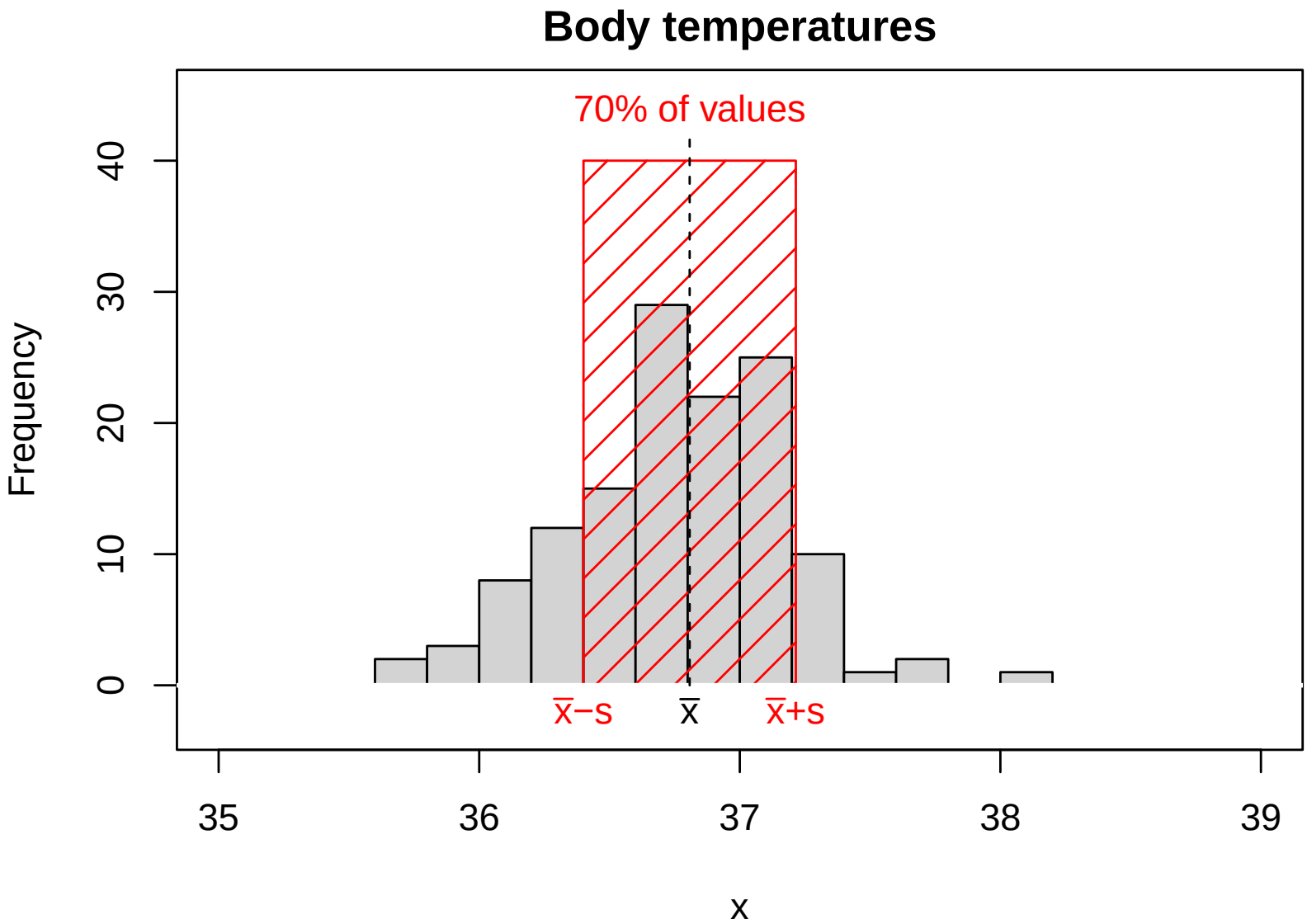
Example: Body Temperatures

Mean $\bar{x} = 36.81$, Standard deviation $s = 0.41$.

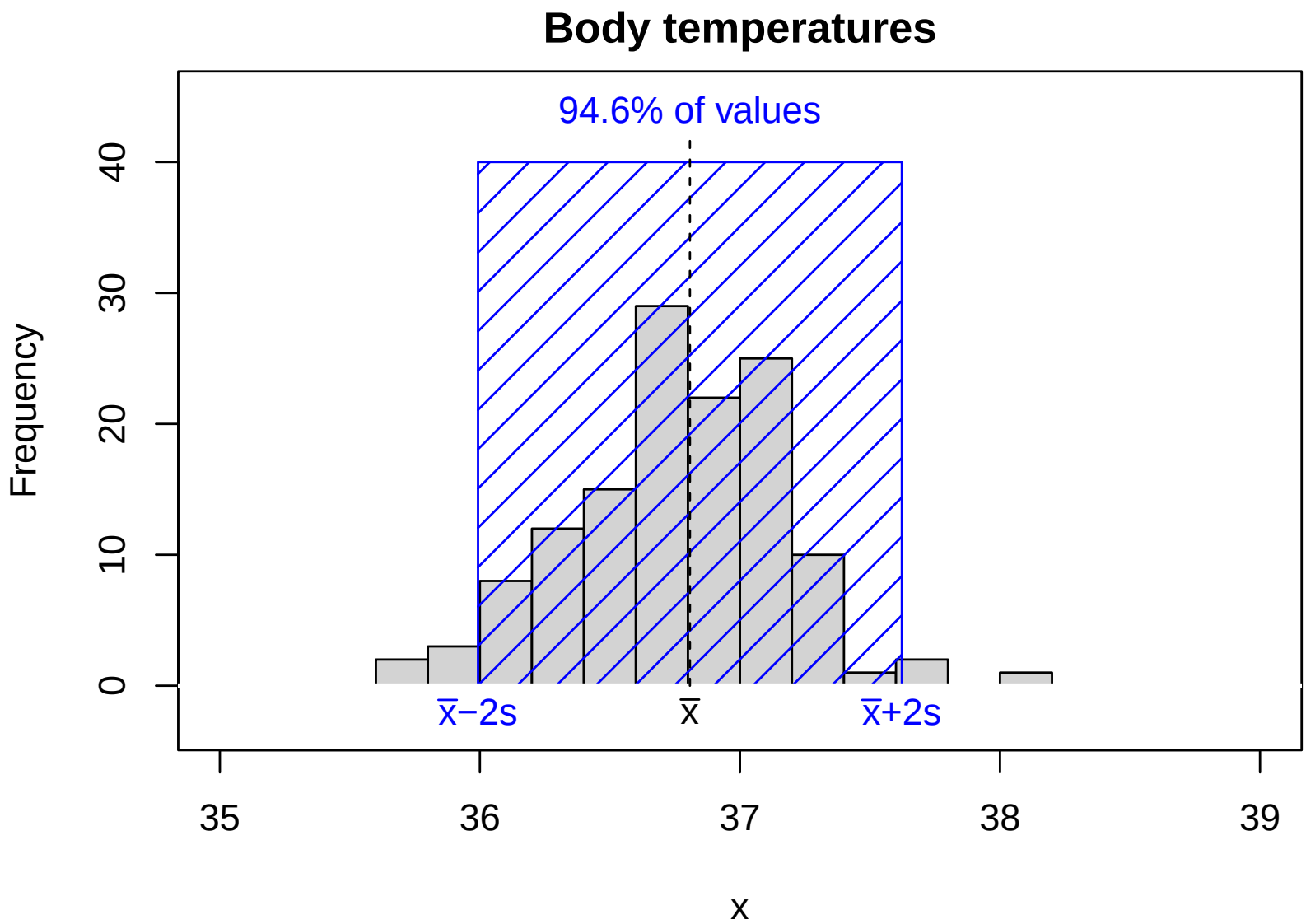
$$(\bar{x} - s, \bar{x} + s) = (36.40, 37.22)$$
$$(\bar{x} - 2s, \bar{x} + 2s) = (35.99, 37.62)$$
$$(\bar{x} - 3s, \bar{x} + 3s) = (35.59, 38.03)$$

Interval	Empirical	Actual
(36.40, 37.22)	68.3%	70.0%
(35.99, 37.62)	95.4%	94.6%
(35.59, 38.03)	99.7%	99.2%

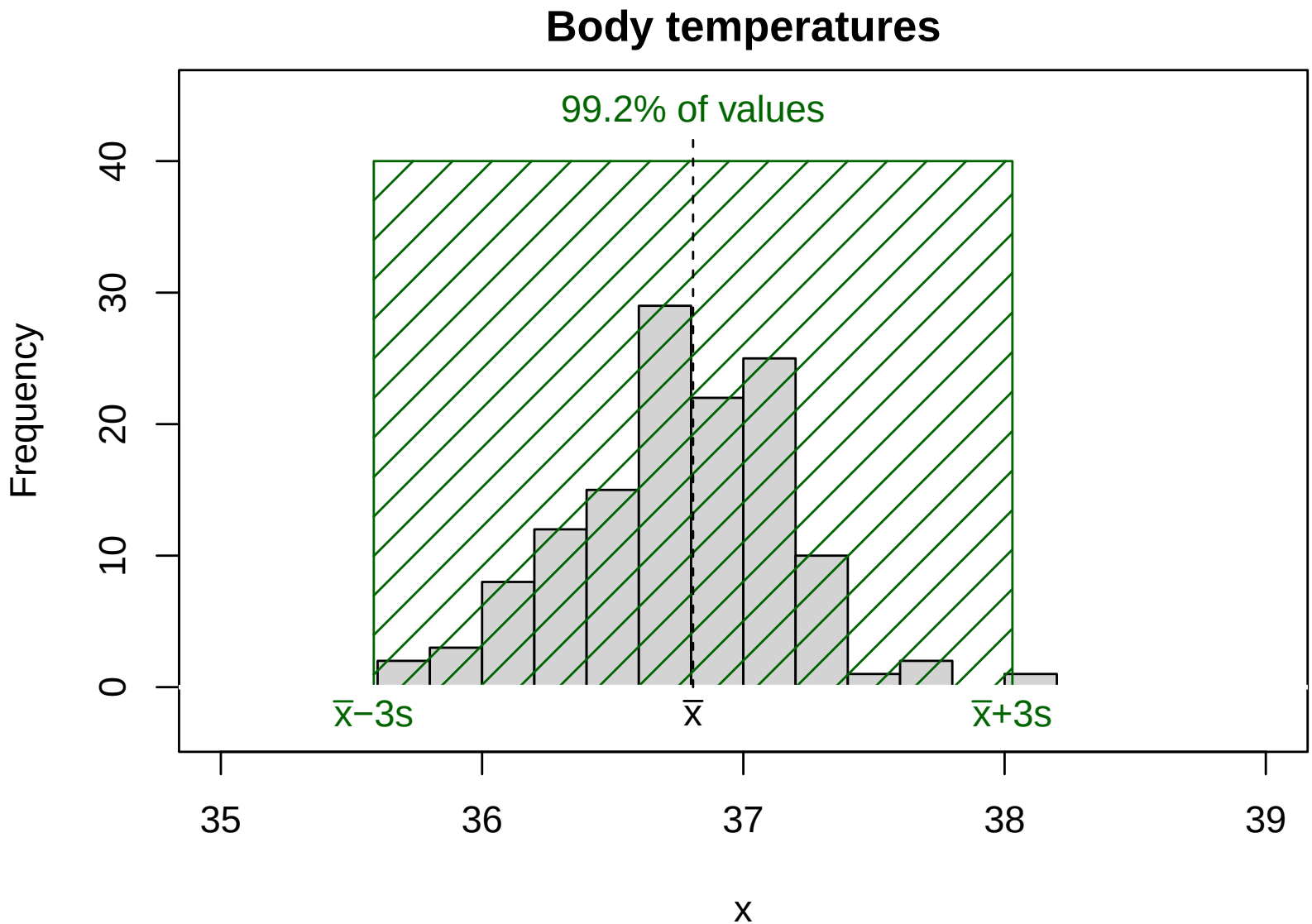
Empirical rule



Empirical rule



Empirical rule



Empirical rule

Example: Heights

In a sample of 300 Canadian adult men, the sample mean height is $\bar{x} = 72$ inches and the sample standard deviation of $s = 3.5$ inches.

Assuming that the heights in this sample have a mound-shape, symmetric distribution (at least approximately), then

- $\approx 68\%$ of men in the sample have a height in the interval

$$(72 - 3.5, 72 + 3.5) = (68.5, 75.5) \text{ inches}$$

- $\approx 95\%$ men in the sample have a height in the interval

$$(72 - 2 \times 3.5, 72 + 2 \times 3.5) = (65, 79) \text{ inches}$$

Empirical rule

Example: Statistics grades

The distribution of grades in a Statistics class is roughly symmetric and mound-shaped, with $\bar{x} = 75$ and $s^2 = 49$. We seek an interval that captures $\approx 95\%$ of the observed data

According to the empirical rule, the interval $(\bar{x} - 2s, \bar{x} + 2s)$ captures $\approx 95\%$ of the observed data: hence

$$(\bar{x} - 2s, \bar{x} + 2s) = (75 - 2 \times 7, 75 + 2 \times 7) = (61, 89)$$

captures $\approx 95\%$ of the observed grades from the class.

Empirical rule

Example: Mathematics grades

A Mathematics exam is graded on a scale of 0 to 100 points. The sample average is 83.5 points and sample standard deviation is 13.6 points.

We seek an interval that captures $\approx$ 95% of the observed grades.

Empirical rule

Example: Mathematics grades

If we assume that the grades in this midterm have a mound-shape, symmetric distribution (at least approximately), from the empirical rule, we obtain an interval

$$\begin{aligned}(\bar{x} - 2s, \bar{x} + 2s) &= (83.5 - 2 \times 13.6, 83.5 + 2 \times 13.6) \\ &= (56.3, 110.7)\end{aligned}$$

Empirical rule

We note that this interval goes above 100, even though students could not score more than 100 on the exam

The interval can be adjusted to $(56.3, 100)$

More importantly, this indicates that the empirical rule is not good to use here: the data probably do not have a mound-shaped histogram

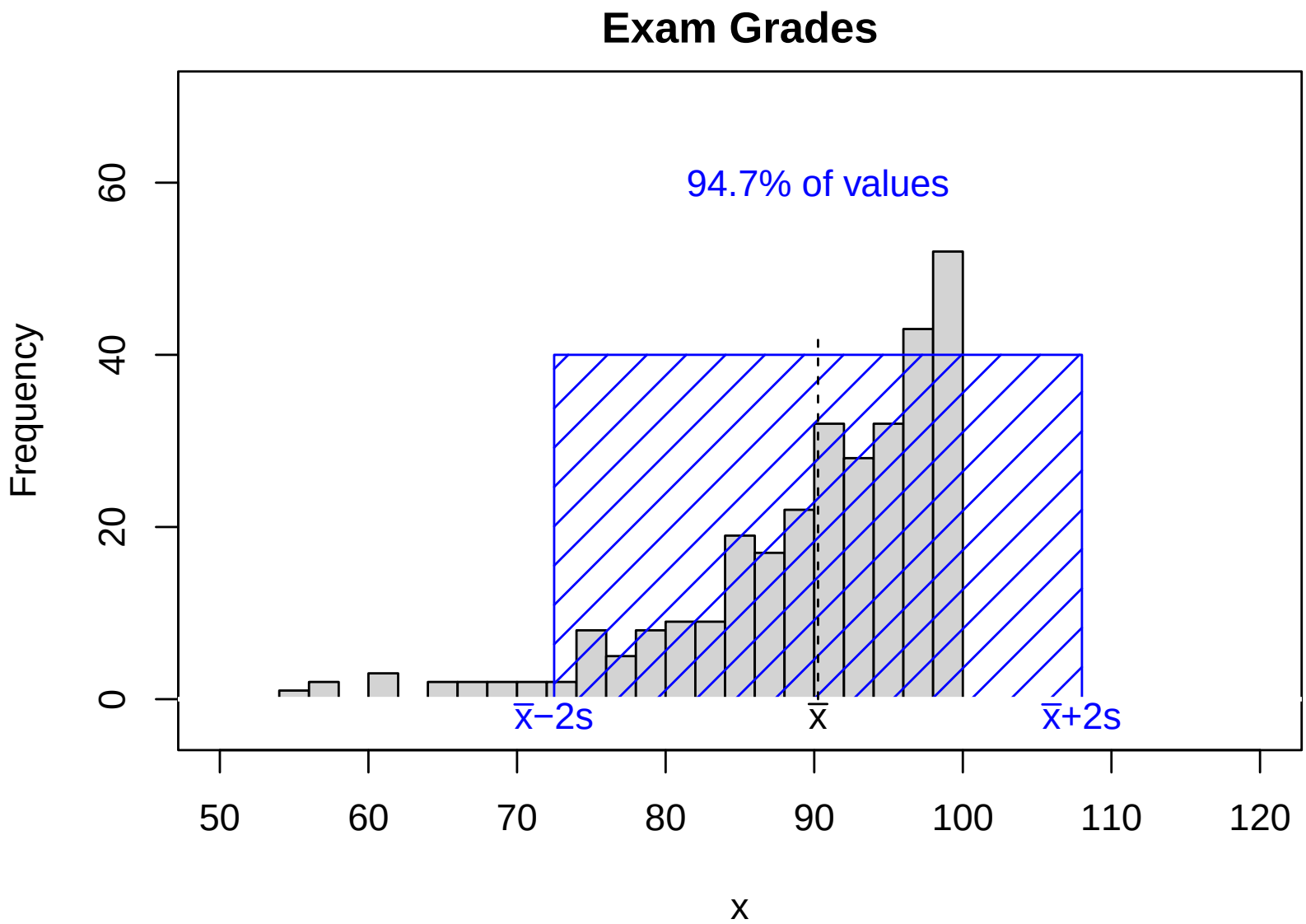
Empirical rule

Example: Exam grades

Sample of 300 exam grades:

```
+    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
+   54.27   86.07   92.28   90.26   96.76   99.99
+ [1] "Sample mean"
+ [1] 90.25576
+ [1] "Sample sd"
+ [1] 8.879329
```

Empirical rule



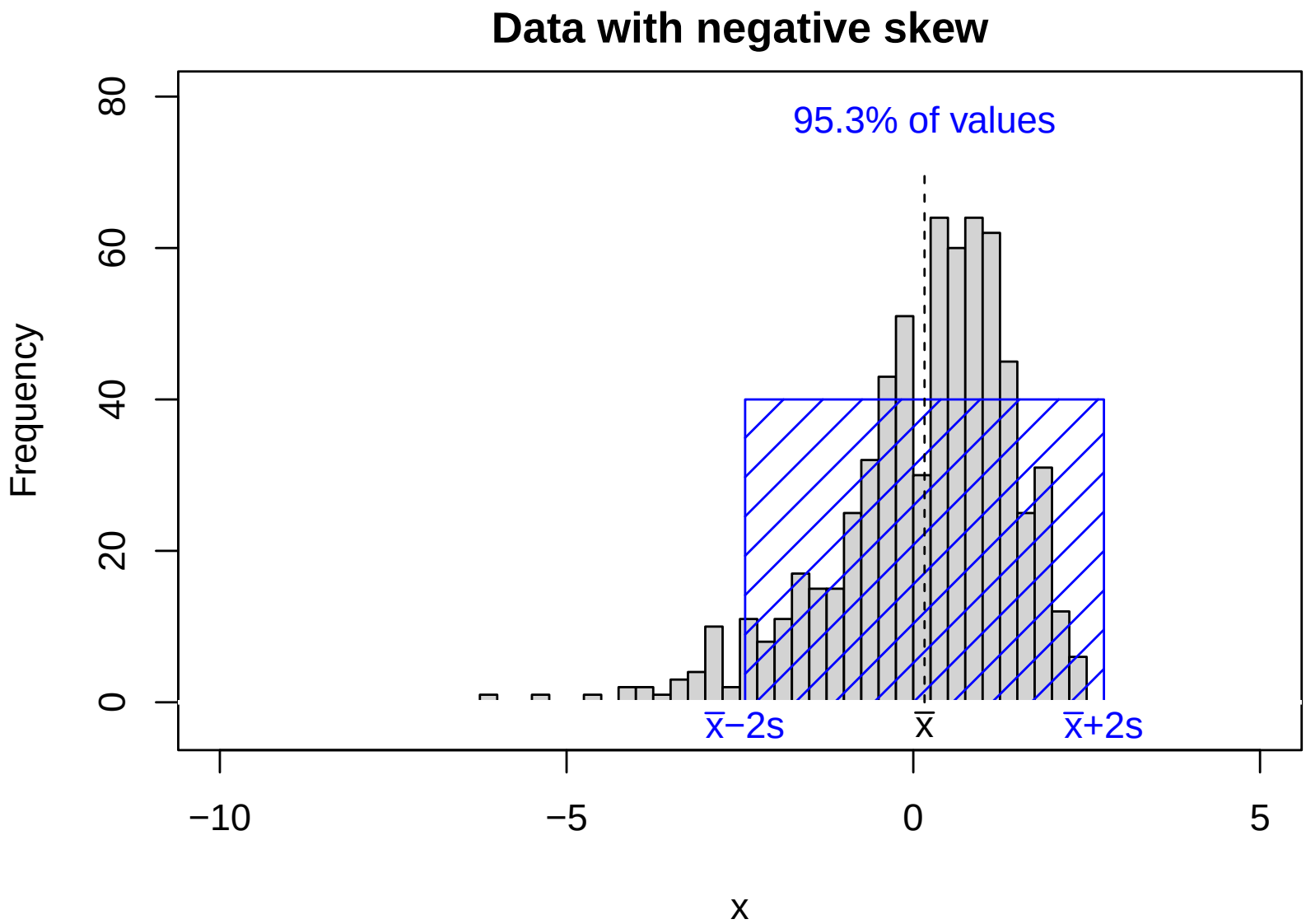
Empirical rule

Example: Negative (left) skewed data

Sample of 654 data with negative skew.

```
+    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
+ -6.1370 -0.4758  0.4034  0.1620  1.0884  2.4717
+ [1] "Sample mean"
+ [1] 0.1620016
+ [1] "Sample sd"
+ [1] 1.293524
```

Empirical rule



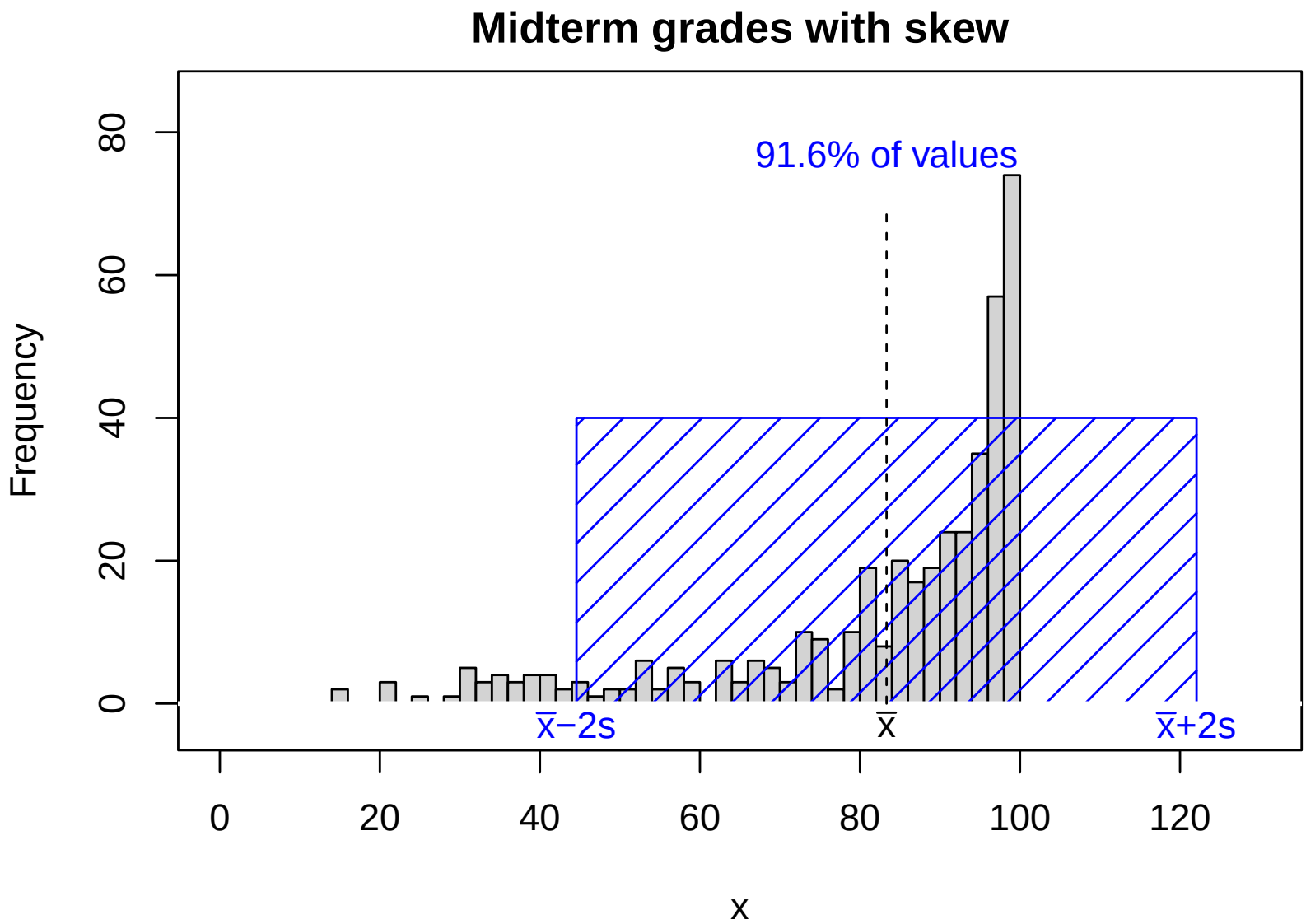
Empirical rule

Example: Midterm grades with skew

407 midterm grades that exhibit skewness.

```
+    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
+   15.10   78.85   90.88   83.32   96.95  100.00
+ [1] "Sample mean"
+ [1] 83.32208
+ [1] "Sample sd"
+ [1] 19.3735
```

Empirical rule



Empirical rule

Here the interval

$$(\bar{x} - 2s, \bar{x} + 2s)$$

does not contain approximately 95% of the data.

The Empirical rule does not always work well if the distribution is not symmetric.

Chebyshev's rule

For any distribution and any number $k > 1$: at least

$$\left(1 - \frac{1}{k^2}\right) \times 100\%$$

of observations will fall into the interval

$$(\bar{x} - ks, \bar{x} + ks)$$

regardless of the "shape" of the histogram.

Chebyshev's rule

- $k = 2$: The percentage of observations falling within two standard deviations of the mean is at least

$$\left(1 - \frac{1}{2^2}\right) \times 100\% = 75\%$$

- $k = 3$: The percentage of observations falling within three standard deviations of the mean is at least

$$\left(1 - \frac{1}{3^2}\right) \times 100\% = 88.9\%$$

We will see that this result applies to samples and populations.

Chebyshev's rule

Example: Mathematics grades

The grades have a sample average of 83.5 points and a standard deviation of 13.6 points

We can apply Chebyshev's rule:

- at least 75% of the grades are in the interval

$$(\bar{x} - 2s, \bar{x} + 2s) = (83.5 - 2 \times 13.6, 83.5 + 2 \times 13.6)$$
$$= (56.3, 110.7)$$

The interval may still be adjusted to $(56.3, 100)$

Chebyshev's rule

Example: Body temperatures

For the sample of $n = 130$ body temperatures

$$\bar{x} = 36.81 \qquad s = 0.41$$

We can apply Chebyshev's rule and say that at least 75% of the temperatures are in the interval

$$(\bar{x} - 2s, \bar{x} + 2s) = (35.99, 37.62)$$

Chebyshev's rule

Example: Body temperatures

However, the temperatures do have an approximately mound-shape and symmetric distribution.

The Empirical rule applies and we can say that roughly 95% of the temperatures are in the interval

$$(\bar{x} - 2s, \bar{x} + 2s) = (35.99, 37.62)$$

Obviously, "roughly 95%" is a better approximation than "at least 75%" in this case

Chebyshev's rule

Conclusion:

- If the data have an approximately mound-shape and symmetric distribution it is better to use the Empirical rule
  - this can give a close approximation to the actual percentage;
- Otherwise, we can only use Chebyshev's rule.

z-scores

We wish to report how ‘extreme’ a value is in a dataset;

- i.e. how far the observation is from the centre on some standard scale.

We can use z-score to answer this question.

z-scores

For a given dataset with sample mean $\bar{x}$ and sample standard deviation s, the z-score for an observation x is

$$z\text{-score} = \frac{x - \bar{x}}{s}$$

We could also define this quantity for a population

$$z\text{-score} = \frac{x - \mu}{\sigma}$$

The z-score can be interpreted as how many standard deviations away from the mean the value x is.

z-scores

Example: Heights

For the sample of 300 Canadian adult men:

- mean height $\bar{x} = 72$ inches,
- standard deviation $s = 3.5$ inches.

z-scores

Example: Heights

- For a man in the sample with height $x = 80$ inches,

$$z\text{-score} = \frac{80 - 72}{3.5} \approx 2.29$$

  i.e. height is 2.29 standard deviations above the average

- For a man in the sample with height $x = 68.5$ inches,

$$z\text{-score} = \frac{68.5 - 72}{3.5} = -1$$

  i.e. height is 1 standard deviation below the average.

z-scores

Note

If the data have a roughly symmetric, mound-shaped distribution, from the Empirical rule

- $\approx 68\%$ of the z-scores will lie between -1 and 1
- $\approx 95\%$ of the z-scores will lie between -2 and 2
- $\approx 99.7\%$ of the z-scores will lie between -3 and 3

Graphical methods: Boxplots

Boxplots are another graphical tool for describing quantitative data, and are useful for assessing

- centre,
- spread,
- skewness,
- outliers.

Graphical methods: Boxplots

Sample quartiles:

- The 1st sample quartile is the number, q_L, such that 25% of the observations are $\leqslant q_L$
  - This the lower quartile
- The 2nd sample quartile is the median, m, such that 50% of the observations are $\leqslant m$
- The 3rd sample quartile is the number, q_U, such that 75% of the observations are $\leqslant q_U$
  - This is the upper quartile

Graphical methods: Boxplots

In general, for $0 \leqslant p \leqslant 1$, the pth sample quantile is the number, q_p, such that $(100 \times p)\%$ of the observations are $\leqslant q_p$.

- $q_L = q_{0.25}$
- $m = q_{0.5}$
- $q_U = q_{0.75}$

etc.

Graphical methods: Boxplots

The interquartile range (IQR) is the distance between the lower quartile and the upper quartile

$$\text{IQR} = q_U - q_L$$

Boxplot construction

Boxplots vary slightly from software to software, but the basic components are

- median line
- box (lower to upper quartile) or lower and upper hinge
- fences and whiskers

Boxplot construction

- Inner fences
  - The upper inner fence is $q_U + 1.5 \times \text{IQR}$
  - The lower inner fence is $q_L - 1.5 \times \text{IQR}$.
- Outer fences
  - The upper outer fence is $q_U + 3 \times \text{IQR}$
  - The lower outer fence is $q_L - 3 \times \text{IQR}$.

Boxplot construction

- Whiskers
  - Placed at data points less extreme than inner fences
  - From `help(boxplot.stats)`:

> ... the whiskers extend to the most extreme data point which is no more than `coef` times the length of the box away from the box.

where `coef` is equal to 1.5.

Boxplot construction

- Outliers that fall outside of the whiskers are also plotted.
  - The outliers are marked by standard point symbols

Boxplot construction

```
x

+  [1] -2.70  0.42  1.93  2.49  0.25  1.28  1.98 -0.42  0.12  1.58  1.34  1.38
+ [13]  1.59  0.54  1.40  1.35  2.99  0.53 -0.14 -0.29  0.64  1.28 -0.07  0.80
+ [25]  0.24  0.97 -0.31 -0.65  1.83 -1.06  0.90  2.16  0.17 -0.19  6.83  5.60
+ [37]  0.38  0.75  3.30  1.19  1.39  0.50  4.56 -0.02  1.36 -0.58  1.36  0.77
+ [49]  1.04 -0.17  3.32  1.19 -0.77  0.57  2.61 -1.76  0.20 -1.14  3.06 -1.79
+ [61] -0.96  0.44

m<-median(x)
qL<-quantile(x,prob=c(0.25))
qU<-quantile(x,prob=c(0.75))
IQR<-as.numeric(qU-qL)
c(qL,qU,IQR)

+     25%     75%
+ -0.0575  1.3975  1.4550

innerL<-qL-1.5*IQR; innerU<-qU+1.5*IQR
outerL<-qL-3*IQR; outerU<-qU+3*IQR
whiskL<-min(x[x > innerL]); whiskU<-max(x[x < innerU])

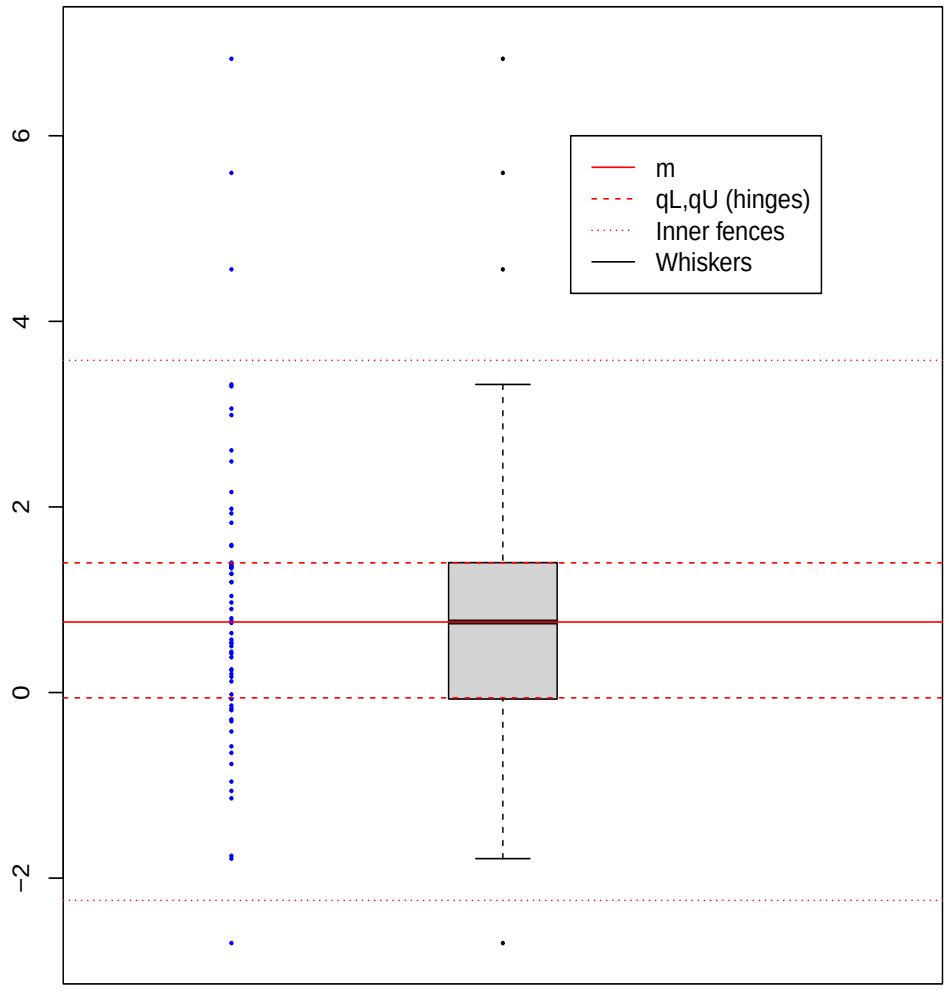
c(m,innerL,innerU,outerL,outerU,whiskL,whiskU)

+             25%     75%     25%     75%
+  0.7600 -2.2400  3.5800 -4.4225  5.7625 -1.7900  3.3200
```

Boxplot construction

```
par(mar=c(2,3,2,1))
boxplot(x,at=0,pch=19,cex=0.25,xlim=range(-1.5,1.5),ylim=range(-2.75,7))
points(rep(-1,n),x,col='blue',pch=19,cex=0.25)
abline(h=c(m,qL,qU,innerL,innerU),col='red',lty=c(1,2,2,3,3))
legend(0.25,6,c('m','qL,qU (hinges)','Inner fences','Whiskers'),lty=c(1,2,3),
       col=c('red','red','red','black'))
```

Boxplot construction



Boxplot construction

```
boxplot(x,plot=FALSE)

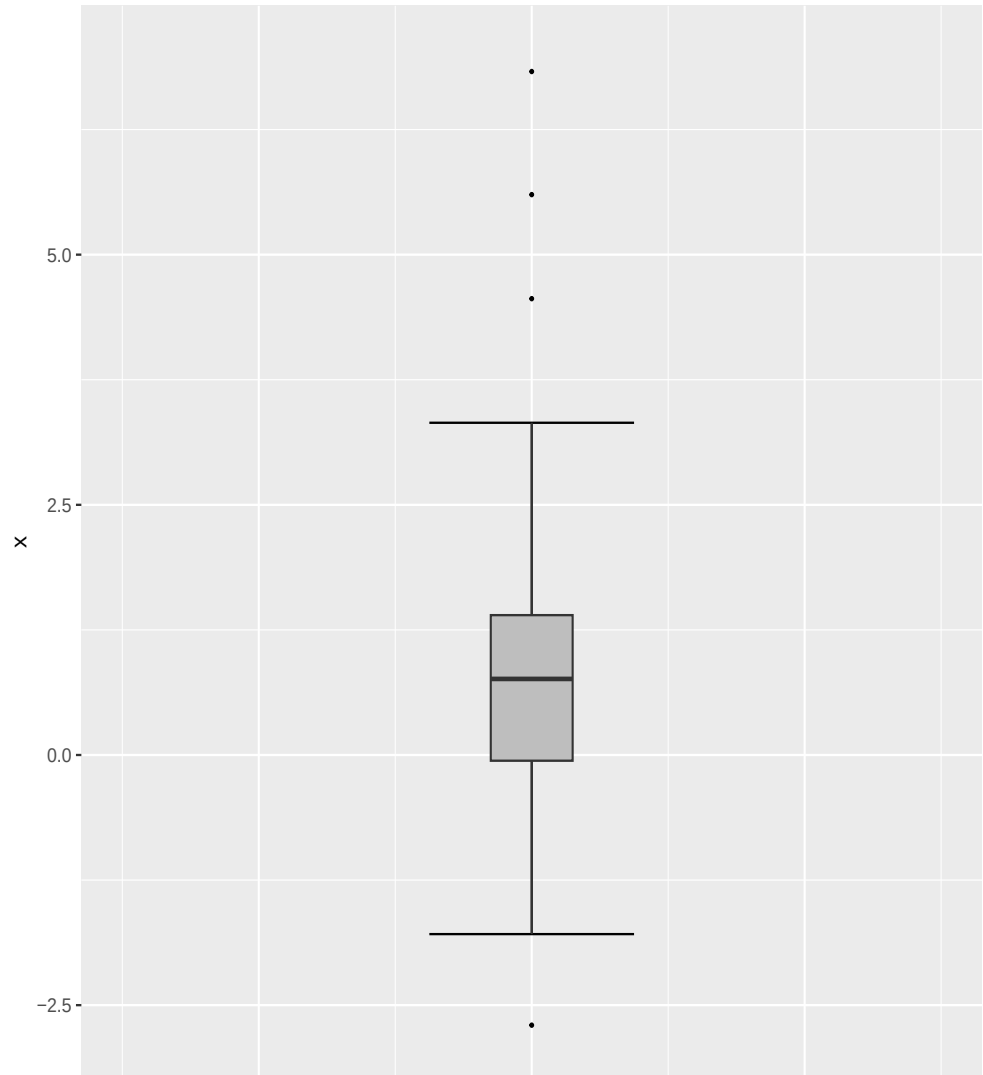
+ $stats
+       [,1]
+ [1,] -1.79
+ [2,] -0.07
+ [3,]  0.76
+ [4,]  1.40
+ [5,]  3.32
+
+ $n
+ [1] 62
+
+ $conf
+           [,1]
+ [1,] 0.4650295
+ [2,] 1.0549705
+
+ $out
+ [1] -2.70  6.83  5.60  4.56
+
+ $group
+ [1] 1 1 1 1
+
+ $names
+ [1] "1"
```

Boxplot construction

Using `ggplot`:

```
ggplot(data.frame(x), aes(y=x)) + stat_boxplot(geom ='errorbar') +
  geom_boxplot(outlier.colour="black", outlier.shape=19, outlier.size=0.5,
                   width=0.3,fill='gray') +
    coord_cartesian(xlim=c(-1.5,1.5),ylim = c(-2.75, 7)) +
    theme(axis.title.x=element_blank(),
        axis.text.x=element_blank(),
        axis.ticks.x=element_blank())
```

Boxplot construction



Interpreting boxplots

- Examine the length of the box to assess variability in the sample
  - The IQR is a measure of variability
- Identification of outliers is straightforward
- It is especially useful for the comparison of two samples

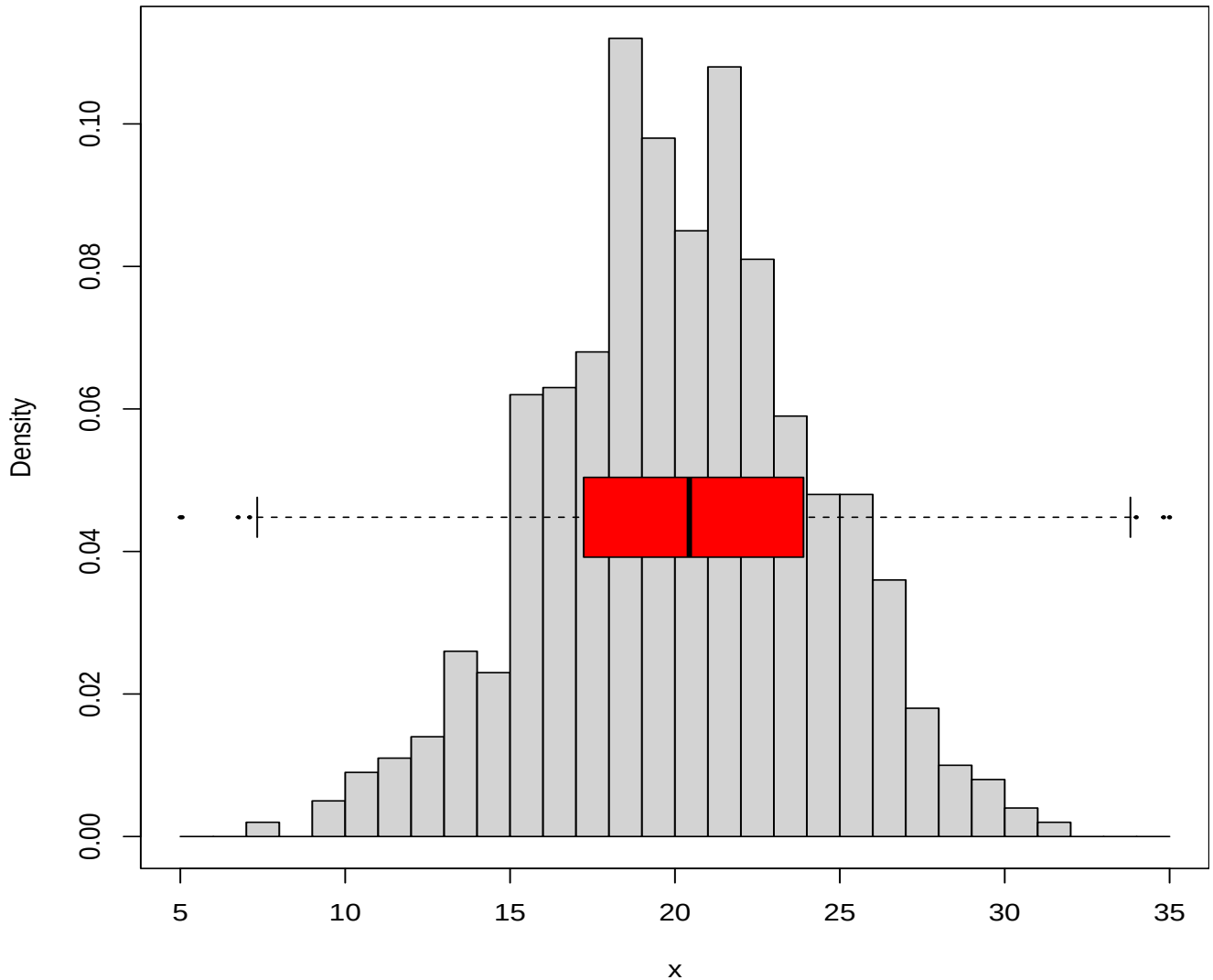
Indication of skewed data

- one whisker clearly extends further then the other
- the median line within the box is not equidistant from the hinges

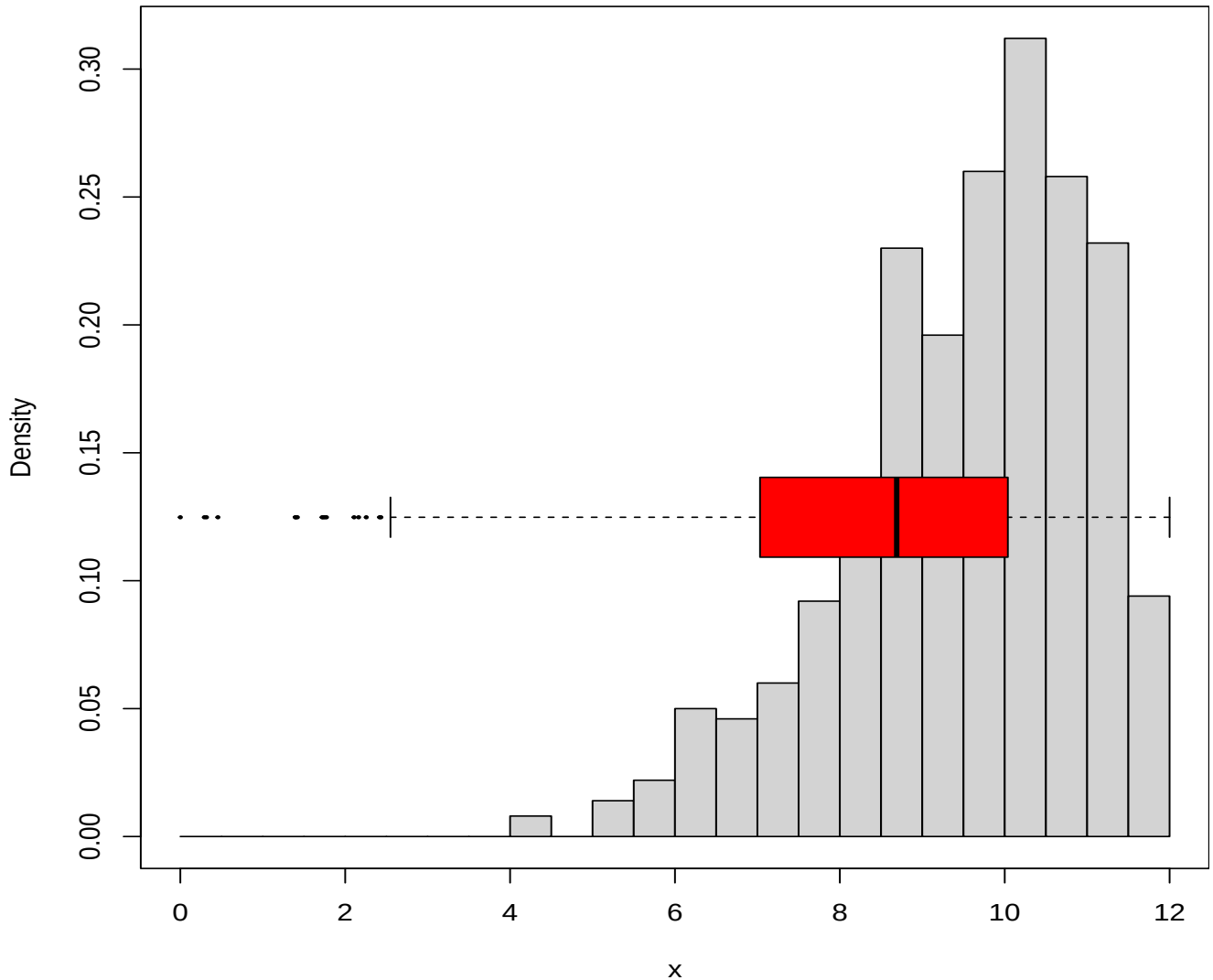
Boxplots vs Histograms

```
xv<-seq(5,35,by=1);par(mar=c(4,4,1,0))
hist(x,breaks=xv,right=FALSE,include.lowest=TRUE,main='',freq=FALSE);box()
par(new=TRUE)
boxplot(x,horizontal=TRUE,at=0.1,pch=19,cex=0.25,xlim=range(-1.5,2.5),axes=FALSE,col='red')
```

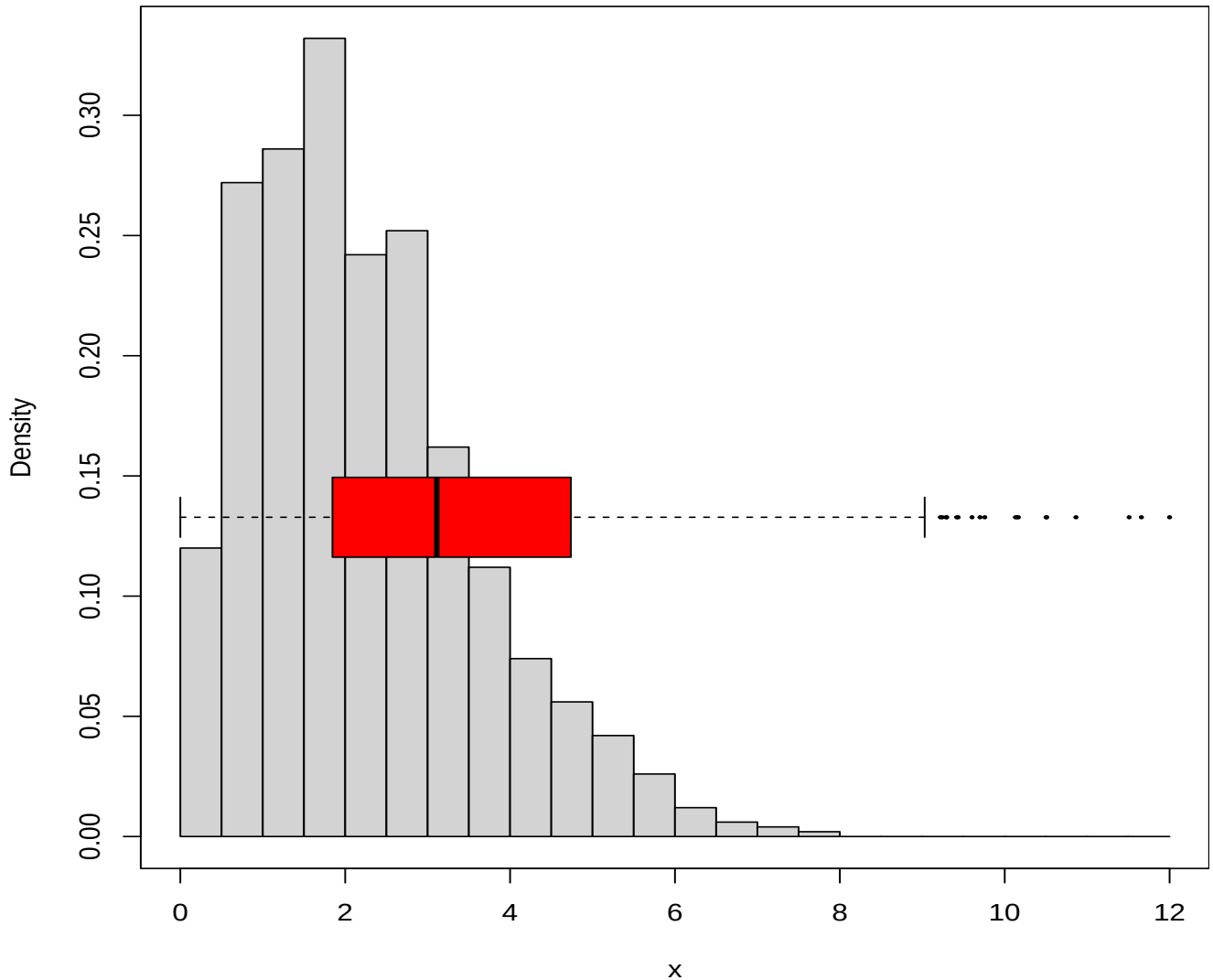
Boxplots vs Histograms



Boxplots vs Histograms



Boxplots vs Histograms



Boxplots vs Histograms

Example: Big Mac Index (The Economist, 2020)

The Economist magazine uses the prices of Big Macs in different countries to illustrate currency exchange concepts.

`https://www.economist.com/big-mac-index`

Boxplots vs Histograms

Example: Big Mac Index (The Economist, 2020)

```
BigMac<-read.csv('https://www.math.mcgill.ca/dstephens/203/Data/BigMac.csv')
head(BigMac)
```

```
+                Country Price
+ 1 United Arab Emirates   4.0
+ 2            Argentina   2.8
+ 3            Australia   4.5
+ 4           Azerbaijan   2.3
+ 5              Bahrain   3.7
+ 6               Brazil   4.8
```

```
n<-nrow(BigMac);n
```

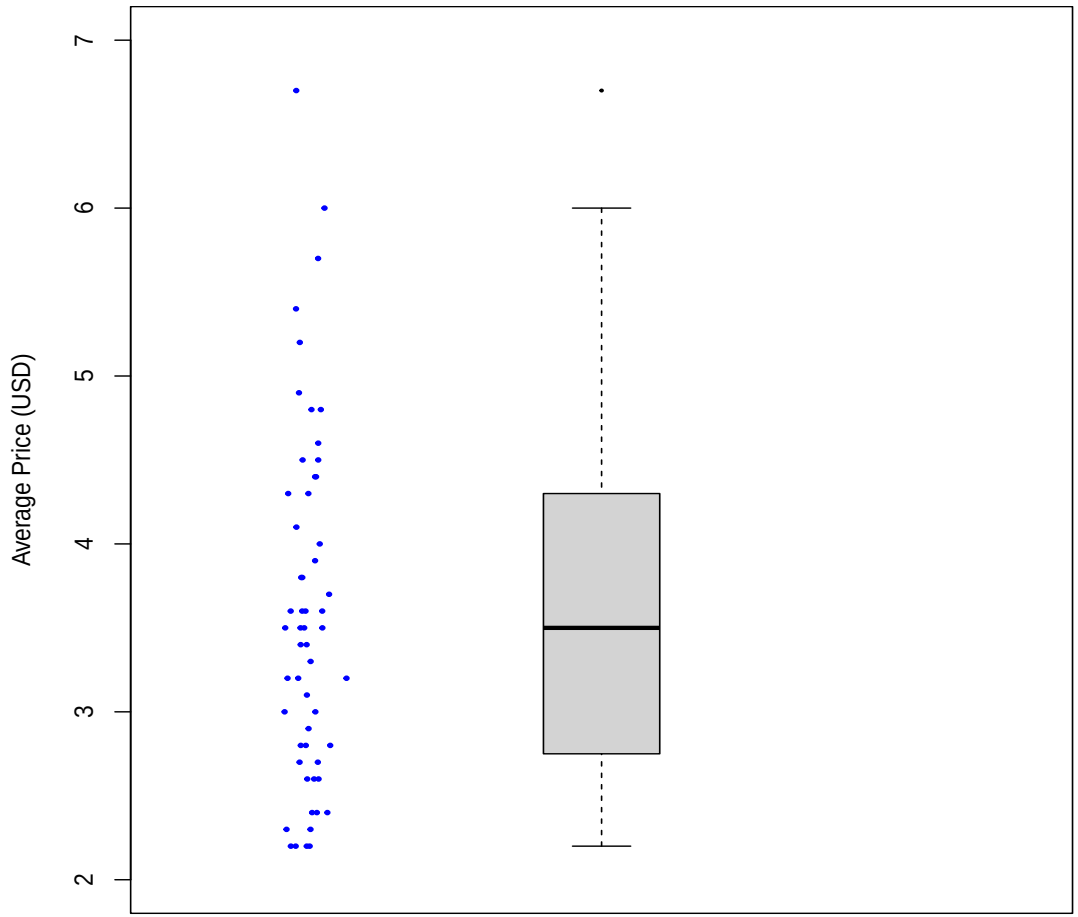
```
+ [1] 56
```

```
summary(BigMac$Price)
```

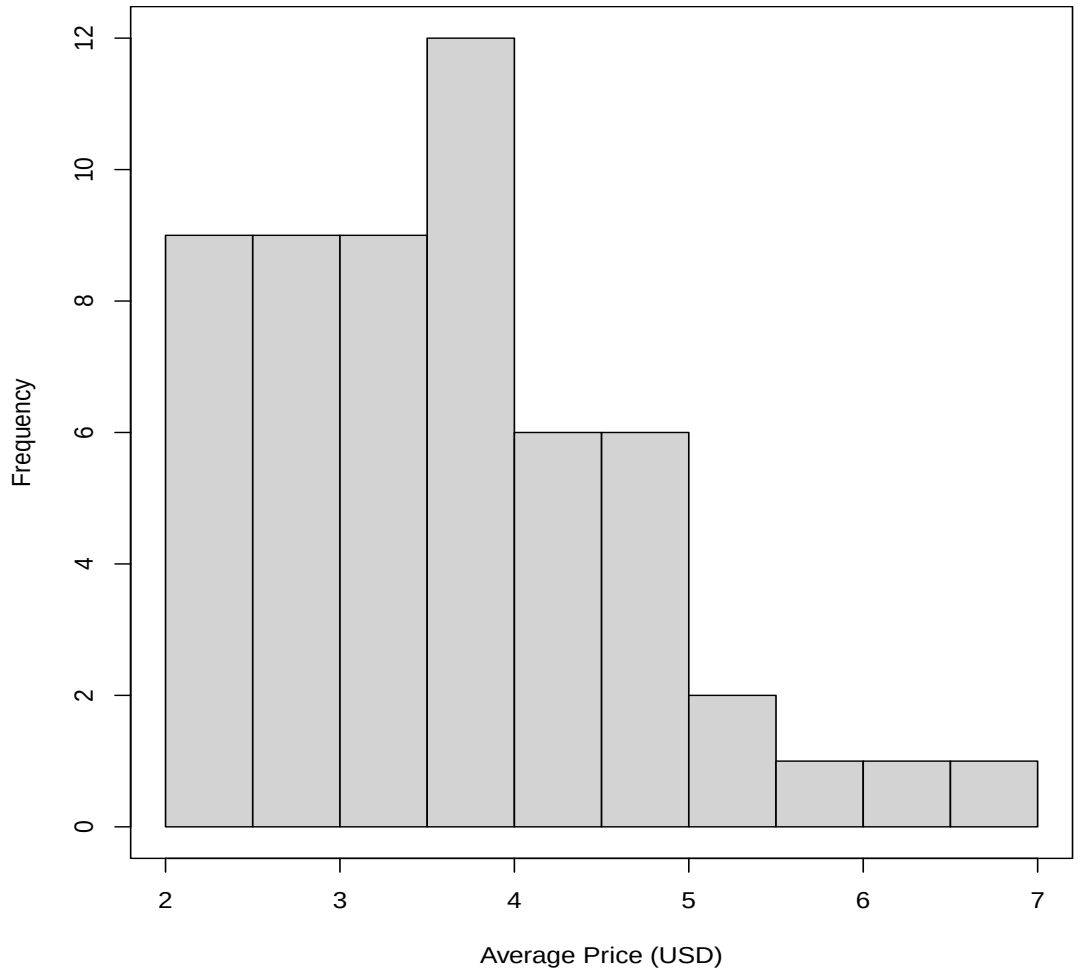
```
+    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
+   2.200   2.775   3.500   3.573   4.300   6.700
```

More data here

Boxplots vs Histograms



Boxplots vs Histograms

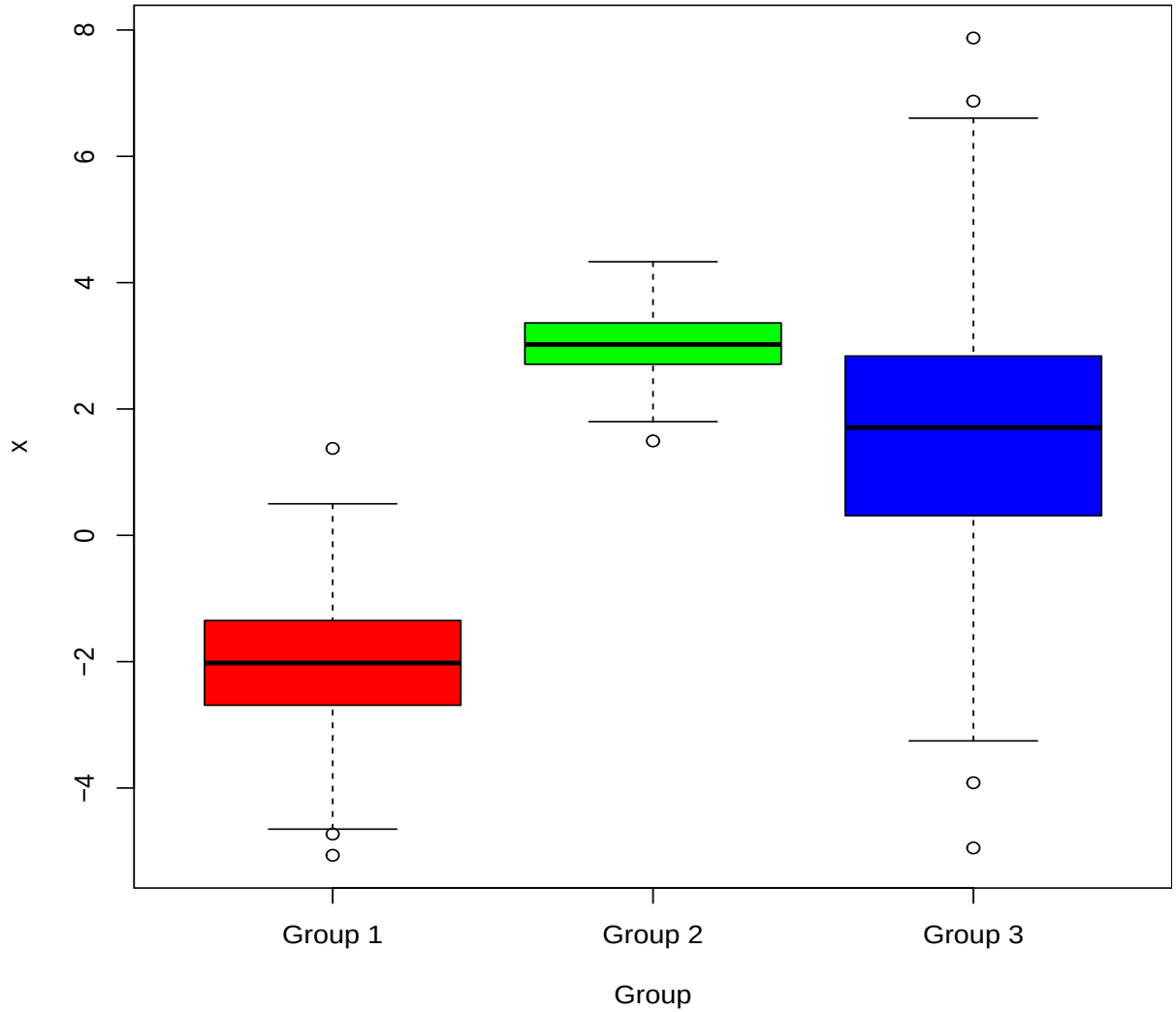


Comparing Distributions

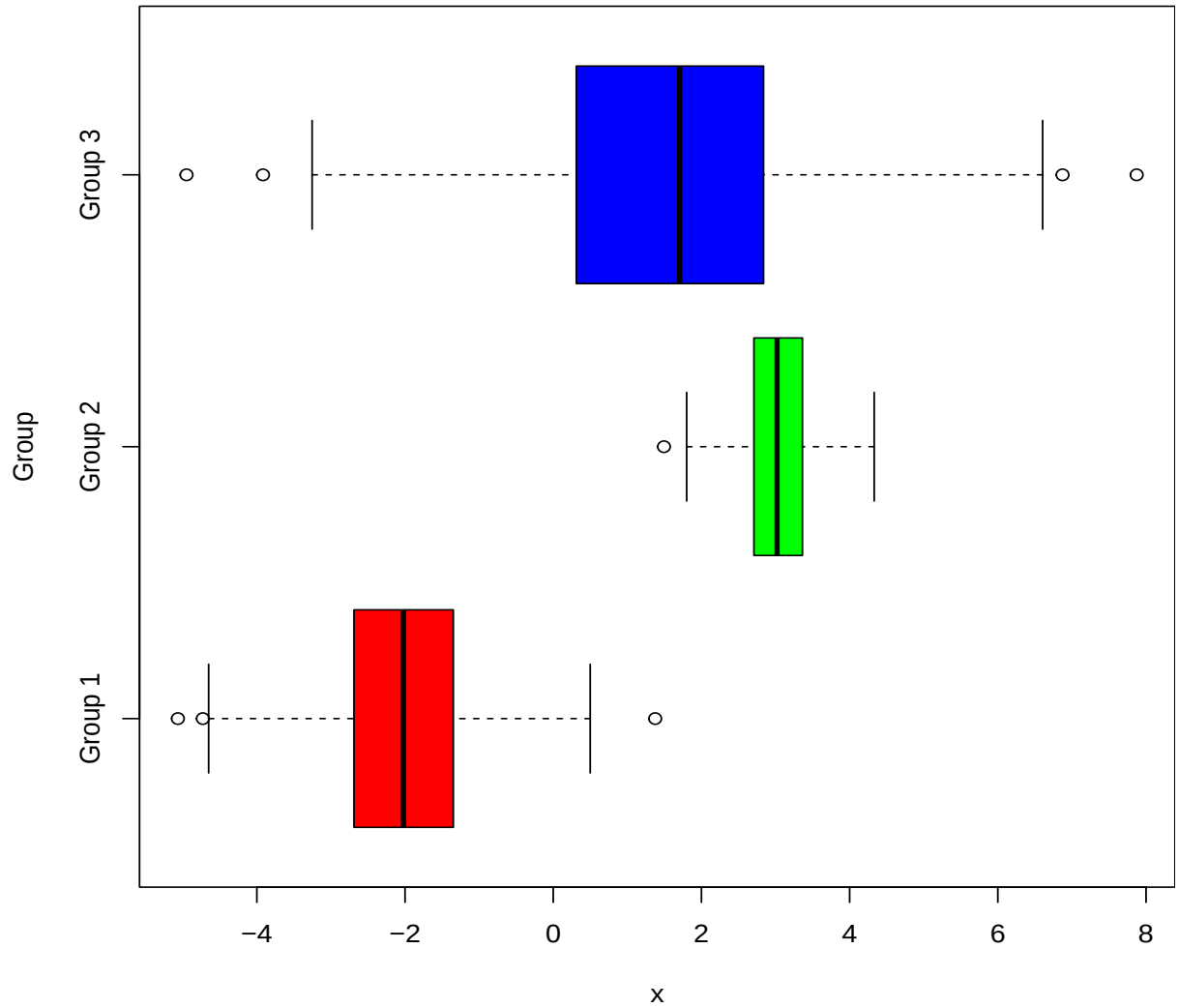
We can plot several boxplots for comparison of groups: we may examine

- central tendency
- spread/variability
- skewness
- outliers

Comparing Distributions



Comparing Distributions



Comparing Distributions

Example: 2014-2016 Ebola outbreak, West Africa

Ebola cases and deaths by country and by date (10 weeks)

Date	1	2	3	4	5	6	7	8	9	10
Guinea										
Cases	648	607	579	543	519	510	506	495	495	485
Deaths	472	460	427	415	410	411	406	409	408	412
Liberia										
Cases	1378	1082	972	834	786	670	599	554	516	468
Deaths	391	329	249	224	196	174	172	142	131	115
Sierra Leone										
Cases	1026	910	907	848	810	783	730	717	691	646
Deaths	574	533	525	454	442	397	386	337	305	252

Comparing Distributions

Example: 2014 Ebola outbreak

```
Guinea<-c(648,607,579,543,519,510,506,495,495,485,472,460,427,415,410,411,406,409,408,412)
Liberia<-c(1378,1082,972,834,786,670,599,554,516,468,391,329,249,224,196,174,172,142,131,115)
Sierra.L<-c(1026,910,907,848,810,783,730,717,691,646,574,533,525,454,442,397,386,337,305,252)
summary(Guinea[1:10])

+    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
+   485.0   497.8   514.5   538.7   570.0   648.0

summary(Liberia[1:10])

+    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
+   468.0   565.2   728.0   785.9   937.5  1378.0

summary(Sierra.L[1:10])

+    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
+   646.0   720.2   796.5   806.8   892.2  1026.0
```

Comparing Distributions

```
#For cases
c(mean(Guinea[1:10]),sd(Guinea[1:10]))

+ [1] 538.70000  55.01323

c(mean(Liberia[1:10]),sd(Liberia[1:10]))

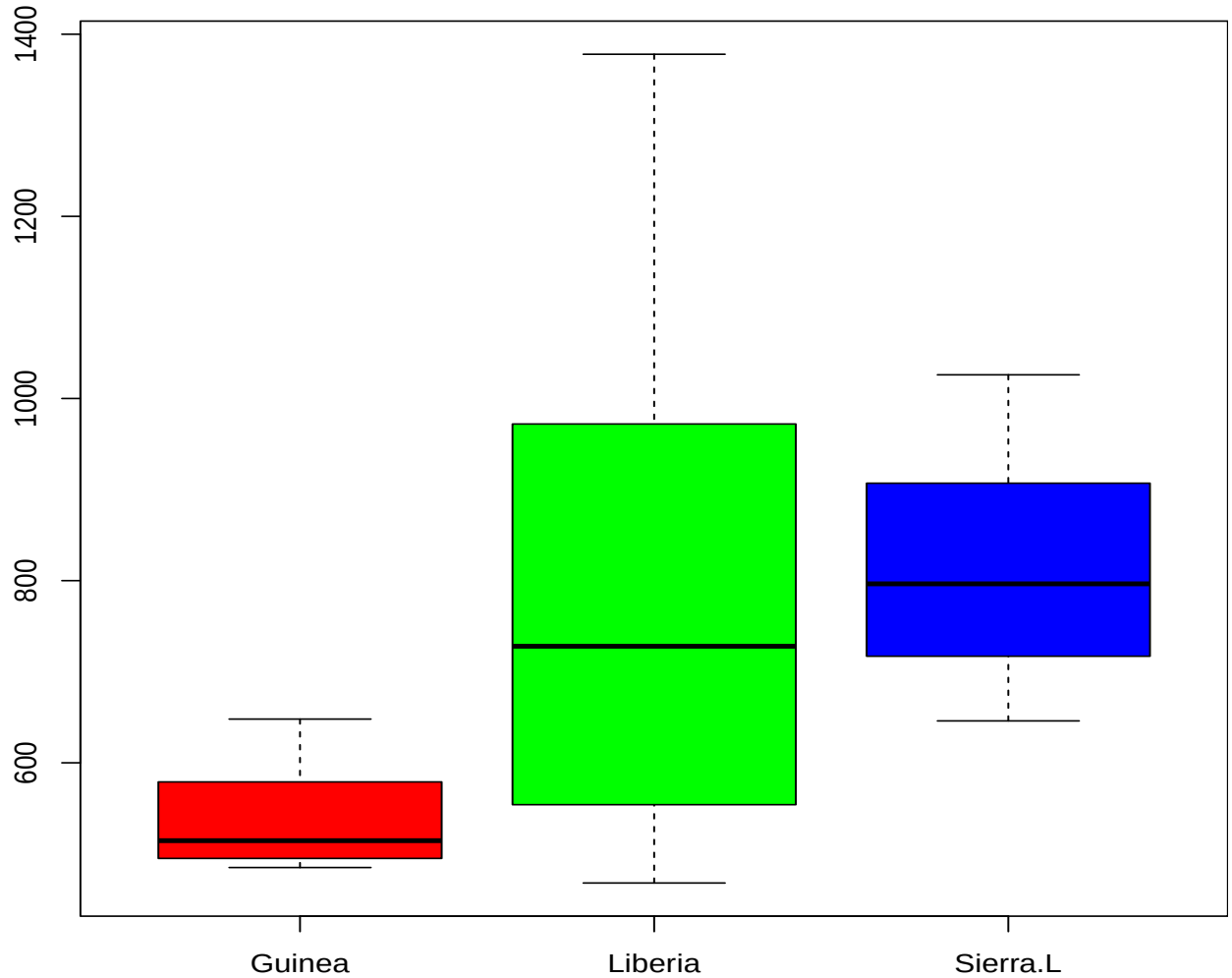
+ [1] 785.9000 288.9162

c(mean(Sierra.L[1:10]),sd(Sierra.L[1:10]))

+ [1] 806.8000 117.4931

par(mar=c(4,4,2,0))
boxplot(cbind(Guinea, Liberia, Sierra.L)[1:10,],col=c('red','green','blue'))
```

Comparing Distributions



Comparing Distributions

Note

There is an important variable that is not being represented in the boxplot: the variation across time:

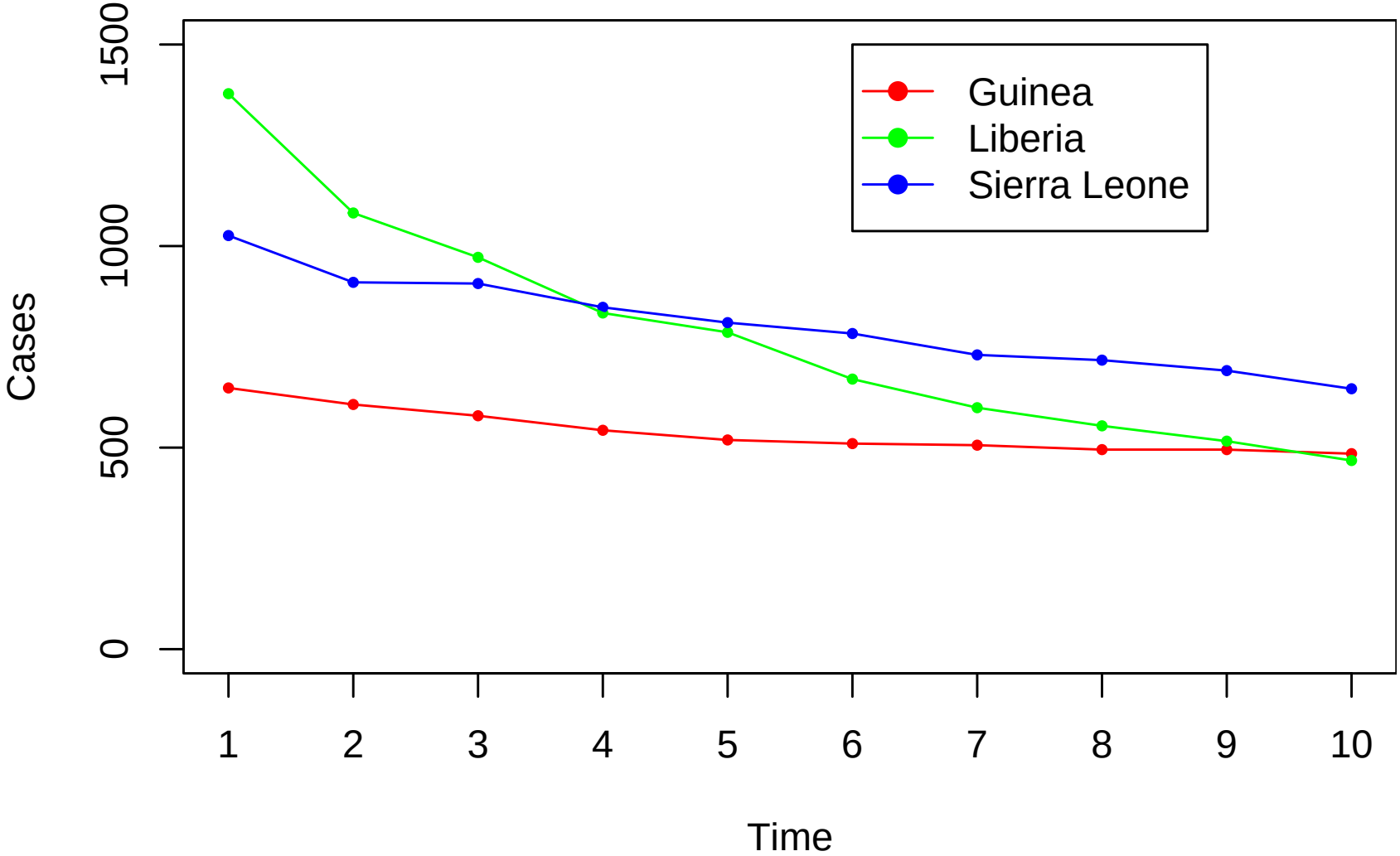
- In this infectious disease setting, we might expect there to be systematic changes across time;
- For each country, the observation at week 1 is not truly comparable with the observation at week 10.

It is also common to report death rate per capita (eg per 10000 individuals) to account for differences in the sizes of the countries.

Comparing Distributions

```
#For cases
par(mar=c(4,4,2,0))
plot(1:10,Guinea[1:10],pch=19,cex=0.5,col='red',ylim=range(0,1500),
        xlab='Time',ylab='Cases',xaxt='n')
axis(1, at = seq(1, 10, by = 1))
points(1:10,Liberia[1:10],pch=19,cex=0.5,col='green')
points(1:10,Sierra.L[1:10],pch=19,cex=0.5,col='blue')
lines(1:10,Guinea[1:10],col='red')
lines(1:10,Liberia[1:10],col='green')
lines(1:10,Sierra.L[1:10],col='blue')
legend(6,1500,c('Guinea','Liberia','Sierra Leone'),
        col=c('red','green','blue'),lty=1,pch=19)
```

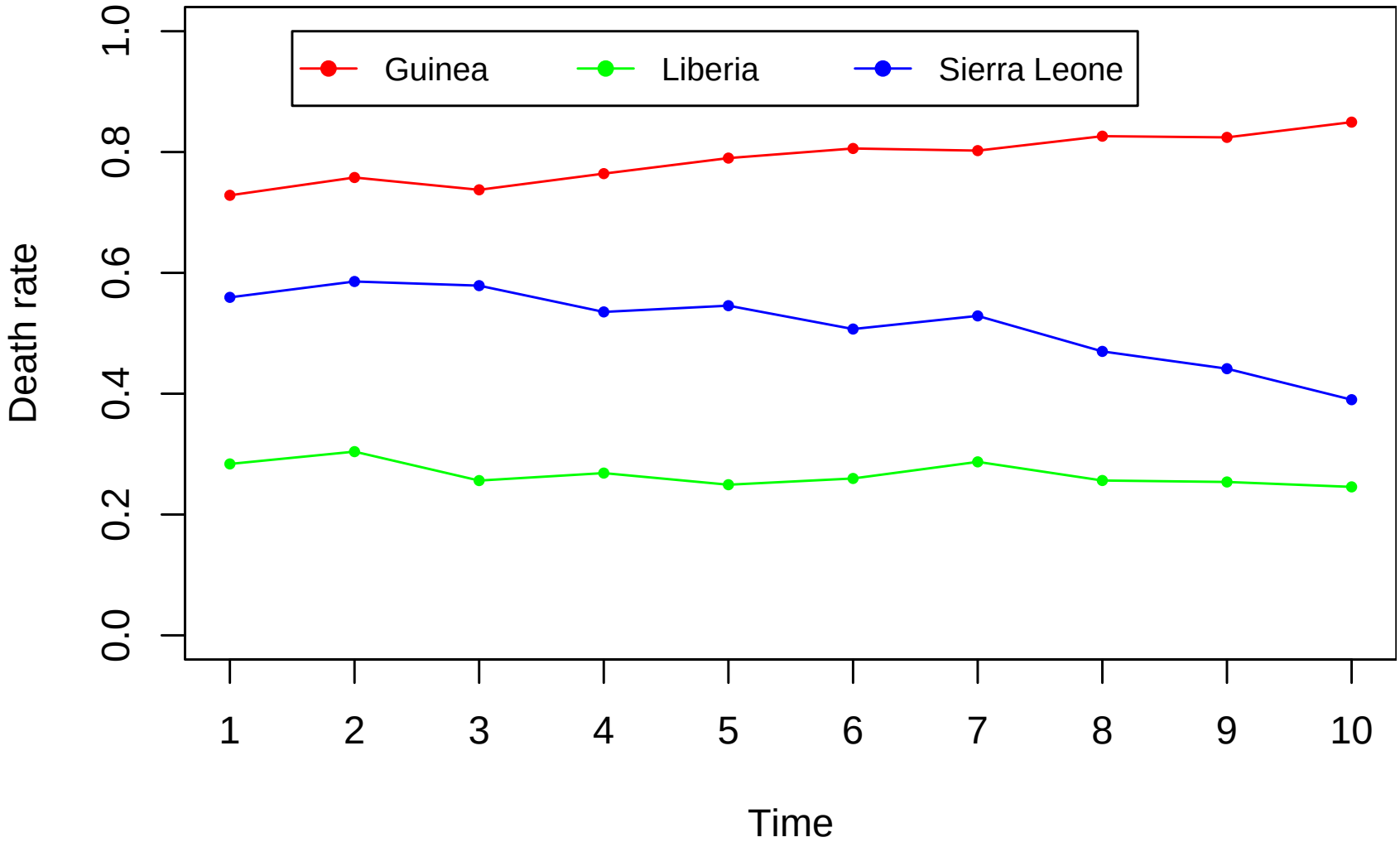
Comparing Distributions



Comparing Distributions

```
par(mar=c(4,4,2,0))
Guinea.rate<-Guinea[11:20]/Guinea[1:10]
Liberia.rate<-Liberia[11:20]/Liberia[1:10]
Sierra.L.rate<-Sierra.L[11:20]/Sierra.L[1:10]
plot(1:10,Guinea.rate,pch=19,cex=0.5,col='red',ylim=range(0,1),
        xlab='Time',ylab='Death rate',xaxt='n')
axis(1, at = seq(1, 10, by = 1))
points(1:10,Liberia.rate,pch=19,cex=0.5,col='green')
points(1:10,Sierra.L.rate,pch=19,cex=0.5,col='blue')
lines(1:10,Guinea.rate,col='red')
lines(1:10,Liberia.rate,col='green')
lines(1:10,Sierra.L.rate,col='blue')
legend(1.5,1,c('Guinea','Liberia','Sierra Leone'),ncol=3,
        col=c('red','green','blue'),lty=1,pch=19,cex=0.8)
```

Comparing Distributions



Chapter 2
Introduction to Probability

Probability: big idea

In statistical studies, we collect data to learn about real-world phenomena.

However, we recognize that as we only collect a finite sample of data, this cannot tell us everything about the system being studied.

Therefore, we need to develop tools to allow us to talk about the "uncertainty" associated with this lack of complete information.

Probability: big idea

Also, we recognize that when we collect data, they are subject to variability

- measurement variability
- variability due to small changes in experimental conditions.

We want to describe the nature of this variability.

Probability: big idea

Example: Repeated coin tosses

Five sequences of 10 tosses of a fair coin

```
+             1   2   3   4   5   6   7   8   9   10
+ Sequence 1 "T" "T" "T" "T" "H" "H" "H" "H" "H" "H"
+ Sequence 2 "H" "T" "T" "T" "H" "T" "H" "H" "H" "H"
+ Sequence 3 "T" "T" "T" "T" "H" "H" "H" "T" "T" "T"
+ Sequence 4 "T" "H" "T" "H" "T" "H" "H" "T" "T" "T"
+ Sequence 5 "T" "H" "T" "T" "H" "T" "T" "T" "T" "H"
```

The number of Heads recorded is different for each sequence; this ‘outcome’ is variable.

Uncertainty

Uncertainty corresponds to a lack of perfect knowledge; it could be a consequence of

- incomplete observation of a system;
- unpredictable variation;
- simple lack of knowledge of the "state of nature", such as
  - the current state that is imperfectly observed, or
  - the future state, that is, the result of a study yet to be carried out.

Uncertainty

Example: Uncertain states of nature

- the outcome of a single coin toss;
- the millionth digit of π;

  $$3.14159265358979323846264338327950288419 71....$$

- the height of the building I am in;
- the number of people on campus now;
- the temperature at noon tomorrow.

Uncertainty

Note

The millionth digit of π is a fixed number: however, unless we know its value, we are still in a state of uncertainty when asked to assess it, as we have a lack of perfect knowledge.

Uncertainty

Example: Coin tossing

If I take a coin and decide to carry out an experiment where I toss the coin once and see which face is turned upward when the coin comes to rest, then I can assess that there are two possible results of the toss:





The outcome is of the toss is uncertain before I toss the coin, and after I toss the coin until I see the result.

Uncertainty

Example: Thumbtack tossing

If I take a thumbtack and decide to carry out an experiment where I toss the thumbtack once, then I can assess that there are two possible results of the toss:





The outcome of the toss is uncertain before I toss the thumbtack, and after I toss the thumbtack until I see the result.

What is probability ?

By probability, we generally mean a numerical assessment of

> *the chance of a particular event occurring, given a particular set of circumstances*.

- *the chance of a particular event occurring ...*
- *... given a particular set of circumstances.*

Constructing the mathematical framework

We will use the term experiment to mean some situation that generates multiple possible outcomes

- we carry out an operation, make a measurement etc.

For a given experiment, we aim to

- identify the possible outcomes;
- assign numerical values to collections of possible values to represent the chance that they will coincide with the actual outcome;
- lay out the rules for how the numerical values can be manipulated.

Basic concepts in set theory

A set S is a collection of individual elements, such as s.

We write

$$s \in S$$

to mean "s is one of the elements of S".

- $S = \{0, 1\}$; $0 \in S, 1 \in S$.
- $S = \{\text{“dog”}, \text{“cat”}, \text{“mouse”}, \text{“rat”}\}$; “rat” $\in S$
- $S = [0, 1]$ (the closed interval from zero to one); then

$$0.3428 \in S$$

Basic concepts in set theory

S may be

- finite if it contains a finite number of elements;
  - e.g. $\{1, 2, 3, 4, 5\}$, $\{\text{Science}, \text{Arts}, \text{Engineering}\}$.

- countable if it contains an infinite number of elements that can be listed;
  - e.g. $\{0, 1, 2, 3, 4, 5, \ldots\}$.

- uncountable if it contains an infinite number of elements that cannot be listed;
  - e.g. $\{\text{All values on a continuum between 0 and 100}\}$.

Basic concepts in set theory

A is a subset of S if it contains some, all, or none of the elements of S. We write

$$\text{some: } A \subset S \qquad \text{some or all: } A \subseteq S.$$

- $S = \{0, 1\}$; $A = \{0\}$, or $A = \{1\}$, or $A = \{0, 1\}$.
- $S = \{\text{“dog”}, \text{“cat”}, \text{“mouse”}, \text{“rat”}\}$; $A = \{\text{“cat”}, \text{“mouse”}\}$
- $S = [0, 1]$; $A = (0.25, 0.3)$.

The empty set, $\varnothing$, is a subset that contains no elements.

Set operations

To manipulate sets, we use three basic operations:

- intersection $\cap$
- union $\cup$
- complement A^c

Set operations

- intersection: the intersection of two sets A and B is the collection of elements that are elements of both A and B.

$$s \in A \cap B \qquad \text{means} \qquad s \in A \text{ and } s \in B.$$

We have for any $A, B \subseteq S$ that

$$A \cap \varnothing = \varnothing$$

$$A \cap S = A$$

$$A \cap B \subseteq A$$

$$A \cap B \subseteq B.$$

Set operations

- union: the union of two sets A and B is the set of distinct elements that are either in A, or in B, or in both A and B.

  $s \in A \cup B$ means $s \in A$, or $s \in B$, or $s \in A \cap B$.

  We have for any $A, B \subseteq S$ that

$$A \cup \varnothing = A$$

$$A \cup S = S$$

$$A \subseteq A \cup B$$

$$B \subseteq A \cup B$$

Set operations

- complement: the complement of set A is the collection of elements of S that are not elements of A.

$$s \in A^c \qquad \text{means} \qquad s \in S \text{ but } s \notin A.$$

We have that

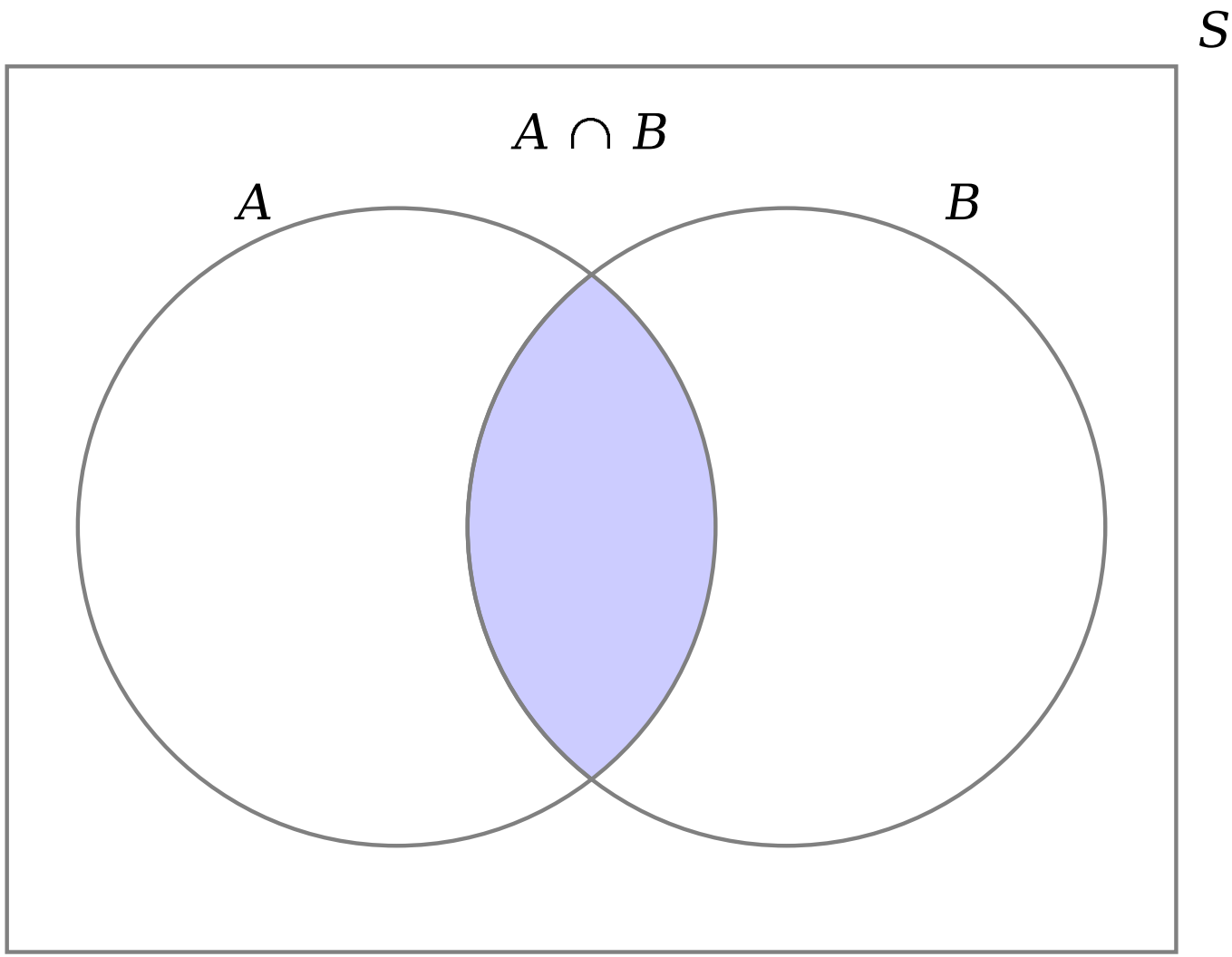
$$A \cap A^c = \varnothing \qquad \text{and} \qquad A \cup A^c = S.$$

We have for any A that

$$(A^c)^c = A.$$

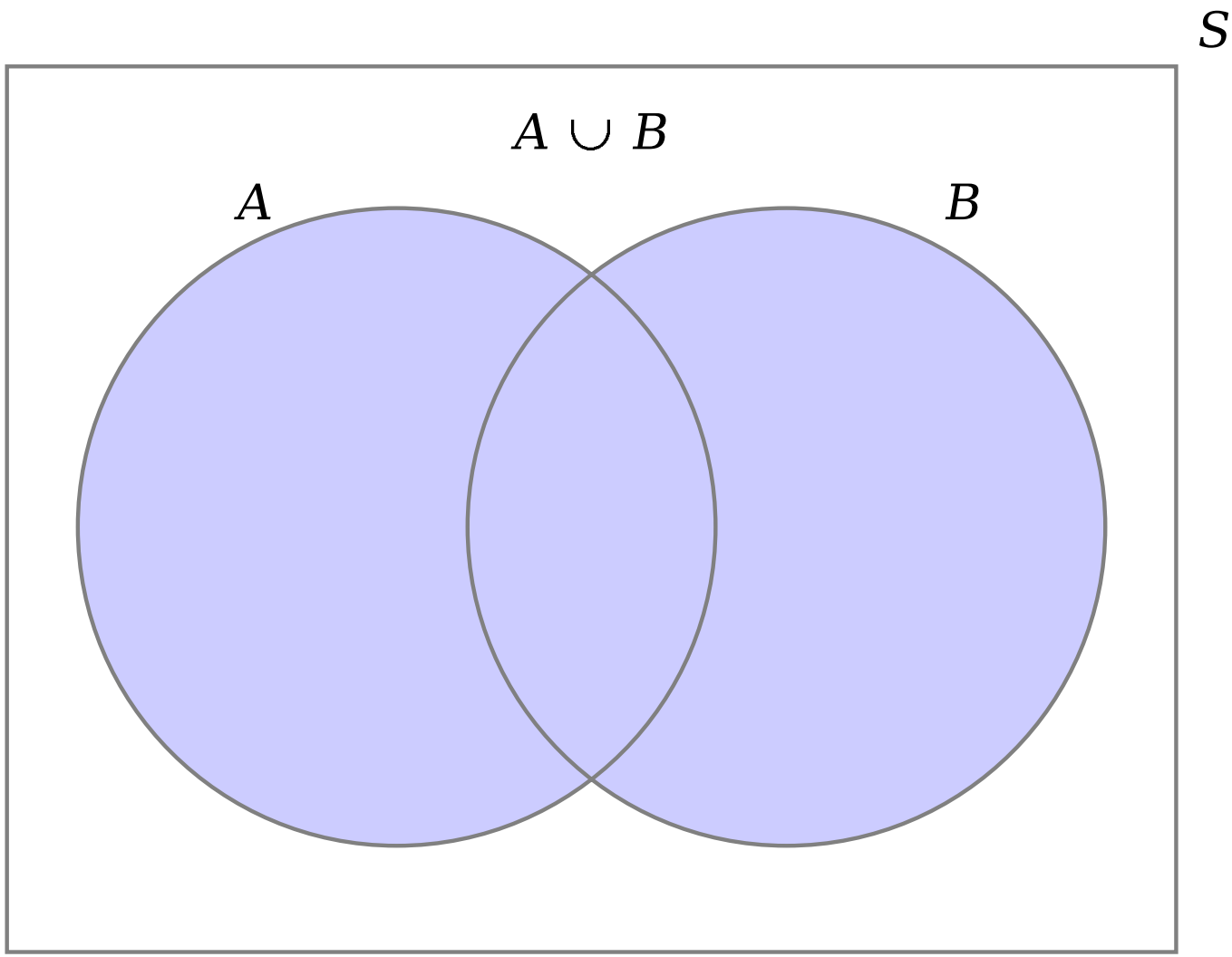
Set operations: in pictures

Intersection: blue shaded area is $A \cap B$.



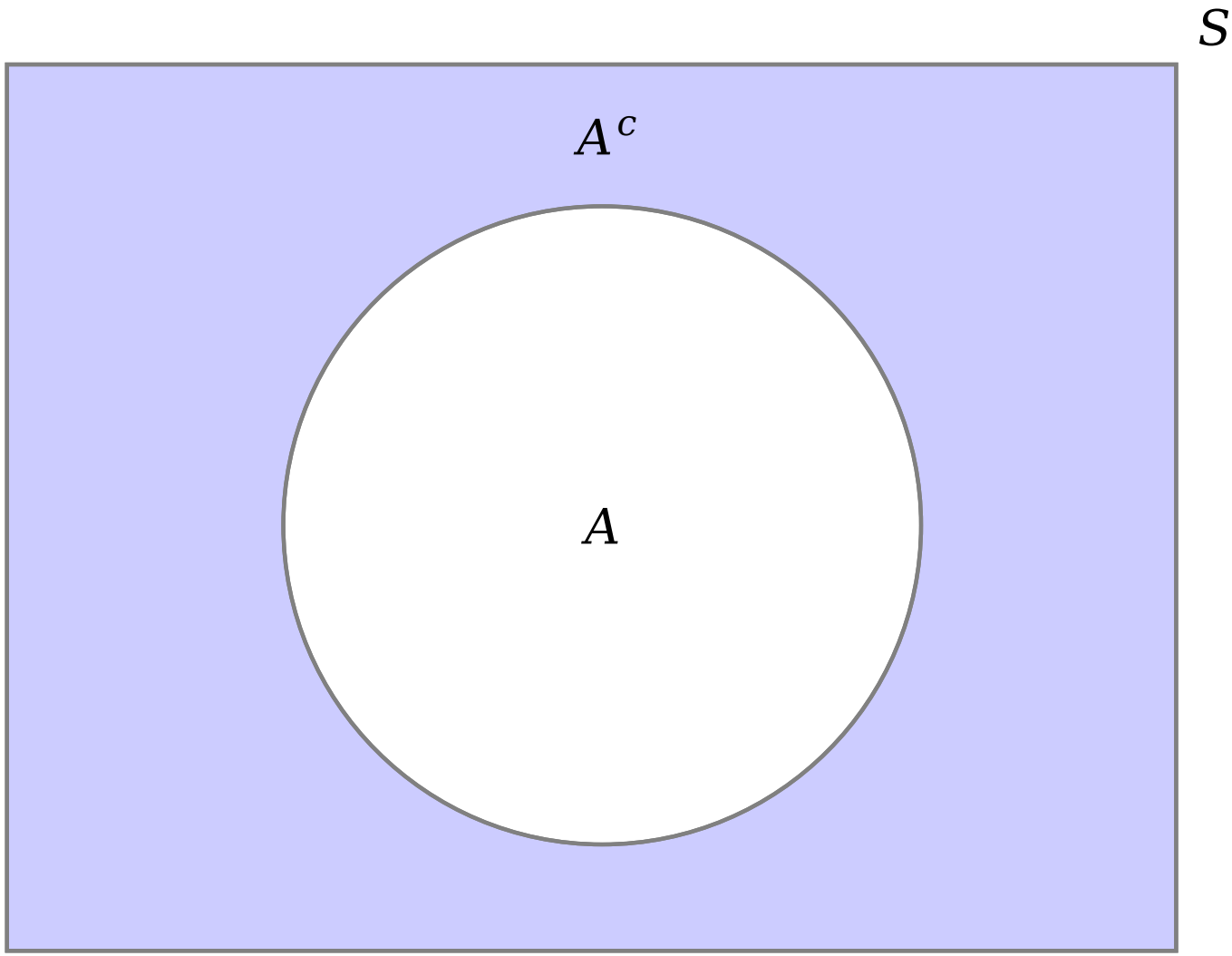
Set operations: in pictures

Union: blue shaded area is $A \cup B$.



Set operations: in pictures

Complement: blue shaded area is A^c



Example

Example: Finite set

- $S = \{1, 2, 3, \ldots, 9, 10\}$
- $A = \{2, 4, 6\}$
- $B = \{1, 2, 5, 7, 9\}$

Then

- $A \cap B = \{2\}$
- $A \cup B = \{1, 2, 4, 5, 6, 7, 9\}$
- $A^c = \{1, 3, 5, 7, 8, 9, 10\}$

Extensions

- Intersections:

$$(A \cap B) \cap C = A \cap B \cap C$$

- Unions:

$$(A \cup B) \cup C = A \cup B \cup C$$

- Intersections and Unions:

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$

These can be extended to any number of events.

Partitions

We call $A_1, A_2, \ldots, A_K$ a partition of S if these sets are

(i) pairwise disjoint (or mutually exclusive):

$$A_j \cap A_k = \varnothing \quad \text{for all } j \neq k$$

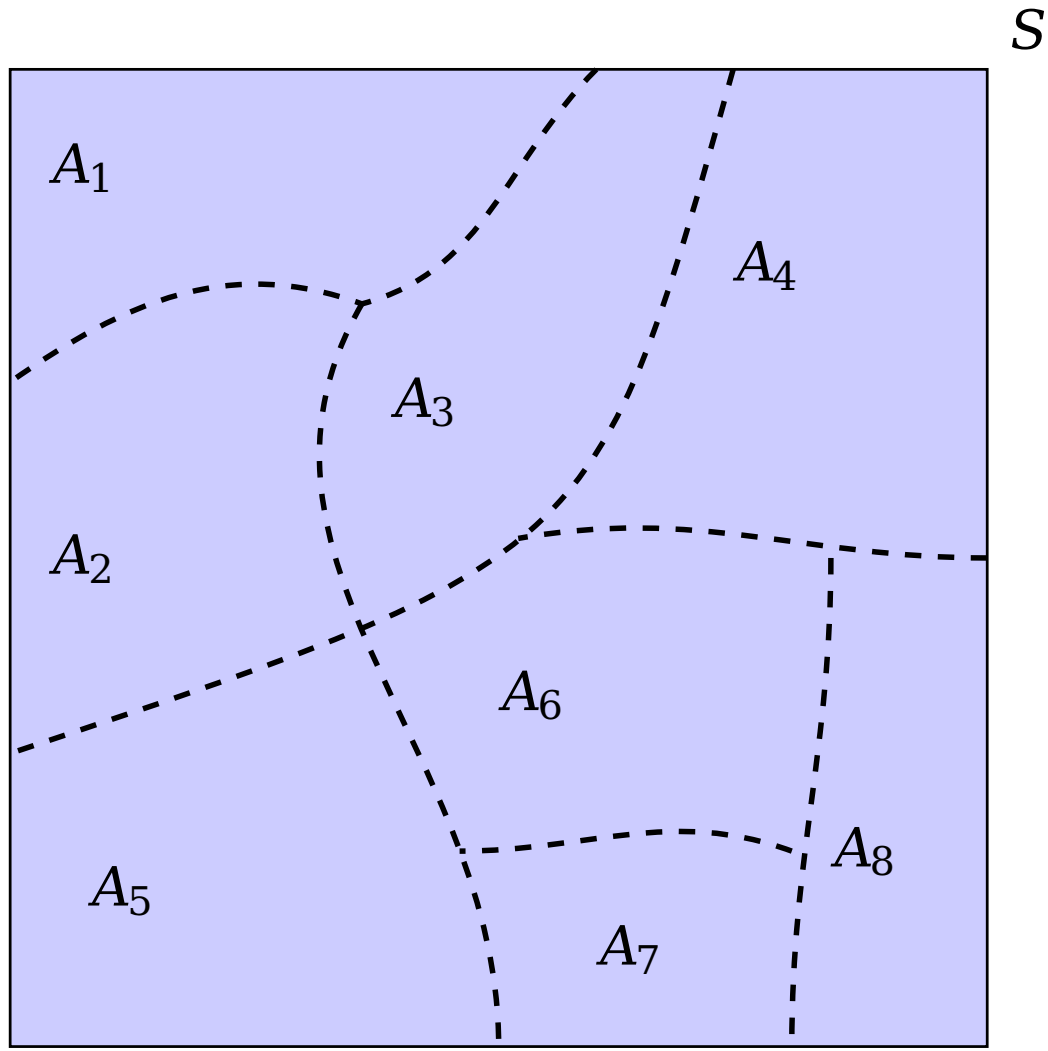
(ii) exhaustive:

$$\bigcup_{k=1}^{K} A_k = S.$$

(i.e. the As cover the whole of S, but do not overlap).

Every $s \in S$ is an element of precisely one of the A_k.

Partitions

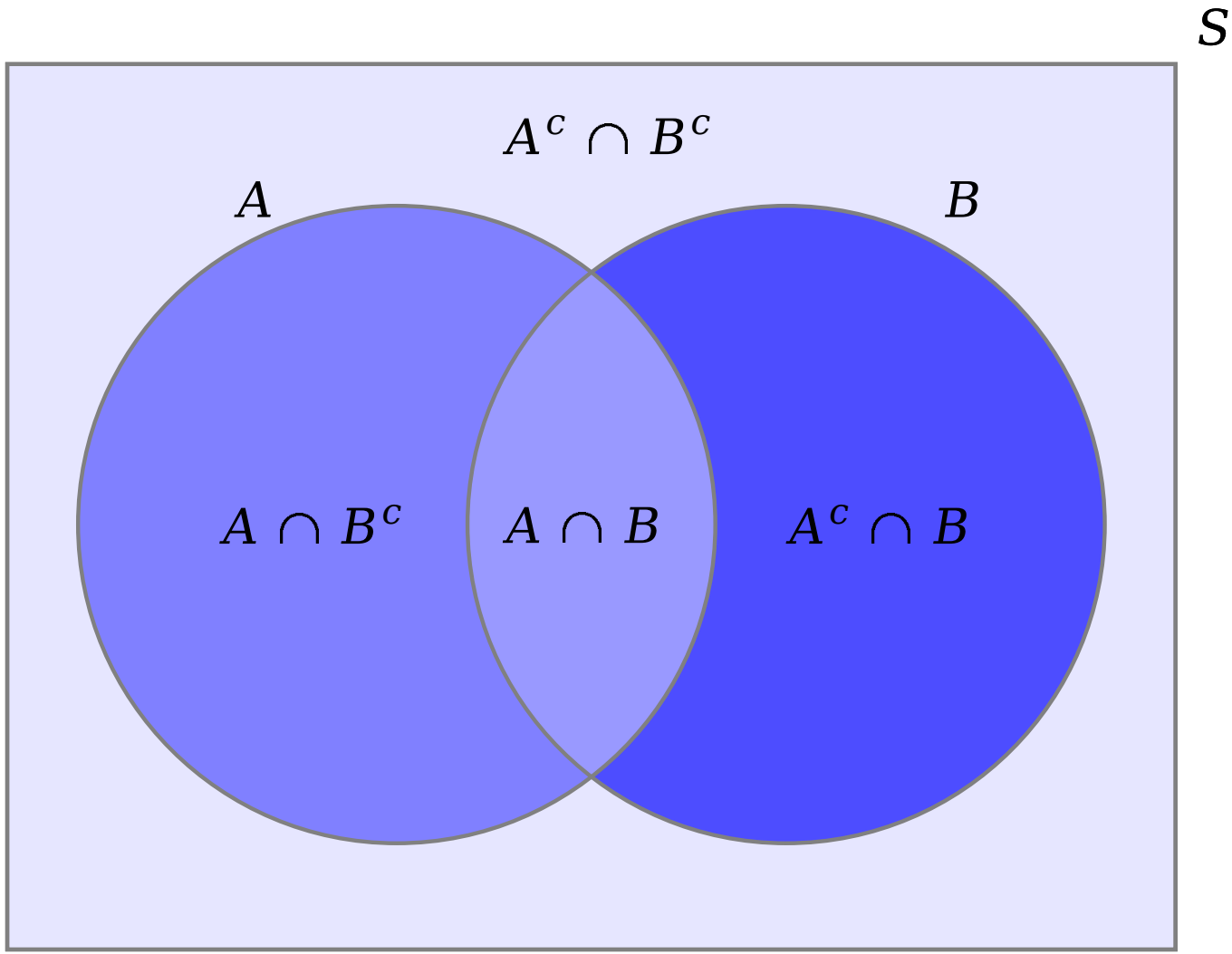


Some simple results

For any two subsets A and B of S, we have the partition of S via the four sets

$$A \cap B \qquad A \cap B^c \qquad A^c \cap B \qquad A^c \cap B^c$$

Some simple results



Some simple results

This picture tells us that

$$(A \cup B)^c = A^c \cap B^c.$$

- $(A \cup B)$: three darker blue regions
- $A^c \cap B^c$: lightest blue region

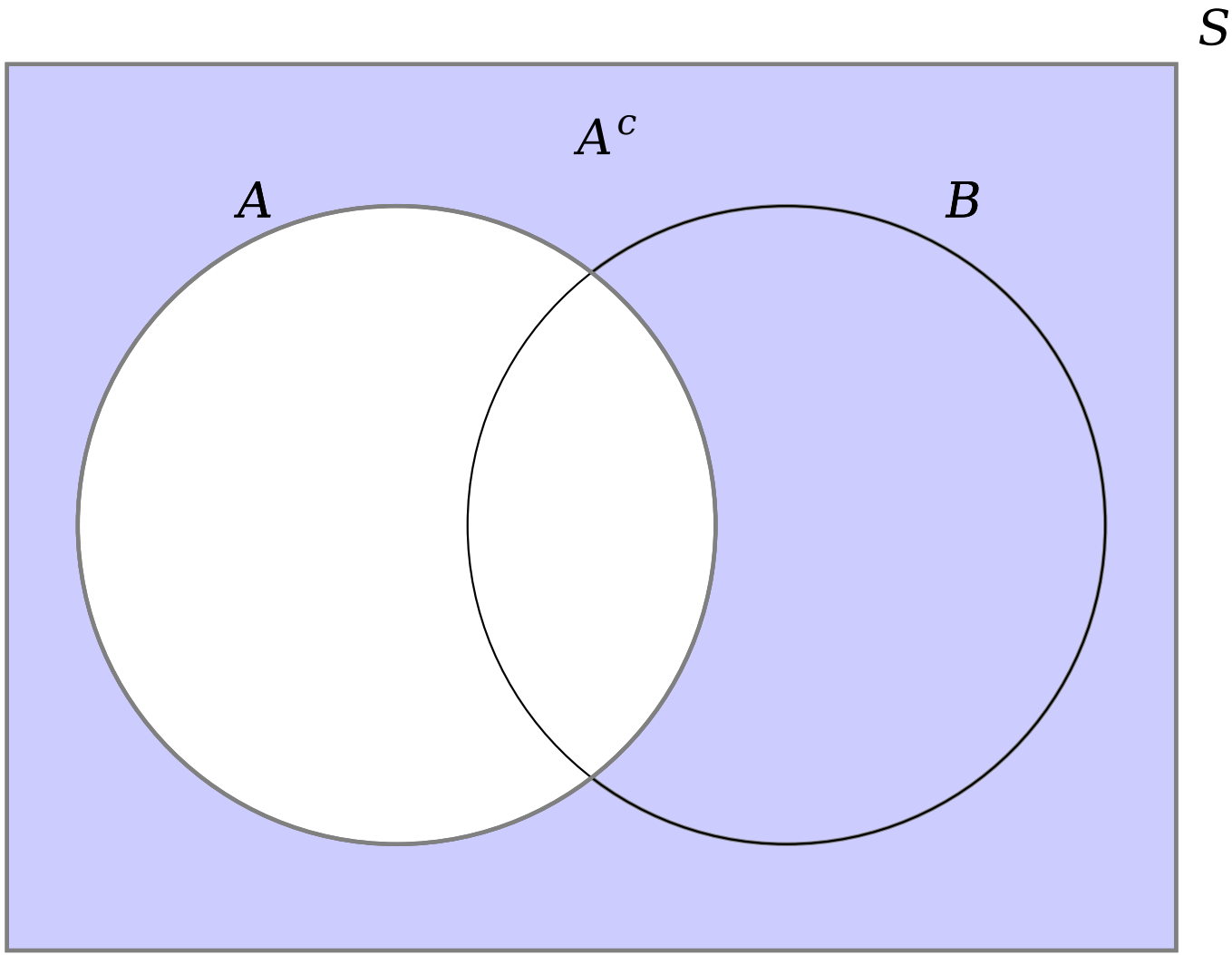
Similarly

$$A^c \cup B^c = (A \cap B)^c$$

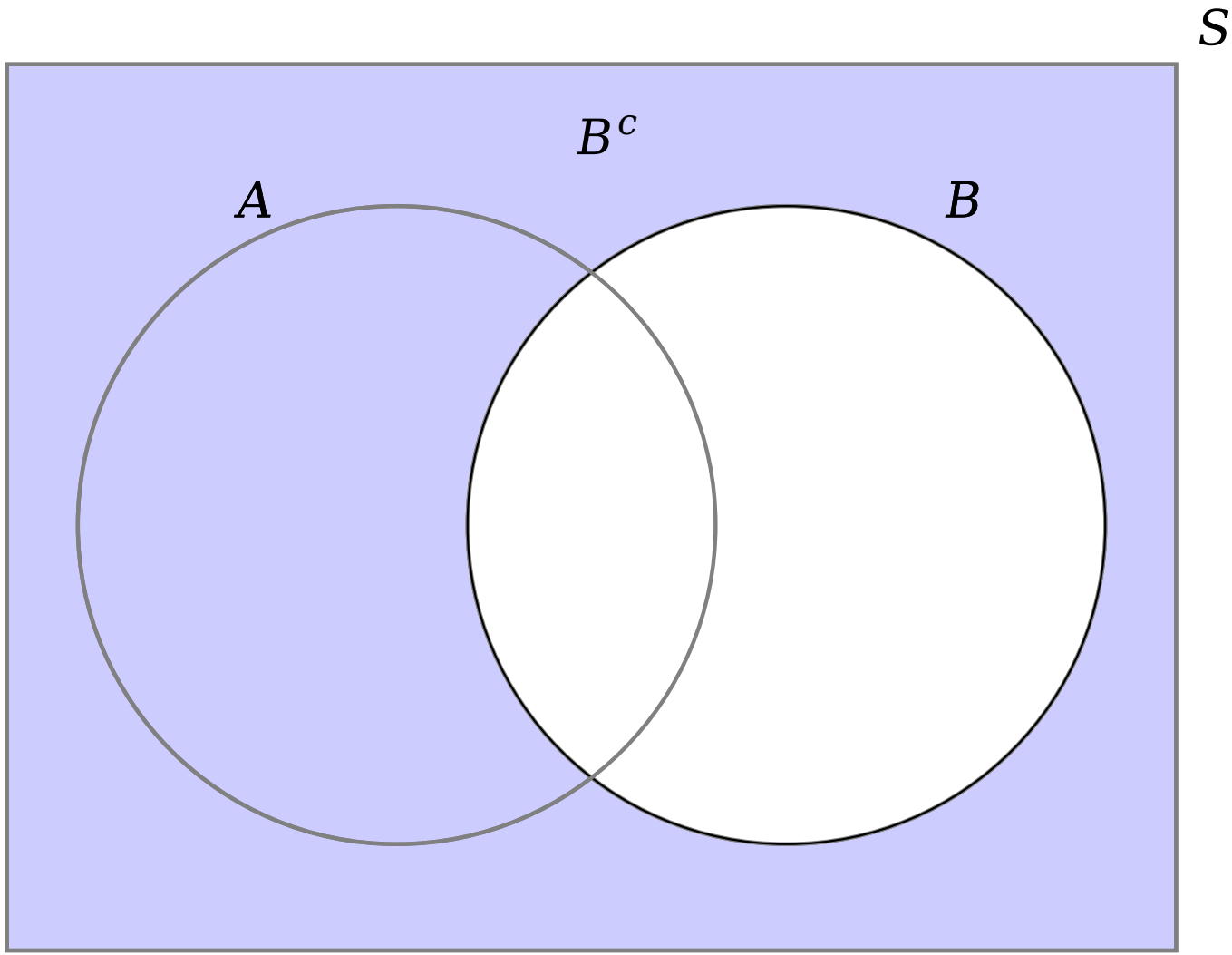
so that

$$(A^c \cup B^c)^c = A \cap B.$$

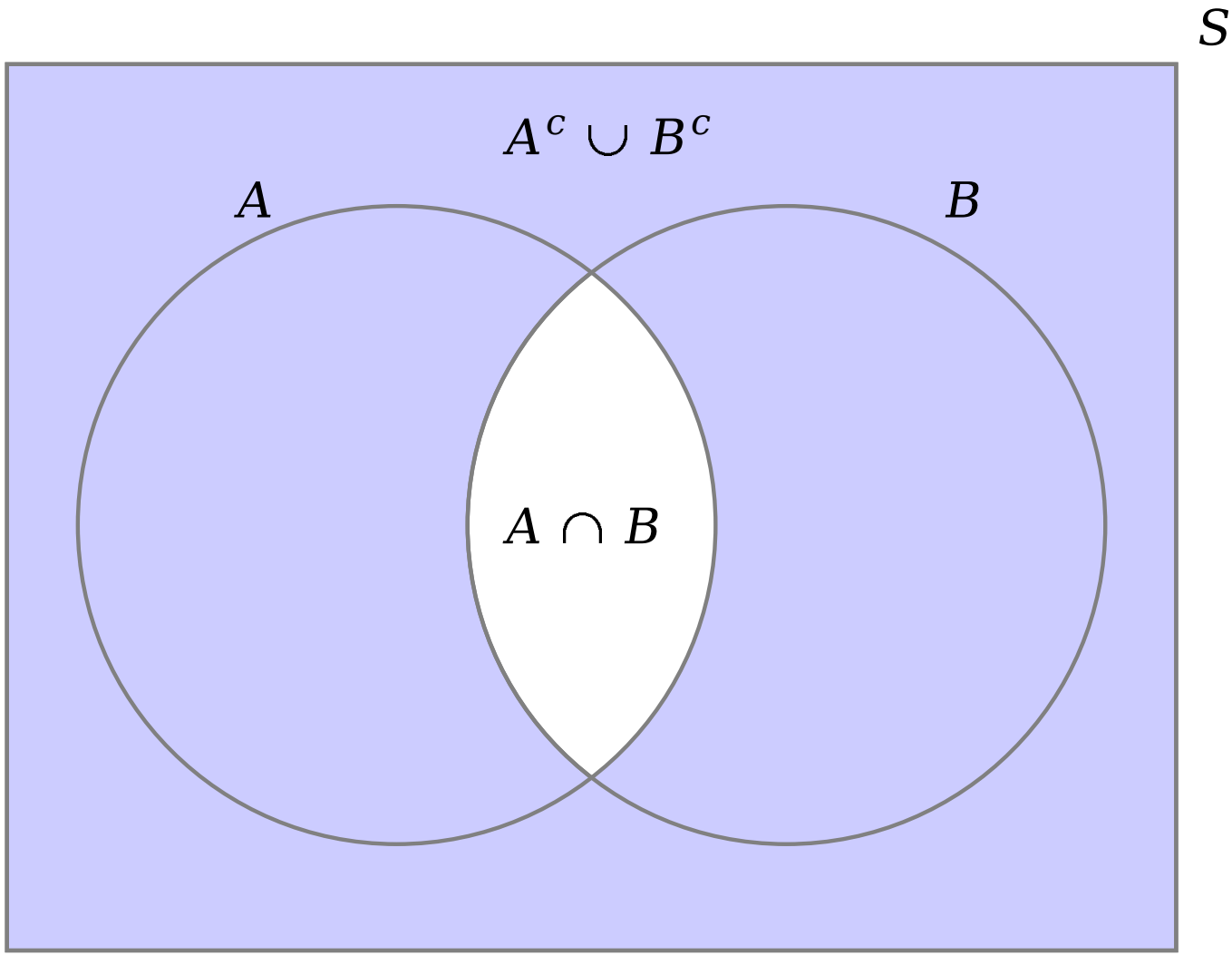
Some simple results



Some simple results



Some simple results



Some simple results

These two results

$$A^c \cap B^c = (A \cup B)^c$$

$$A^c \cup B^c = (A \cap B)^c$$

are sometimes known as de Morgan's Laws.

Sample spaces and events

Recall the earlier informal definition:

> *By probability, we generally mean the chance of a particular event occurring, given a particular set of circumstances. The probability of an event is generally expressed as a quantitative measurement.*

We need to carefully define what an 'event' is, and what constitutes a 'particular set of circumstances'.

Sample spaces and events

We will consider the general setting of an experiment:

- this can be interpreted as any setting in which an uncertain consequence is to arise;
- could involve observing an outcome, taking a measurement etc.

Constructing the framework

1. **Consider the possible outcomes of the experiment:** Identify all the outcomes that can arise, and denote the corresponding set by S.
   - $S = \{0, 1\}$;
   - $S = \{\text{“head”}, \text{“tail”}\} \equiv \{\text{H}, \text{T}\}$;
   - $S = \{\text{“cured”}, \text{“not cured”}\}$;
   - $S = \{\text{“Arts”}, \text{“Engineering”}, \text{“Medicine”}, \text{“Science”}\}$;
   - $S = \{1, 2, 3, 4, 5, 6\}$;
   - $S = \mathbb{R}^+$.

   The set S is termed the sample space of the experiment. The individual elements of S are termed sample points (or sample outcomes).

Constructing the framework

2. **Events**: An event A is a collection of sample outcomes. That is, A is a subset of S, $A \subseteq S$.

   For example

   - $A = \{0\}$;
   - $A = \{\text{“tail”}\} \equiv \{\text{T}\}$;
   - $A = \{\text{“cured”}\}$;
   - $A = \{\text{“Arts”}, \text{“Engineering”}\}$;
   - $A = \{1, 3, 5\}$;
   - $A = [2, 3)$.

   The individual sample outcomes are termed simple (or elementary) events, and may be denoted

$$E_1, E_2, \ldots, E_K, \ldots.$$

Constructing the framework

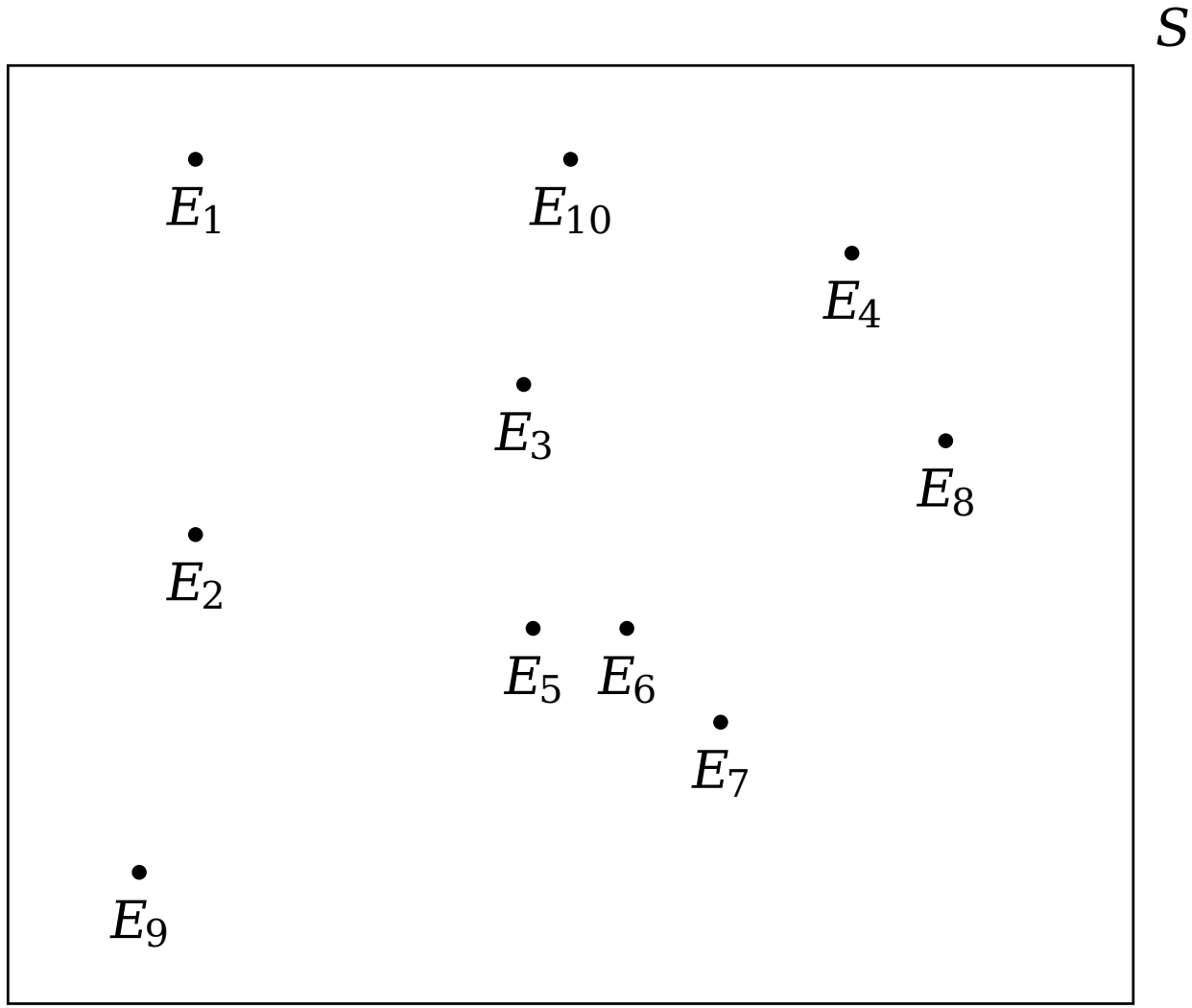
3. **Terminology**: We say that event A occurs if the actual outcome, s, is an element of A.

   For two events A and B

   - $A \cap B$ occurs if and only if A occurs and B occurs, that is
   $$s \in A \cap B.$$
   - $A \cup B$ occurs if A occurs or if B occurs, or if both A and B occur, that is
   $$s \in A \cup B$$
   - if A occurs, then A^c does not occur.

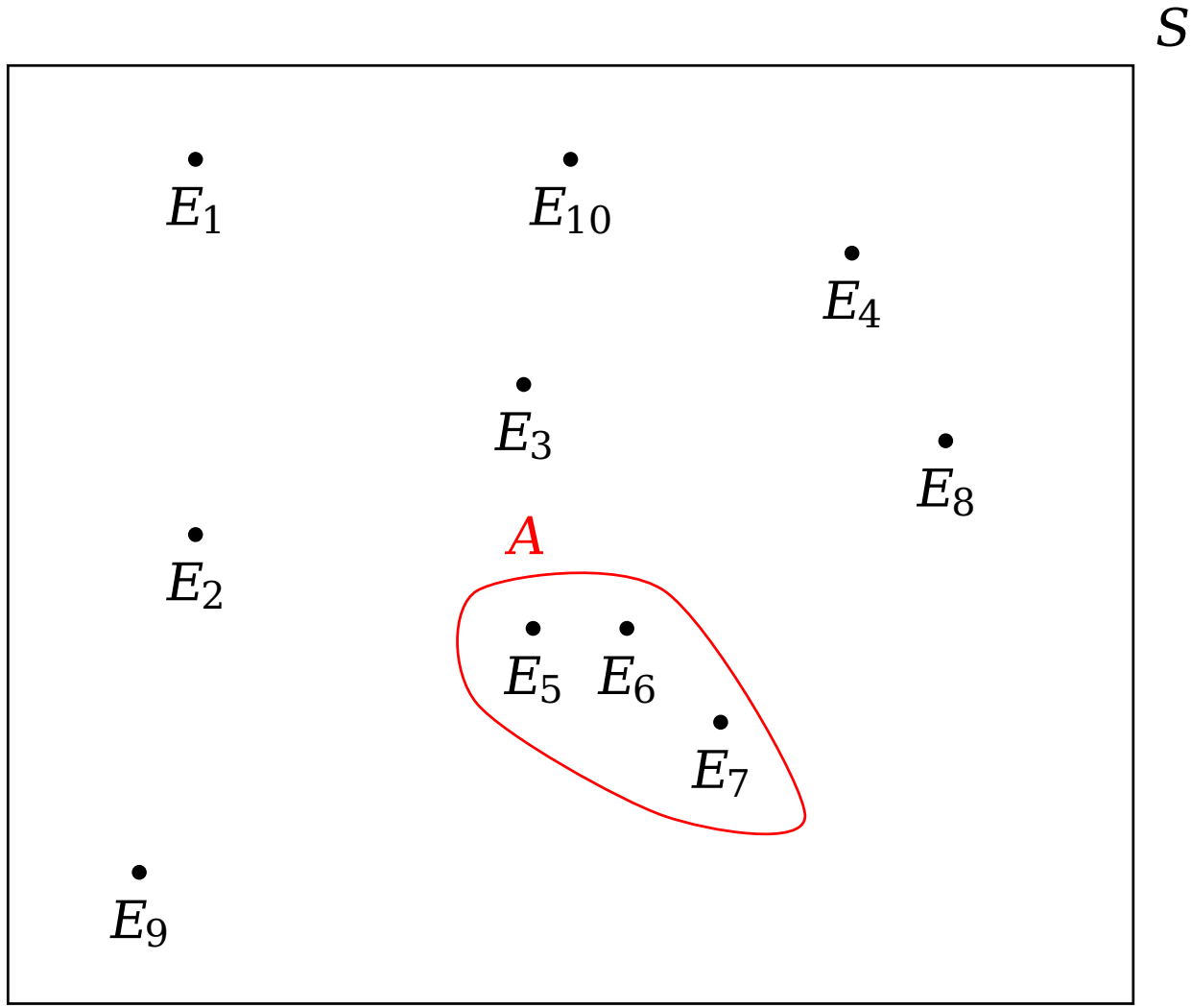
Constructing the framework

$$S = E_1 \cup E_2 \cup \cdots \cup E_{10}$$



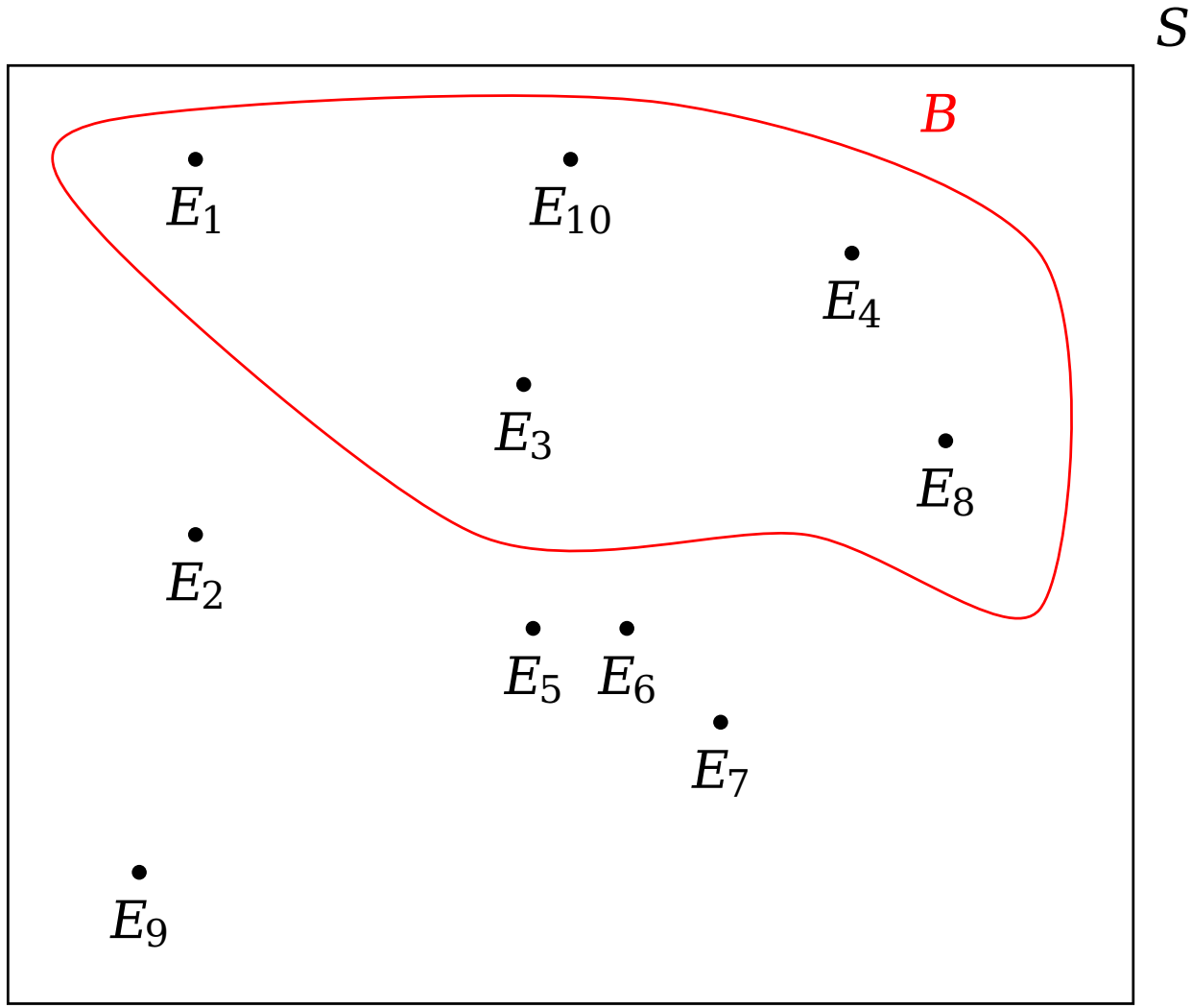
Constructing the framework

$$A = E_5 \cup E_6 \cup E_7$$



Constructing the framework

$$B = E_1 \cup E_3 \cup E_4 \cup E_8 \cup E_{10}$$



Constructing the framework

Note

- The event S is termed the certain event;
- The event $\varnothing$ is termed the impossible event.

Mathematical definition of probability

For event A, P is the function that assigns

$$P(A) = p$$

where p is a numerical value.

However, this definition is too general:

- what properties does $P(.)$ have ?
- how does $P(.)$ assign numerical values; that is, how do we compute

  $$P(A)$$

  for a given event A in sample space S ?

The rules of probability

Suppose S is a sample space for an experiment, A and B are events (i.e. subsets of S).

Then $P(A)$ obeys the following rules:

(I) $P(A) \geqslant 0$.

(II) $P(S) = 1$.

(III) If $A \cap B = \varnothing$, then

$$P(A \cup B) = P(A) + P(B).$$

The rules of probability

Informally, P measures the 'content' of A. The rules say

(I) The content cannot be negative.

(II) The total content of S is standardized to be 1.

(III) If A and B do not overlap, then their total content is the sum of the their individual contents.

The rules of probability

Rule (III) extends to any number of events:

If $A_1, A_2, \ldots, A_K$ are a collection of events such that

$$A_j \cap A_k = \varnothing \qquad \text{for all } j \neq k$$

then

$$P\left(A_1 \cup A_2 \cup \cdots \cup A_K\right) = P(A_1) + P(A_2) + \cdots + P(A_K).$$

The rules of probability

Example: Partitions

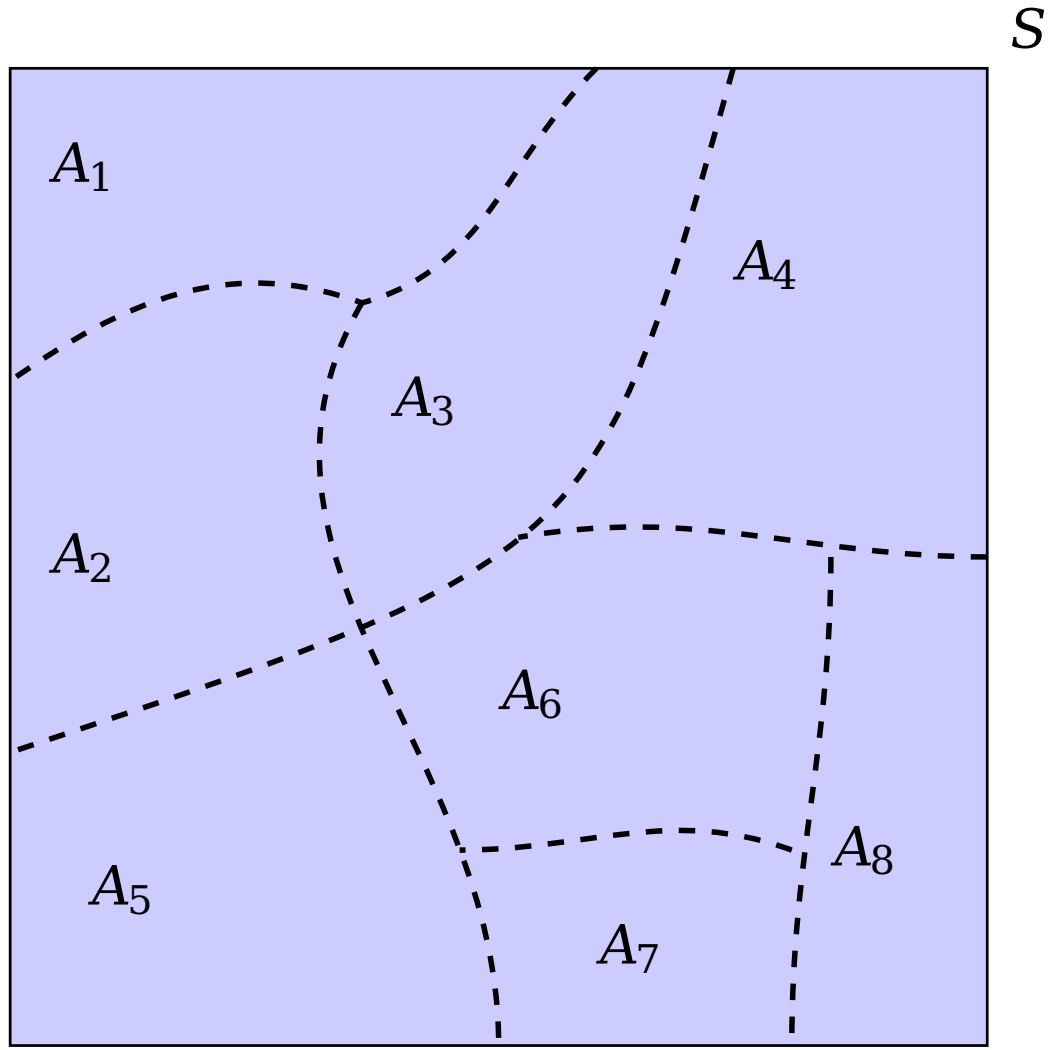
If $A_1, A_2, \ldots, A_K$ form a partition of S, and

$$P(A_k) = p_k \qquad k = 2, \ldots, K$$

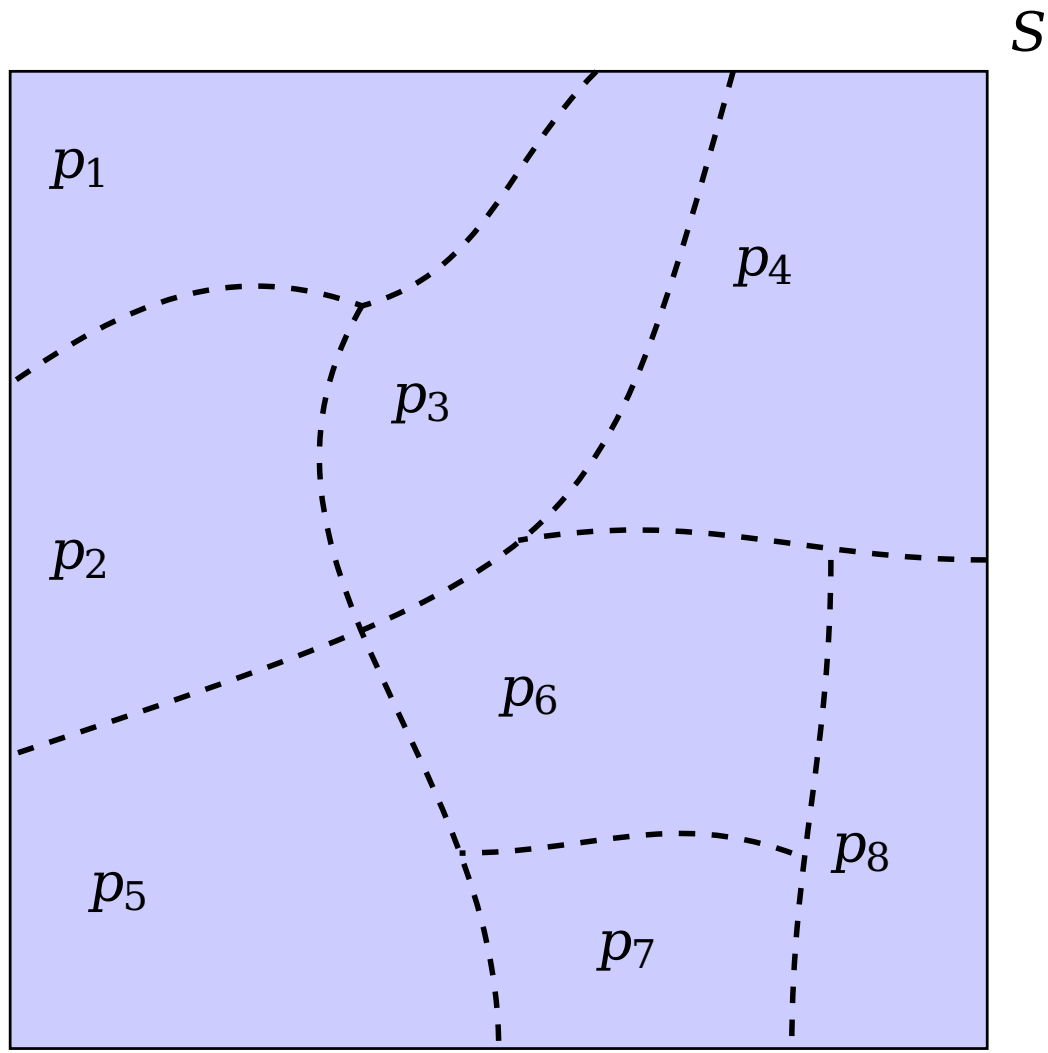
say, then

$$\sum_{k=1}^{K} P(A_k) = \sum_{k=1}^{K} p_k = 1.$$

The rules of probability



The rules of probability



The rules of probability

Some immediate consequences of the rules:

(i) For any A, $P(A^c) = 1 - P(A)$.

(ii) $P(\varnothing) = 0$.

(iii) For any A, $P(A) \leqslant 1$.

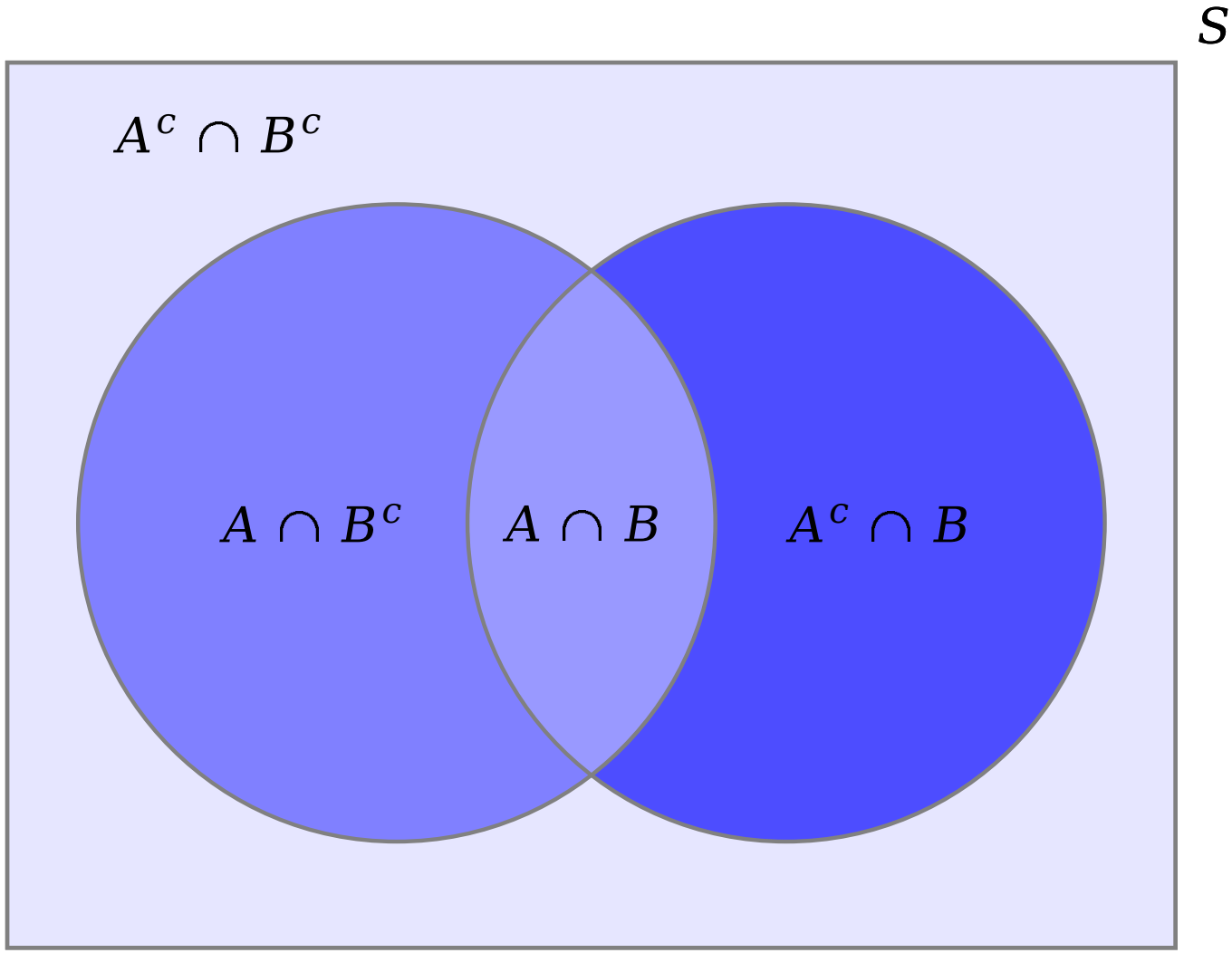
(iv) For two events A and B, if $A \subseteq B$, then

$$P(A) \leqslant P(B)$$

(v) General Addition Rule: For two events A and B,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

The rules of probability



The rules of probability

We can see from the diagram that if we simply compute

$$P(A) + P(B)$$

we are double counting the intersection, so we need to subtract $P(A \cap B)$ to obtain that

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

We may deduce from this that

$$P(A \cup B) \leqslant P(A) + P(B)$$

as $P(A \cap B) \geqslant 0$.

The rules of probability

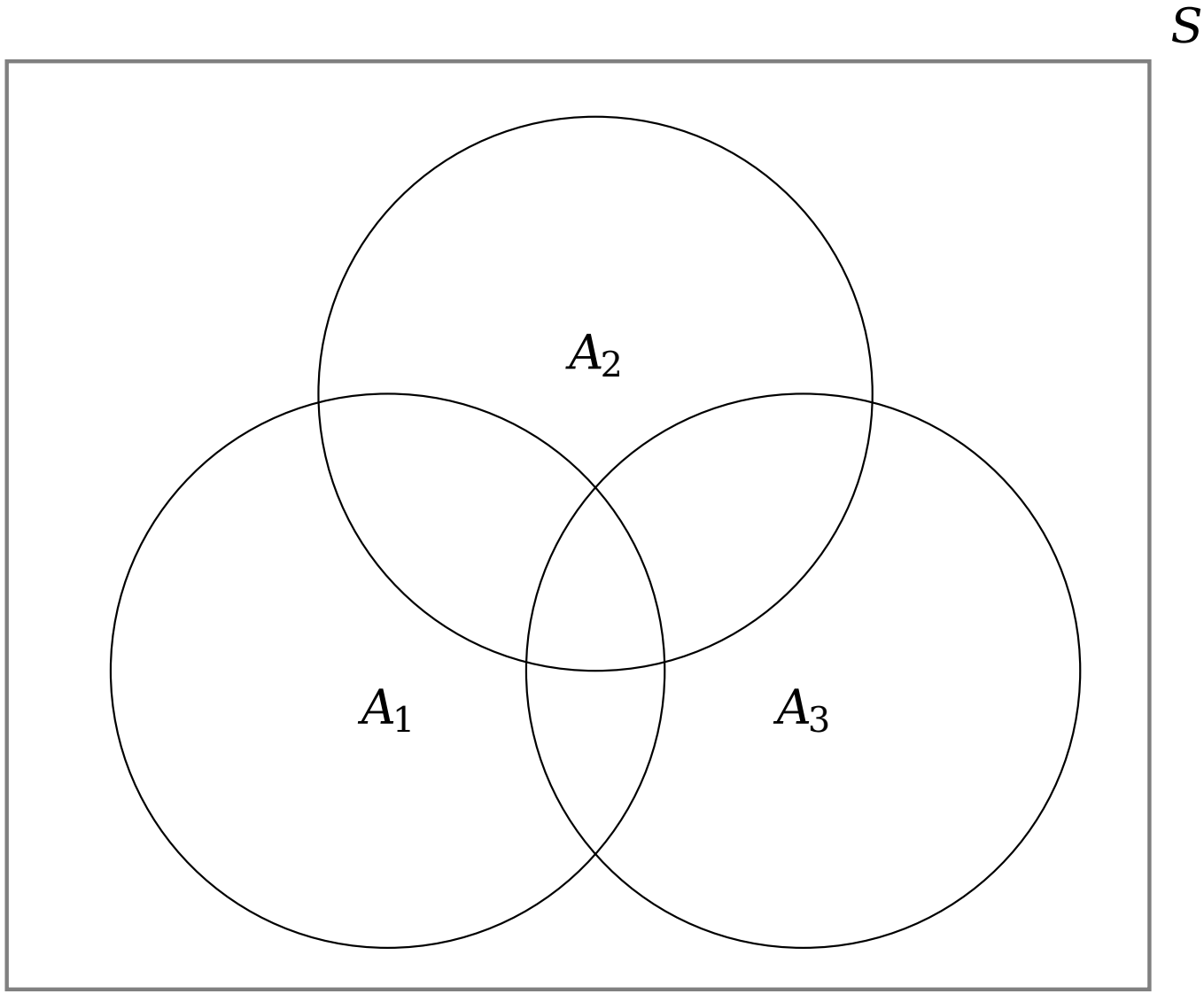
For events $A_1, A_2, \ldots, A_K$, we can construct a formula for

$$P\left(\bigcup_{k=1}^{K} A_k\right)$$

For example, with $K = 3$, we have that

$$\begin{aligned} P(A_1 \cup A_2 \cup A_3) & \\ &= P(A_1) + P(A_2) + P(A_3) \\ &\quad - P(A_1 \cap A_2) - P(A_1 \cap A_3) - P(A_2 \cap A_3) \\ &\quad + P(A_1 \cap A_2 \cap A_3). \end{aligned}$$

The rules of probability



The rules of probability

A more straightforward general result is Boole's Inequality

$$P\left(\bigcup_{k=1}^{K} A_k\right) \leqslant \sum_{k=1}^{K} P\left(A_k\right).$$

Specifying probabilities

We have defined how the probability function must behave mathematically.

We now consider two ways via which we could specify the numerical probability of an event.

Equally likely sample outcomes

Suppose that S is finite, with N sample outcomes in total that are considered to be equally likely to occur.

Then for the elementary events $E_1, E_2, \ldots, E_N$, we have

$$P(E_i) = \frac{1}{N} \qquad i = 1, \ldots, N.$$

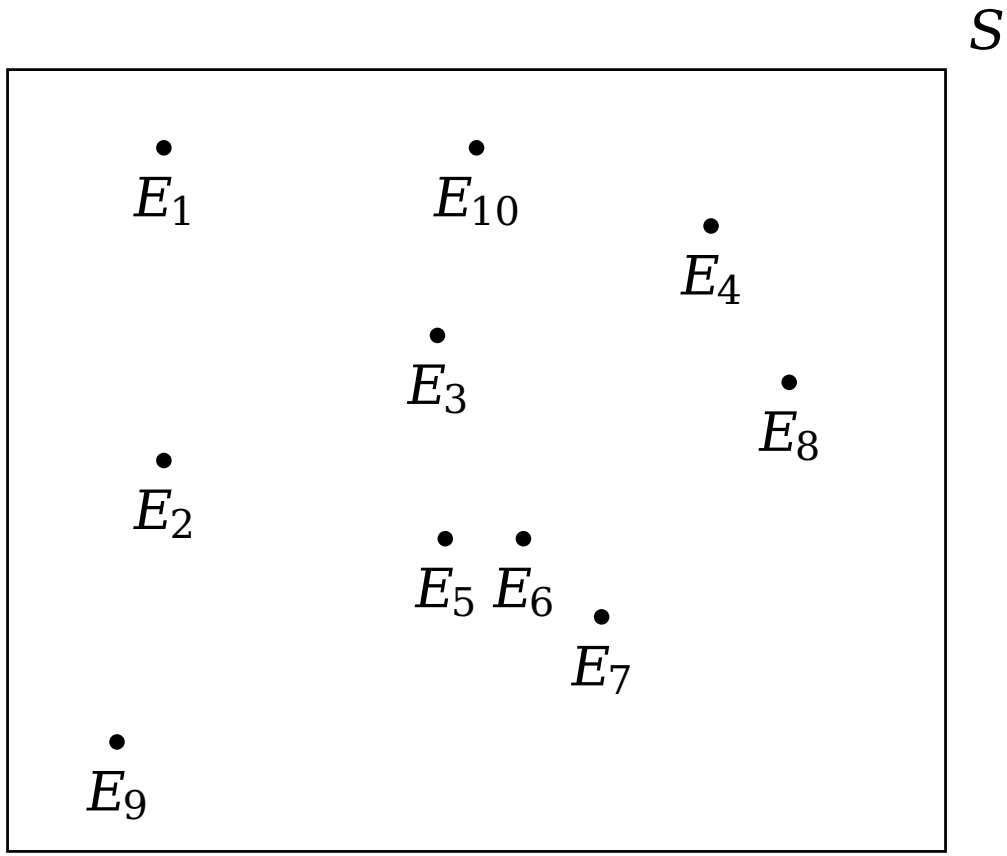
For $A \subseteq S$

$$P(A) = \frac{\text{Number of sample outcomes in } A}{\text{Number of sample outcomes in } S} = \frac{n}{N}$$

say.

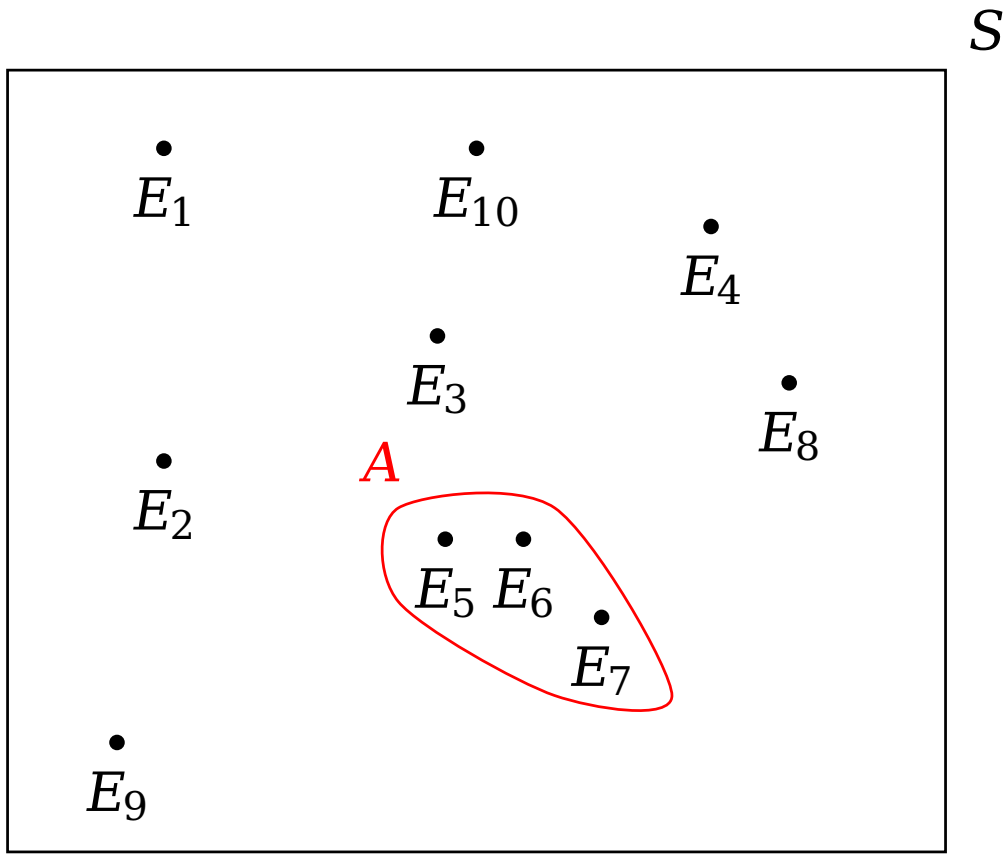
Equally likely sample outcomes

Suppose we have $N = 10$.



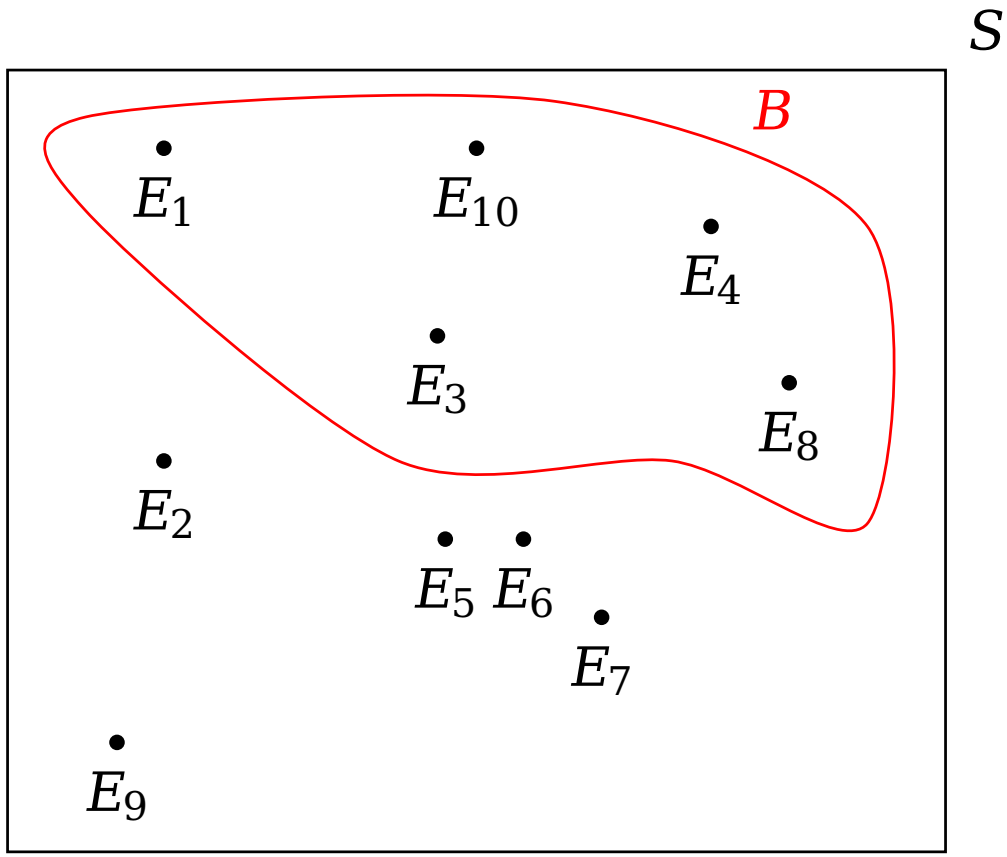
Equally likely sample outcomes

$P(A) = 3/10.$



Equally likely sample outcomes

$P(B) = 5/10.$



Equally likely sample outcomes

We have "equally likely outcomes" where we can appeal to some form of symmetry argument.

(a) **Coins:** For a single coin flip, the sample space is

$$S = \{\text{Head}, \text{Tail}\}$$

with elementary events $E_1 = \{\text{Head}\}$ and $E_2 = \{\text{Tail}\}$, and it is usual to assume that

$$P(E_1) = P(E_2) = \frac{1}{2}.$$

Equally likely sample outcomes

(b) **Dice:** For six-sided die, the sample space is

$$S = \{1, 2, 3, 4, 5, 6\}$$

with elementary events $E_i = \{i\}$ for $i = 1, 2, 3, 4, 5, 6$, and it is usual to assume that

$$P(E_i) = \frac{1}{6}.$$

If A is the event that the outcome is an even number,

$$A = E_2 \cup E_4 \cup E_6$$

and

$$P(A) = \frac{3}{6} = \frac{1}{2}.$$

Relative frequency

Example: Thumbtack tossing





- $S = \{\text{Up}, \text{Down}\}$.
- $A = \{\text{Up}\}$.

What is $P(A)$?

Relative frequency

We may define $P(A)$ by considering the relative frequency with which event A occurs in a long sequence of repeated identical experiments.

- Consider a finite sequence of N repeat experiments.
- Let n be the number of times (out of N) that A occurs.
- Define

$$P(A) = \lim_{N \longrightarrow \infty} \frac{n}{N}.$$

This is the frequentist definition of probability;

Relative frequency

The frequentist definition covers the thumbtack example:



Is it always possible to consider an infinite sequence of repeats ?

Rules for counting outcomes

A sequence of k operations, in which operation i can result in n_i possible outcomes, can result in

$$n_1 \times n_2 \times \cdots \times n_k$$

possible sequences of outcomes.

Rules for counting outcomes

Example: Two dice

Two dice are rolled: there are $6 \times 6 = 36$ possible outcomes.

	1	2	3	4	5	6
1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

Rules for counting outcomes

Example:

We are to pick $k = 5$ numbers sequentially from the set

$$\{1, 2, 3, \ldots, 100\},$$

where every number is available on every pick. The number of ways of doing that is

$$100 \times 100 \times 100 \times 100 \times 100 = 100^5.$$

Rules for counting outcomes

Example:

We are to pick $k = 11$ players sequentially from a squad of 25 players. The number of ways of doing that is

$$25 \times 24 \times 23 \times \cdots \times 15.$$

as players cannot be picked twice.

Rules for counting outcomes

When selecting repeatedly from a finite set $\{1, 2, \ldots, N\}$ we may select

- with replacement, that is,
  - each successive selection can be one of the original set, irrespective of previous selections,

  or
- without replacement, that is,
  - the set is depleted by each successive selection,

Permutations

An ordered arrangement of r objects is called a permutation.

The number of ways of ordering r objects selected, without replacement, from n is denoted P_r^n, and we have that

$$P_r^n = n \times (n-1) \times (n-2) \times \cdots (n-r+2) \times (n-r+1) = \frac{n!}{(n-r)!}$$

where

$$n! = n \times (n-1) \times \cdots 3 \times 2 \times 1.$$

Combinations

The number of combinations of n objects taken r at a time is the number of subsets of size r that can be formed.

This number is denoted

$$C_r^n = \binom{n}{r}$$

where

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}$$

i.e. we "choose r from n".

Combinations

The number of possible selections of r objects (in sequence, without replacement) is

$$n \times (n-1) \times \ldots \times (n-r+1) = \frac{n!}{(n-r)!} = P_r^n$$

leaving $(n-r)$ objects unselected.

If the order of selected objects is not important in identifying a combination, we must have that

$$P_r^n = r! \times C_r^n .$$

as there are $r!$ equivalent combinations that yield the same permutation.

Binomial coefficients

Note

Binomial theorem:

$$(a+b)^n = \sum_{j=0}^{n} \binom{n}{j} a^j b^{n-j}$$

where

$$\binom{n}{j} = \frac{n!}{j!(n-j)!}$$

Binary sequences

Binary sequences

$$101000100101010$$

arise in many probability settings; if we take ‘1’ to indicate inclusion and ‘0’ to indicate exclusion, then we can identify combinations with binary sequences.

- the number of binary sequences of length n containing r 1s is

$$\binom{n}{r}$$

Binary sequences

The above results are used to compute probabilities in the case of equally likely outcomes.

- S: complete list of possible sequences of selections.
- A: sequences having property of interest.
- We have

$$P(A) = \frac{\text{Number of elements in } A}{\text{Number of elements in } S} = \frac{n_A}{n_S}$$

Conditional probability

In the case of 'ordinary' probability, the outcome must be an element in the set S.

- the only 'certain' knowledge we have is that the outcome, s, is in S.
- we can then assess $P(A)$ for event A.

Conditional probability

Suppose we know that, in fact,

$$s \in B \subseteq S$$

for some B; that is, we know that event B occurs.

In light of the information that B occurs, what do we now think about the probability that A occurs also ?

Conditional probability

With the knowledge that event B occurs, we restrict the parts of the sample space that should be considered;

- we are certain that the sample outcome must lie in B,
- we can forget about those outcomes that are not in B.

Conditional probability

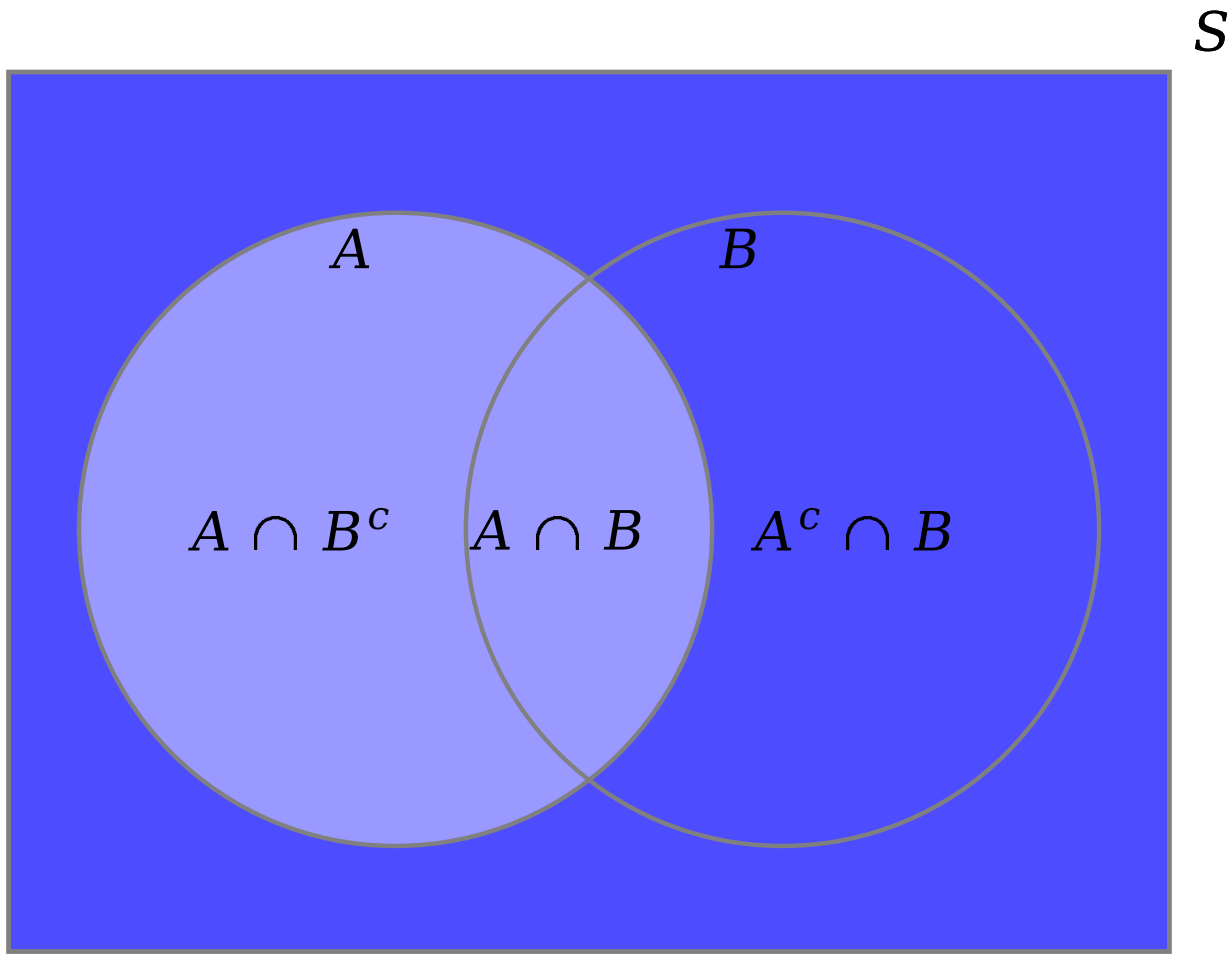
For two events A and B, the conditional probability of A given B is denoted $P(A|B)$, and is defined by

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Note that we consider this only in cases where $P(B) > 0$.

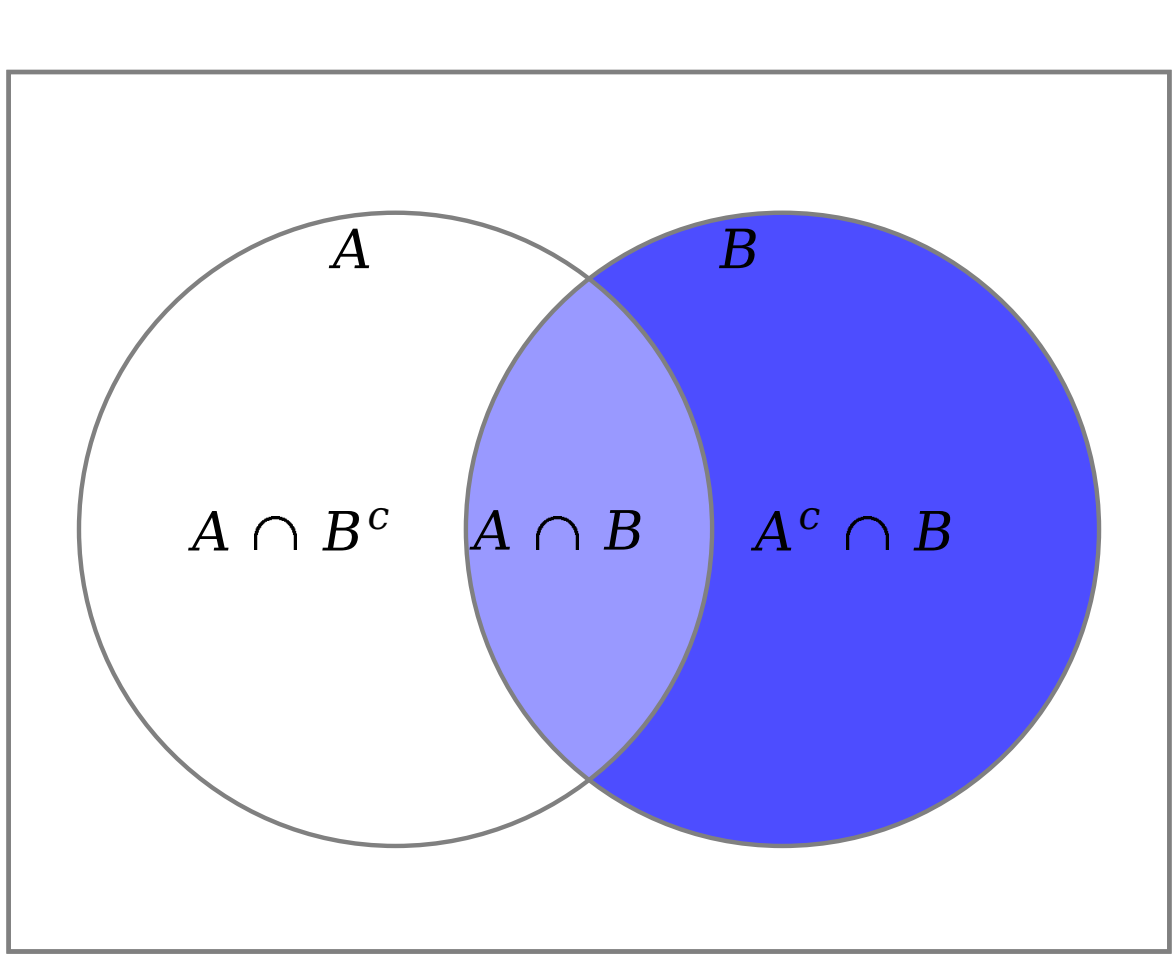
Conditional probability

Ordinary Probability: $P(A)$.



Conditional probability

Conditional Probability: $P(A|B) = P(A \cap B)/P(B)$.



Conditional probability

Example:

Experiment: roll a fair die, record the score.

- $S = \{1, 2, 3, 4, 5, 6\}$, all outcomes equally likely.
- $A = \{1, 3, 5\}$ (score is odd).
- $B = \{4, 5, 6\}$ (score is more than 3).
- $A \cap B = \{5\}$.

$$P(A) = \frac{3}{6} = \frac{1}{2} \qquad P(B) = \frac{3}{6} = \frac{1}{2} \qquad P(A \cap B) = \frac{1}{6}$$

so

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/6}{1/2} = \frac{1}{3}.$$

Conditional probability

Example:

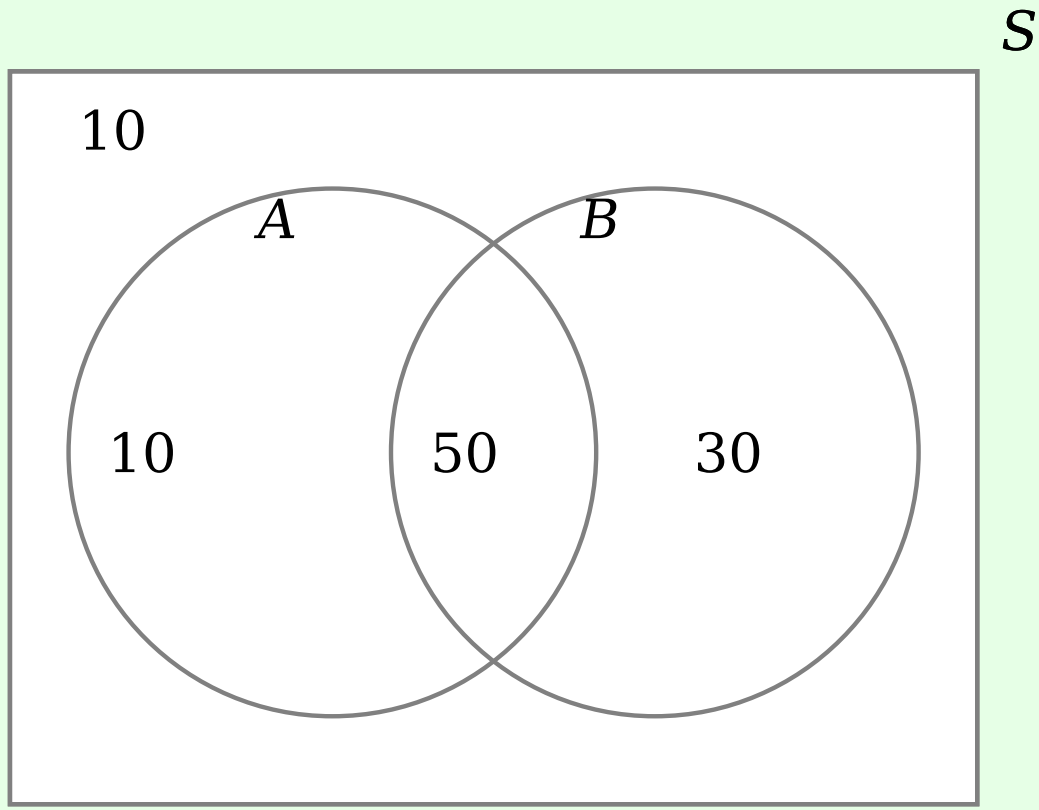
In a class of 100 students:

- 60 are from Science, 40 are from Arts.
- 80 are in a Major program, 20 are in another program.
- 50 of the Science students are in a Major program.

Conditional probability

Example:

A – Science; B – Major program.



Conditional probability

Example:

A student is selected from the class, with all students equally likely to be selected.

- What is the probability that the selected student is from Science, $P(A)$?:

$$P(A) = \frac{60}{100} = \frac{3}{5}$$

Conditional probability

Example:

- What is the probability that the selected student is from Science and in a Major program, $P(A \cap B)$?:

$$P(A \cap B) = \frac{50}{100} = \frac{1}{2}$$

Conditional probability

Example:

- Given that the selected student is known to be in a Major program, what is the probability that the student is from Science, $P(A|B)$?

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{50/100}{80/100} = \frac{50}{80} = \frac{5}{8}.$$

Conditional probability

Note

- If $B \equiv S$, then

$$P(A|S) = \frac{P(A \cap S)}{P(S)} = \frac{P(A)}{1} = P(A).$$

- If $B \equiv A$, then

$$P(A|A) = \frac{P(A \cap A)}{P(A)} = \frac{P(A)}{P(A)} = 1.$$

- We have

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \leqslant \frac{P(B)}{P(B)} = 1.$$

Conditional probability

Direct from the definition, we have that

$$P(A \cap B) = P(B)P(A|B)$$

(recall $P(B) > 0$).

Conditional probability

Note

It is important to understand the distinction between

$$P(A \cap B) \qquad \text{and} \qquad P(A|B).$$

- $P(A \cap B)$ is the chance of A and B occurring relative to S. A and B are treated symmetrically in the calculation.
- $P(A|B)$ records the chance of A and B occurring relative to B. A and B are not treated symmetrically; from the definition, we see that in general

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \neq \frac{P(A \cap B)}{P(B)} = P(A|B).$$

Conditional probability

Note

From the above calculation, we see that we could have

$$P(A|B) \leqslant P(A) \quad \text{or} \quad P(A|B) \geqslant P(A) \quad \text{or} \quad P(A|B) = P(A).$$

That is, knowledge that B occurs could decrease, increase or leave unchanged the probability that A occurs.

Recall: Simpson's Paradox

Conditional probability

Note

If $P(B) > 0$,

(I) Non-negativity:

$$P(A|B) = P(A \cap B)/P(B) \geqslant 0.$$

(II) For the event S:

$$P(S|B) = P(S \cap B)/P(B) = P(B)/P(B) = 1$$

(III) Additivity: If A_1 and A_2 are mutually exclusive, then

$$P(A_1 \cup A_2|B) = P(A_1|B) + P(A_2|B)$$

i.e. conditional probability obeys the Probability Rules !

Conditional probability

Example:

A box of 500 light bulbs is purchased from each of two factories (labelled 1 and 2).

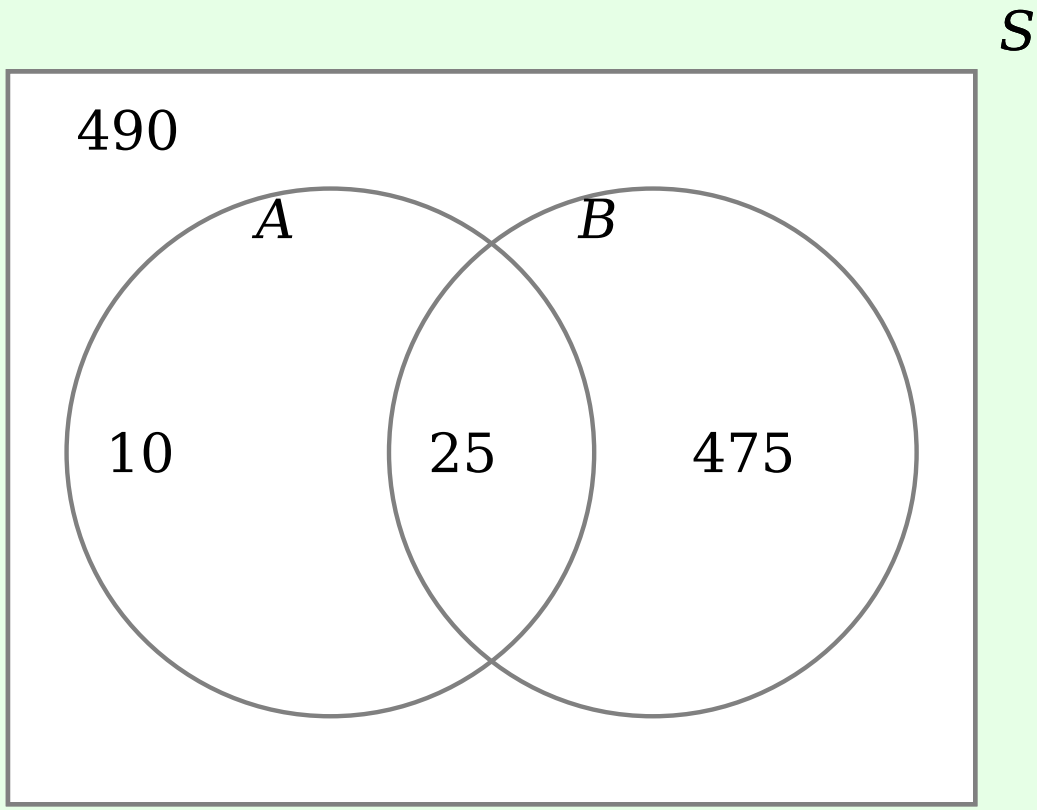
- In the box from Factory 1, there are 25 defective bulbs.
- In the box from Factory 2, there are 10 defective bulbs.

The bulbs are unpacked and placed in storage.

Conditional probability

Example:

A – Defective; B – Factory 1.



Conditional probability

Example:

At the next installation, a bulb is selected from the store.

- What is the probability that the bulb is defective, $P(A)$?:

$$P(A) = \frac{35}{1000} = 0.035.$$

Conditional probability

Example:

- What is the probability that the bulb came from Factory 1, $P(B)$?:

$$P(B) = \frac{500}{1000} = 0.5.$$

Conditional probability

Example:

- Given that the selected bulb came from Factory 1, what is the probability it is defective $P(A|B)$?:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{25/1000}{500/1000} = \frac{25}{500} = 0.05.$$

Conditional probability

Example:

- Given that the selected bulb is defective, what is the probability it came from Factory 1, $P(B|A)$?:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{25/1000}{35/1000} = \frac{25}{35} = \frac{5}{7}.$$

Conditional probability

Example:

We measure the failure time of an electrical component.

- $S = \mathbb{R}^+$.
- $A = [10, \infty)$ (fails after 10 hours or more).
- $B = [5, \infty)$ (fails after 5 hours or more).

Suppose we assess

$$P(A) = 0.25 \qquad P(B) = 0.4.$$

Now here $A \subset B \subset S$, so therefore $P(A \cap B) \equiv P(A)$, and

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)}{P(B)} = \frac{0.25}{0.4} = 0.625.$$

Conditional probability

Example: Drug testing

A drug testing authority tests 100000 athletes to assess whether they are using performance enhancing drugs (PEDs).

The drug test is not perfect as it produces

- False positives: declares an athlete to be using PEDs, when in reality they are not.
- False negatives: declares an athlete not to be using PEDs, when in reality they are.

Conditional probability

Example: Drug testing

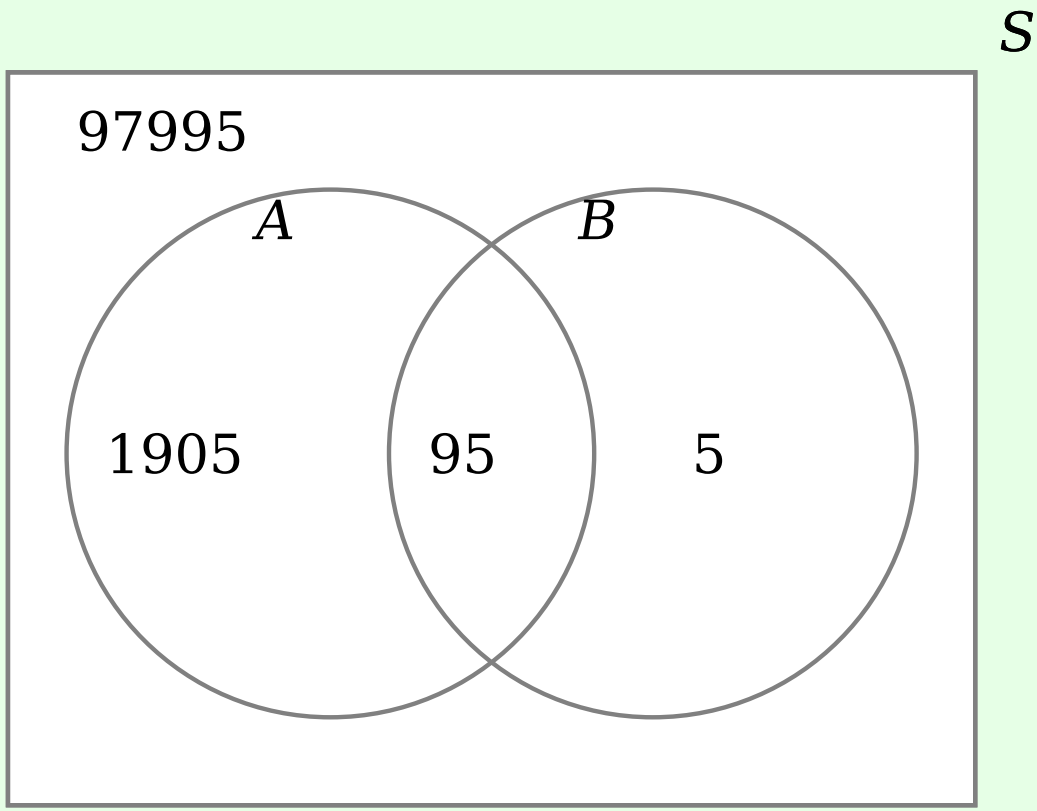
Suppose that after detailed investigation it is discovered that

- 2000 athletes gave positive test results.
- 100 athletes were using PEDs.
- 95 athletes who were using PEDs tested positive.

Conditional probability

Example: Drug testing

A – Drug test is positive; B – Athlete is using PED.



Conditional probability

Example: Drug testing

An athlete is selected at random from the 100000, and is observed to test positive.

What is the probability that they actually were using PEDs ?

Conditional probability

Example: Drug testing

The required conditional probability is

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{95}{2000} = 0.0475.$$

Note that is very different from

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{95}{100} = 0.95.$$

Independence

Two events A and B in sample space S are independent if

$$P(A|B) = P(A)$$

equivalently, they are independent if

$$P(A \cap B) = P(A)P(B).$$

Note that if $P(A|B) = P(A)$, then

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A)P(B)}{P(A)} = P(B).$$

Independence

Note

- Independence is not a property of the events A and B themselves, it is a property of the probabilities assigned to them.
- Independence is not the same as mutual exclusivity.

$$\text{Independent: } P(A \cap B) = P(A)P(B)$$

$$\text{Mutually exclusive: } P(A \cap B) = 0.$$

Independence for multiple events

For events A_1, A_2, A_3 in sample space S, we can consider independence pairwise: we might assume that

$$P(A_1 \cap A_2) = P(A_1)P(A_2)$$

$$P(A_1 \cap A_3) = P(A_1)P(A_3)$$

$$P(A_2 \cap A_3) = P(A_2)P(A_3).$$

Can we deduce from this

$$P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3)?$$

In general, no.

Independence for multiple events

Example: Two dice

Suppose we roll two dice with outcomes independent.

- A_1 – first roll outcome is odd.
- A_2 – second roll outcome is odd.
- A_3 – total score is odd.

We have, for instance,

$$P(A_1) = P(A_2) = P(A_1|A_3) = P(A_2|A_3) = \frac{1}{2}$$

(identify all 36 pairs of equally likely scores and compute).

Thus A_1, A_2 and A_3 are pairwise independent.

Independence for multiple events

Example: Two dice

However,

$$P(A_1 \cap A_2 \cap A_3) = 0$$

and

$$P(A_1 | A_2 \cap A_3) = 0$$

etc.

Conditional independence

Events A_1, A_2 and B in sample space S, with $P(B) > 0$, are conditionally independent given B if

$$P(A_1 | A_2 \cap B) = P(A_1 | B)$$

or equivalently

$$P(A_1 \cap A_2 | B) = P(A_1 | B) P(A_2 | B).$$

General Multiplication Rule

For events $A_1, A_2, \ldots, A_K$, we have the general result that

$$\begin{aligned} P(A_1 \cap A_2 \cap \cdots \cap A_K) = P(A_1) \\ \times P(A_2|A_1) \\ \times P(A_3|A_1 \cap A_2) \\ \vdots \\ \times P(A_K|A_1 \cap A_2 \cap \cdots \cap A_{K-1}) \end{aligned}$$

Also known as the Chain Rule for probabilities.

General Multiplication Rule

If the events are mutually independent, then we have that

$$P(A_1 \cap A_2 \cap \cdots \cap A_K) = \prod_{k=1}^{K} P(A_k)$$

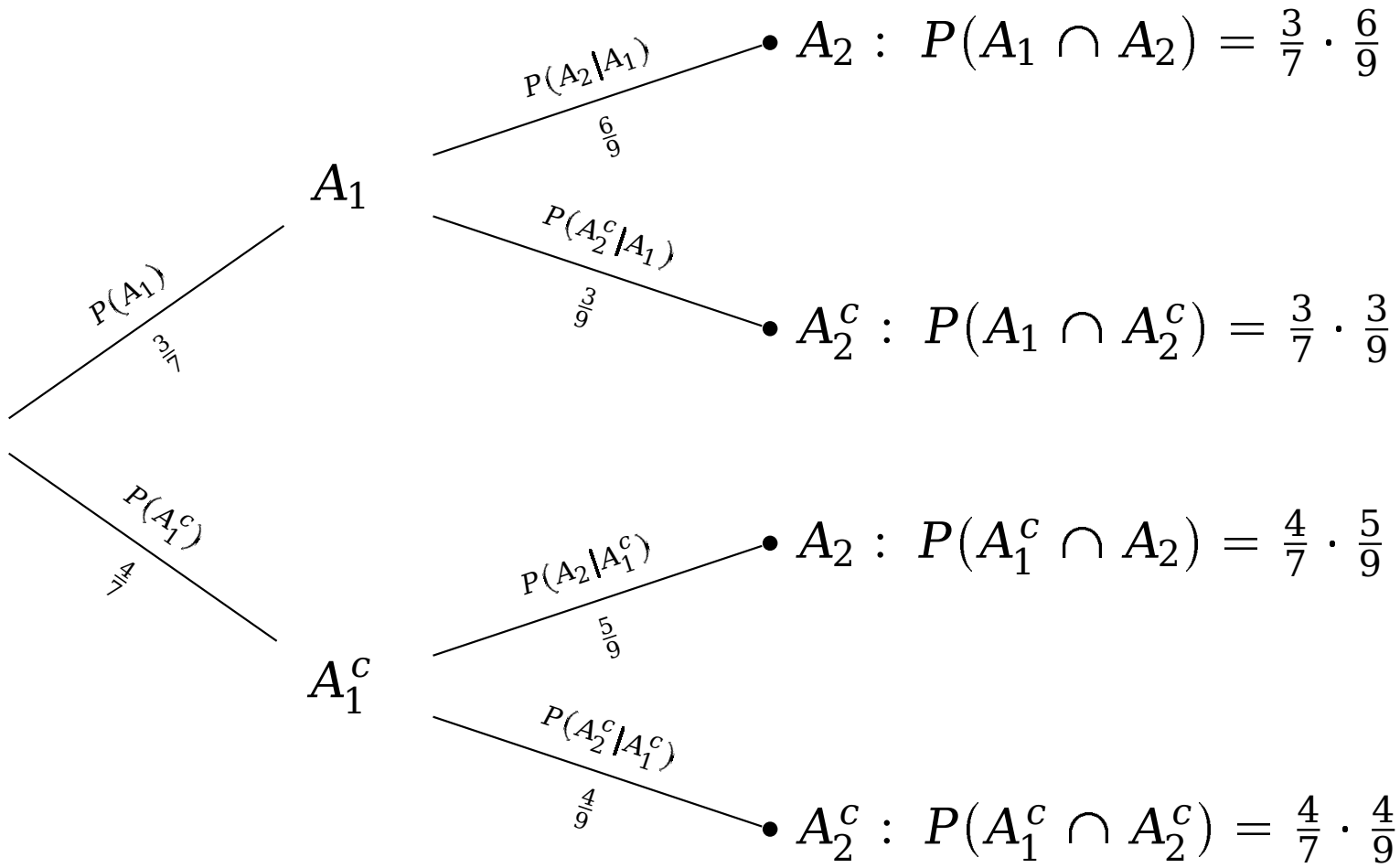
Probability Trees

Probability trees are simple ways to display joint probabilities of multiple events. They comprise

- Junctions: correspond to the multiple events
- Branches: correspond to the sequence of (conditional) choices of events, given the previous choices.

We multiply along the branches to get the joint probabilities.

Probability Trees



Probability Trees

Note that each junction has two possible branches coming from it, corresponding to

$$A_k \quad \text{and} \quad A_k^c$$

respectively.

Such a tree extends to as many events as we need.

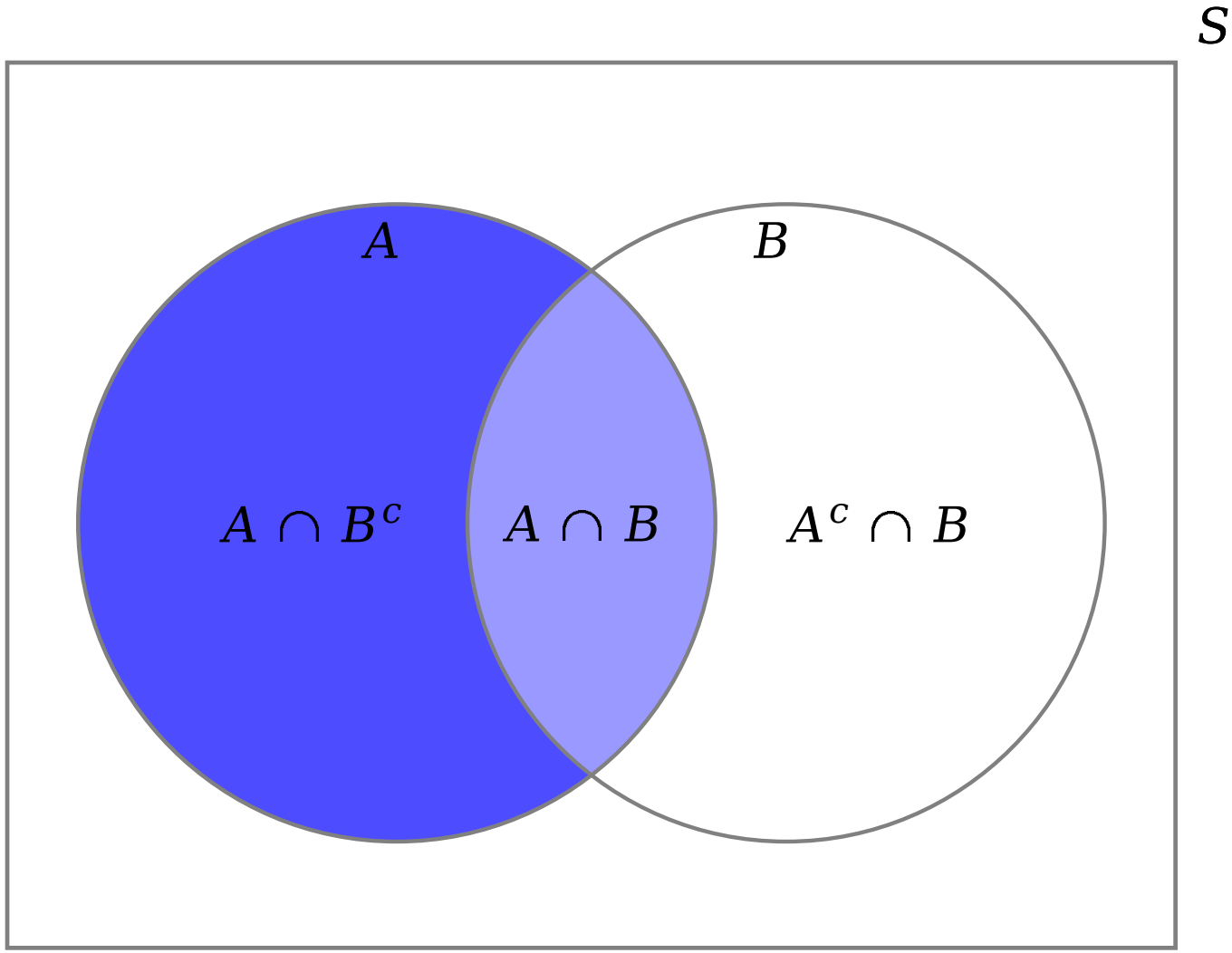
The Theorem of Total Probability

For two events A and B in sample space S, we have the partition of A as

$$A = (A \cap B) \cup (A \cap B^c).$$

- $(A \cap B)$: "in A and in B"
- $(A \cap B^c)$: "in A and not in B"

The Theorem of Total Probability



The Theorem of Total Probability

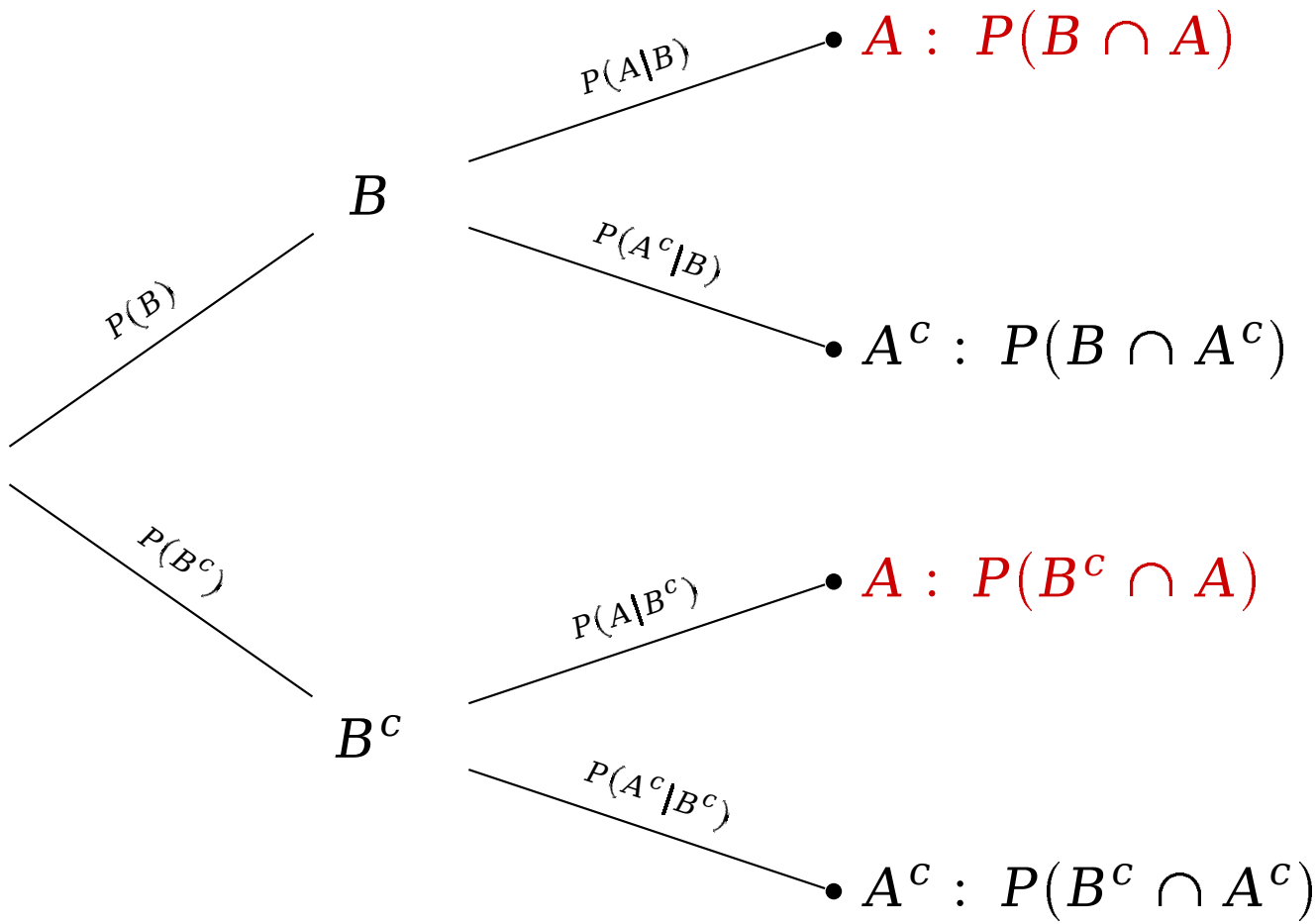
Therefore

$$P(A) = P(A \cap B) + P(A \cap B^c)$$

and using the multiplication rule, we may rewrite this as

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c).$$

The Theorem of Total Probability



The Theorem of Total Probability

That is, there are two ways to get to A:

- via B;
- via B^c.

To compute $P(A)$ we add up all the probabilities on paths that end up at A.

Note that B and B^c together form a partition of S.

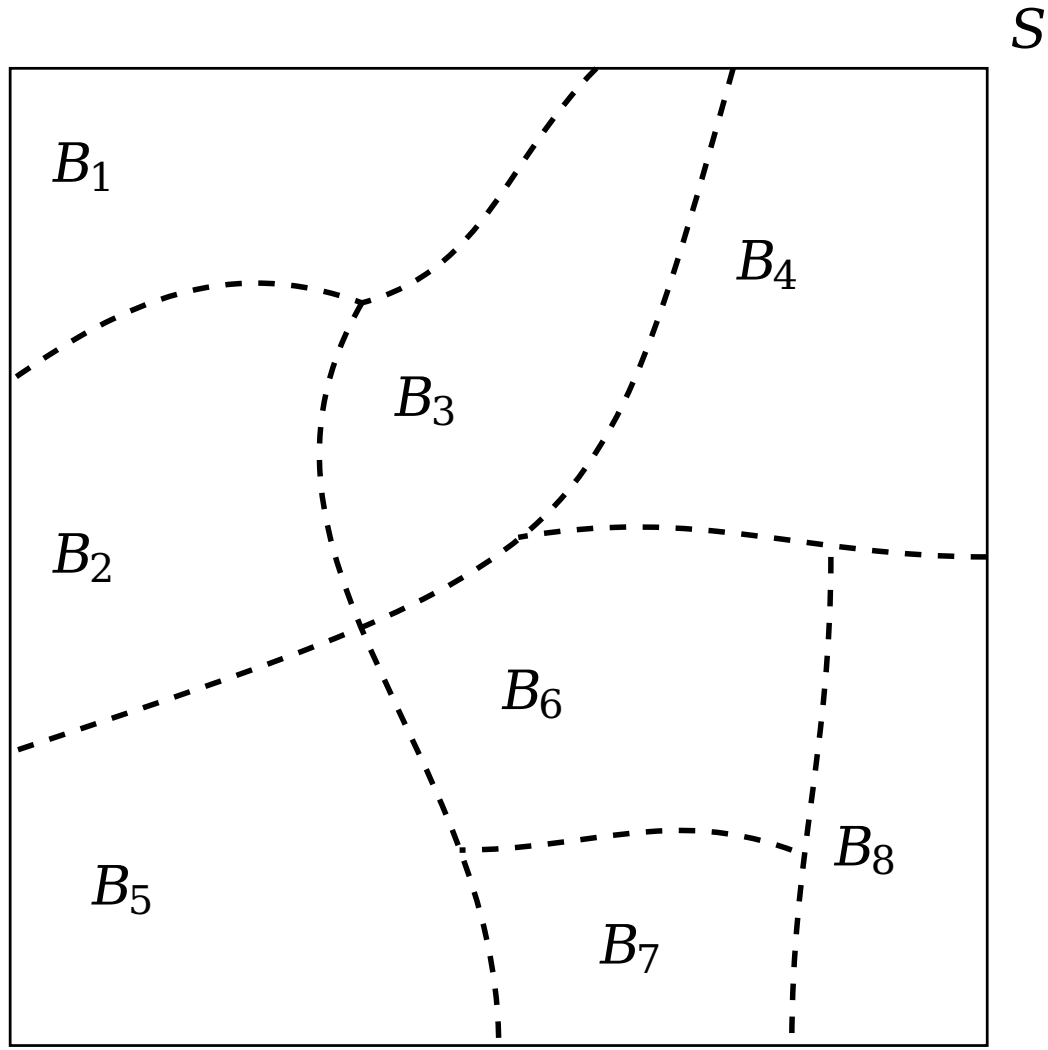
The Theorem of Total Probability

Now suppose we have a partition into K subsets

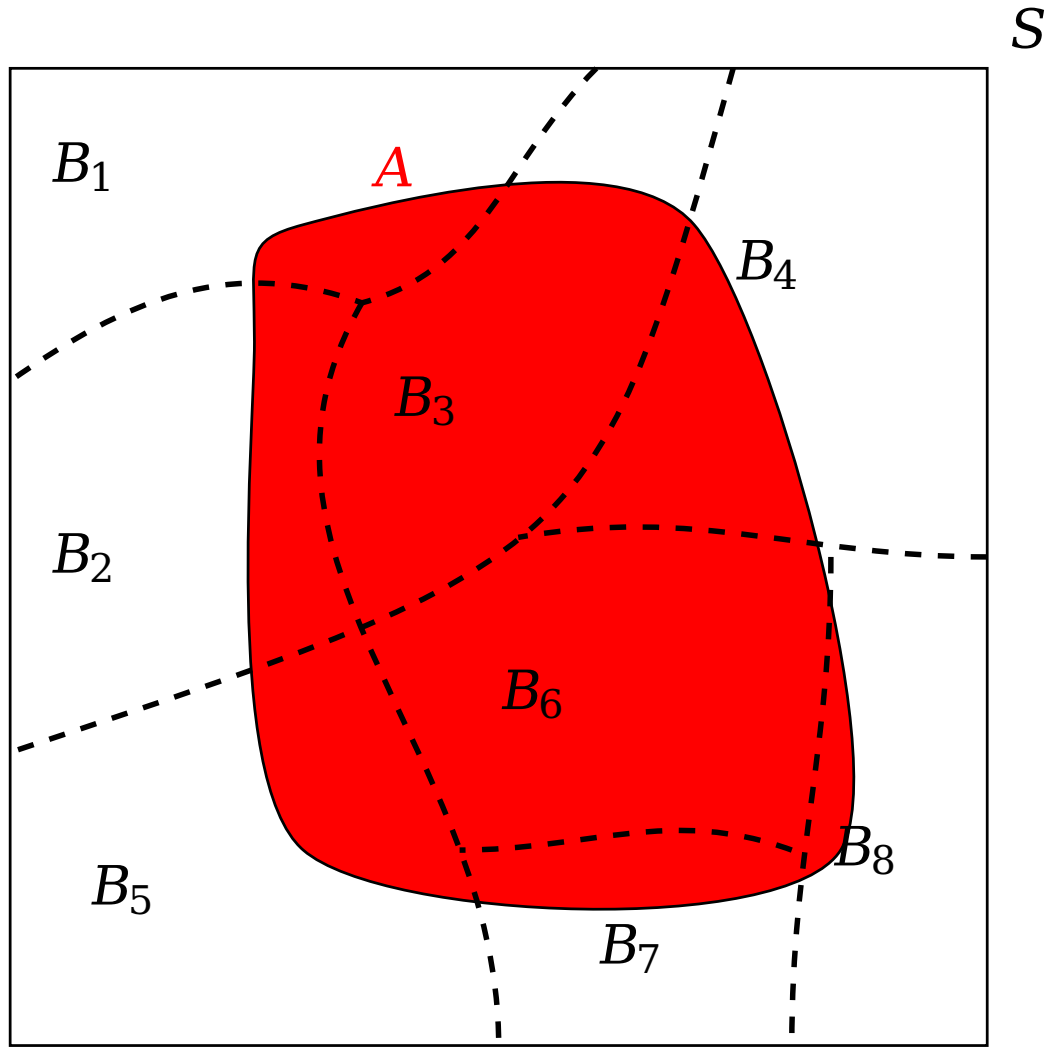
$$B_1, B_2, \ldots, B_K.$$

Again, these events also partition A.

The Theorem of Total Probability



The Theorem of Total Probability



The Theorem of Total Probability

That is, we have that

$$A = (A \cap B_1) \cup (A \cap B_2) \cup \cdots \cup (A \cap B_K)$$

and

$$P(A) = P(A \cap B_1) + P(A \cap B_2) + \cdots + P(A \cap B_K).$$

The Theorem of Total Probability

Using the definition of conditional probability, we have

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \cdots + P(A|B_K)P(B_K).$$

that is

$$P(A) = \sum_{k=1}^{K} P(A|B_k)P(B_k).$$

This result is known as the Theorem of Total Probability.

The Theorem of Total Probability

Note

- The formula assumes that $P(B_k) > 0$ for $k = 1, \ldots, K$.
- It might be that

$$P(A \cap B_k) = P(A|B_k) = 0.$$

  for some k.

- This formula represents how probability is used to compute $P(A)$.

The Theorem of Total Probability

Example: Three bags

Suppose that an experiment involves selecting a ball from one of three bags. Let

- Bag 1: 4 red and 4 white balls.
- Bag 2: 1 red and 10 white balls.
- Bag 3: 7 red and 11 white balls.

A bag is selected (with all bags equally likely), and then a ball is selected from that bag (with all balls equally likely).

What is the probability that the ball selected is red ?

The Theorem of Total Probability

Example: Three bags

Let

- S: all possible selections of balls.
- B_1: bag 1 selected.
- B_2: bag 2 selected.
- B_2: bag 3 selected.
- A: ball selected is red.

The Theorem of Total Probability

Example: Three bags

We have that

$$
\begin{aligned}
P(A) &= P(A \cap B_1) + P(A \cap B_2) + P(A \cap B_3) \\
&= P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + P(A|B_3)P(B_3) \\
&= \left(\frac{4}{8} \times \frac{1}{3}\right) + \left(\frac{1}{11} \times \frac{1}{3}\right) + \left(\frac{7}{18} \times \frac{1}{3}\right) \\
&= \frac{97}{297}
\end{aligned}
$$

The Theorem of Total Probability

Example: Three bags

Note that this is not equal to

$$\frac{\text{Total number of red balls}}{\text{Total number of balls}}$$

that is

$$\frac{4+1+7}{8+11+18} = \frac{12}{37}$$

as the red balls are not equally likely to be selected – this is only true conditional on the bag selection.

Bayes Theorem

The second great theorem of probability is Bayes Theorem.

- Named after Reverend Thomas Bayes (1701 – 1761)

  `https://en.wikipedia.org/wiki/Thomas_Bayes`

- Could be written Bayes's Theorem.

- It is not really a theorem.

Bayes Theorem

Theorem: Bayes Theorem

For two events A and B with $P(A) > 0$ and $P(B) > 0$,

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}.$$

If $0 < P(B) < 1$, then we may expand the denominator

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)}$$

using the Theorem of Total Probability.

Bayes Theorem

"Proof": By the definition of conditional probability

$$P(A|B)P(B) = P(A \cap B) = P(B|A)P(A).$$

As $P(A) > 0$ and $P(B) > 0$ we can cross multiply and write

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}.$$

If $0 < P(B) < 1$, then we can write

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)$$

and substitute this expression in the denominator.

Bayes Theorem

If $B_1, B_2, \ldots, B_K$ form a partition of S, with $P(B_k) > 0$ for $k = 1, \ldots, K$, and A is an event with $P(A) > 0$, then for $k = 1, \ldots, K$

$$P(B_k|A) = \frac{P(A|B_k)P(B_k)}{P(A)} = \frac{P(A|B_k)P(B_k)}{\sum_{j=1}^{K} P(A|B_j)P(B_j)}.$$

Probability tree interpretation: Given that we end up at A, what is the probability that we got there via branch B_k ?

Bayes Theorem

Example: Three bags

Suppose that an experiment involves selecting a ball from one of three bags. Let

- Bag 1: 4 red and 4 white balls.
- Bag 2: 1 red and 10 white balls.
- Bag 3: 7 red and 11 white balls.

A bag is selected (with all bags equally likely), and then a ball is selected from that bag (with all balls equally likely).

The ball selected is red. What is the probability it came from Bag 2 ?

Bayes Theorem

Example: Three bags

Let

- S: all possible selections of balls.
- B_1: bag 1 selected.
- B_2: bag 2 selected.
- B_2: bag 3 selected.
- A: ball selected is red.

Bayes Theorem

Example: Three bags

Recall that

$$
\begin{aligned}
P(A) &= P(A \cap B_1) + P(A \cap B_2) + P(A \cap B_3) \\
&= P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + P(A|B_3)P(B_3) \\
&= \left(\frac{4}{8} \times \frac{1}{3}\right) + \left(\frac{1}{11} \times \frac{1}{3}\right) + \left(\frac{7}{18} \times \frac{1}{3}\right) \\
&= \frac{97}{297}
\end{aligned}
$$

Bayes Theorem

Example: Three bags

Then

$$P(B_2|A) = \frac{P(A|B_2)P(B_2)}{P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + P(A|B_3)P(B_3)}$$

$$= \frac{\frac{1}{11} \times \frac{1}{3}}{\frac{4}{8} \times \frac{1}{3} + \frac{1}{11} \times \frac{1}{3} + \frac{7}{18} \times \frac{1}{3}}$$

$$= \frac{1/33}{97/297}$$

$$= \frac{9}{97} \doteq 0.09278.$$

Bayes Theorem

Example: Three bags

We can also compute that

$$P(B_1|A) = \frac{4/24}{97/297} = \frac{297}{582} \simeq 0.51031$$

and

$$P(B_3|A) = \frac{21/54}{97/297} = \frac{1746}{2079} \simeq 0.39691.$$

Bayes Theorem

Note

We have that

$$P(B_1|A) + P(B_2|A) + P(B_3|A) = 1$$

but note that

$$P(A|B_1) + P(A|B_2) + P(A|B_3) \neq 1.$$

In the first formula, we are conditioning on A everywhere; in the second, we have different conditioning sets.

Using Bayes Theorem

Bayes theorem is often used to make probability statements concerning an event B that has not been observed, given an event A that has been observed.

Using Bayes Theorem

Example: Medical screening

A health authority tests individuals from a population to assess whether they are sufferers from some disease. The screening test is not perfect as it produces

- False positives: declares someone to be a sufferer, when in reality they are a non-sufferer.
- False negatives: declares someone to be a non-sufferer, when in reality they are a sufferer.

Using Bayes Theorem

Example: Medical screening

Suppose that we denote by

- A – Screening test is positive;
- B – Person is actually a sufferer.

Using Bayes Theorem

Example: Medical screening

Suppose that

- $P(B) = p$;
- $P(A|B) = 1 - \alpha$ (true positive rate);
- $P(A|B^c) = \beta$ (false positive rate).

for probabilities p, α, β.

Using Bayes Theorem

Example: Medical screening

We have for the rate of positive tests

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c) = (1-\alpha)p + \beta(1-p)$$

by the Theorem of Total Probability, and by Bayes theorem

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)}$$

$$= \frac{(1-\alpha)p}{(1-\alpha)p + \beta(1-p)}.$$

Using Bayes Theorem

Example: Medical screening

Similarly, for the false negative rate, we have

$$P(B|A^c) = \frac{P(A^c|B)P(B)}{P(A^c|B)P(B) + P(A^c|B^c)P(B^c)}$$

$$= \frac{\alpha p}{\alpha p + (1-\beta)(1-p)}.$$

Using Bayes Theorem

Example: Medical screening

We have

$$P(A|B) = (1-\alpha) \qquad P(B|A) = (1-\alpha)\frac{p}{(1-\alpha)p + \beta(1-p)}$$

and these two values are potentially very different.

Using Bayes Theorem

Example: Medical screening

That is, in general,

$$P(\text{“Spots”}|\text{“Measles”}) \neq P(\text{“Measles”}|\text{“Spots”})$$

Using Bayes Theorem

Note

This is sometimes known as the Prosecutor's Fallacy:

$$P(\text{“Evidence”}|\text{“Guilt”}) \neq P(\text{“Guilt”}|\text{“Evidence”})$$

Bayes Theorem with multiple events

If A_1 and A_2 are two events in a sample space S so that

$$P(A_1 \cap A_2) > 0$$

and there is a partition $B_1, \ldots, B_K$, with $P(B_k) > 0$, then

$$P(B_k|A_1 \cap A_2) = \frac{P(A_1 \cap A_2|B_k)P(B_k)}{\sum_{j=1}^{K} P(A_1 \cap A_2|B_j)P(B_j)}$$

Bayes Theorem with multiple events

- we use the previous version of Bayes Theorem with

$$A \equiv A_1 \cap A_2.$$

- this extends to conditioning events $A_1, A_2, \ldots, A_K$.

By earlier results

$$P(A_1 \cap A_2 | B_k) = P(A_1 | B_k) P(A_2 | A_1 \cap B_k)$$

that is, the chain rule applies to conditional probabilities also.

Chapter 3
Random variables and probability distributions

Random variables and distributions: big idea

We seek a way to

- efficiently describe the sample outcomes of experiments in numerical form
- describe how probability is spread around the sample outcomes using mathematical functions.

Random variables

Up to this point we have dealt with probability assignments for events defined on a general sample space S.

We now focus on the specific case where S is a subset of the real numbers, $\mathbb{R}$. We could have S

- finite or countable $\{0, 1, 2, \ldots\}$;
- uncountable.

Events in a sample space that is a subset of $\mathbb{R}$ can be termed numerical events.

Random variables

For most experiments defined on an arbitrary space S, we can usually consider a simple mapping, Y, defined on S that maps the sample outcomes onto $\mathbb{R}$.

The mapping Y is termed a random variable; it is just a way of recording the outcome in a numerical fashion.

Random variables

Example: Two coins

Two coins are tossed with the outcomes being independent. The sample space is the list of sequences of possible outcomes

$$S = \{HH, HT, TH, TT\}.$$

with all sequences equally likely.

Define the random variable Y by

$$Y(s) = \text{“the number of heads in sequence } s\text{”}$$

Random variables

Example: Two coins

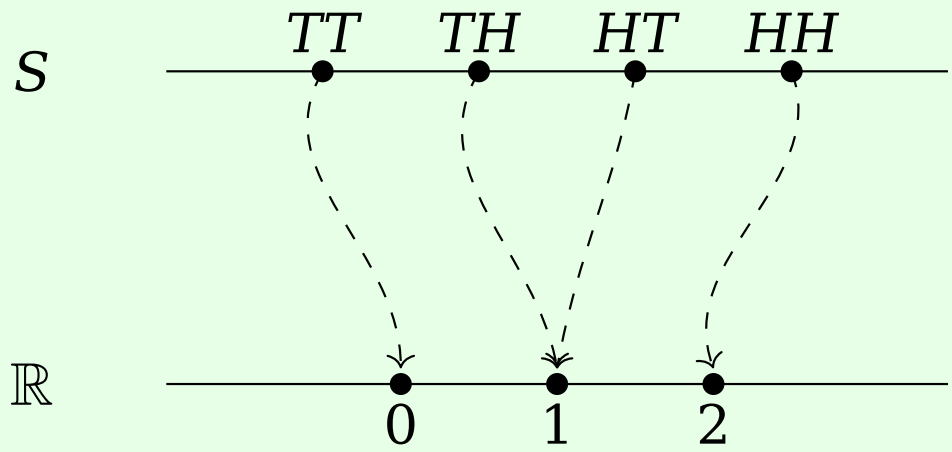
In the original probability specification for events in S,

$$P(\{HH\}) = P(\{HT\}) = P(\{TH\}) = P(\{TT\}) = \frac{1}{4}$$

Each $s \in S$ carries its probability with it through the mapping via $Y(.)$.

Random variables

Example: Two coins



Random variables

Example: Two coins

Then

$$Y(HH) = 2 \qquad Y(HT) = Y(TH) = 1 \qquad Y(TT) = 0$$

and $Y(.)$ maps S onto the set $\{0, 1, 2\}$. We have that

$$P(Y = 0) \equiv P(s \in \{TT\}) = \frac{1}{4}$$

$$P(Y = 1) \equiv P(s \in \{HT, TH\}) = \frac{2}{4}$$

$$P(Y = 2) \equiv P(s \in \{HH\}) = \frac{1}{4}$$

with $P(Y = y) = 0$ for all other values of y.

Random variables

Note

- $Y(.)$ can be a many-to-one mapping.
- We must consider the possible values that Y can take.
- We usually simply write

$$P(Y = y)$$

Random variables

Note

We typically use upper case letters

$$X, Y, Z$$

etc. to denote random variables, and lower case letters

$$x, y, z$$

etc. to denote the numerical values that the random variable can take.

Discrete Random variables

A random variable Y is termed discrete if the possible values that Y can take is a countable set:

$$y_1, y_2, \ldots, y_n, \ldots$$

i.e. we can make a list of the possible values.

Could be

- a finite set: e. g. $\{1, 2, \ldots, 100\}$
- an infinite set: e.g the integers

Write the set of values that carry positive probability $\mathcal{Y}$.

Discrete Random variables

The probabilities assigned to the individual points in the set $\mathcal{Y}$ in must carry all the probability:

$$\sum_{y \in \mathcal{Y}} P(Y = y) = 1.$$

This follows by the probability rules, as all the probability from S has been mapped into $\mathcal{Y}$.

Probability mass function

The probability mass function (pmf) for Y, $p(.)$, is the mathematical function that records how probability is distributed.

That is,

$$p(y) = P(Y = y)$$

- If $y \in \mathcal{Y}$, $p(y) > 0$,
- otherwise $p(y) = 0$.

Probability mass function

The pmf could be defined pointwise: for example

$$p(y) = \begin{cases} 0.2 & y = -2 \\ 0.1 & y = 0 \\ 0.2 & y = 3 \\ 0.5 & y = 5 \\ 0 & \text{otherwise} \end{cases}$$

e.g. The probability that the random variable takes the value 5 is 0.5 etc.

Probability mass function

Alternatively it could be defined using some functional form:

$$p(y) = \begin{cases} \frac{y}{15} & y \in \{1, 2, 3, 4, 5\} \\ \\ 0 & \text{otherwise} \end{cases}$$

Probability mass function

The pmf must satisfy two conditions: it

(i) must specify probabilities: that is

$$0 \leqslant p(y) \leqslant 1 \qquad \text{for all } y$$

(ii) must sum to 1 when evaluated at points in $\mathcal{Y}$:

$$\sum_{y \in \mathcal{Y}} P(Y = y) = 1.$$

Probability mass function

Note

The collection of y values for which $p(y) > 0$ may be a countably infinite set, say

$$\mathcal{Y} = \{y_1, y_2, \ldots\}.$$

In this case we would require the infinite sum

$$\sum_{i=1}^{\infty} p(y_i)$$

be equal to 1.

Expectations for discrete random variables

The expectation (or expected value) of a discrete random variable Y, denoted $E(Y)$, is defined as

$$E(Y) = \sum_{y \in \mathcal{Y}} y\, p(y)$$

whenever this sum is finite; if it is not finite, we say that the expectation does not exist.

Expectations for discrete random variables

Note

- $E(Y)$ is the centre of mass of the probability distribution.
- $E(Y)$ is also termed the mean of the distribution.
- If $\mathcal{Y} = \{y_1, y_2, \ldots, y_n, \ldots\}$, we have that

$$E(Y) = y_1 p(y_1) + y_2 p(y_2) + \cdots + y_n p(y_n) + \cdots$$

Expectations for discrete random variables

Example: Expectation and Mean

Suppose that $\mathcal{Y} = \{y_1, y_2, \ldots, y_n\}$, and

$$p(y) = \frac{1}{n} \qquad y \in \mathcal{Y}$$

and zero otherwise. Then

$$E(Y) = y_1 p(y_1) + y_2 p(y_2) + \cdots + y_n p(y_n)$$

$$= \frac{1}{n}(y_1 + y_2 + \cdots + y_n) = \overline{y}.$$

This is an example of a discrete uniform distribution.

Expectations for discrete random variables

The discrete uniform distribution is a distribution with constant $p(y)$, that is,

$$p(y) = c \qquad y \in \mathcal{Y}$$

for some finite set $\mathcal{Y}$. It follows that

$$\frac{1}{c}$$

equals the number of elements in $\mathcal{Y}$.

Variance

We can also compute

$$E(Y^2) = \sum_{y \in \mathcal{Y}} y^2 p(y)$$

and interpret it as the expectation of Y^2. That is

$$E(Y^2) = y_1^2 p(y_1) + y_2^2 p(y_2) + \cdots$$

the sum of the squares of the possible values, weighted by their probabilities.

Variance

The variance of random variable Y is denoted $\text{Var}(Y)$, and is defined by

$$\text{Var}(Y) = E[(Y - \mu)^2]$$

where $\mu = E(Y)$. The standard deviation of Y is

$$\text{sd}(Y) = \sqrt{\text{Var}(Y)}.$$

Variance

The variance is a measure of how "dispersed" the probability distribution is. Note that

$$\text{Var}(Y) = E[(Y-\mu)^2] = \sum_y (y-\mu)^2 p(y) \geqslant 0$$

as $p(y) \geqslant 0$ and $(y-\mu)^2 \geqslant 0$ for all y.

In fact, $\text{Var}(Y) > 0$ unless the distribution of Y places all its probability at a single value, in which case $\text{Var}(Y) = 0$.

Variance

Now

$$\begin{aligned} E[(Y-\mu)^2] &= E(Y^2 - 2\mu Y + \mu^2) \\ &= E(Y^2) - 2\mu E(Y) + \mu^2 \qquad \text{by linearity} \\ &= E(Y^2) - 2\mu^2 + \mu^2 \\ &= E(Y^2) - \mu^2. \end{aligned}$$

so

$$\text{Var}(Y) = E(Y^2) - \mu^2$$

which gives an alternative method for calculation.

The binomial distribution

We can now consider specific pmfs that correspond to particular experimental settings.

On simple pmf is obtained if we consider an experiment with two possible outcomes, which we can refer to as ‘success’ and ‘failure’ respectively. We can define the random variable Y by

$Y = 1$ if outcome is ‘success’ $\qquad$ $Y = 0$ if outcome is ‘failure’

and then consider

$$P(Y = 0) \text{ and } P(Y = 1).$$

The binomial distribution

Note however that

$$P(Y = 0) + P(Y = 1) = 1$$

so if we specify

$$P(Y = 1) = p$$

then automatically

$$P(Y = 0) = 1 - p = q$$

say.

The binomial distribution

We may then write

$$p(y) = p^y(1-p)^{1-y} = p^y q^{1-y} \qquad y \in \{0, 1\}$$

with $p(y) = 0$ for all other values of y.

This distribution is known as the Bernoulli distribution with parameter p.

Note

To avoid the trivial case, we typically assume that $0 < p < 1$.

The binomial distribution

This model covers many experimental settings:

- coin tosses;
- thumbtack tosses;
- pass/fail;
- cure/not cure

etc.

The binomial distribution

We have for this model that

$$\mu \equiv E(Y) = \sum_{y=0}^{1} yp(y) = (0 \times p(0) + 1 \times p(1))$$

$$= (0 \times q + 1 \times p) = p$$

The binomial distribution

Similarly

$$E(Y^2) = \sum_{i=0}^{1} y^2 p(y) = (0^2 \times p(0) + 1^2 \times p(1))$$

$$= (0 \times q + 1 \times p) = p$$

so therefore

$$\text{Var}(Y) = E(Y^2) - \mu^2 = p - p^2 = p(1-p).$$

The binomial distribution

We now consider a version of the experiment where a sequence of n identical binary experiments with success probability p whose outcomes are independent is carried out.

$$0101011101$$

We consider the total number of successes in the sequence.

Provided $0 < p < 1$, we have that

$$\mathcal{Y} = \{0, 1, 2, \ldots, n\}$$

The binomial distribution

Now if Y is the random variable that records the number of successes in the sequence, we have that

$$(Y = y) \quad \text{means} \quad \text{“}y \text{ successes in } n \text{ trials”}$$

Consider a sequence that results in y successes and $n - y$ failures: for example, if $n = 8$ and $y = 3$, we might have

10101000
01010001
11100000

etc.

The binomial distribution

All such sequences are equally likely, under the 'identical and independent' assumption. The probability of any such sequence is, for $y \in \{0, 1, 2, \ldots, n\}$,

$$p^y q^{n-y}$$

that is, in the examples above

$$pqpqpqqq$$
$$qpqpqqqp$$
$$pppqqqqq$$

etc. all yield $p^3 q^5$.

The binomial distribution

However for the total number of successes, we do not care about the order of the outcomes; we must count the number of all such sequences, as the individual sequences partition the event of interest.

By the previous combinatorial arguments, we see that the number of such sequences is

$$C_y^n = \binom{n}{y} = \frac{n!}{y!(n-y)!}.$$

The binomial distribution

Therefore we have that

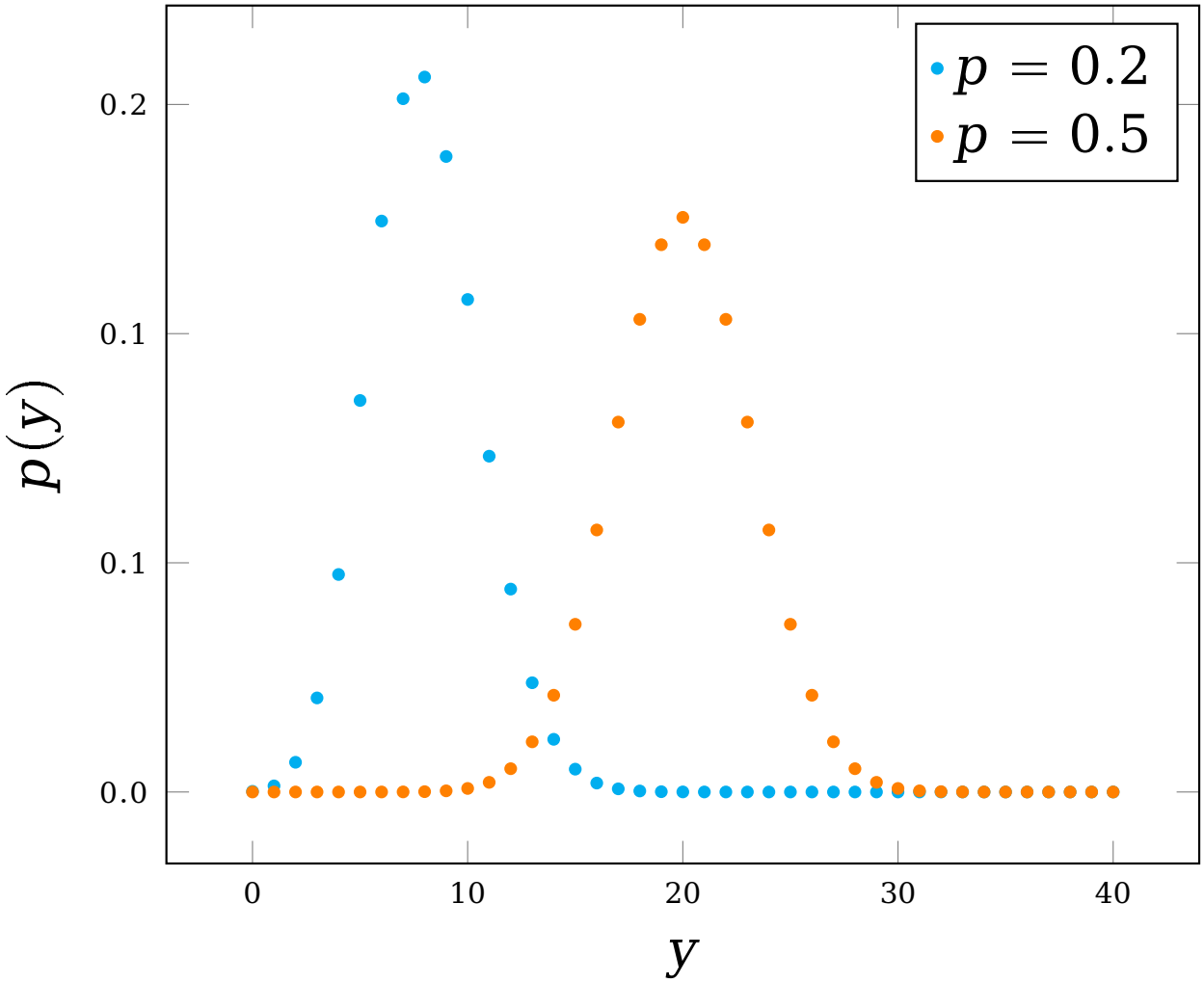
$$p(y) \equiv P(Y = y) = \binom{n}{y} p^y q^{n-y} \qquad y \in \{0, 1, 2, \ldots, n\}$$

and $p(y) = 0$ for all other y.

This is the pmf for the Binomial distribution.

The binomial distribution

Two examples: $n = 40$, with $p = 0.2$ and $p = 0.5$.



The binomial distribution

For the binomial distribution, we have that if

$$Y \sim Binomial(n, p)$$

then

$$\begin{aligned} \text{Expectation:} \quad E(Y) &= np \\ \text{Variance:} \quad \text{Var}(Y) &= npq \end{aligned}$$

The binomial distribution

Note

Recall that for a single binary experiment, for the total number of successes Y we have that

$$E(Y) = p \qquad \text{Var}(Y) = pq$$

whereas for n identical and independent binary experiments, for the total number of successes Y we have that

$$E(Y) = np \qquad \text{Var}(Y) = npq.$$

The binomial distribution

If $n = 1$, we have that

$$p(y) = p^y q^{1-y} \qquad y \in \{0, 1\}$$

and $p(y) = 0$ for all other y, which coincides with Bernoulli distribution with parameter p.

The geometric distribution

We consider again identical and independent binary trials, but rather than counting the total number of successes, we count the number of trials carried out until the first success.

Using the 0/1 notation, we consider sequences such as

0001
01
000000001
1
000001

etc.

The geometric distribution

If random variable Y records the number of trials carried out up to an including the first success, then

$$p(y) = P(Y = y) = q^{y-1}p \qquad y \in \{1, 2, 3, \ldots\}$$

and $p(y) = 0$ for other values of y.

- $y - 1$ failures, followed by a success.
- e.g. the sequence 00001 occurs with probability

$$q \times q \times q \times q \times p = q^4 p.$$

This is the pmf of the geometric distribution.

The geometric distribution

Note

- Sometimes the geometric distribution is written

$$p(y) = q^y p \qquad y \in \{0, 1, 2, \ldots\}$$

– this is the same distribution, but for the random variable that records the number of failures observed up to the first success, that is, for the random variable

$$Y^* = Y - 1.$$

The geometric distribution

For the geometric distribution, we have that if

$$Y \sim Geometric(p)$$

then

$$\begin{aligned} \text{Expectation:} \quad E(Y) &= 1/p \\ \text{Variance:} \quad \text{Var}(Y) &= (1-p)/p^2 \end{aligned}$$

Y takes values on the whole number set $\{1, 2, 3, \}$; this distribution does not place probability anywhere else.

Cumulative probability

We now study another way to describe how probability is distributed. Rather than

$$P(Y = y)$$

for a single value y, we consider

$$P(Y \in A)$$

for some set of values A.

Cumulative probability

For example if $A = \{2, 3, 4, 5\}$, we have

$$\begin{aligned} P(Y \in A) &= P(Y = 2 \text{ or } Y = 3 \text{ or } Y = 4 \text{ or } Y = 5) \\ &= P(2 \leqslant Y \leqslant 5) \end{aligned}$$

which we can compute as

$$P(Y \in A) = P(Y = 2) + P(Y = 3) + P(Y = 4) + P(Y = 5).$$

Such probabilities are called cumulative probabilities.

Cumulative probability

For the geometric distribution, consider the probability

$$P(Y > y)$$

for $y \in \{1, 2, 3, \ldots\}$. We see that

$$\begin{aligned} P(Y > y) &= P(Y = y + 1) + P(Y = y + 2) + \cdots \\ &= \sum_{j=y+1}^{\infty} P(Y = j). \end{aligned}$$

Cumulative probability

Using the geometric pmf

$$
\begin{aligned}
P(Y > y) &= \sum_{j=y+1}^{\infty} q^{j-1} p \\
&= pq^y \sum_{j=y+1}^{\infty} q^{j-y-1} \qquad \text{a geometric progression} \\
&= pq^y \frac{1}{1-q}.
\end{aligned}
$$

Cumulative probability

Therefore

$$P(Y > y) = q^y \qquad y \in \{1, 2, 3, \ldots\}$$

and consequently, as

$$P(Y \leqslant y) = 1 - P(Y > y)$$

we have that

$$P(Y \leqslant y) = 1 - q^y = 1 - (1 - p)^y \qquad y \in \{1, 2, 3, \ldots\}.$$

Cumulative probability

We can define the function

$$F(y) = P(Y \leqslant y) = 1 - (1-p)^y \qquad y \in \{1, 2, 3, \ldots\}.$$

$F(y)$ is the cumulative distribution function (cdf).

Cumulative probability

We can define all probabilities such as

$$P(Y \leqslant a) \qquad P(Y \geqslant b) \qquad P(c \leqslant Y \leqslant d)$$

etc. via $F(y)$. This is because for discrete distributions

$$P(Y \in A) = \sum_{y \in A} p(y).$$

for set A.

Cumulative probability

Note

We can compute

$$F(y) = P(Y \leqslant y)$$

for any y. However, we must remember that, e.g.

$$P(Y \leqslant 2.635) = P(Y = 1) + P(Y = 2) = P(Y \leqslant 2)$$

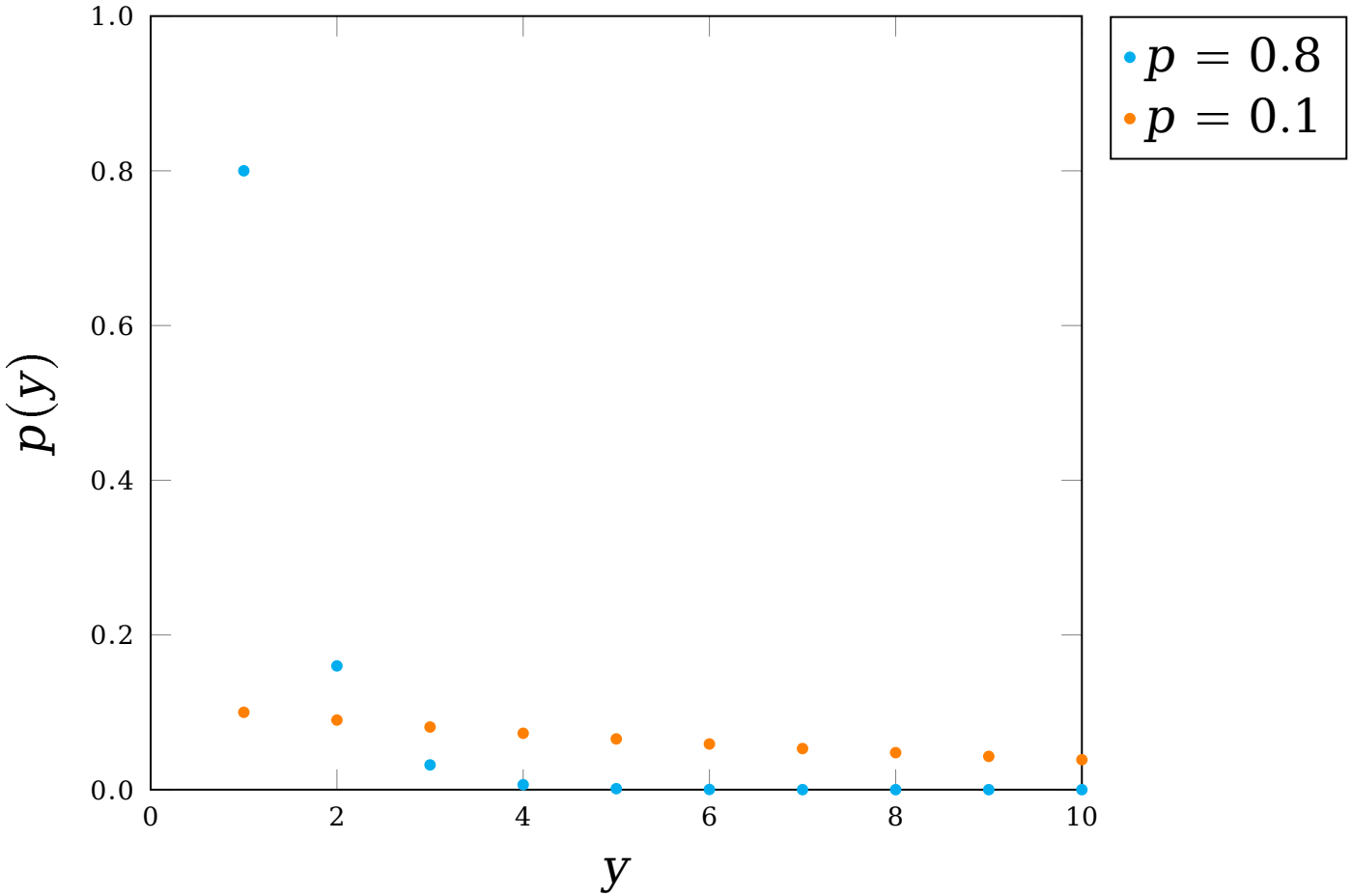
as the probability is only assigned to whole number values. Therefore

$$F(2.635) = F(2) \qquad F(1.234) = F(1) \qquad F(4.789) = F(4)$$

etc.

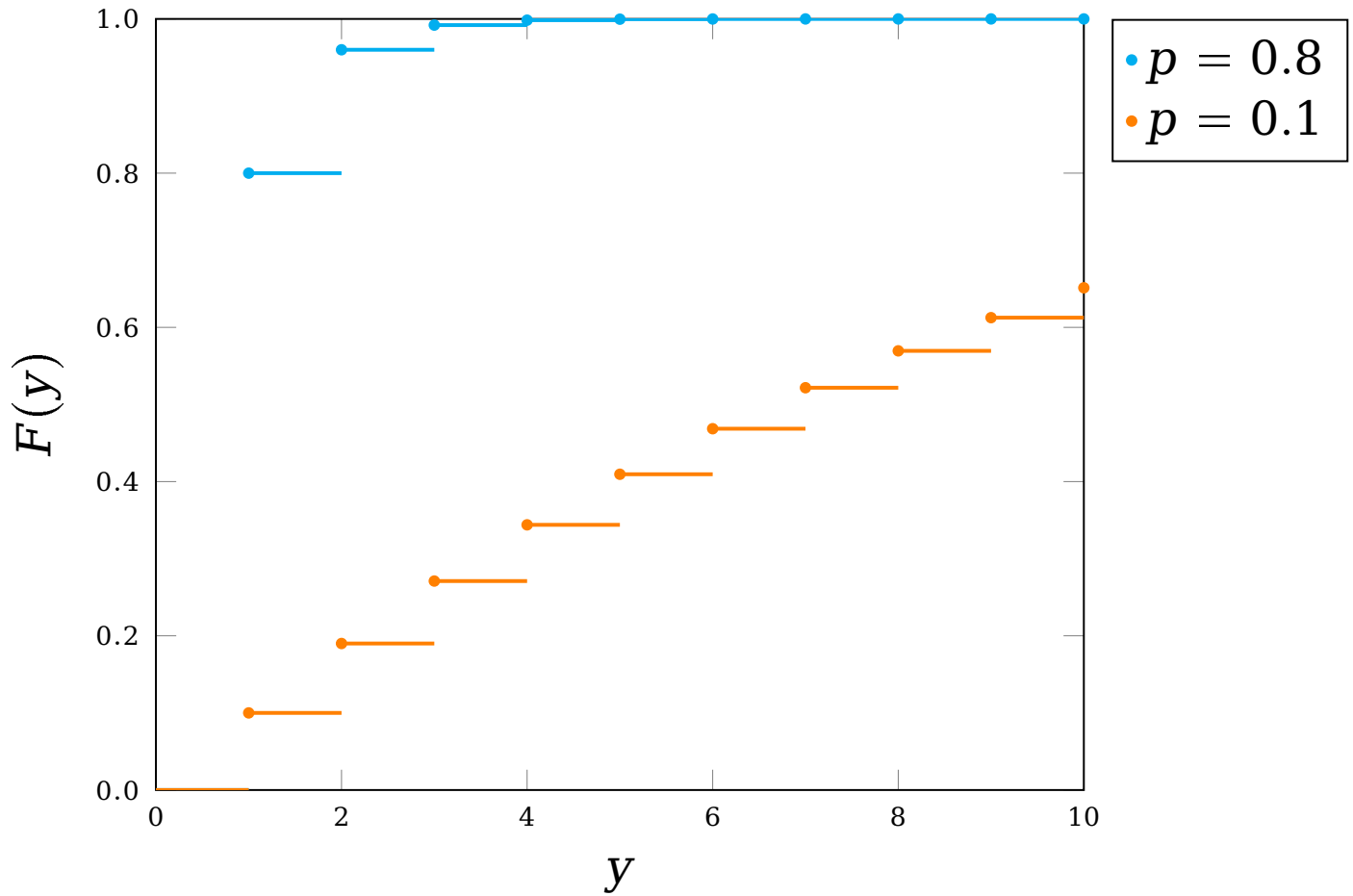
Cumulative probability

Geometric $p(y)$: with $p = 0.8$ and $p = 0.1$.



Cumulative probability

Geometric $F(y)$: with $p = 0.8$ and $p = 0.1$.



Cumulative probability

We can define the $F(y)$ function for any discrete distribution; for example, for the binomial distribution

$$F(y) = P(Y \leqslant y) = \sum_{j=0}^{y} \binom{n}{j} p^j q^{n-j} \qquad y \in \{0, 1, 2, 3, \ldots, n\}$$

but there is no simple formula for this.

Cumulative probability

Note that

$$\begin{aligned} p(y) &= P(Y = y) \\ &= P(Y \leqslant y) - P(Y \leqslant y - 1) \\ &= F(y) - F(y - 1) \end{aligned}$$

so that we can compute $p(y)$ from $F(y)$ and vice versa.

Cumulative probability

$F(y)$ must have the following properties:

1. 'Starts at 0'
$$F(-\infty) = P(Y \leqslant -\infty) = 0$$
2. 'Ends at 1'
$$F(\infty) = P(Y \leqslant \infty) = 1$$
3. 'Non-decreasing'
$$\text{If } y_1 < y_2 \text{ then } F(y_1) \leqslant F(y_2)$$
4. 'Lies in the range $[0, 1]$'
$$0 \leqslant F(y) \leqslant 1 \qquad \text{for all } y.$$

The Poisson distribution

The binomial distribution arises when we consider n identical and independent trials each with success probability p:

$$10001000101010101010101$$

etc. We have

$$p(y) = \binom{n}{y} p^y (1-p)^{n-y} \qquad y \in \{0, 1, 2, \ldots, n\}$$

with $p(y) = 0$ for all other values of y.

The Poisson distribution

If $n \longrightarrow \infty$ with $p = \lambda/n$ for fixed λ, we have that the binomial pmf converges to

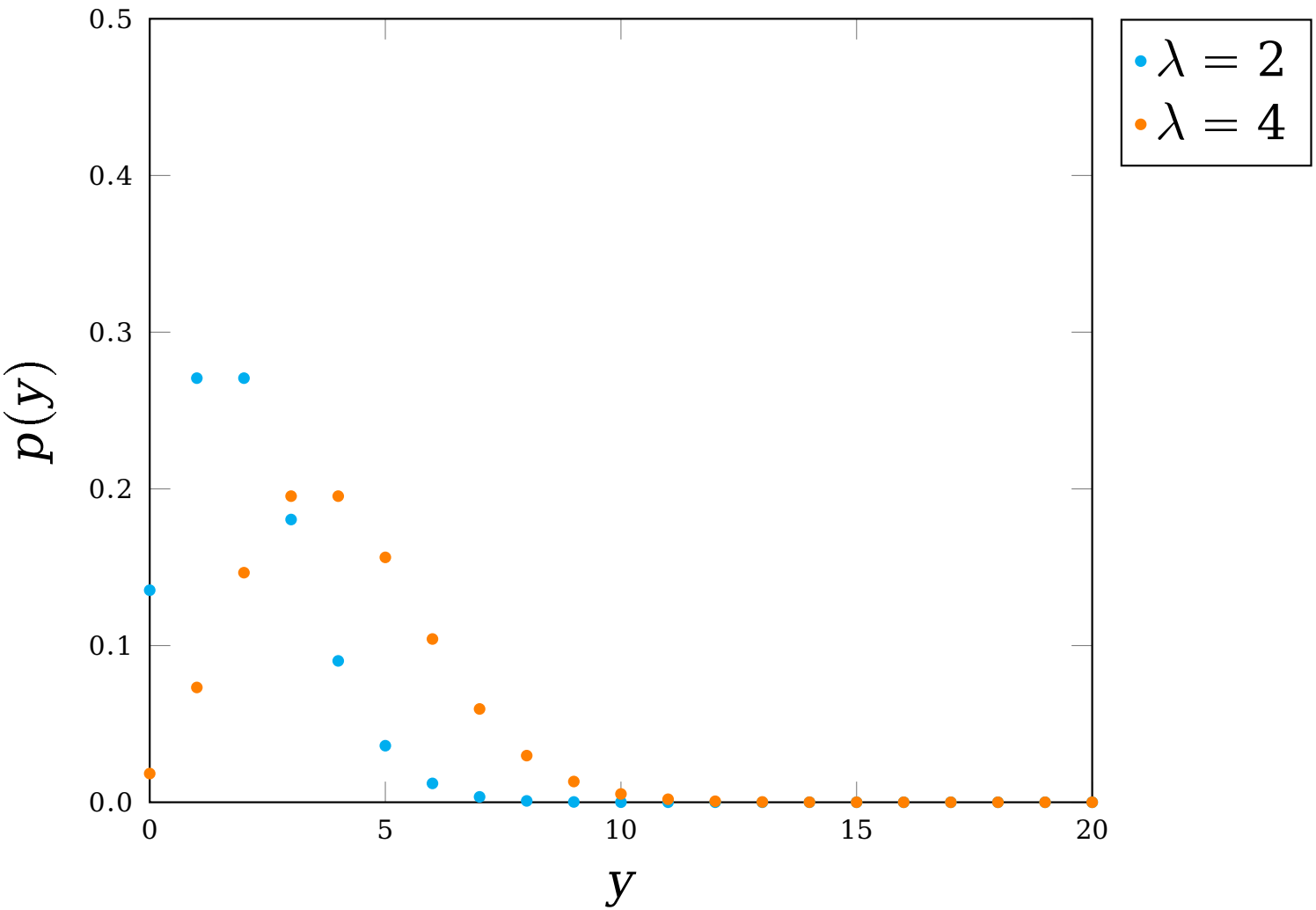
$$p(y) = \frac{\lambda^y}{y!} e^{-\lambda} \qquad y \in \{0, 1, 2, \ldots\}$$

with $p(y) = 0$ for other values of y.

This is the pmf for the Poisson distribution.

The Poisson distribution

Poisson $p(y)$: with $\lambda = 2$ and $\lambda = 4$.



The Poisson distribution

For the Poisson distribution, we have that if

$$Y \sim Poisson(\lambda)$$

then

$$\begin{aligned} \text{Expectation:} \quad E(Y) &= \lambda \\ \text{Variance:} \quad \text{Var}(Y) &= \lambda \end{aligned}$$

Y takes values on the set $\{0, 1, 2, 3, \}$; this distribution does not place probability anywhere else.

The Poisson distribution

Note

The Poisson distribution is widely used to model the distribution of count data, and acts as an approximation to the binomial when

- n is large,
- p is small.

e.g.

$$Binomial(10000, 0.002) \approx Poisson(20)$$

Continuous Random variables

Recall the previous definition of the cumulative distribution function (or cdf), $F(y)$, for a random variable:

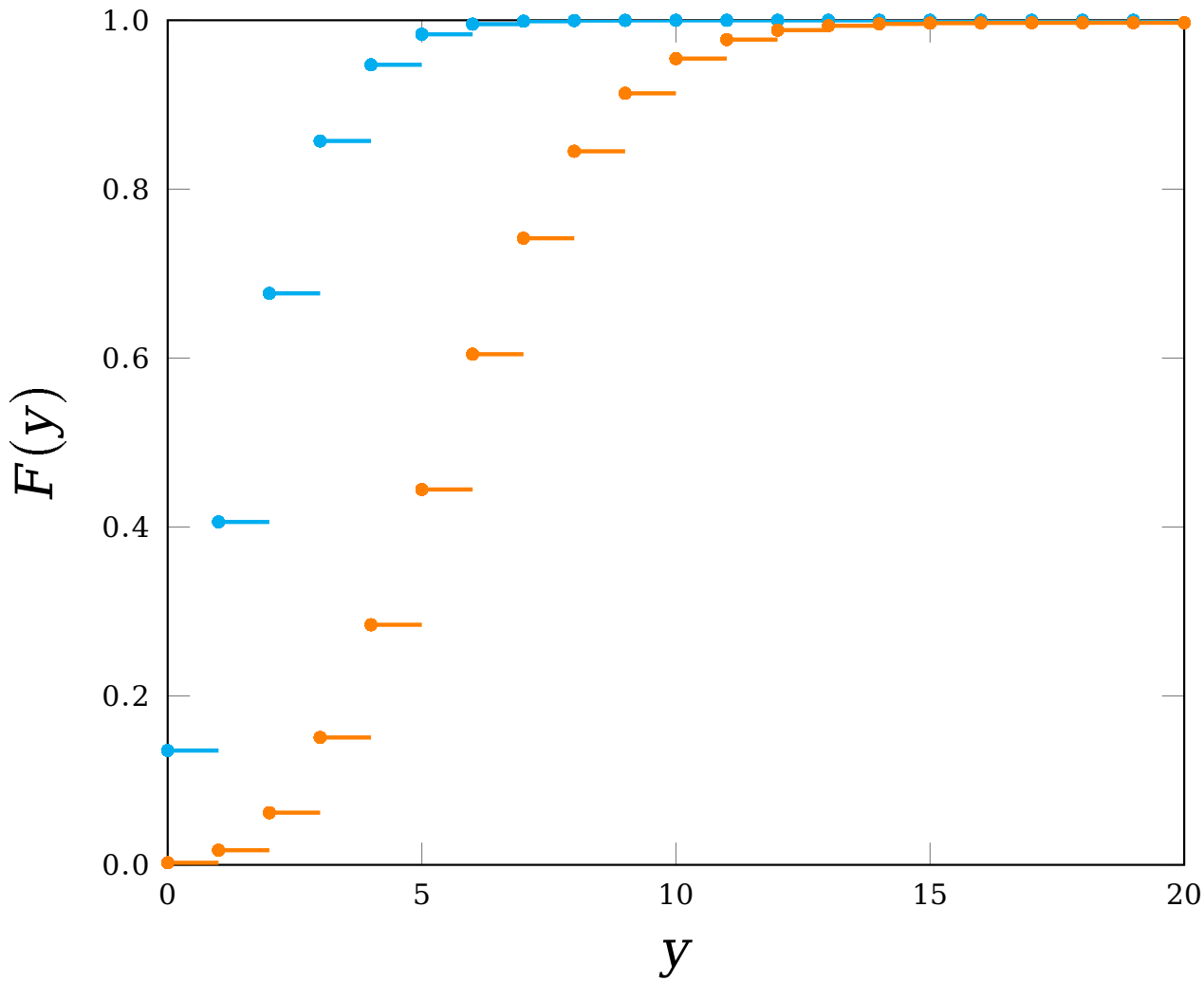
$$F(y) = P(Y \leqslant y)$$

which can be considered for any value y.

For a discrete random variable, this function is termed a step function, with steps at only a countable number of points.

Continuous Random variables

Poisson $F(y)$: with $\lambda = 2$ and $\lambda = 6$.



Continuous Random variables

Now suppose that the random variable Y is most accurately thought of as varying continuously, that is, can potentially take any value in an interval

- height/weight,
- temperature,
- time,

and so on.

Continuous Random variables

A random variable Y with distribution function $F(y)$ is called continuous if $F(y)$ is continuous for all y.

- we forgo the formal mathematical definition here;
- we can interpret to mean 'has no discontinuities'

The probability accumulates in a continuous fashion.

Probability density functions

Let $f(y)$ be the derivative of $F(.)$ at y

$$f(y) = \frac{dF(y)}{dy}$$

wherever this derivative exists.

$f(y)$ is known as the probability density function (pdf).

Probability density functions

It follows that

$$F(y) = \int_{-\infty}^{y} f(t)\, dt$$

that is, $f(y)$ is the function you need to integrate to get $F(y)$.

i.e. you compute probability by performing an "area under the pdf" calculation.

Probability density functions

The pdf is equivalent to the pmf for discrete distributions

$$F(1) = P(Y \leqslant 1) = \int_{-\infty}^{1} f(y)\, dy.$$

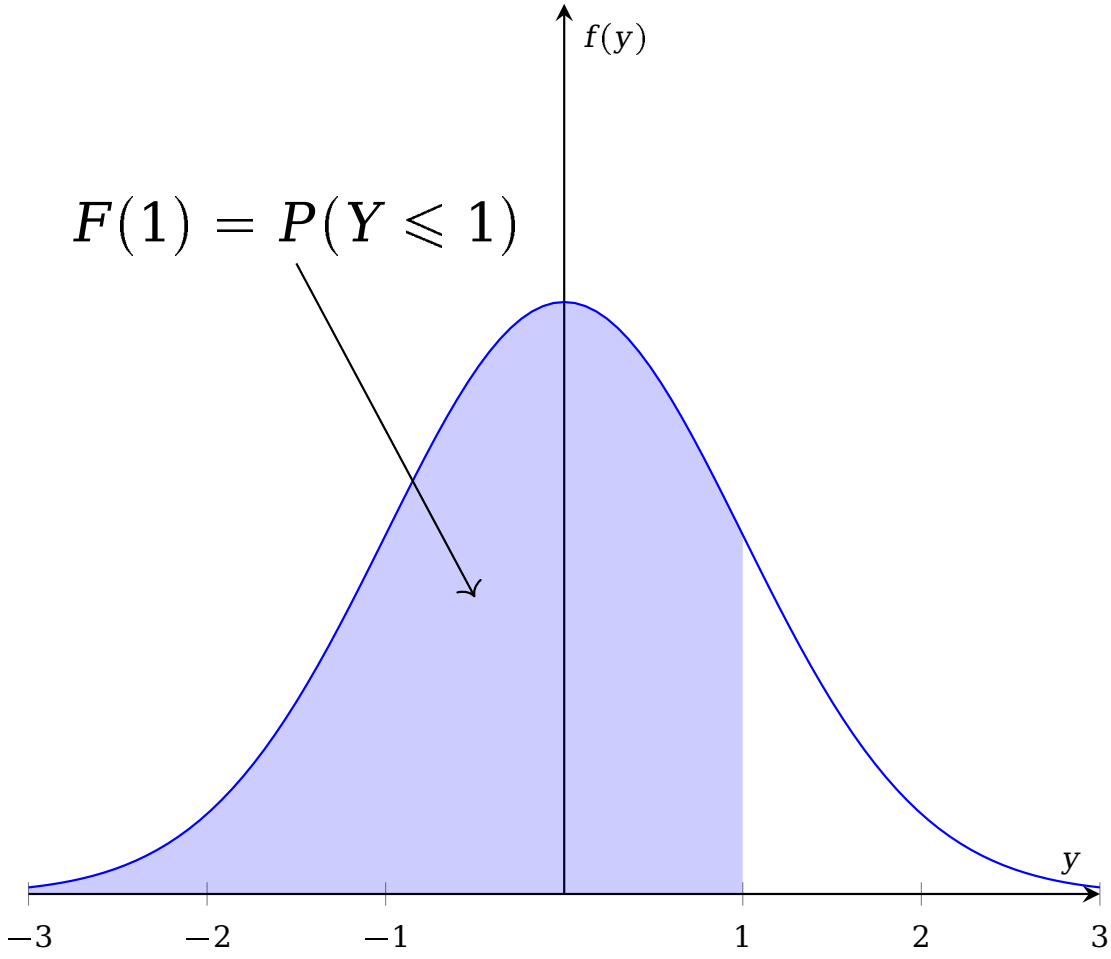
means that to compute $P(Y \leqslant 1)$, we "add up" all the contributions given by $f(y)$ for all the y lying in the range $(-\infty, 1]$.

Similarly

$$P(Y \in (3,5)) = \int_{3}^{5} f(y) dy$$

and so on.

Probability density functions



Probability density functions

We compute probabilities as "area under $f(y)$" calculations:

- the pdf measures 'density', not probability, that is

$$f(y) \neq P(Y = y)$$

- you can think of $f(y)$ as similar to what is measured in the bars in a histogram
- $f(y)$ displays the "shape" of the distribution.

Properties of density functions

From the definition, we have two properties of the pdf $f(y)$:

1. Non-negative:

$$f(y) \geqslant 0 \qquad -\infty < y < \infty$$

   – this follows as $F(y)$ is non-decreasing.

2. Integrates to 1:

$$\int_{-\infty}^{\infty} f(y)\, dy = 1$$

   – this follows as

$$F(\infty) \equiv \lim_{y \longrightarrow \infty} F(y) = 1.$$

Interval probabilities

To compute

$$P(a < Y \leqslant b)$$

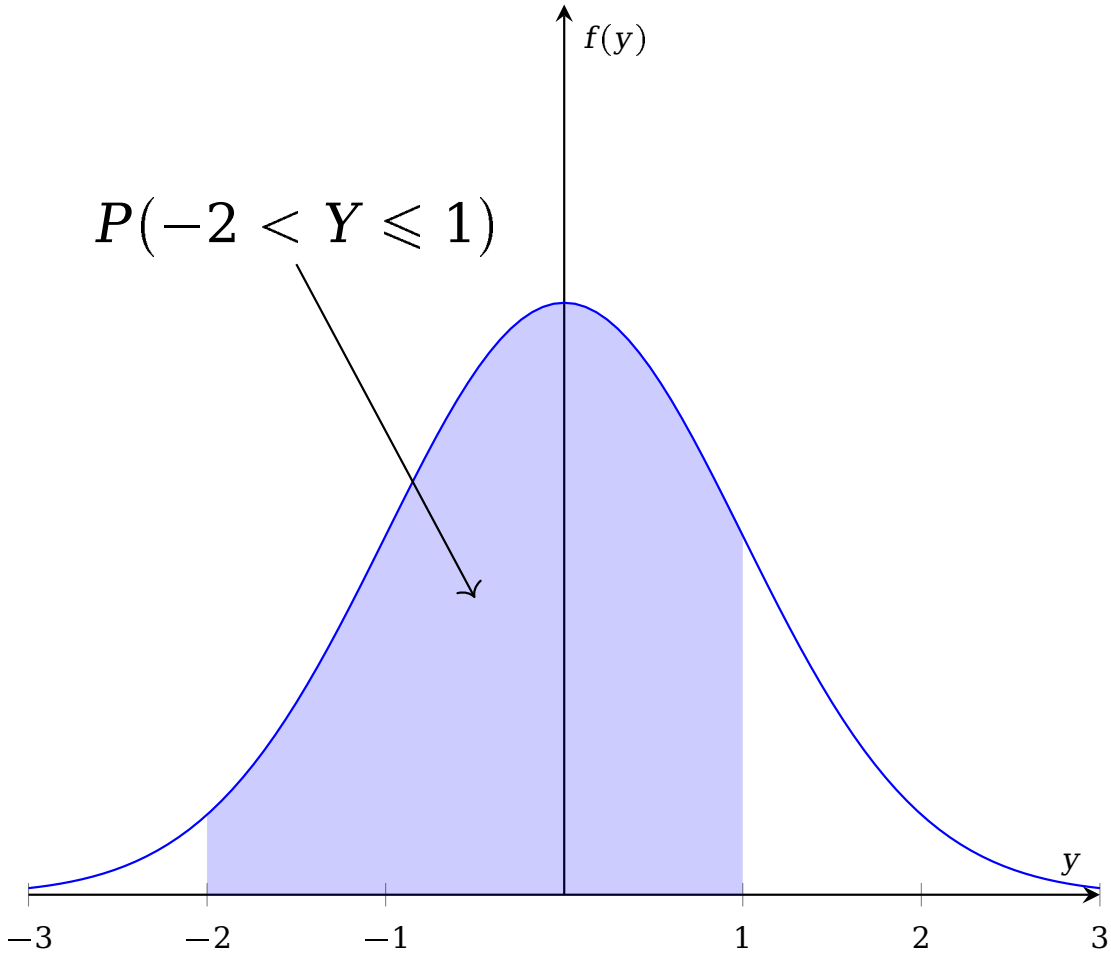
we integrate $f(y)$ over the interval (a, b) as

$$\begin{aligned} P(a < Y \leqslant b) &= P(Y \leqslant b) - P(Y \leqslant a) \\ &= \int_{-\infty}^{b} f(t)\, dt - \int_{-\infty}^{a} f(t)\, dt \\ &= \int_{a}^{b} f(t)\, dt. \end{aligned}$$

e.g.

$$P(-2 < Y \leqslant 1) = \int_{-2}^{1} f(y)\, dy = F(1) - F(-2).$$

Interval probabilities



Interval probabilities

Note

For a continuous random variable

$$\begin{aligned} P(a < Y \leqslant b) &= P(a \leqslant Y \leqslant b) \\ &= P(a \leqslant Y < b) \\ &= P(a < Y < b) \end{aligned}$$

Interval probabilities

Note

We often have that for some $y \in \mathbb{R}$

$$f(y) = 0$$

and so we again denote by $\mathcal{Y}$ the values for which $f(y) > 0$, that is

$$\mathcal{Y} = \{y : f(y) > 0\}$$

Expectation and Variance

The expectation of continuous random variable Y is

$$E(Y) = \int_{-\infty}^{\infty} y\, f(y)\, dy = \int_{\mathcal{Y}} y\, f(y)\, dy = \mu$$

say, provided this integral exists.

For the variance

$$\text{Var}(Y) = E\left[(Y-\mu)^2\right] = \int_{-\infty}^{\infty} (y-\mu)^2 f(y)\, dy$$

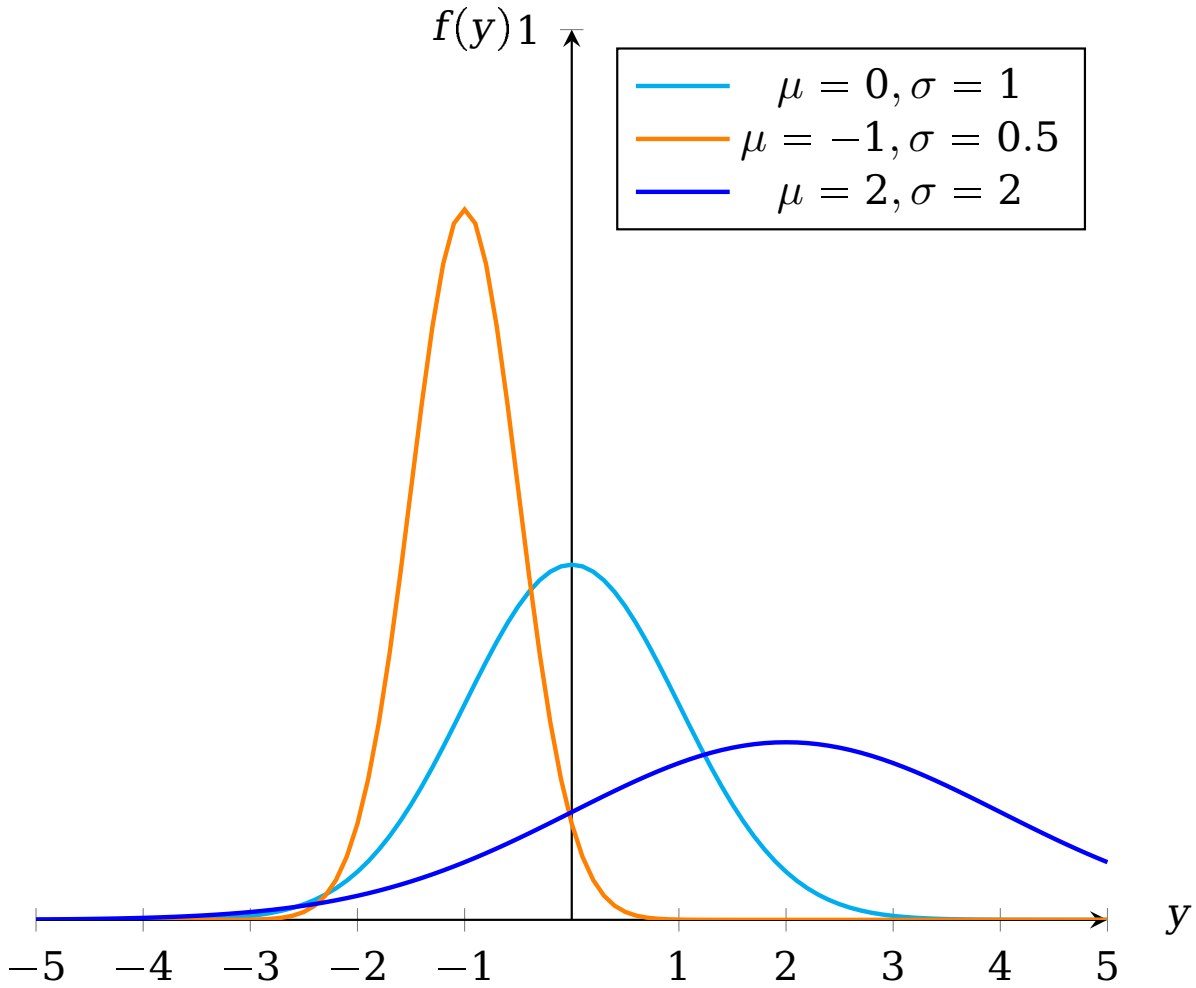
The Normal distribution

The Normal (or Gaussian) distribution has pdf defined on the whole real line by

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\} \qquad -\infty < y < \infty$$

where μ and $\sigma > 0$ are parameters.

The Normal distribution



The Normal distribution

Note

- μ is the centre of the distribution;
  - $f(y)$ is symmetric about μ.
- σ measures the spread of the distribution;
  - as σ decreases, $f(y)$ becomes more concentrated around μ.
- It can be shown that

$$E(Y) = \mu \qquad \text{Var}(Y) = \sigma^2.$$

The Normal distribution

If $\mu = 0$ and $\sigma = 1$ then we have the standard Normal pdf

$$f(y) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{y^2}{2}\right\} \qquad -\infty < y < \infty$$

The Normal distribution

The cdf, $F(y)$, in the general case is given by

$$F(y) = \int_{-\infty}^{y} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(t-\mu)^2\right\}\, dt$$

but this cannot be computed analytically.

Often, this cdf is written using the special notation $\Phi(y)$.

The Normal distribution

If Y has a Normal distribution with parameters μ and σ, and we consider the random variable Z where

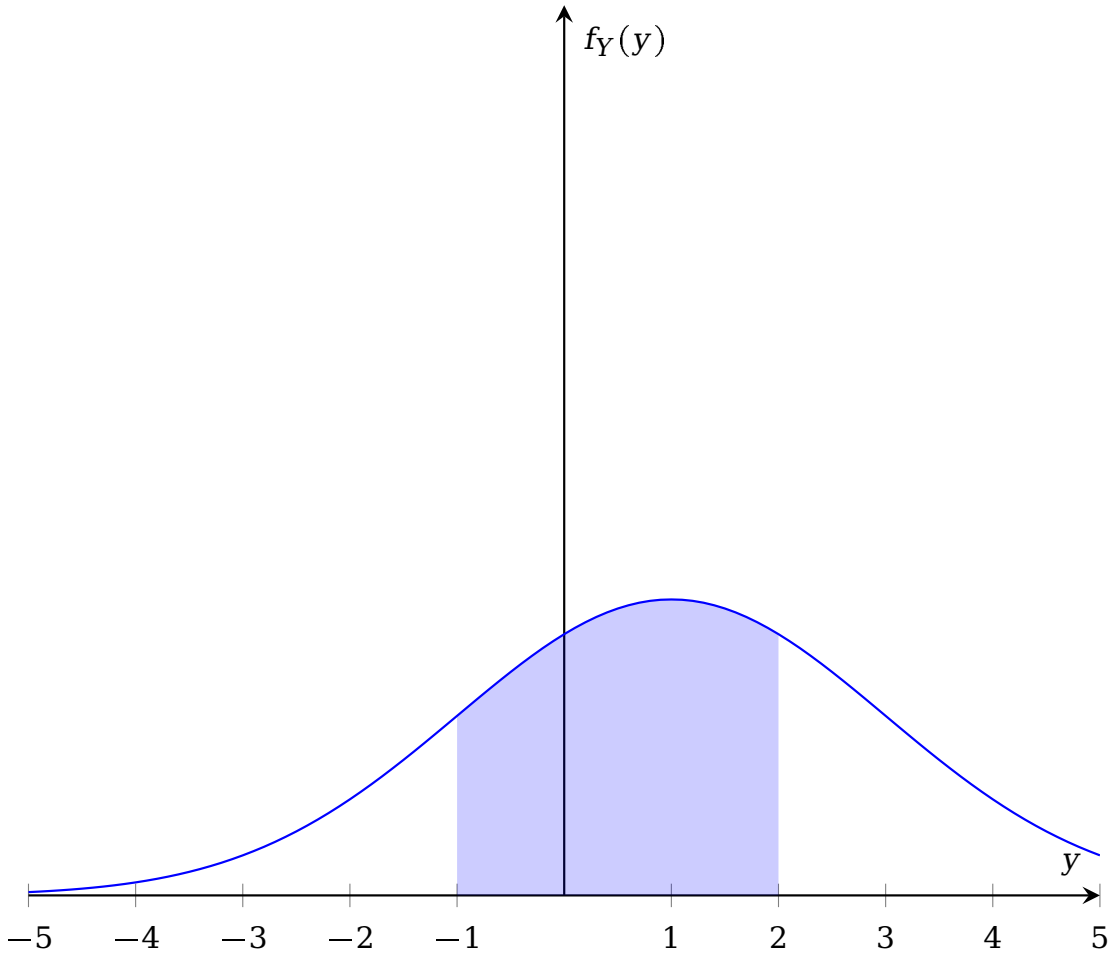
$$Z = \frac{Y - \mu}{\sigma}$$

then Z has a standard Normal distribution.

This result tells us that we can always transform a general Normal random variable Y into a standard Normal random variable Z.

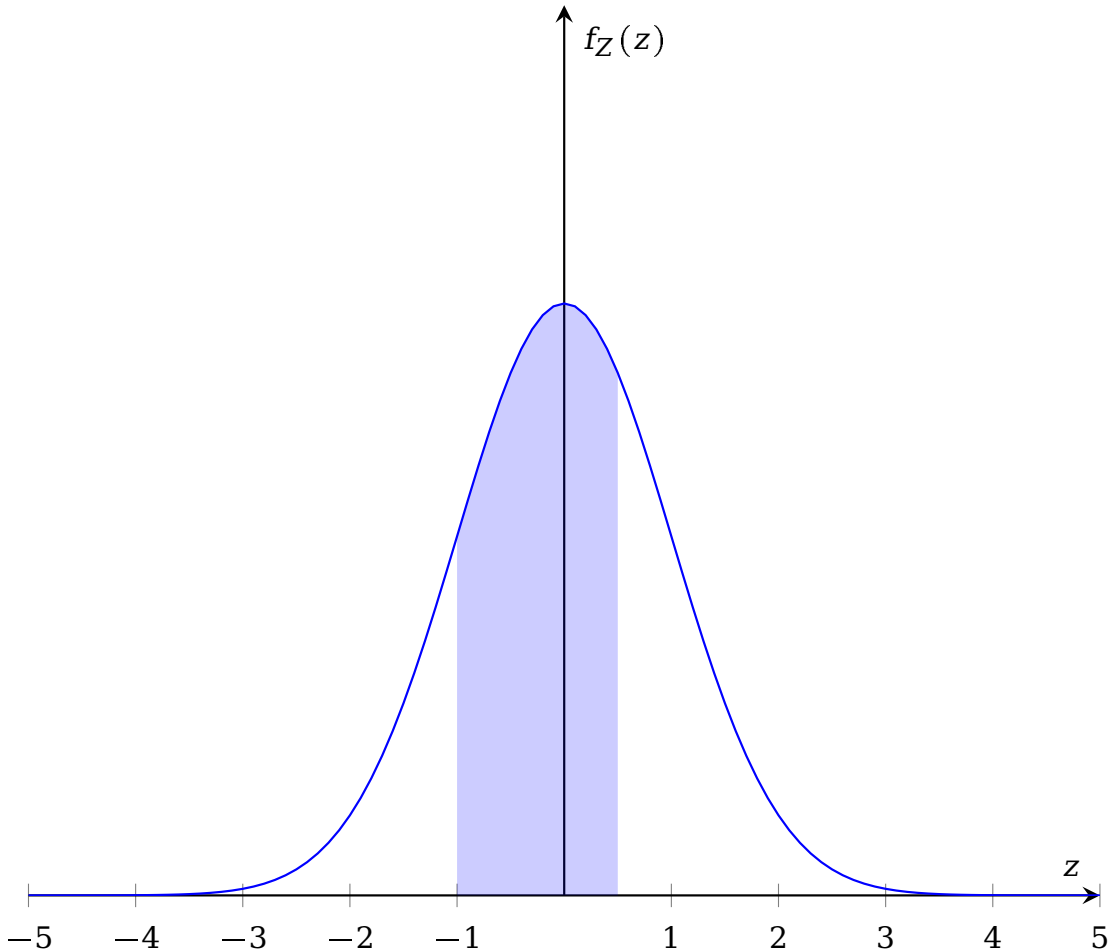
The Normal distribution

$$Y : \mu = 1, \sigma = 2 : P(-1 \leqslant Y \leqslant 2)$$



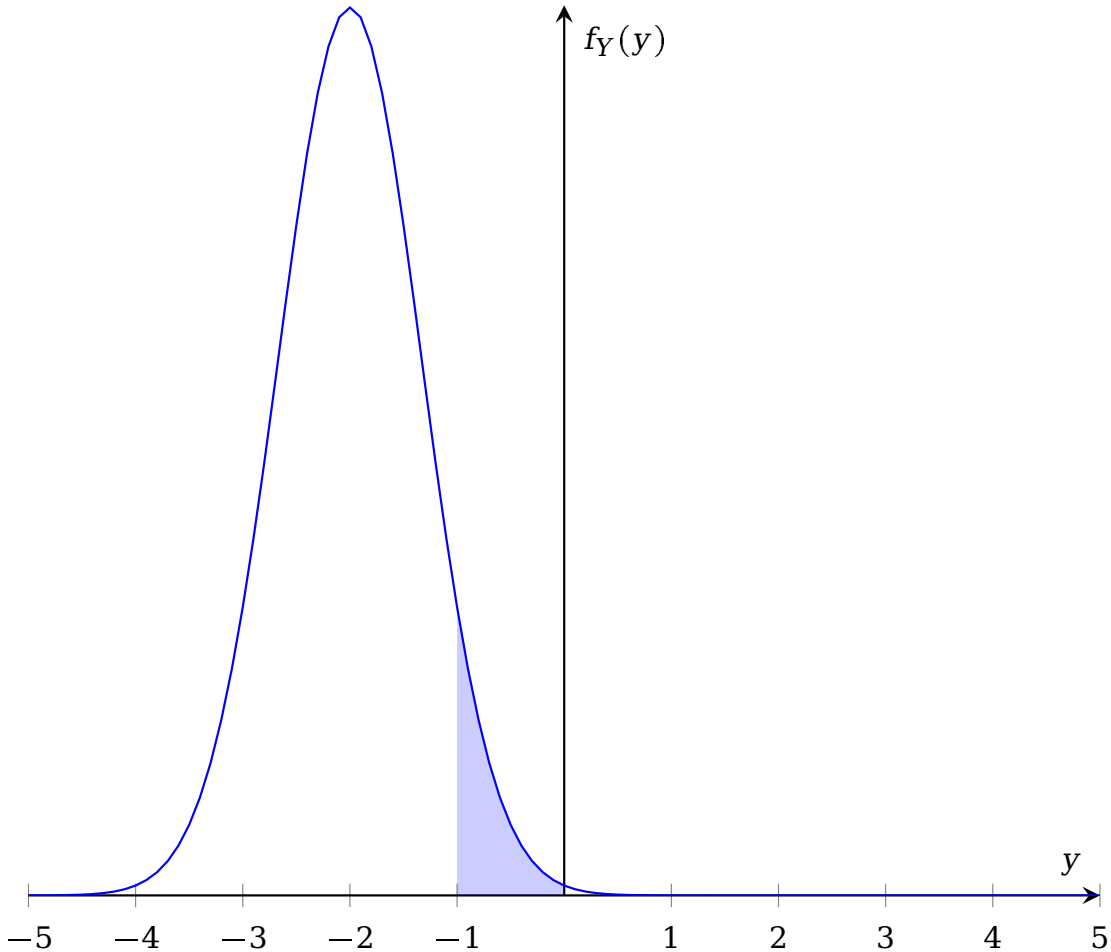
The Normal distribution

$$Z : \mu = 0, \sigma = 1 : P(-1 \leqslant Z \leqslant 1/2)$$



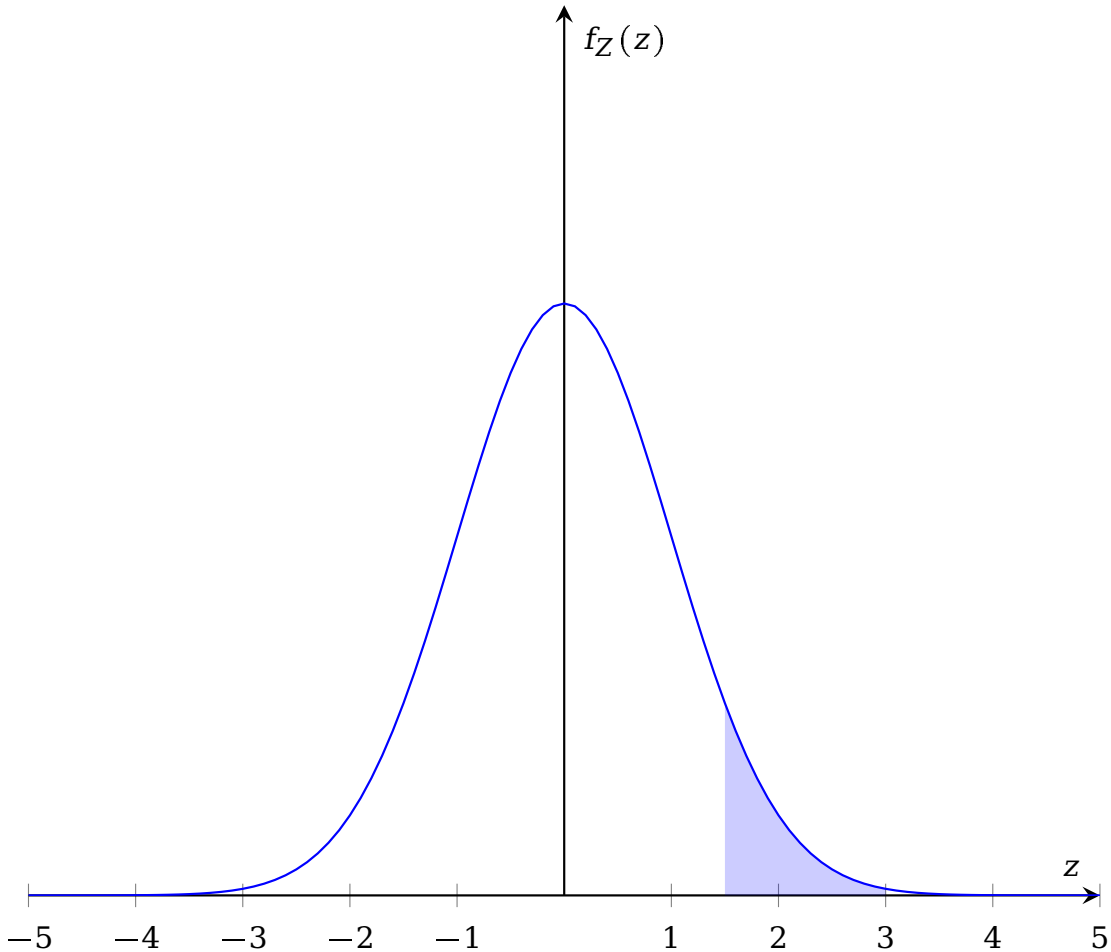
The Normal distribution

$$Y : \mu = -2, \sigma = 2/3 : P(-1 \leqslant Y \leqslant 0)$$



The Normal distribution

$$Z : \mu = 0, \sigma = 1 : P(3/2 \leqslant Z \leqslant 3)$$



Continuous distributions

There is a large number of continuous distributions that can be used to model data. The details of the

- $f(y)$
- $F(y)$

functions change, but the principles of all the calculations remain the same.

In this course, we will meet only two continuous distributions

- Normal distribution
- t-distribution.

Distributions in R

There are a lot of functions in R for computing with probability distributions.

All the standard distributions are available, specifically

- Binomial, Poisson, Geometric
- Normal (and related distributions)

Distributions in R

There are four types of function

1. pmf/pdf functions
2. cdf functions
3. inverse cdf functions
4. random number generating functions

Distributions in R

Example: Normal distribution

For the Normal distribution:

1. pdf function: `dnorm`
2. cdf function: `pnorm`
3. inverse cdf function: `qnorm`
4. random number generating function: `rnorm`.

Distributions in R

Example: Normal distribution

Suppose $X \sim Normal(3.2, 1.5^2)$: to compute

- $f(2.7)$
- $f(-1.2)$

```
dnorm(2.7,3.2,1.5)
```

```
+ [1] 0.2515888
```

```
dnorm(-1.2,3.2,1.5)
```

```
+ [1] 0.003600704
```

The second and third arguments to `dnorm` are $\mu = 3.2$ and $\sigma = 1.5$.

Distributions in R

Example: Normal distribution

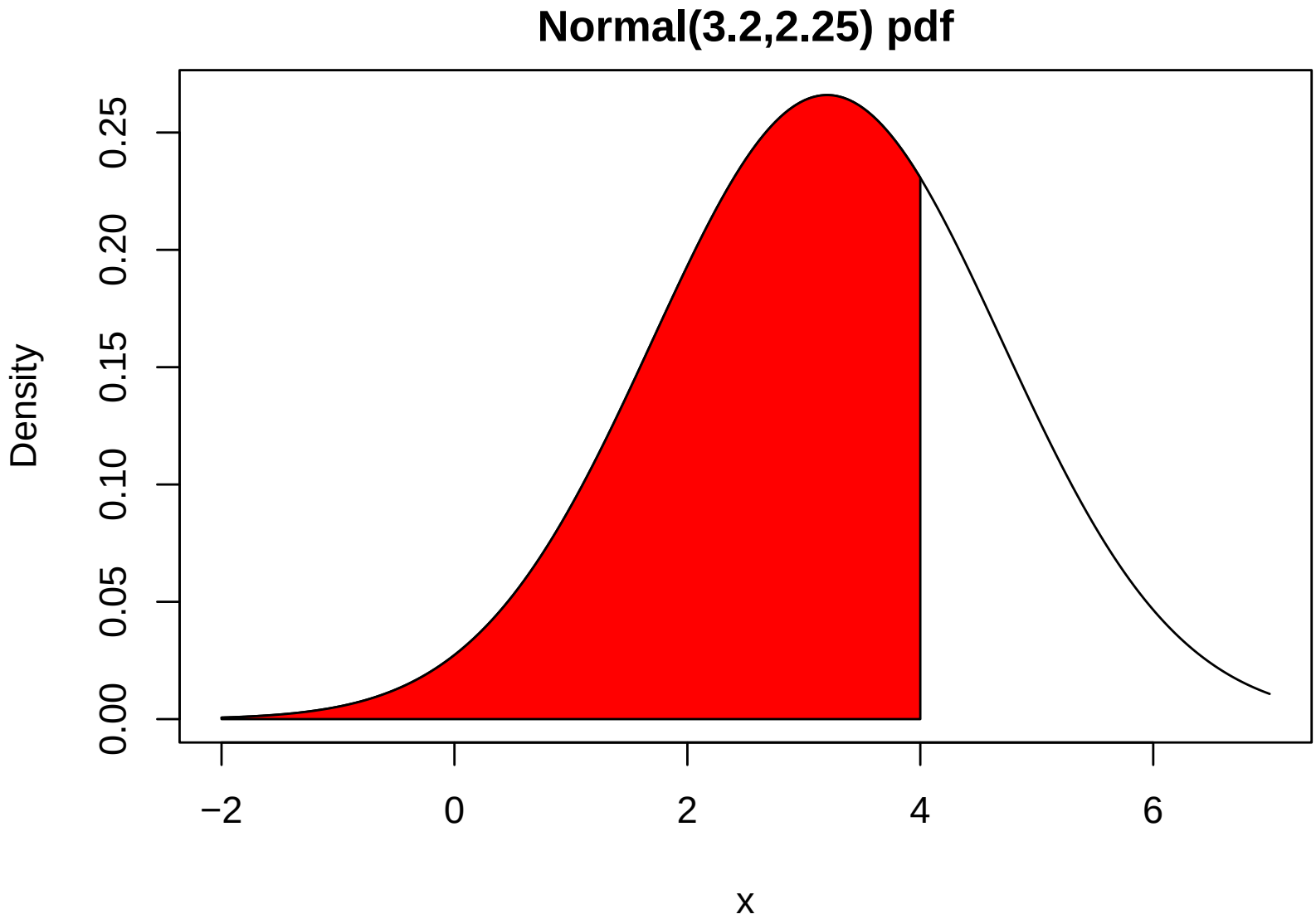
To compute $P(X \leqslant 4.0)$, we use `pnorm`:

```
pnorm(4.0,3.2,1.5)
```

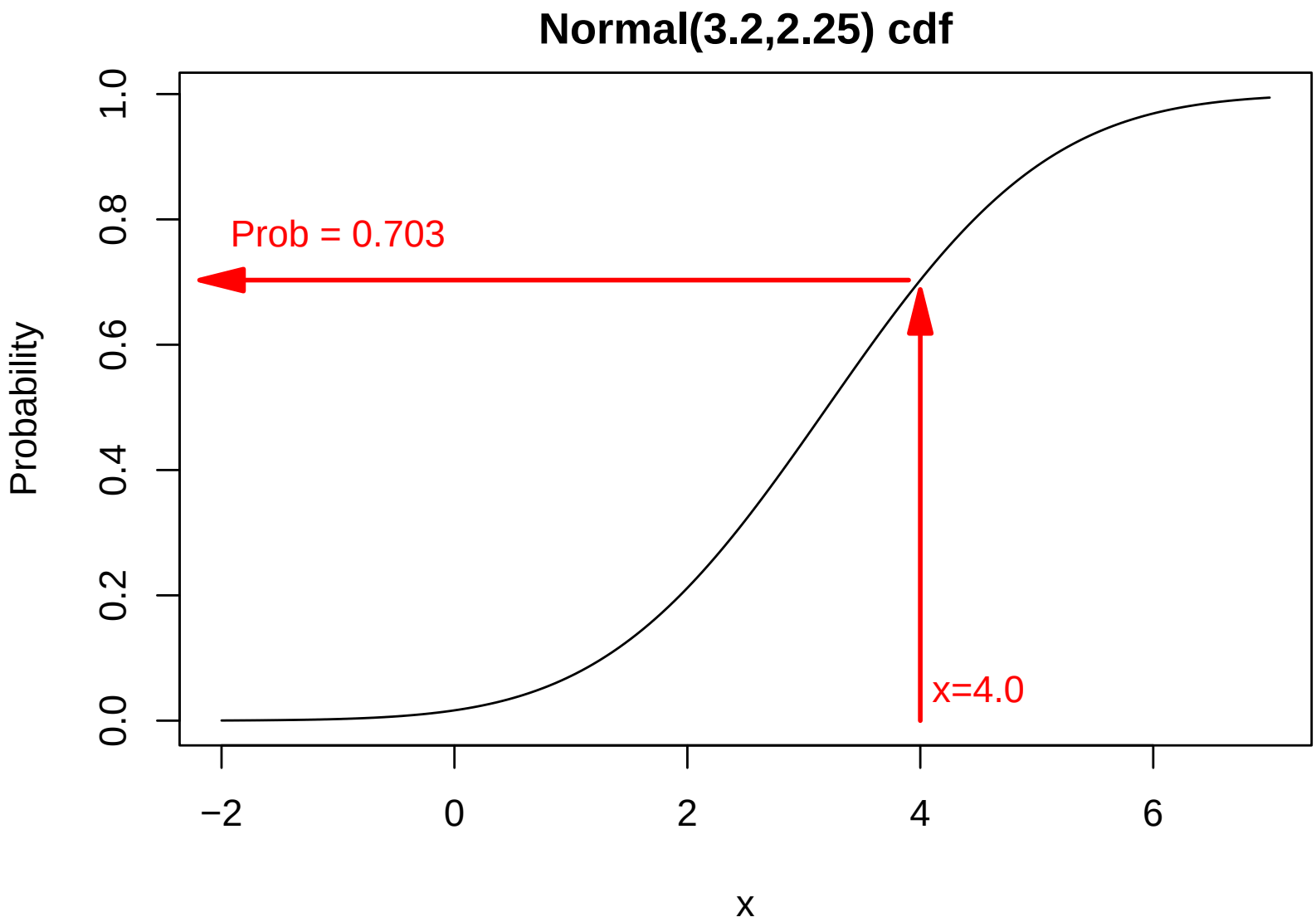
```
+ [1] 0.7030986
```

Recall that this is the area under the Normal pdf from negative infinity up to 4.0.

Distributions in R



Distributions in R



Distributions in R

Example: Normal distribution

We can also compute the x value that solves

$$P(X \leqslant x) = p$$

for some specified p.

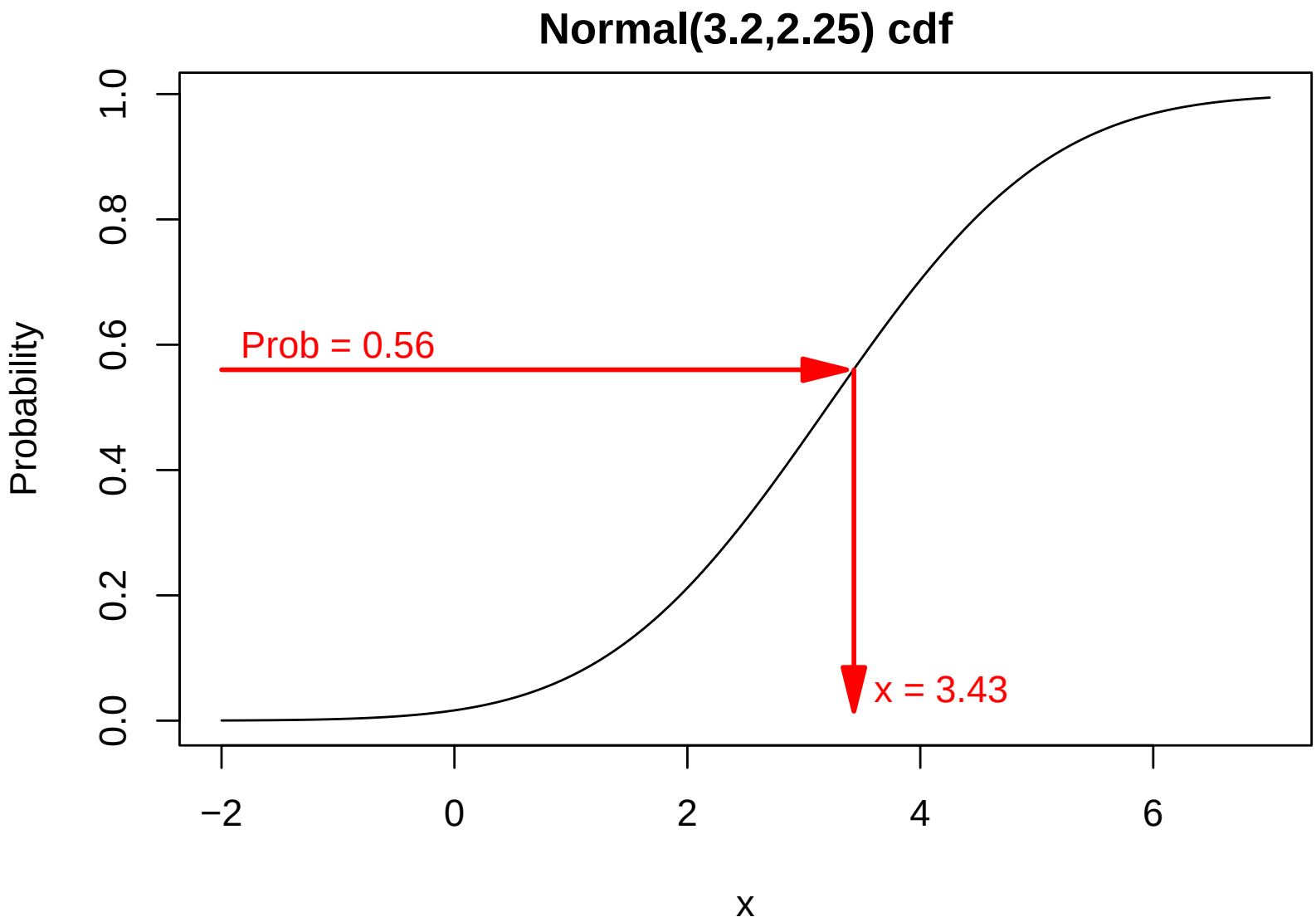
For example, if $p = 0.56$, we use the function `qnorm`.

```
qnorm(0.56,3.2,1.5)
```

```
+ [1] 3.426454
```

This value represents how far we need to go to get an area under the curve to be 0.56.

Distributions in R



Distributions in R

Example: Normal distribution

The function `qnorm` computes the `inverse cdf`. It can be used to compute probability tables.

```
pvec<-c(0.8,0.9,0.95,0.99,0.999)
qvec<-qnorm(pvec)
cbind(pvec,qvec)
```

```
+       pvec      qvec
+ [1,] 0.800 0.8416212
+ [2,] 0.900 1.2815516
+ [3,] 0.950 1.6448536
+ [4,] 0.990 2.3263479
+ [5,] 0.999 3.0902323
```

Distributions in R

Example: Normal distribution

The function `rnorm` simulates values of a Normal random variable

```
n<-10
rnorm(n)
```

```
+  [1] -0.7204912 -0.7894890 -1.1909317 -1.4324848  0.6493333  1.8838497
+  [7] -0.9629458  1.1761699 -0.3092402  1.0042386
```

Distributions in R

Example: Binomial distribution

For the Binomial distribution, we have

- pmf: `dbinom`
- cdf: `pbinom`
- inverse cdf: `qbinom`
- simulated values: `rbinom`

Distributions in R

Example: Binomial distribution

If $X \sim Binomial(300, 0.25)$: for $x = 75, 76, \ldots, 85$

```
n<-300
p<-0.25
x<-c(75:85)
prob<-dbinom(x,n,p)
rbind(x,round(prob,3))
```

```
+     [,1]   [,2]   [,3]   [,4]   [,5]   [,6]   [,7]   [,8]   [,9]  [,10]  [,11]
+ x 75.000 76.000 77.000 78.000 79.000 80.000 81.000 82.000 83.000 84.000 85.000
+    0.053  0.052  0.051  0.048  0.045  0.042  0.038  0.034  0.029  0.025  0.022
```

Distributions in R

Example: Binomial distribution

To compute

$$P(X > 70) = \sum_{j=71}^{300} \binom{300}{j} \left(\frac{1}{4}\right)^{j} \left(1 - \frac{1}{4}\right)^{300-j} = 1 - P(X \leqslant 70)$$

```
1-pbinom(70,n,p)
```

```
+ [1] 0.7233254
```

```
sum(dbinom(71:n,n,p))
```

```
+ [1] 0.7233254
```

Distributions in R

Example: Binomial distribution

For the special case of $Bernoulli(p) \equiv Binomial(1, p)$:

```
rbinom(20,1,p)

+  [1] 0 1 0 0 0 1 0 0 0 0 1 1 1 1 0 0 0 0 0 0
```

Chapter 4
Statistical analysis

Statistical analysis

Using the framework of probability and distributions, we now have a framework to talk about aspects of a statistical analysis that allows us to learn about underlying population characteristics.

Statistical analysis

We will investigate

- sampling distributions
  - probability distributions of summary statistics
- estimates and estimators
  - summary statistics used as best guesses at population parameters
- interval estimates
  - representations of uncertainty about population parameters
- hypothesis tests
  - decision procedures for specific proposed properties of the population.

Random samples

In our analyses, we will assume that we have access to one or more random samples of data

$$x_1, \ldots, x_n$$

where we presume

- each data point is collected under the same conditions
- the data points are collected such that they do not influence one another.

Random samples

In terms of the underlying probability model, the corresponding random variables

$$X_1, \ldots, X_n$$

are assumed to be independent and to have the same probability distribution.

We shorten this to i.i.d

- independent and identically distributed.

Sampling distributions: big idea

In data collection

- we collect data $x_1, \ldots, x_n$ in an attempt to learn about a population and its parameters
- we summarize the data using a statistic, such as $\overline{x}$
- the finite sample of size n is just one data set
- if we repeated the data collection, we would get a slightly different collection of data, and a different value of $\overline{x}$

We wish to understand how $\overline{x}$ is going to vary over repeated experiments.

Sampling distributions

Example: Sample mean for a discrete distribution

A large "population" which consists of the numbers 1,2 and 3, such that 30% are 1s, 40% are 2s and 30% are 3s

- Let X be the random variable that denotes the result obtained from selecting at random one element from this population
- The probability distribution of X is given by

$X = x$	1	2	3
$P(X = x)$	0.3	0.4	0.3

We will say that the population has the distribution given by the table above

Sampling distributions

Example: Sample mean for a discrete distribution

The expected value of X is

$$E(X) = 1 \times 0.3 + 2 \times 0.4 + 3 \times 0.3 = 2$$

- The population variance and standard deviation of X are

$$\text{Var}(X) = 0.6 \qquad \text{sd}(X) = \sqrt{0.6} = 0.77.$$

- We will say that the population mean is $\mu = 2$ and the population standard deviation $\sigma = 0.77$.

The population quantities represent what would be obtained for the sample quantities if you had a massive sample.

Sampling distributions

Example: Sample mean for a discrete distribution

Now suppose we have a random sample of size n

$$X_1, \ldots, X_n$$

from this distribution. The sample mean random variable is

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Sampling distributions

Example: Sample mean for a discrete distribution

$\overline{X}$ is also a random variable

- before we see the observed data, we do not know what value $\overline{X}$ will take
- we can see it will take a value in

$$\left\{ \frac{n}{n}, \frac{(n+1)}{n}, \frac{(n+2)}{n}, \ldots, \frac{3n}{n} \right\}.$$

but it is not immediately obvious what the value of

$$P(\overline{X} = x)$$

is for each x.

Sampling distributions

Note

- The “distribution” for a population is the probability distribution of the random variable X, the result of selecting one unit from the population
- We will interchangeably use the terms ‘distribution of a population’ and ‘distribution of a random variable’, defined by sampling at random one unit for the population
- The distribution of a population can be
  - discrete (Bernoulli, Binomial) or
  - continuous (Normal)

$\overline{X}$ as a Random Variable

- Suppose that we select two numbers at random and independently from a population
- Thus, we have now two random variables, say X_1 and X_2
- Because they are coming from the same population, X_1 and X_2 have the same probability distribution.

$\overline{X}$ as a Random Variable

- X_1 and X_2 are a independent and identically distributed (i.i.d.) random variables from the same population
  - Independence:

    $$P\left(X_1 = x_1 \text{ and } X_2 = x_2\right) = P\left(X_1 = x_1\right) \times P\left(X_2 = x_2\right)$$

    for all values $\left(x_1, x_2\right)$
  - Identically distributed:

    $$P\left(X_1 = x\right) = P\left(X_2 = x\right)$$

    for any x

$\overline{X}$ as a Random Variable

In the previous example

$X = x$	1	2	3
$P(X = x)$	0.3	0.4	0.3

this implies that

$$E(X_1) = E(X_2) = \mu = 2 \qquad \text{Var}(X_1) = \text{Var}(X_2) = 0.6$$

$\overline{X}$ as a Random Variable

Let us say now that after we sample the two numbers, we plan compute their mean.

$$\overline{X} = \frac{X_1 + X_2}{2}.$$

As we do not know what values we will get before we do the sampling, $\overline{X}$ is a random variable.

After we have sampled the two numbers, say $(1, 3)$ then we will have an observed value for $\overline{X}$, specifically

$$\overline{x} = \frac{x_1 + x_2}{2} = \frac{1 + 3}{2} = 2.$$

$\overline{X}$ as a Random Variable

What is the probability distribution of $\overline{X}$?

The possible values of $\overline{X}$ are

$(1,1)$	implies $\overline{X} = 1$
$(1,2)$ or $(2,1)$	imply $\overline{X} = 3/2$
$(1,3)$ or $(3,1)$ or $(2,2)$	imply $\overline{X} = 4/2 = 2$
$(2,3)$ or $(3,2)$	imply $\overline{X} = 5/2$
$(3,3)$	implies $\overline{X} = 6/2 = 3$

$\overline{X}$ as a Random Variable

The probability of each value of $\overline{X}$ is

$$
\begin{aligned}
P(\overline{X} = 1) &= 0.3 \times 0.3 = 0.09 \\
P\left(\overline{X} = 3/2\right) &= (0.3 \times 0.4) + (0.4 \times 0.3) = 0.24 \\
P(\overline{X} = 2) &= (0.3 \times 0.3) + (0.3 \times 0.3) + (0.4 \times 0.4) = 0.34 \\
P\left(\overline{X} = 5/2\right) &= (0.4 \times 0.3) + (0.3 \times 0.4) = 0.24 \\
P(\overline{X} = 3) &= (0.3 \times 0.3) = 0.09
\end{aligned}
$$

$\overline{X}$ as a Random Variable

We can say that the possible values of $\overline{X}$ form a population with distribution

$X = \overline{x}$	1	1.5	2	2.5	3
$P(X = x)$	0.09	0.24	0.34	0.24	0.09

The probability distribution of $\overline{X}$ is determined by the probability distribution of X_1 and X_2

$\overline{X}$ as a Random Variable

Simulating in R:

```
prob.vec<-c(0.3,0.4,0.3)
xvals<-c(1,2,3)
n<-2
nreps<-100000
xsamp<-replicate(nreps,mean(sample(xvals,prob=prob.vec,size=n,rep=TRUE)))
xsamp[1:10]

+  [1] 2.5 2.5 2.0 3.0 2.5 2.0 1.0 1.5 1.0 2.0

table(xsamp)

+ xsamp
+     1   1.5     2   2.5     3
+  9097 23939 33961 23851  9152

table(xsamp)/nreps

+ xsamp
+       1     1.5       2     2.5       3
+ 0.09097 0.23939 0.33961 0.23851 0.09152
```

$\overline{X}$ as a Random Variable

Consider a sequence of $X_1, X_2 \ldots, X_n$ i.i.d. random variables from this population, and define the random variable

$$\overline{X} = \frac{X_1 + X_2 \ldots + X_n}{n}$$

We can compute the probability distribution of $\overline{X}$ using the same procedure.

- the calculation will be more complicated
- we may seek to approximate the distribution

$\overline{X}$ as a Random Variable

- We have a plan for data collection:
  - we will make n observations, independently and under identical conditions,
  - once we observe our sample data we will sum our observations and divide by n.
- We will store the observations as the quantities

$$X_1, X_2, \ldots, X_n$$

- These quantities are random variables
- We use them to a new quantity: $\overline{X}$
  - $\overline{X}$ is itself a random variable,
  - its distribution is defined by the that of the X_i.

The Central Limit Theorem

Sample averages play a big part in probability and statistics

- We want to make probability statements about sample averages
- It is often difficult to say what the probability distribution of an average or mean of random variables is directly.

The Central Limit Theorem

- A theorem called the Central Limit Theorem which can come to our rescue, if the number of items we're adding or averaging is "large"
- As long as the random variables being added are independent and have the same probability distribution, if n is large enough, then their sum and their average have an approximate Normal distribution.

The Central Limit Theorem

Theorem: Central Limit Theorem

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed random variables with population expectation μ and population standard deviation σ. Then if

$$S_n = X_1 + X_2 \ldots + X_n,$$

we have that

$$S_n \dot{\sim} N\left(n\mu, n\sigma^2\right)$$

when n is "large".

The Central Limit Theorem

The notation

$$\dot{\sim} \quad \text{means} \quad \text{“is approximately distributed as”}$$

that is, the distribution function of S_n is very close to that of the Normal.

We can use the Normal distribution function to compute interval probabilities for S_n: e.g.

$$P(l < S_n \leqslant u) \approx \Phi\left(\frac{u - n\mu}{\sqrt{n\sigma^2}}\right) - \Phi\left(\frac{l - n\mu}{\sqrt{n\sigma^2}}\right)$$

The Central Limit Theorem

Alternatively, if

$$\overline{X} = \frac{1}{n}\left(X_1 + X_2 + \ldots + X_n\right)$$

then

$$\overline{X} \dot{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$$

In practice, “large” is often taken to be $n \geqslant 30$.

The Central Limit Theorem

By results for the Normal distribution, we have

$$Z_n = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \stackrel{\cdot}{\sim} Normal(0, 1)$$

The Central Limit Theorem

Example: Previous example

If

$X = x$	1	2	3
$P(X = x)$	0.3	0.4	0.3

and $n = 50$.

The Central Limit Theorem

Example: Previous example

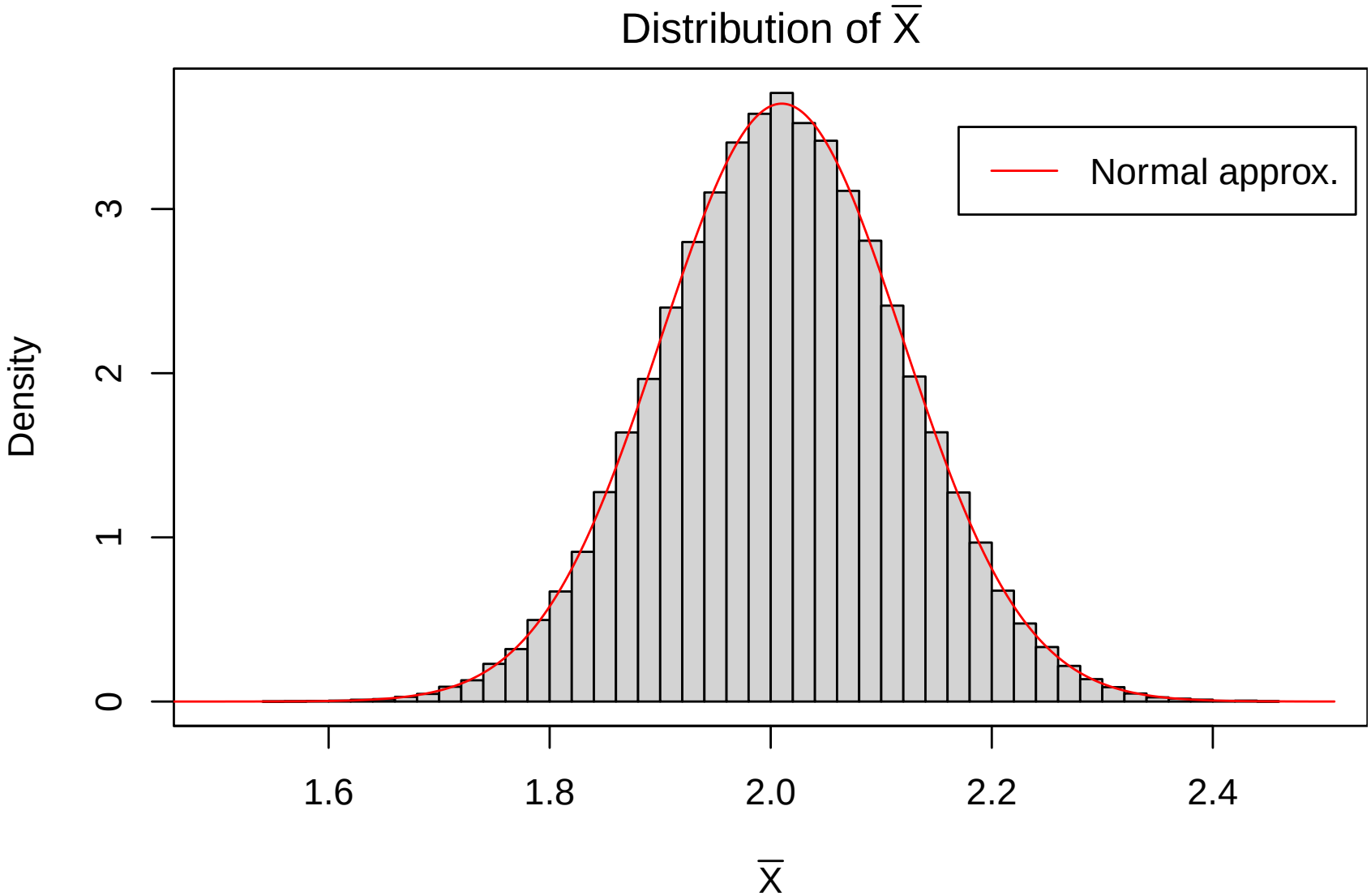
Then by the Central Limit Theorem

$$\overline{X} \dot{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$$

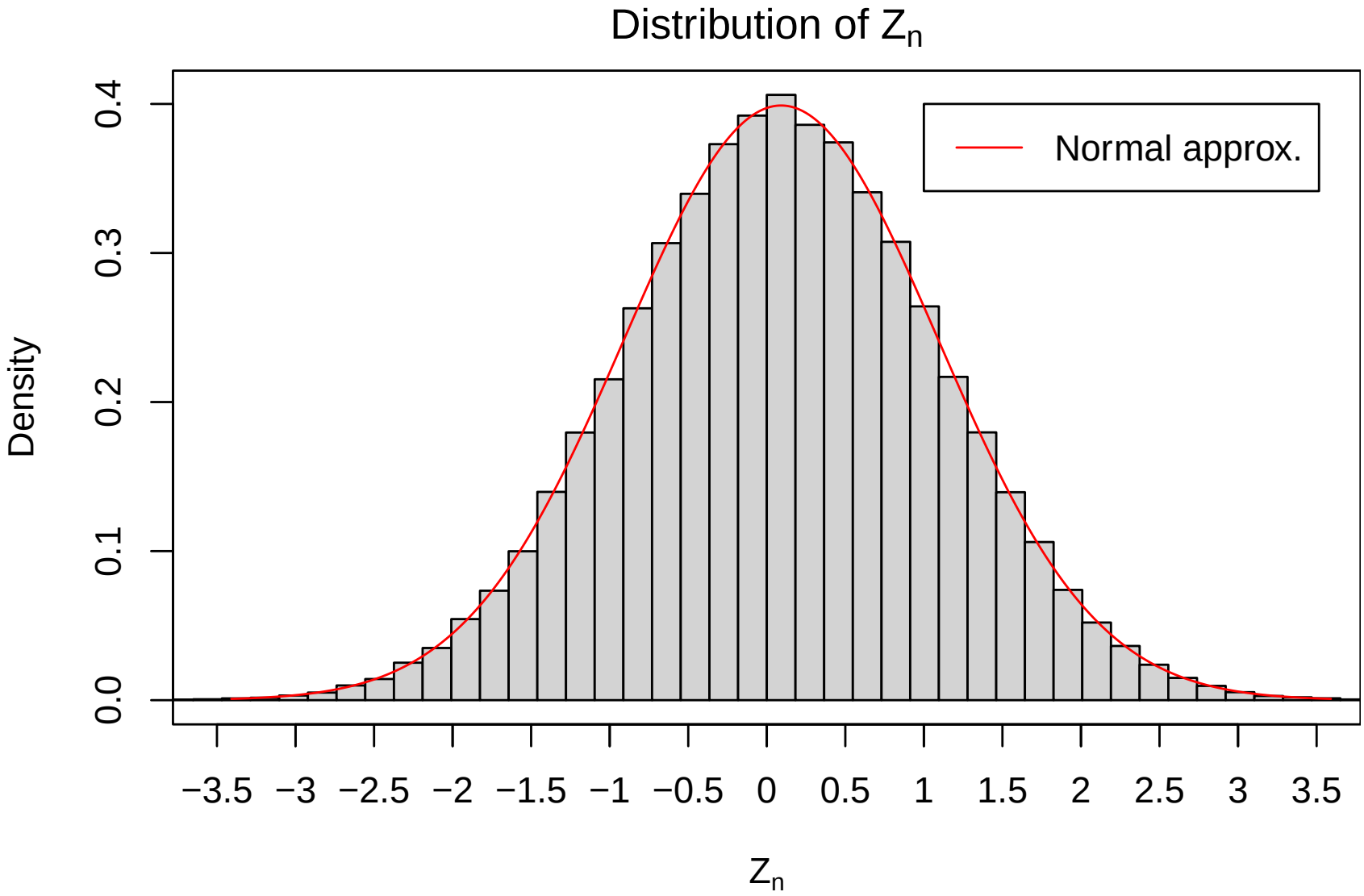
where $\mu = 2$ and $\sigma = \sqrt{0.6}$.

In the following plots, $n = 50$.

The Central Limit Theorem



The Central Limit Theorem



The Central Limit Theorem

Note

- If the individual random variables $X_1, \ldots, X_n$ are themselves Normally distributed, then $\overline{X}$ and S_n are in fact exactly Normally distributed, no matter the value of n
- The Central Limit Theorem does not mean that a large number of independent random variables magically become Normal themselves - it's their sum/average that becomes (approximately) Normal

The Central Limit Theorem

The Central Limit Theorem is important in statistics, since $\overline{X}$ plays a crucial role in many areas of statistical inference

The Central Limit Theorem

Problem: CD4 counts

We are interested in looking at CD4 count in people with HIV. Suppose the CD4 count follows a Normal distribution, with mean 300 and standard deviation 80.

1. For a single randomly selected person with HIV, what is the probability that their CD4 count will be less than 200?
2. If we sample, at random and independently, 50 people with HIV, what is the probability that the sample average of their CD4 counts will be less than 200?

The Central Limit Theorem

Solution: CD4 counts

1. Let X be the CD4 count of a person with HIV: X is Normal with $\mu = 300$ and $\sigma = 80$

$$X \sim Normal(300, 80^2)$$

Note that the z-score transformation implies

$$Z = \frac{X - 300}{80} \sim Normal(0, 1)$$

The Central Limit Theorem

Solution: CD4 counts

We want to find $P(X < 200)$: using tables or software

$$P(X < 200) = P\left(\frac{X - 300}{80} < \frac{200 - 300}{80}\right)$$

$$= P(Z < -1.25)$$

$$= 0.1056$$

That is, the probability that a randomly chosen person with HIV has a CD4 count of less than 200.

The Central Limit Theorem

Solution: CD4 counts

2. We must find $P(\overline{X} < 200)$: we assume that we have a sequence $X_1, X_2, \ldots, X_{50}$, where X_i represents the CD4 count of subject $i, i = 1, 2, \ldots, 50$.

   We know the each $X_i \sim Normal(300, 80^2)$, and it can be shown that

   $$\overline{X} \sim Normal(300, 80^2/50)$$

   that is, based on our assumptions about the individual X_i, we can show that $\overline{X}$ is also normally distributed, but with

   $$\text{Mean} = 300 \qquad \text{s.d.} = \frac{80}{\sqrt{50}} = 11.3137$$

Solution: CD4 counts

Again using the z-score transform.

$$P(\overline{X} < 200) = P\left(\frac{\overline{X} - 300}{80/\sqrt{n}} < \frac{200 - 300}{80/\sqrt{n}}\right)$$

$$= P(Z < -8.8388)$$

$$= 0.0000$$

So, the probability that the average CD4 cell count will be less than 200 is approximately zero (to four decimal places).

The Central Limit Theorem

Problem: Book lengths

Suppose that the mean book length for science fiction books from a certain publisher is 350 pages, with a standard deviation of 42 pages. However, we do not know anything else about the distribution of the book lengths.

If we sample 75 books, what is the probability that their sample mean length will be greater than 360 pages?

Solution: Book lengths

Let

$$\overline{X} = \frac{1}{75}\sum_{i=1}^{75} X_i$$

be the sample mean of the book lengths. We have

$$\mu = E\left(X_i\right) = 350 \qquad \sigma = sd\left(X_i\right) = 42$$

for any $i = 1, \ldots, 75$.

We want to find $P(\overline{X} > 360)$.

The Central Limit Theorem

Solution: Book lengths

$n = 75$ is large enough, so by the Central Limit Theorem

$$\overline{X} \dot{\sim} \mathit{Normal}(\mu, (\sigma/\sqrt{n})^2)$$

Again using the z-score transform.

$$P(\overline{X} > 360) = P\left(\frac{\overline{X} - 350}{42/\sqrt{n}} > \frac{360 - 350}{42/\sqrt{n}}\right)$$

$$= P(Z > 2.0620)$$

$$= 1 - P(Z \leqslant 2.0620)$$

$$= 1 - 0.9804 = 0.0196$$

Solution: Book lengths

Thus

$$P(\overline{X} > 360) = 0.0196$$

Point estimators: big idea

We want to use data to summarize aspects of the population from which the data are drawn.

The population can be described by the parameters of the probability distribution that generates the data.

Therefore, we wish to use the data to construct our best representations of those parameters.

Point estimators

Suppose we have a population which is described through a distribution that depends on a parameter θ.

In this course we will take θ to be

- μ (population mean), or
- p in the Binomial set-up (population proportion)

We wish to "guess" at target parameter θ by observing a random sample of independent and identically distributed (i.i.d.) observations from this population

Point estimators

Let $X_1, X_2, \ldots, X_n$ be the random variables which comprise this sample. We call any function of $X_1, \ldots, X_n$ which does not depend on any unknown quantities an estimator of θ.

- We denote the estimator function by $\widehat{\theta}$, i.e.

$$\widehat{\theta} = \widehat{\theta}\,(X_1, \ldots, X_n)$$

  - $\widehat{\theta}$ must be independent of any unknown parameters, otherwise we could not use it as an estimator of θ
- An estimator is a random variable since it depends on the random variables $X_1, \ldots, X_n$
- θ itself is a constant (despite being unknown)

Point estimators

Once we conduct our experiment and have values $x_1, \ldots, x_n$ for our random variables $X_1, \ldots, X_n$.

The result $\widehat{\theta}(x_1, \ldots, x_n)$ is a numerical value, and will be considered an estimate of θ.

Properties of Estimators

Estimators are random variables and therefore have expected value and variance

For a given parameter θ, there may be many estimators

- Statisticians have agreed on two properties that indicate a "good" estimator:
  - Unbiasedness
  - Minimum Variance

Properties of Estimators

Unbiasedness: Estimator $\widehat{\theta}$ is unbiased of θ for if and only if

$$E(\widehat{\theta}) = \theta \quad \text{for all } \theta$$

- $\widehat{\theta}$ is a random variable
- $\widehat{\theta}$ has its own distribution
- this distribution has a population mean
- for unbiasedness, this population mean must be θ

Properties of Estimators

Minimum Variance: if $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are two unbiased estimators for the same target parameter θ,

$$E(\widehat{\theta}_1) = E(\widehat{\theta}_2) = \theta$$

we will always prefer the estimator with the smaller variance.

Estimating a population mean

Let $X_1, \ldots, X_n$ be a random sample of i.i.d. random variables from a distribution with unknown expected value μ

It turns out that in almost all cases, the "best" estimator for μ (i.e. unbiased with the minimum variance among all the unbiased estimators) is

$$\widehat{\mu} = \overline{X}$$

Estimating a population mean

Example: Painkiller duration

A drug manufacturer is estimating the duration (time from administration to relief of the pain) of their new over-the-counter painkiller.

- In this experiment we have a hypothetical population of potential users of the painkiller.
- We are interested in an estimate of the mean painkiller duration, say μ, in this population.
- The parameter μ is a constant, but is unknown to us.

Estimating a population mean

Example: Painkiller duration

Suppose that 18 subjects are taken at random and independently and the duration of the painkiller is measured on each of them. We have a sample of (i.i.d.) random variables

$$X_1, X_2, \ldots X_{18}$$

where X_i is the duration of the painkiller for subject $i, i = 1, 2, \ldots 18$.

Estimating a population mean

Example: Painkiller duration

The population mean duration μ will be estimated using the estimator $\overline{X}$.

If sample mean duration of these 18 subjects is $\overline{x} = 5.2$ hours, we say that μ is estimated to be 5.2 hours, and write

$$\widehat{\mu} = 5.2$$

Estimating a population proportion

Let $X_1, \ldots X_n$ be i.i.d. Bernoulli random variables

$$(X_i = 1) = p \qquad P(X_i = 0) = 1 - p$$

with unknown "success" probability p.

The best estimator for p (i.e. unbiased with the minimum variance among all the unbiased estimators) is

$$\hat{p} = \frac{X}{n} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

where X is the total number of successes in the sample

- $\hat{p}$ is the observed proportion of successes in the sample

Estimating a population proportion

Example: Student Stress

McGill is interested in estimating the stress levels of undergraduate student at the end of the semester.

A survey is to be conducted where the stress level will be ascertained as either high or low.

The population of interest in this experiment is all McGill undergraduate students

Estimating a population proportion

Example: Student stress

We are interested in estimating the proportion p of all McGill undergraduate students with high stress levels.

112 students were sampled randomly and independently, and asked to report their stress levels as either high or low.

We have a sample of (i.i.d. Bernoulli) random variables $X_1, \ldots, X_{112}$; X_i is either

- a "success" (high) or
- a "failure" (low)

for student $i, i = 1, 2, \ldots, 112$.

Estimating a population proportion

Example: Student stress

Suppose that 87 of these 112 students reported that their stress level is high.

The proportion p of all McGill undergraduate students with high stress levels is estimated to be the proportion of "successes" in the sample, i.e.

$$\hat{p} = \frac{87}{112} \approx 0.78$$

We say that an estimated 78% of McGill undergraduate students have high stress levels at the end of the semester.

Interval Estimation

Goal: to "guess" the (unknown) population parameter θ with an interval that, when used repeatedly, will capture the unknown θ on a high proportion of occasions.

Interval Estimation

We seek two random variables

$$L = L\left(X_1, \ldots, X_n\right)$$

and

$$R = R\left(X_1, \ldots, X_n\right)$$

such that for a specified probability α

$$P\big(L < \theta < R\big) = 1 - \alpha.$$

Interval Estimation

Both

$$L\left(X_1, \ldots, X_n\right)$$

and

$$R\left(X_1, \ldots, X_n\right)$$

are random variables, so have their own (sampling) probability distributions.

Interval Estimation

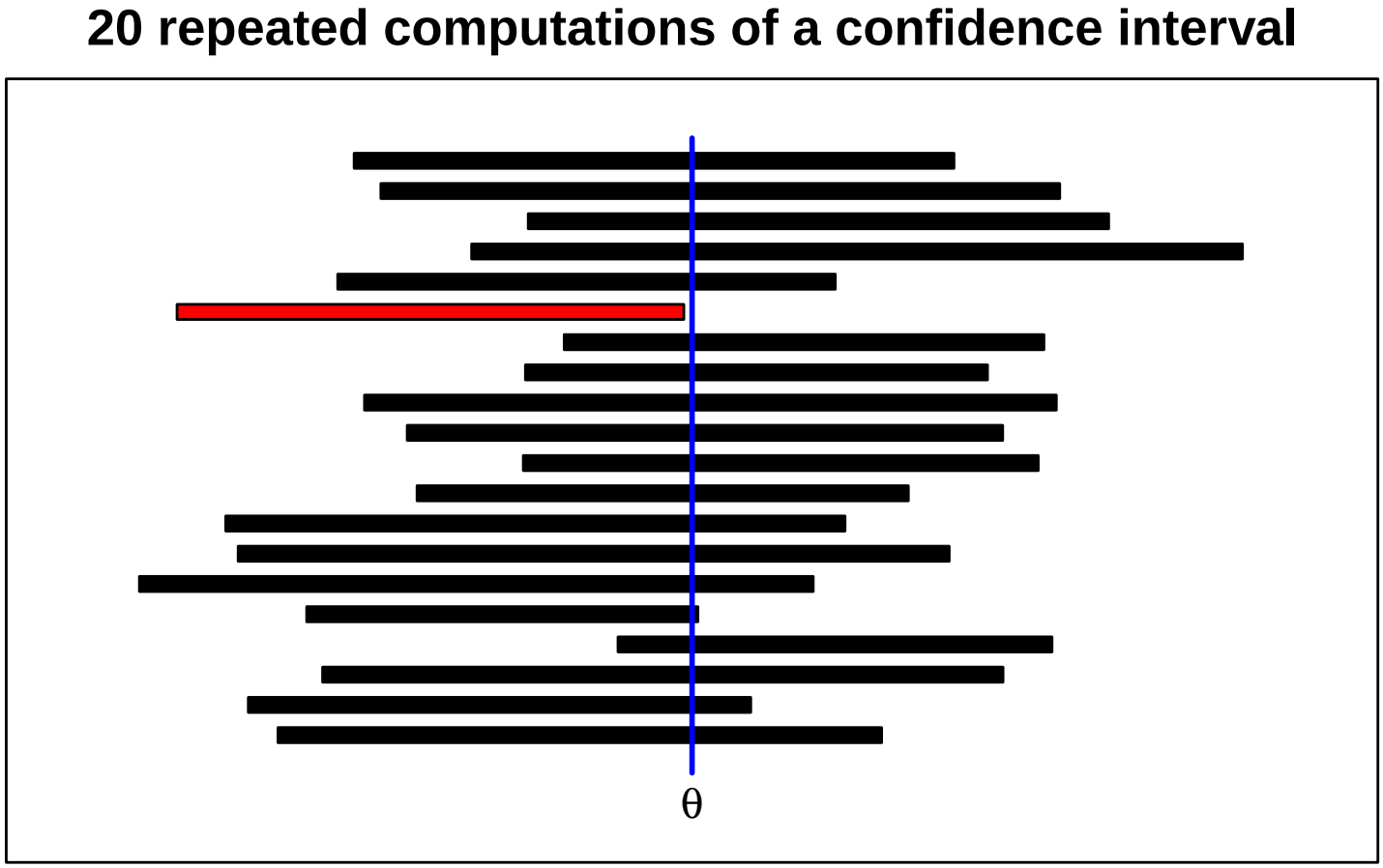
The interval (L, R) is called a $100(1-\alpha)\%$ confidence interval (C. I.) for θ.

- This random interval (L, R) will 'cover' the true value θ with probability $1 - \alpha$

Once we use the observed data $x_1, \ldots, x_n$ to compute the observed interval (l, r), this interval has fixed endpoints and is no longer random

- e.g. $(3.1, 8.4)$ is not a random interval;
- this interval either contains θ or it does not

Interval Estimation



Interval Estimation

We proceed as follows:

1. Objective: find a $100(1-\alpha)\%$ C.I. for the parameter of interest: either μ or p.
    - The value of α will be given; if it is not, take $\alpha = 0.05$
2. Identify the setting from the wording of the problem: the formula for (L, R) depends on the setting
    - Once we know the setting, we look up formulas for L and R.
3. Substitute the known information into the formulas to obtain the observed interval (l, r)

Interval Estimation

We interpret the interval (L, R) as follows:

Textbook: "We are $100(1-\alpha)\%$ confident that θ falls in (L, R)"

- This is imprecise ...
- We add: "To be more precise, if we were to carry out the same study infinitely often under the same conditions, and each time compute a $100(1-\alpha)\%$ C.I. using the same procedure, then $100(1-\alpha)\%$ of these intervals will contain θ."

Interval Estimation

Ideally we would like to be able to say something like

$$\text{“}\theta \text{ lies in } (l, r) \text{ with probability } (1 - \alpha)\text{”}$$

e.g.

$$\text{“}\theta \text{ lies in } (19.7, 22.4) \text{ with probability } 0.95\text{”}.$$

However, this would require us to regard θ itself as random.

The t-distribution

We will see that construction of confidence intervals depends on understanding the sampling distribution of different statistical summaries.

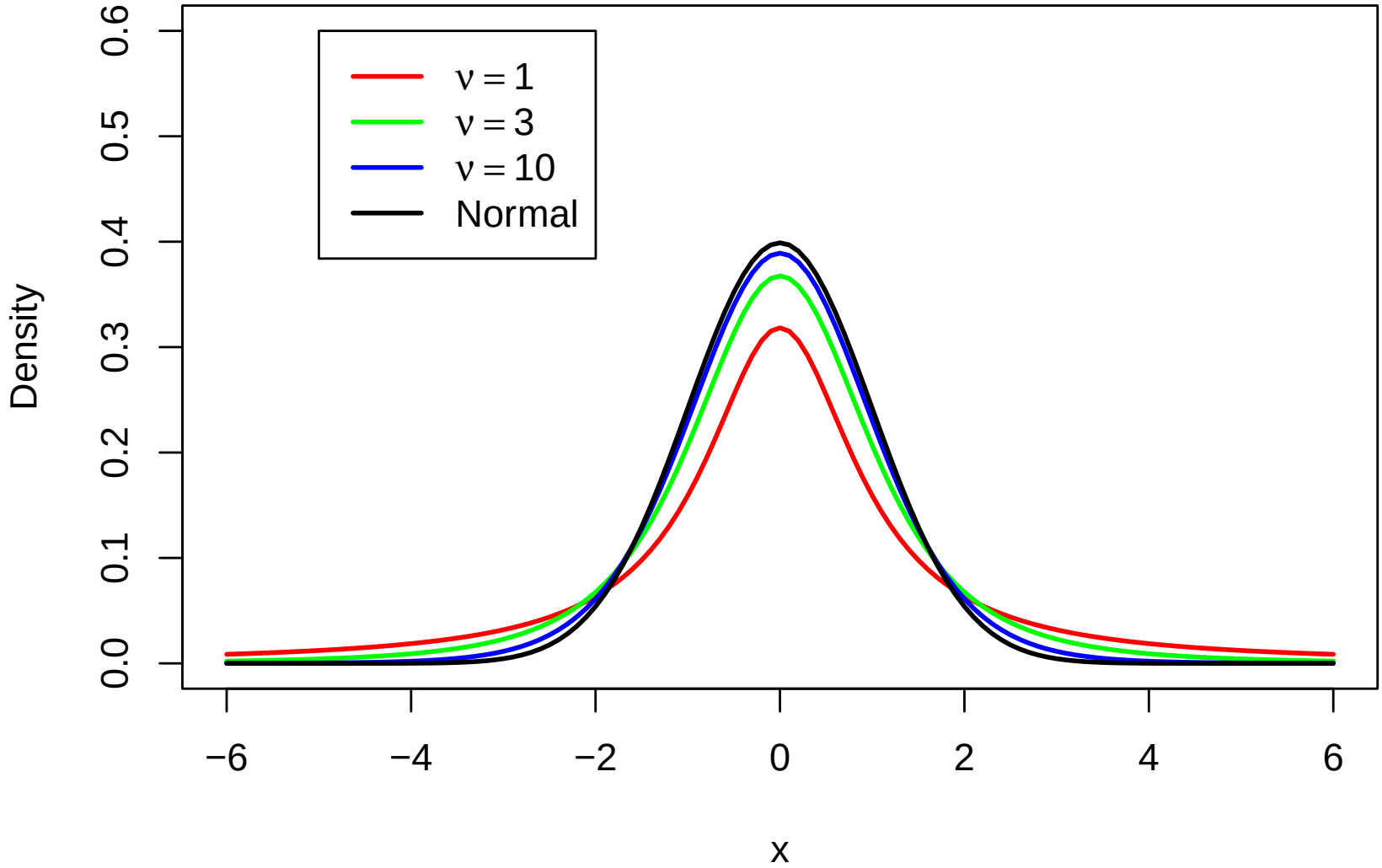
- the Normal distribution plays an important role
- the Student-t-distribution (or just t-distribution) is also needed
  - The t-distribution is specified through a parameter ν called its degrees of freedom (d.f.)
  - Each ν gives a different curve, but this curve is always symmetric about 0.

The t-distribution

The t-distribution has heavier tails than the Normal

- the t-distribution allocates more probability to tail regions
- the standard Normal and the t-distribution get more similar as ν gets larger
- as $\nu \longrightarrow \infty$, the t-distribution becomes identical to the standard Normal.

The t-distribution

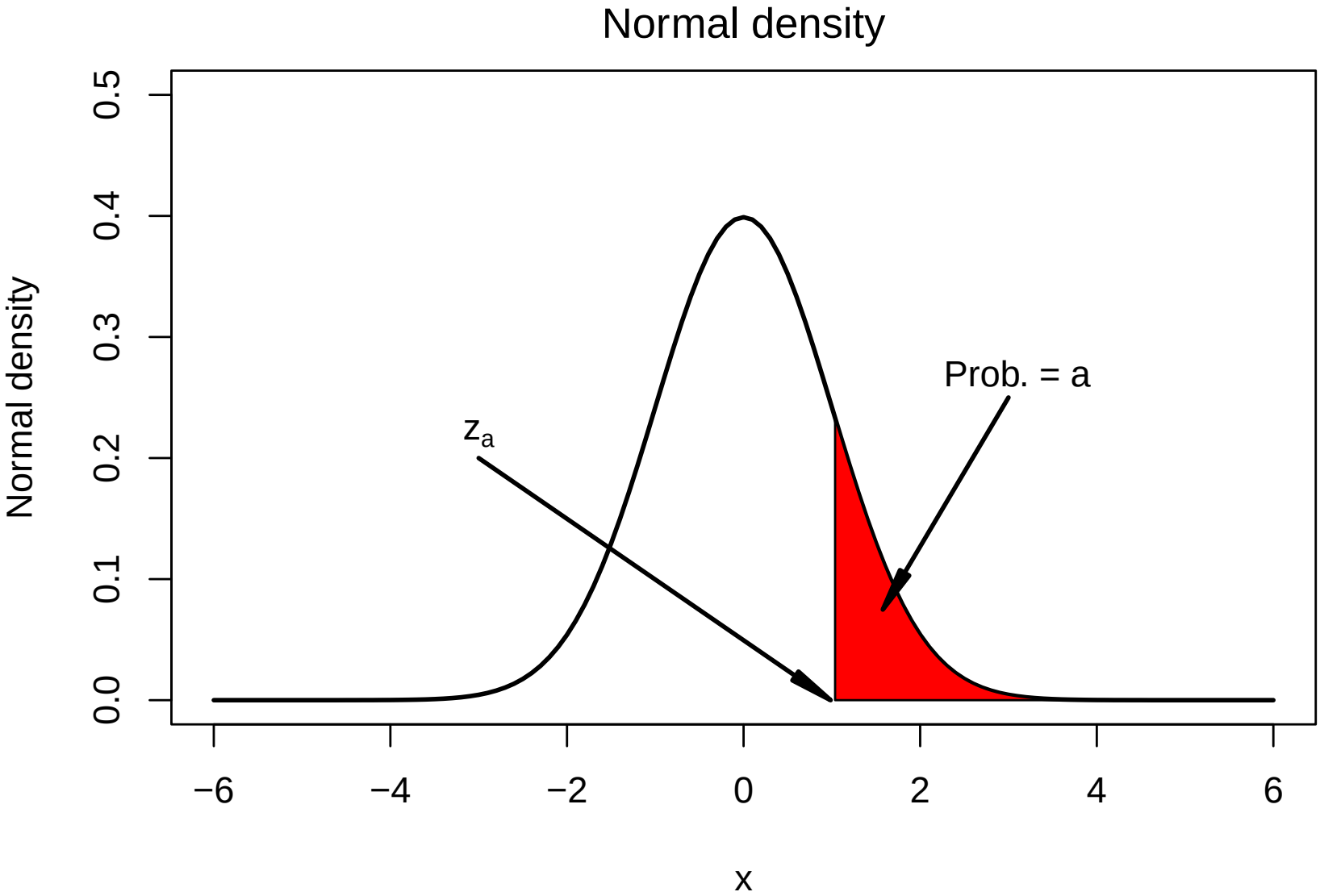


The t-distribution

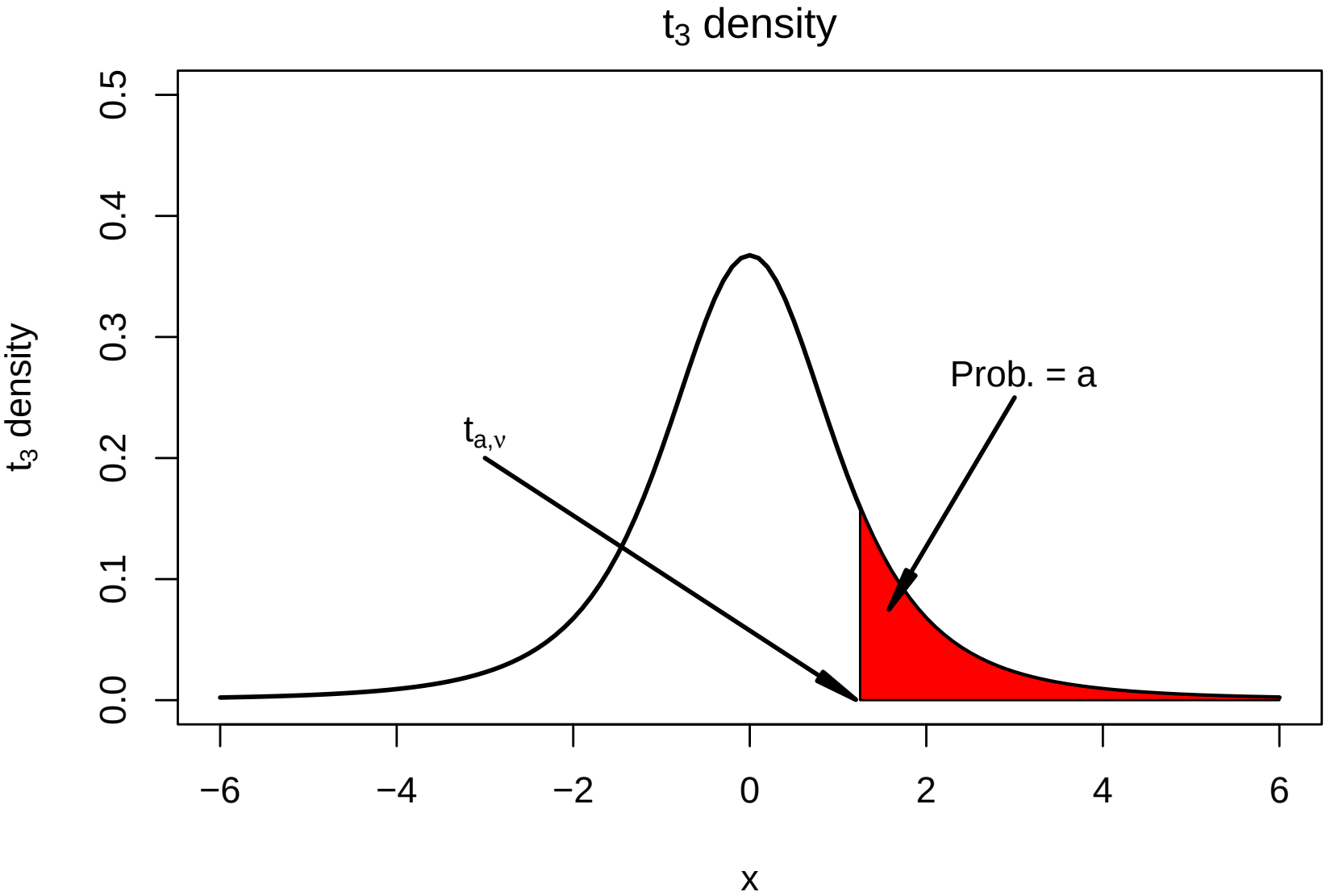
In order to calculate confidence intervals we first define the notation z_a and $t_{a,\nu}$, the "upper a point" of these distributions

- The upper a point of a standard Normal distribution is the point on the horizontal axis, z_a such that $P\left(Z > z_a\right) = a$.
- The upper a point of a t-distribution with ν degrees of freedom is the point on the horizontal axis, $t_{a,\nu}$ such that $P\left(T > t_{a,\nu}\right) = a$, where the random variable T has a t-distribution with ν degrees of freedom.

The t-distribution



The t-distribution



The t-distribution

In these plots: $a = 0.15$ (shaded red)

- Normal case: $z_a = 1.0364$
- t_3 case: $t_{a,3} = 1.2498$

so note that

$$t_{a,3} > z_a.$$

The t-distribution

	a				
ν	0.1	0.05	0.025	0.01	0.001
1	3.0777	6.3138	12.7062	31.8205	318.3088
3	1.6377	2.3534	3.1824	4.5407	10.2145
10	1.3722	1.8125	2.2281	2.7638	4.1437
30	1.3104	1.6973	2.0423	2.4573	3.3852
50	1.2987	1.6759	2.0086	2.4033	3.2614
100	1.2901	1.6602	1.9840	2.3642	3.1737
∞	1.2816	1.6449	1.9600	2.3263	3.0902

Note: $\nu = \infty$ corresponds to $Normal(0, 1)$.

Experiment Scenarios

The relevant experiment scenario can be determined by establishing

- is the parameter of interest, a population mean μ, or a population proportion p ?
- is the sample size (n) large or small?

Experiment Scenarios

To determine the correct strategy, we need to identify

(A) the assumed probability model/parameter of interest
- Normal, μ
- Binomial, p

(B) the sample size assumption
- n small ($n < 30$)
- n large ($n \geqslant 30$)

(C) the variance assumption for the Normal case
- σ known
- σ unknown

Parameter μ, Small n

Theorem: Scenario 1

Suppose that $X_1, \ldots, X_n$ are a random sample from a Normal distribution with unknown μ and σ, and $n < 30$.

A $100(1-\alpha)\%$ confidence interval for μ is (L, R), where:

$$L = \overline{X} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}$$

$$R = \overline{X} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}$$

Parameter μ, Small n, σ known

Theorem: Scenario 2

Suppose that $X_1, \ldots, X_n$ are a random sample from a Normal distribution with unknown μ, with σ known.

A $100(1-\alpha)\%$ confidence interval for μ is (L, R), where:

$$L = \overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$R = \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Parameter μ, Large n

Theorem: Scenario 3

Suppose that $X_1, \ldots, X_n$ are a random sample from a distribution with unknown μ and σ, where $n \geqslant 30$.

A $100(1-\alpha)\%$ confidence interval for μ is (L, R), where:

$$L = \overline{X} - z_{\alpha/2}\frac{S}{\sqrt{n}}$$

$$R = \overline{X} + z_{\alpha/2}\frac{S}{\sqrt{n}}$$

Note: Normality of the Xs is not required.

Parameter p

Theorem: Scenario 4

Let $X_1, \ldots, X_n$ be a random sample of Bernoulli random variables with unknown probability of success p.

A $100(1-\alpha)\%$ confidence interval for p is (L, R), where:

$$L = \hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$R = \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where

$$\hat{p} = \frac{\text{Number of “successes” in the sample}}{n}$$

Parameter p

Theorem: Scenario 4

Note: "Large" sample size in this scenario means that

$$n\hat{p} \geqslant 15 \text{ and } n(1-\hat{p}) \geqslant 15$$

Confidence Interval: Examples

Problem: Painkiller duration

A drug manufacturer is testing the duration of pain-relief of their new over-the-counter painkiller.

Duration is measured on 18 subjects, with sample mean of $\overline{x} = 5.2$ hours, and sample standard deviation of $s = 1.1$ hours.

Calculate a 95% confidence interval for the true mean duration μ; state any assumptions you make.

Confidence Interval: Examples

Solution: Painkiller duration

We are in Scenario 1: $n < 30$ and we are trying to find a 95%CI for μ, with $\alpha = 0.05$.

In order to calculate the CI , we have to assume that the X_i s are Normally distributed: then the 95%CI for the unknown μ.

We have

- $n = 18 < 30$.
- $\overline{x} = 5.2$
- $s = 1.1$
- $a = 0.05$
- $t_{a/2,n-1} = t_{0.025,17} = 2.110$

Confidence Interval: Examples

Solution: Painkiller duration

Hence the CI is

$$
\begin{aligned}
(l, r) &: \left(\overline{x} - t_{0.025,17} \frac{s}{\sqrt{n}}, \overline{x} + t_{0.025,17} \frac{s}{\sqrt{n}} \right) \\
&= \left(5.2 - 2.110 \times \frac{1.1}{\sqrt{18}}, 5.2 + 2.110 \times \frac{1.1}{\sqrt{18}} \right) \\
&= (4.654, 5.746)
\end{aligned}
$$

measured in hours.

Confidence Interval: Examples

Solution: Painkiller duration

Interpretation: We are 95% confident that μ, the mean population duration of relief, is between 4.65 and 5.75 hours

To be more precise, if this study were repeated infinitely often under identical conditions, and on each occasion a 95% confidence interval were calculated using the technique described here, then 95% of these intervals would include the true, but unknown, mean duration μ.

Confidence Interval: Examples

Problem: Painkiller duration

Now suppose that the drug is tested on 65 subjects, with a sample average of $\overline{x} = 5.8$ hours and a sample standard deviation of $s = 0.8$ hours

Calculate a 99% confidence interval for the true mean duration μ.

Confidence Interval: Examples

Solution: Painkiller duration

We are now in Scenario 3: $n > 30$ and we are finding a 99%Cl for μ, with $\alpha = 0.01$.

We have

- $n = 65 > 30$.
- $\overline{x} = 5.8$
- $s = 0.8$
- $a = 0.01$
- $z_{a/2} = z_{0.005} = 2.576$

Confidence Interval: Examples

Solution: Painkiller duration

$$
\begin{aligned}
&\left(\overline{x} - z_{0.005}\frac{s}{\sqrt{n}}, \overline{x} + z_{0.005}\frac{s}{\sqrt{n}}\right) \\
&\quad = \left(5.8 - 2.576 \times \frac{0.8}{\sqrt{65}}, 5.8 + 2.576 \times \frac{0.8}{\sqrt{65}}\right) \\
&\quad = (5.545, 6.055)
\end{aligned}
$$

Confidence Interval: Examples

Problem: Student stress

McGill is conducting a survey on undergrad student stress levels at the end of the semester.

- 112 students are selected, and asked to report their stress levels;
- 87 of these students report that their stress level is high.

Calculate a 90% confidence interval for the proportion of McGill students with a high stress level.

Confidence Interval: Examples

Solution: Student stress

We are in Scenario 4: a Binomial setup, and we are finding a 90%Cl for p, with $\alpha = 0.1$.

From the data,

$$\hat{p} = \frac{87}{112} = 0.777$$

and

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.777(1-0.777)}{112}} = 0.039$$

rounding to three decimal places. We also have

$$z_{0.05} = 1.645$$

Confidence Interval: Examples

Solution: Student stress

Therefore

$$\left(\hat{p} - z_{0.05}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{0.05}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

$$= (0.777 - 1.645 \times 0.039, 0.777 + 1.645 \times 0.039)$$

$$= (0.713, 0.841)$$

Confidence Interval: Examples

Solution: Student stress

Checking the large sample assumption:

$$n\hat{p} = 87 \geqslant 15$$

and

$$n(1-\hat{p}) = 27 \geqslant 15$$

so the large sample assumption holds.

The assumption of Normality

When estimating a population mean μ if the sample size is "small" $(n < 30)$, it can only be done under the assumption that the population has a Normal distribution

- Note that we cannot really "prove" the assumption, as it is an assumption on the population.
- However, in practice, it is not enough to state the assumption of Normality of the population, we need to show some evidence to support this assumption

The assumption of Normality

- One way to do this is to use the descriptive methods to assess Normality discussed in previous lectures:
  - histograms
  - boxplots
- If the assumption holds for a random sample, then it may hold for the population
- We just need to show, from the sample data, that we have at least an approximation to Normality

Assessing Normality

Many wide-spread techniques for statistical inference require data that are (approximately) Normally distributed.

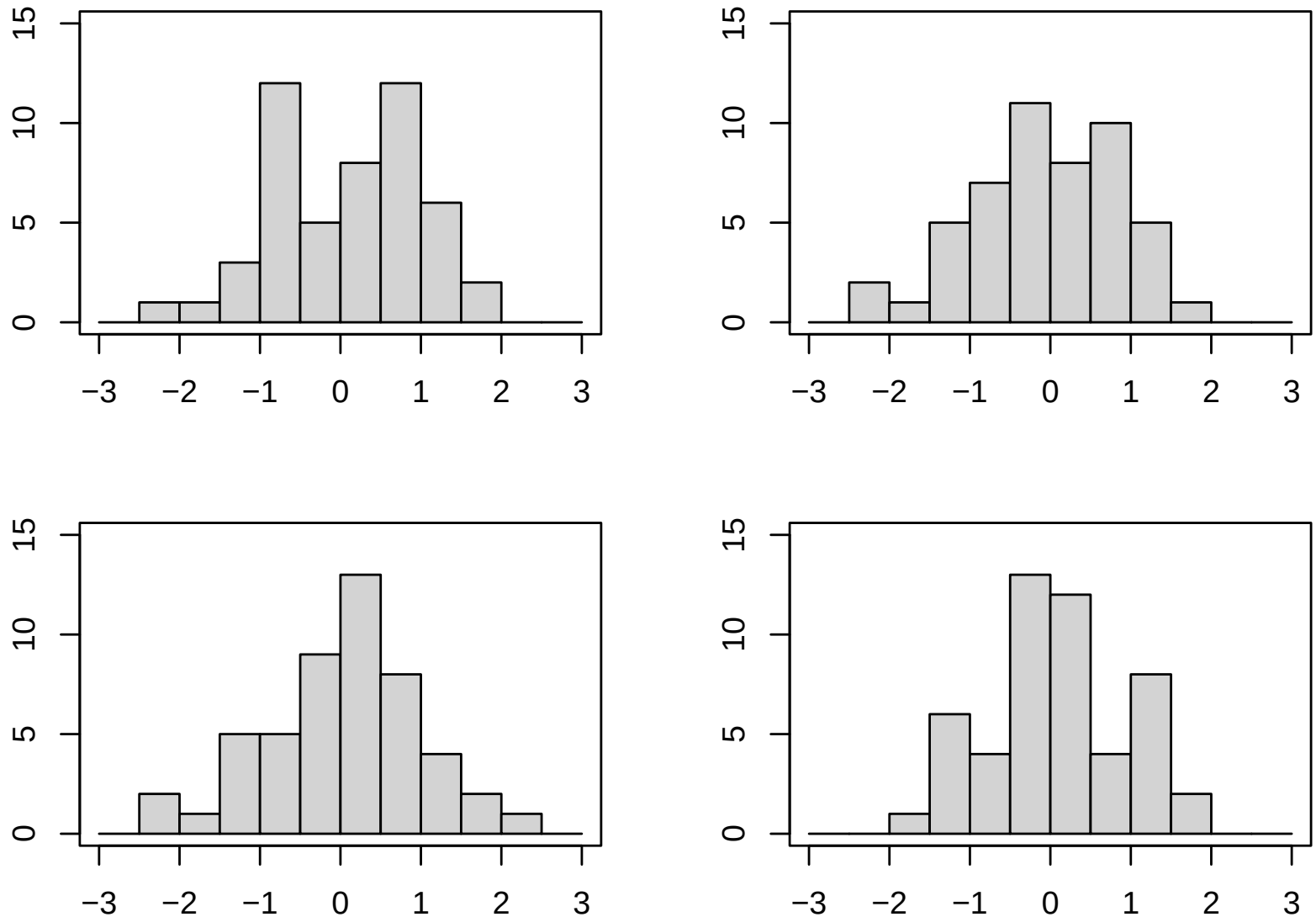
- Normally distributed random variables are found in many settings.
- Important question: How can we tell if our data are (approximately) Normally distributed?

Assessing Normality

If the data are approximately Normal, the shape of the histogram will be mound shape and symmetric, similar to the Normal probability density function

However, this is not always easy to assess from a histogram:

Assessing Normality



Q-Q Plots

A more reliable method for checking normality: normal probability plot (also called a quantile-quantile (Q-Q) plot.)

A Q-Q plot compares the quantiles from the sample with the theoretical quantiles from the normal distribution.

Q-Q Plots

In the plot:

- the x coordinates are the values

$$F^{-1}\left(j/(n+1)\right) \qquad j = 1, 2, \ldots, n$$

that is, the theoretical quantile corresponding to probability $j/(n+1)$, where F^{-1} is the inverse of the appropriate Normal distribution function.

Recall that if $X \sim Normal(\mu, \sigma^2)$, then the distribution function for X can be computed

$$F(\mathrm{x}) = \Phi\left(\frac{\mathrm{x} - \mu}{\sigma}\right).$$

For the Q-Q plot, we estimate μ and σ from the data.

Q-Q Plots

- the y coordinates are the ordered data values, sorted into ascending order

  $$y_j = x_{(j)}$$

  that is, the sample quantiles corresponding to probabilities j/n.

Q-Q Plots

We often transform the data onto a standardized (z-score) scale

$$x \longrightarrow z = \frac{x - \overline{x}}{s}$$

before plotting the Q-Q plot.

We expect to see the points in the plot with

x-coordinate: theoretical Normal quantile

y-coordinate: sample quantile for z

lie on the line $y = x$; if they do not, then the Normal assumption is questionable.

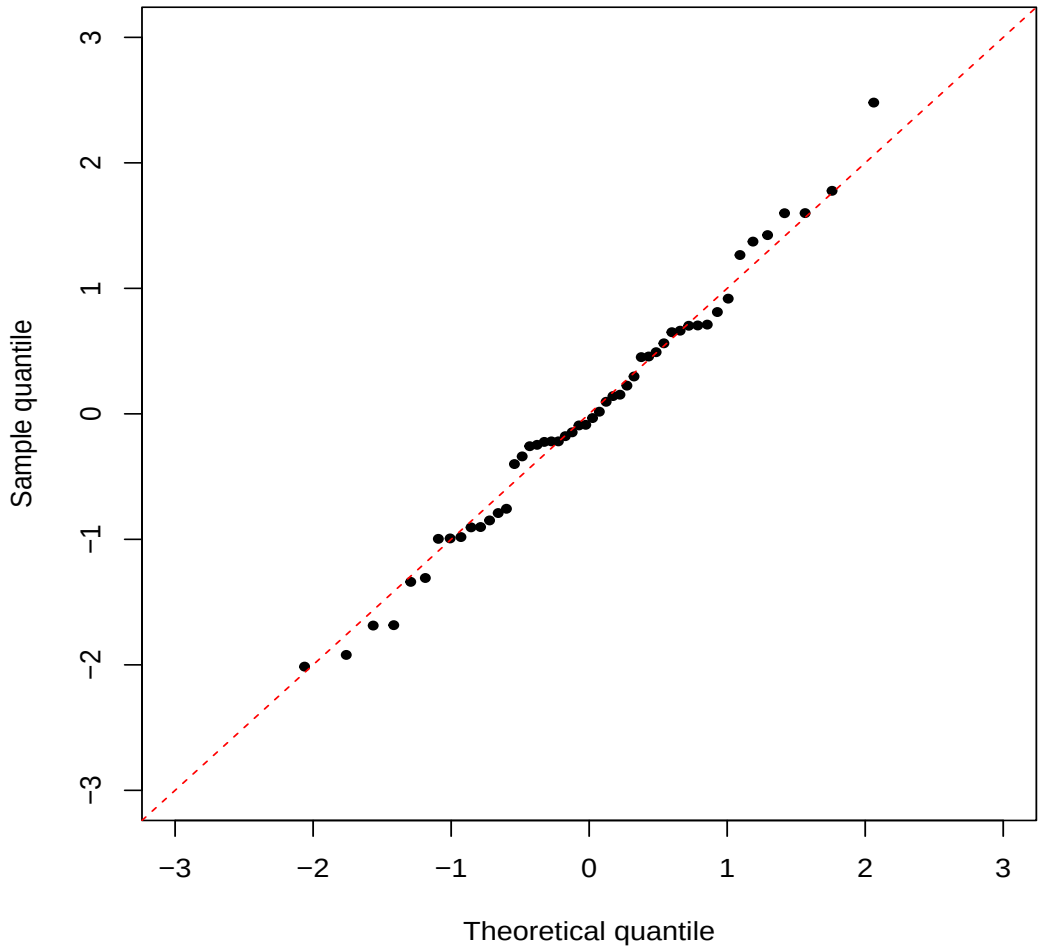
Q-Q Plots

```
n<-50
x<-rnorm(n,10,2)
xbar<-mean(x)
sval<-sd(x)
c(xbar,sval)

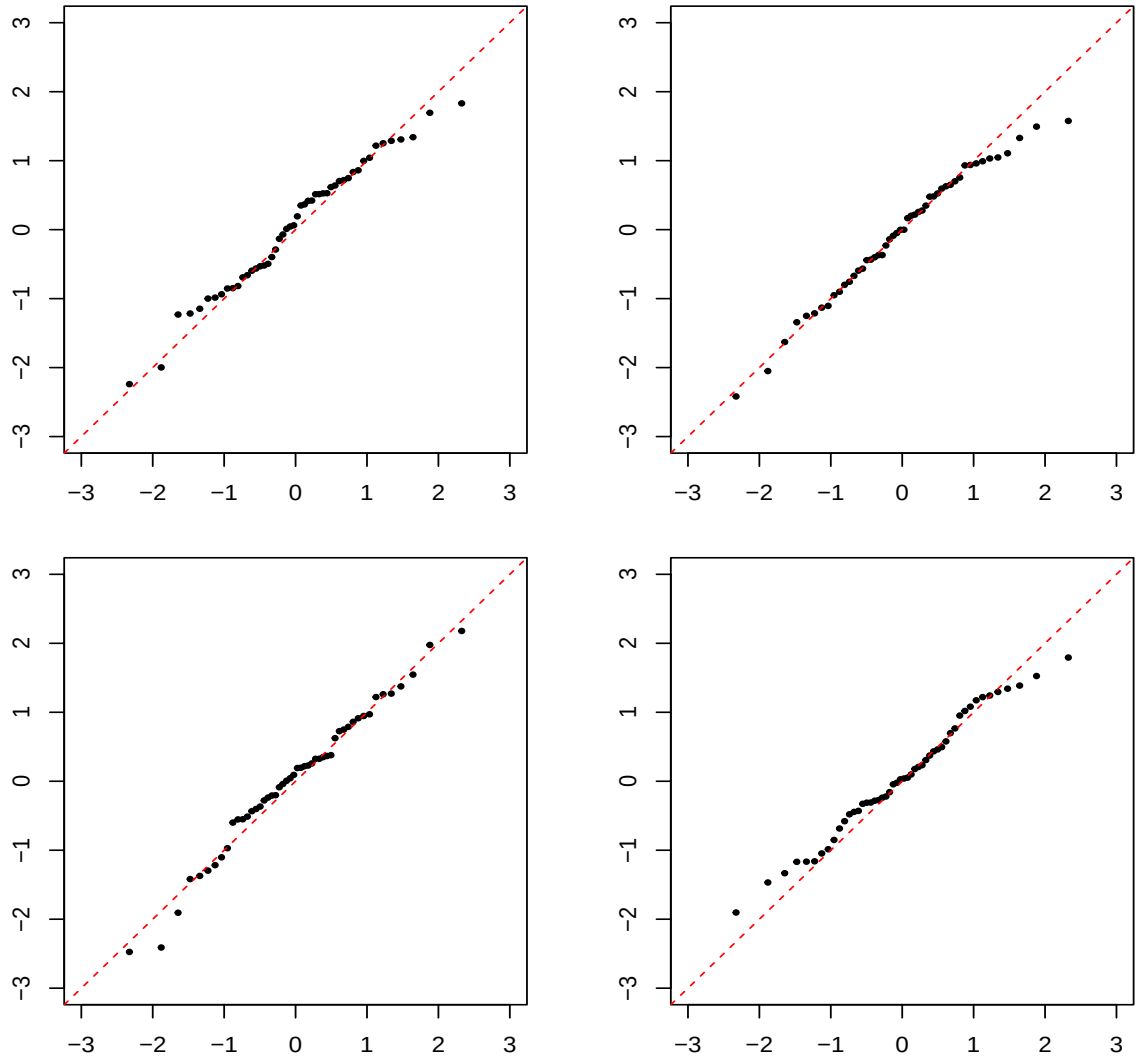
+ [1] 9.652594 2.044429

z<-(x-xbar)/sval
qq.x<-qnorm(c(1:n)/(n+1))
qq.y<-sort(z)
par(pty='s')
plot(qq.x,qq.y,pch=19,cex=0.7,xlim=range(-3,3),ylim=range(-3,3),
    xlab='Theoretical quantile', ylab='Sample quantile')
abline(0,1,lty=2,col='red')
```

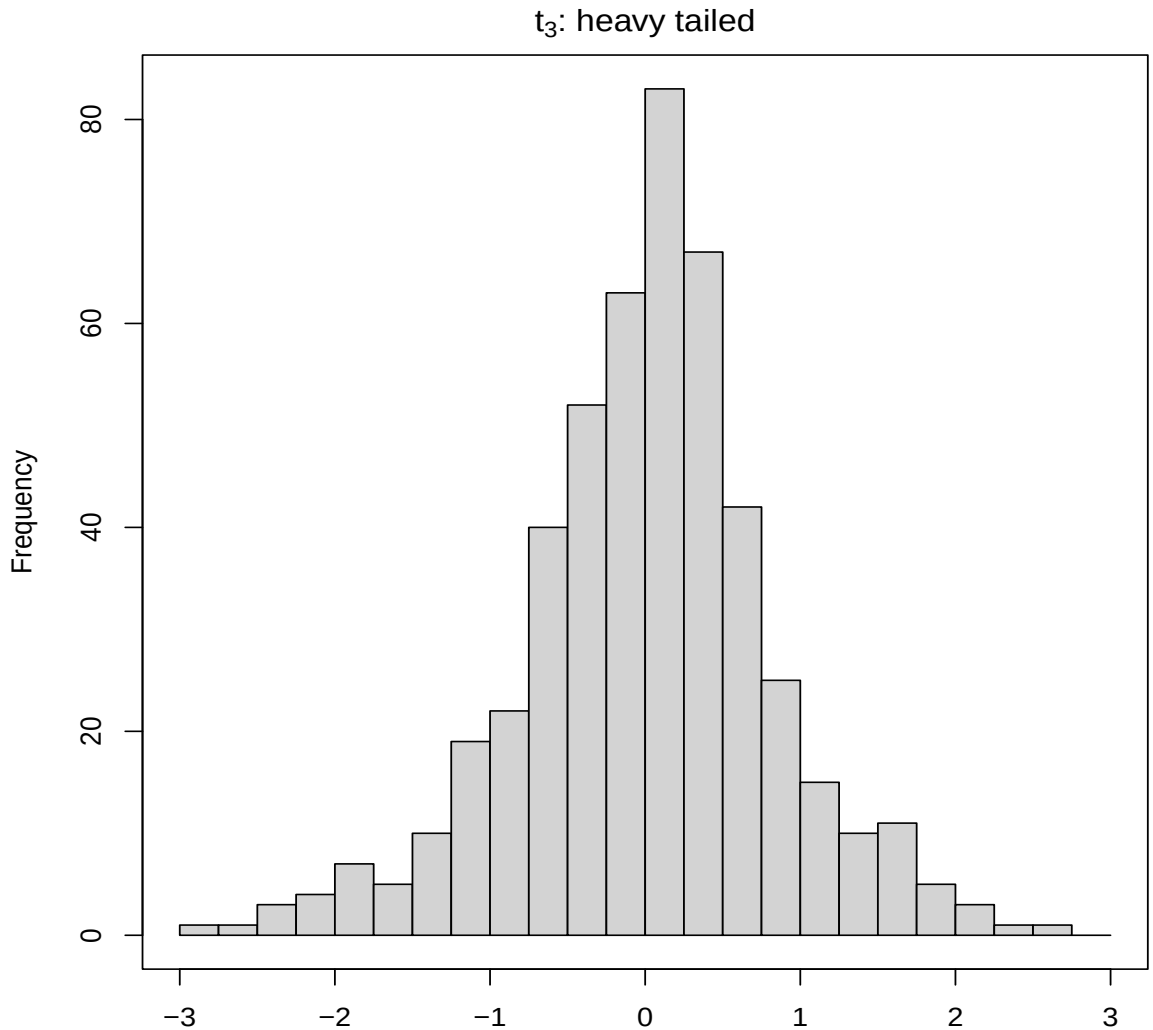
Q-Q Plots



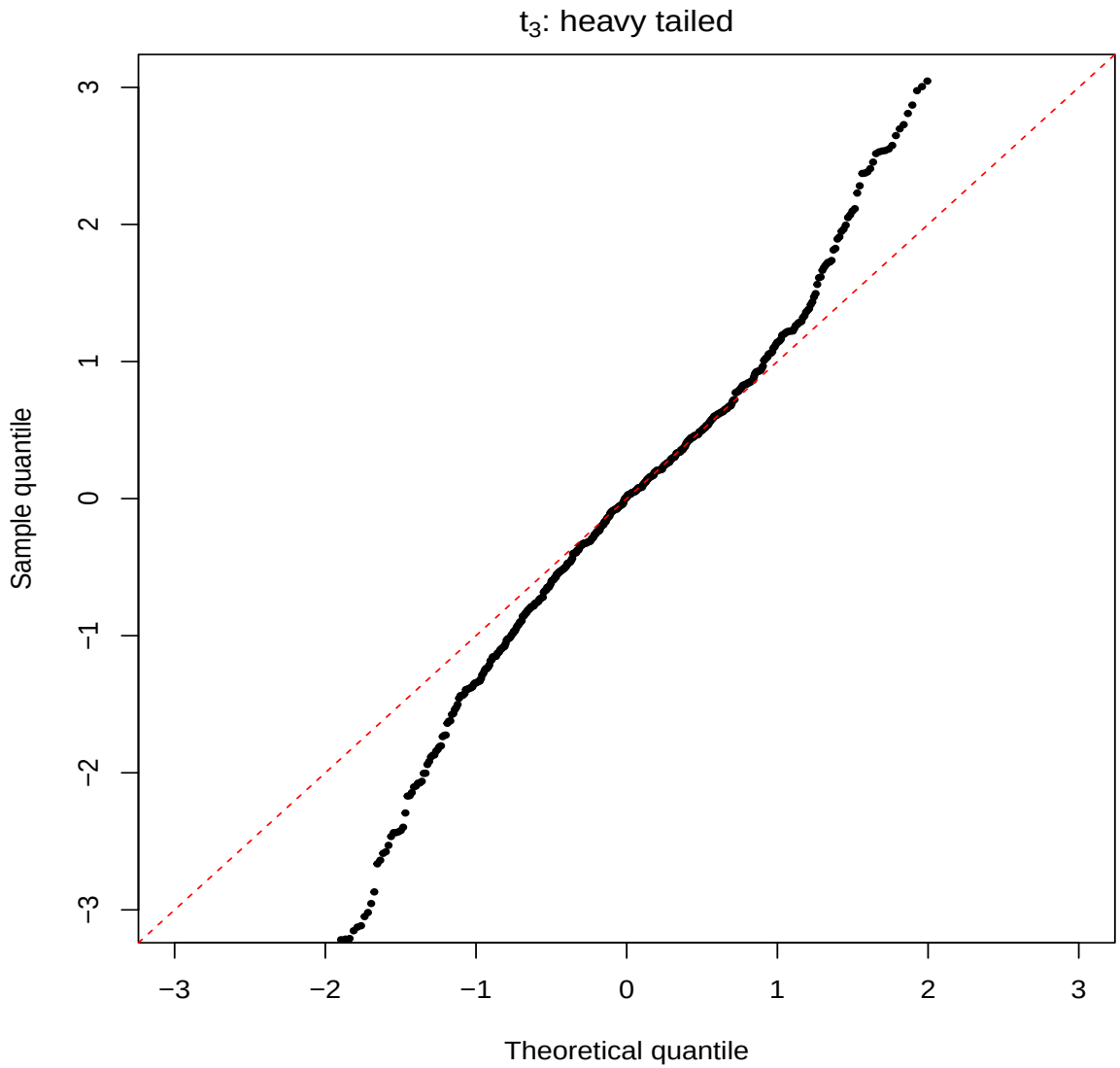
Q-Q Plots



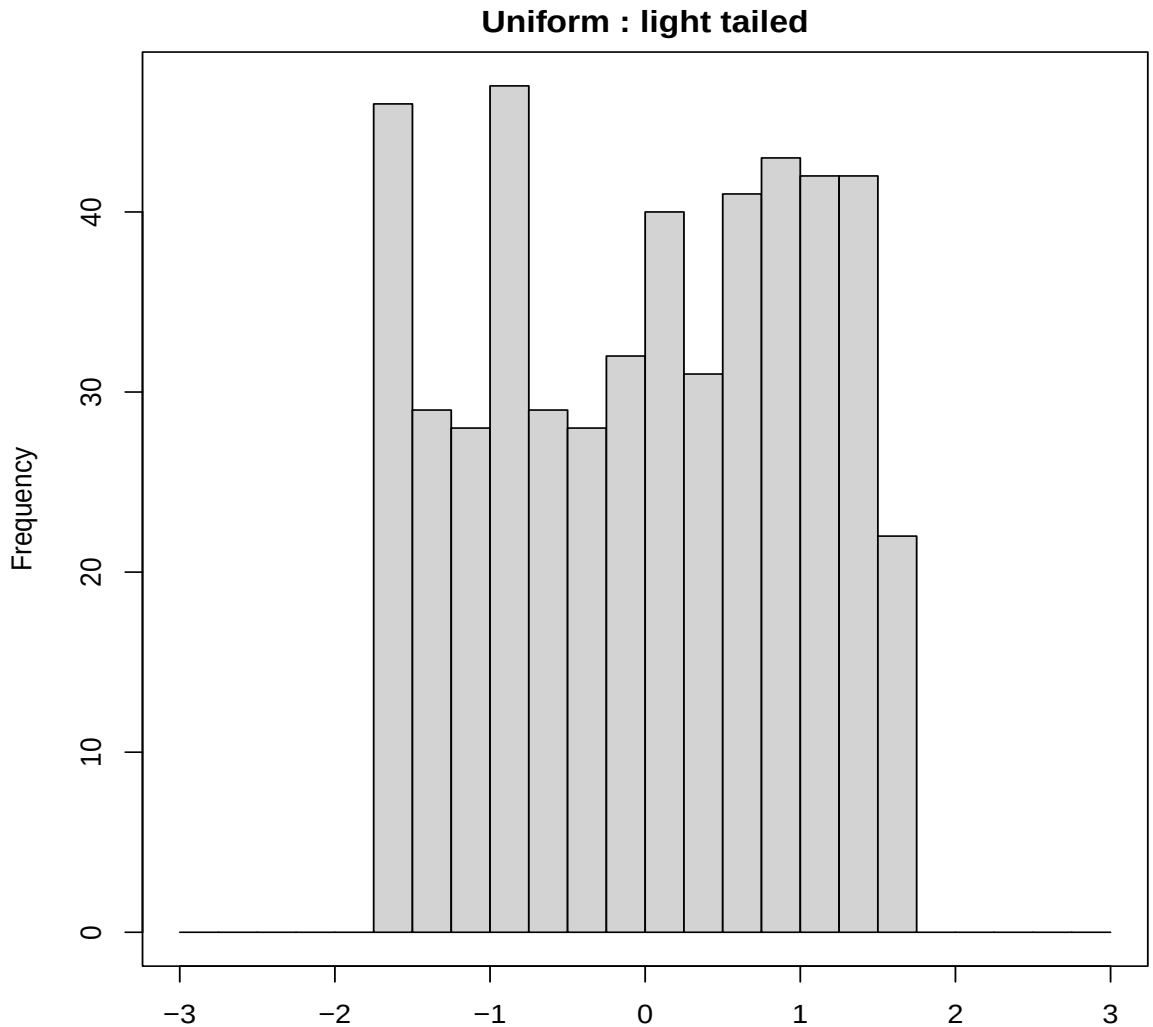
Q-Q Plots



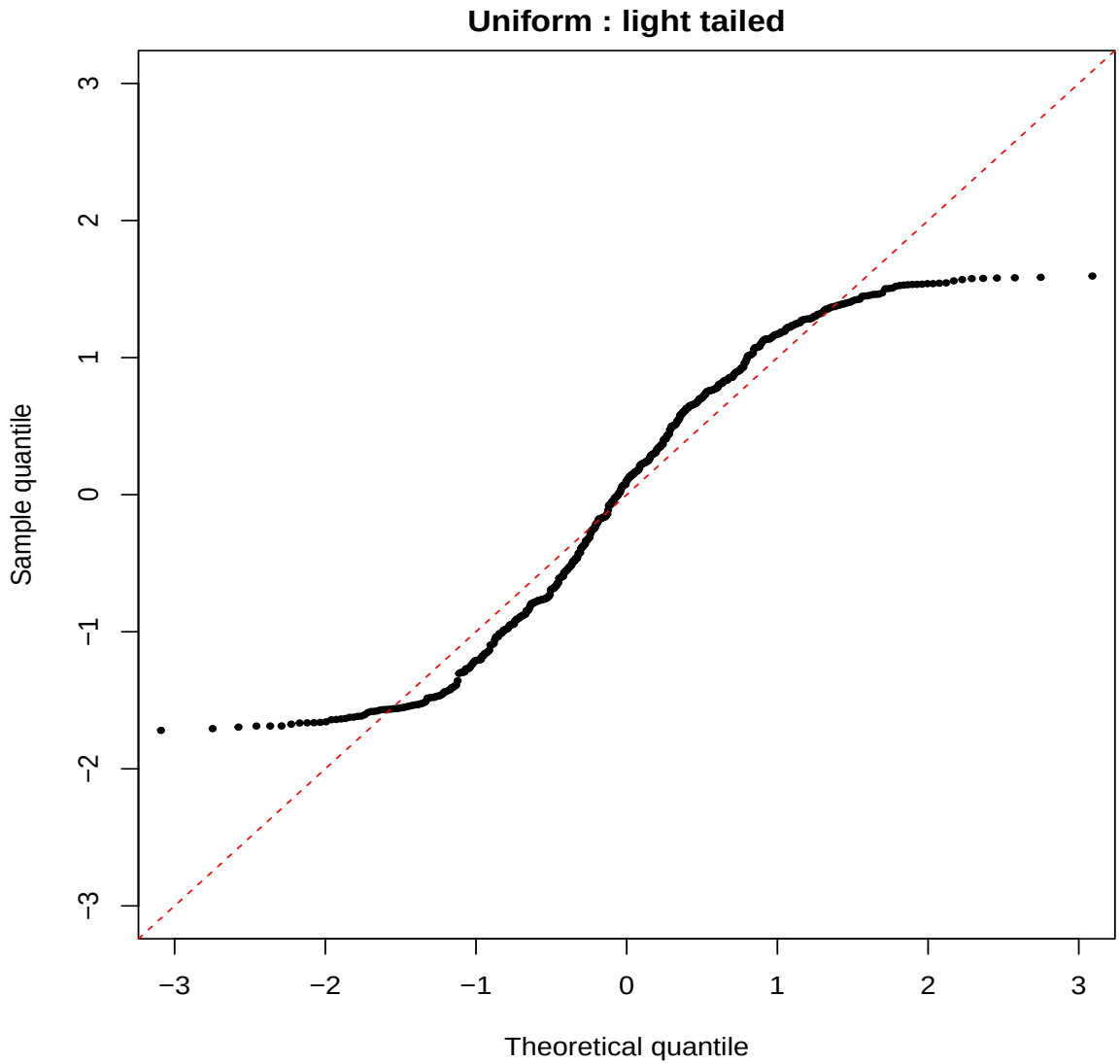
Q-Q Plots


t_3: heavy tailed
Sample quantile
Theoretical quantile
−3
−2
−1
0
1
2
3

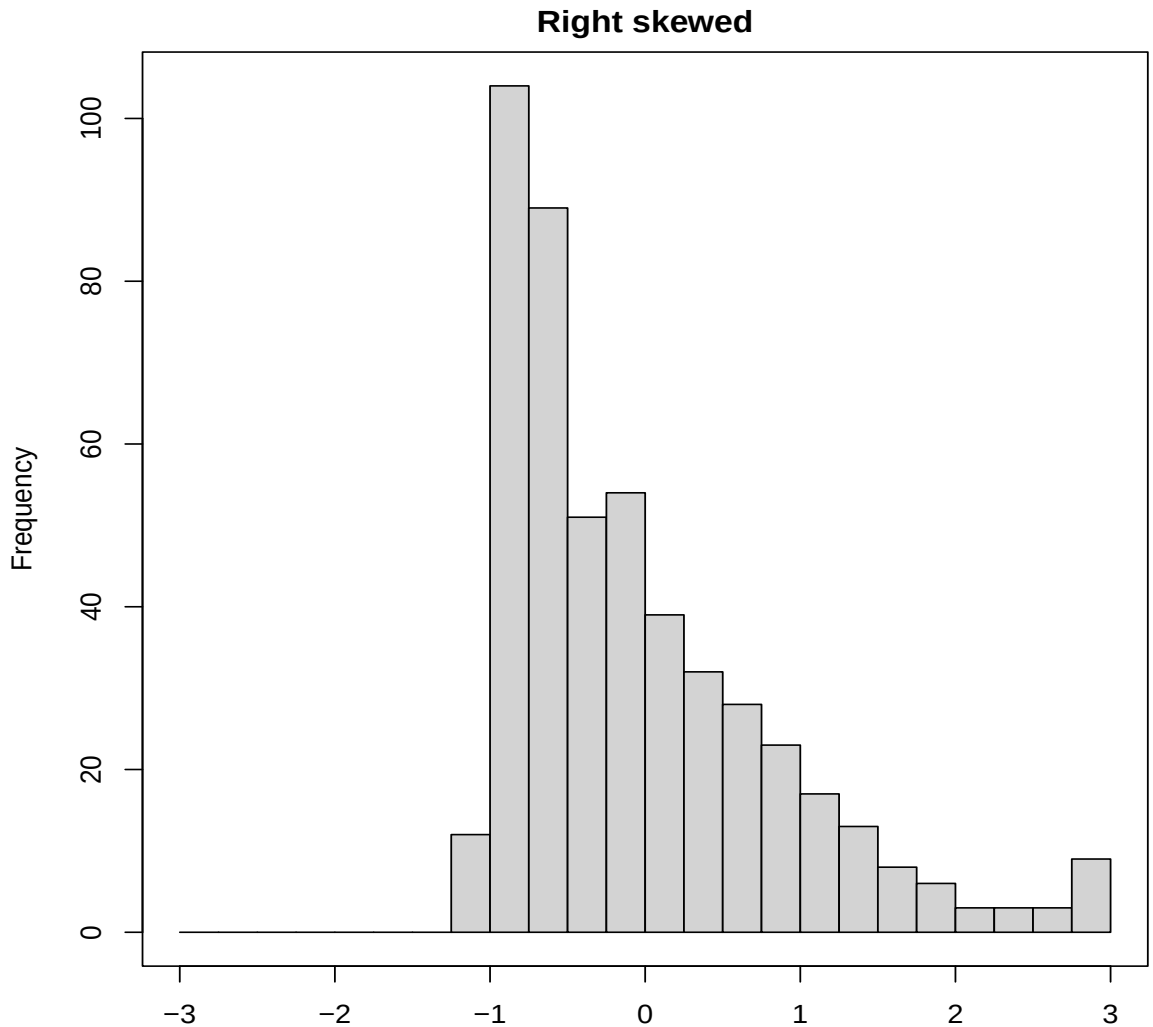
Q-Q Plots



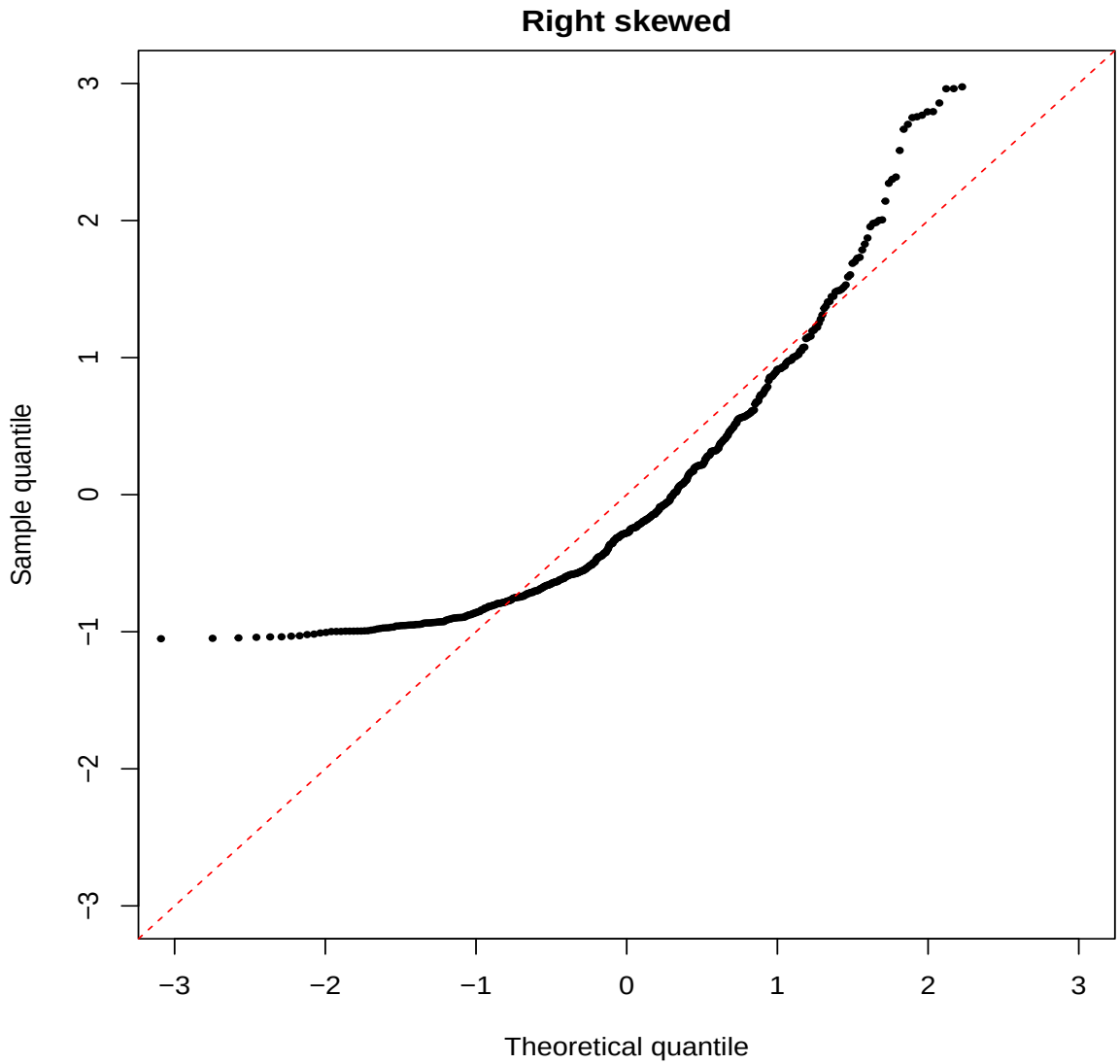
Q-Q Plots


Uniform : light tailed
Sample quantile
Theoretical quantile
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3

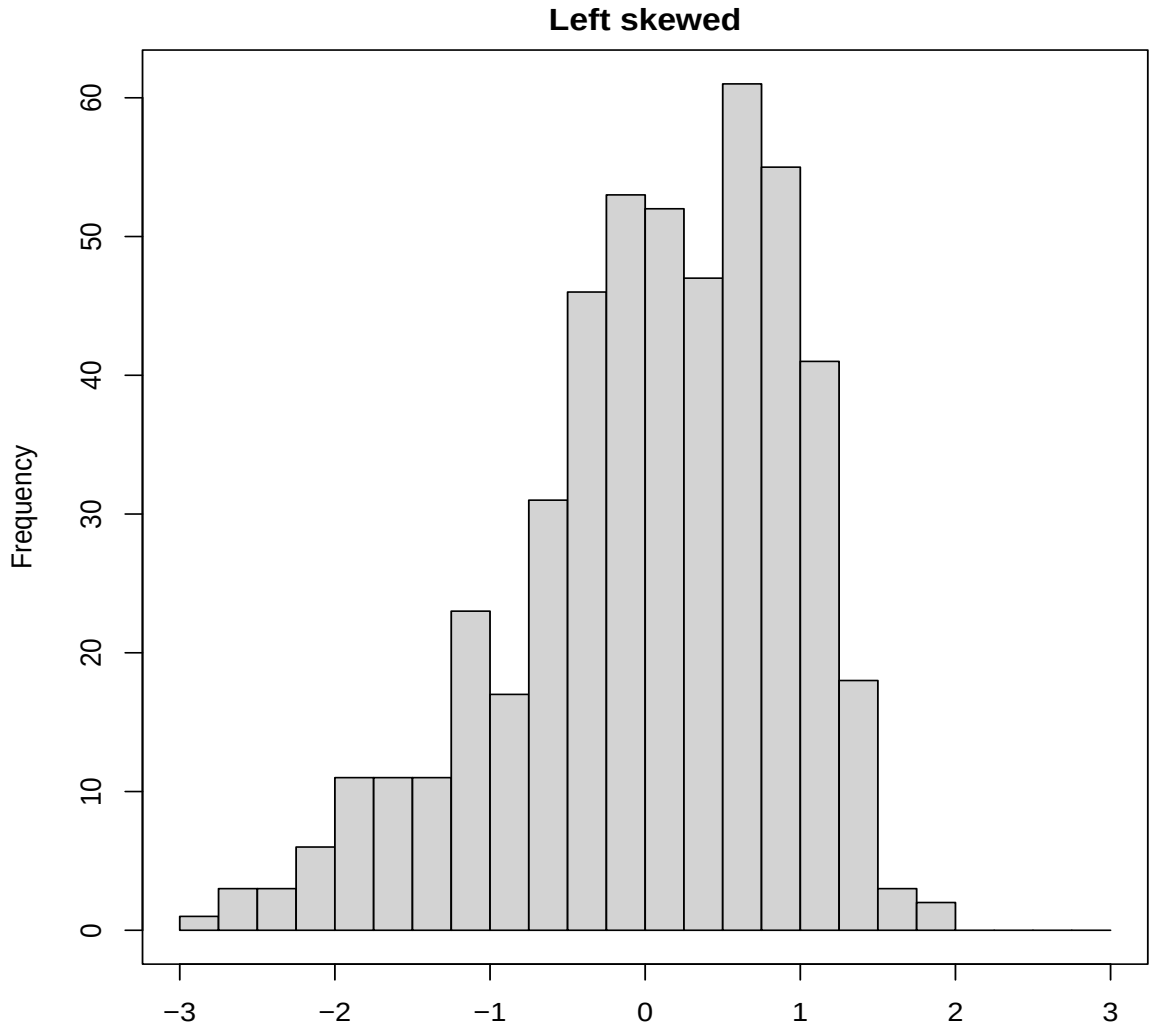
Q-Q Plots



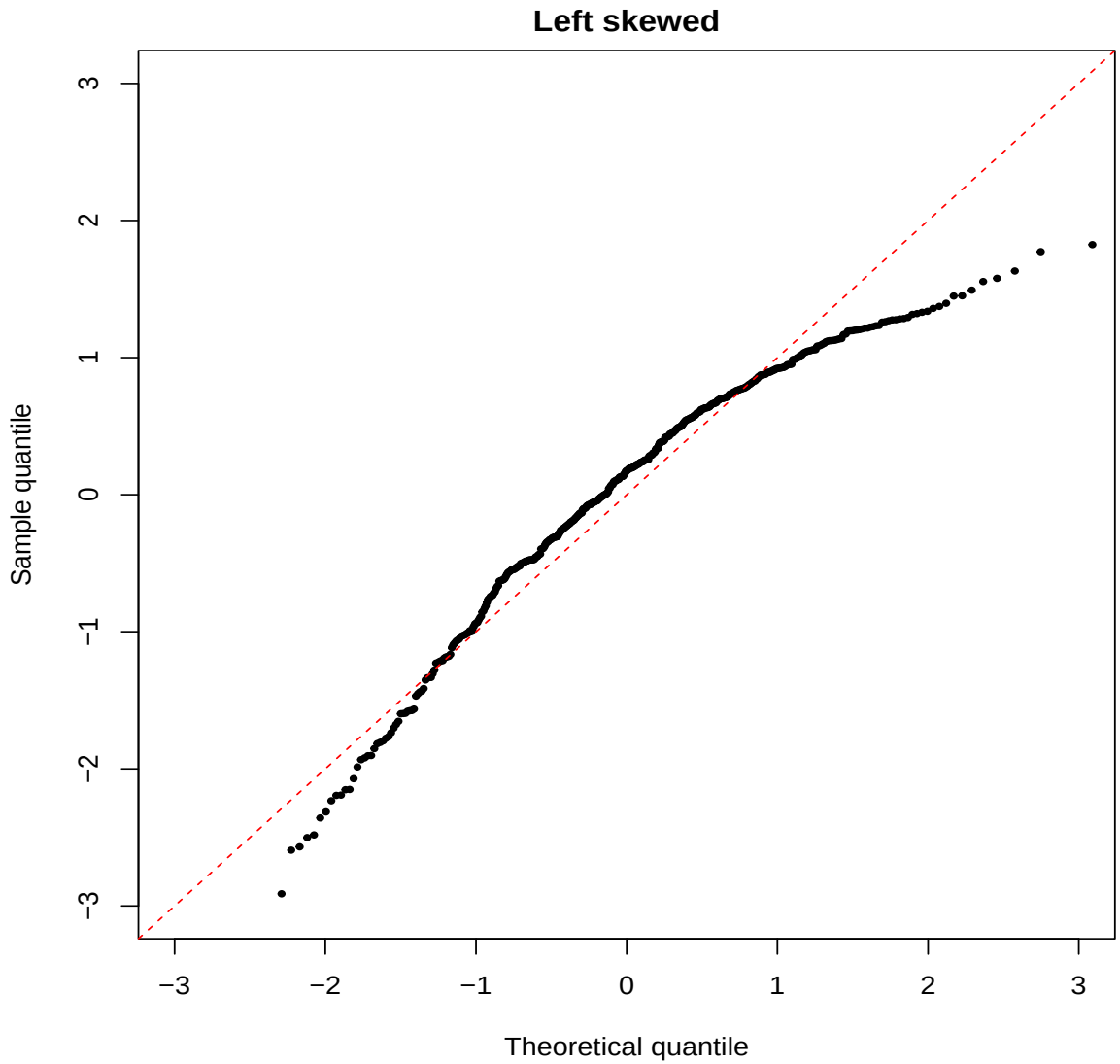
Q-Q Plots


Right skewed
Sample quantile
Theoretical quantile
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3

Q-Q Plots



Q-Q Plots


Left skewed
Sample quantile
Theoretical quantile
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3

Hypothesis Testing: big idea

Previously:

- Estimated θ through a single point $\widehat{\theta}$
- Estimated θ through an interval (L, R)
- Now: Want to make a decision relating to θ

In order to make such decisions, based on data, in a rigorous manner, we have to use hypothesis testing.

Hypothesis Testing: big idea

Idea: We will have two competing hypotheses about a population parameter θ (for example, μ and p)

- We will use our sample data to decide which of the two hypotheses is true

NB: The goal is not to know which hypothesis is true (We will never know unless we could observe the full population data)

Hypothesis Testing: big idea

The two hypotheses are called

- the null hypothesis, H_0,
- the alternative hypothesis, H_a.

As for θ in confidence intervals, the truth of H_0 versus H_a is a fixed state.

We are trying to assess which is true, and our "guess" will be either correct or not.

Hypothesis Testing: null hypothesis

H_0 generally represents the perceived "status quo":

- 70% of McGill students have high stress levels
- Two lakes have the same pollution levels
- A new treatment has no effect on recovery time
- Students in two schools have, on average, the same test scores

Hypothesis Testing: alternative hypothesis

H_a generally represents a deviation from the "status quo":

- More than 70% of McGill students have high stress levels
- One lake has higher mean pollution levels than another
- A new treatment reduces mean surgery recovery time
- Students in two schools have appreciably different average test scores

H_a is always based only on the a priori belief about the question of interest, never on the data themselves.

Hypothesis Testing: steps

1. Determine H_0 and H_a from the question (null hypothesis and alternative hypothesis)
2. Determine an appropriate Test Statistic T:
   - T is a random variable whose value is determined by the individual sample random variables
3. Identify the (sampling) distribution of T assuming H_0 is true.
4. Determine the Rejection Region:
   - this will be a set of possible values of the Test Statistic T for which we will "reject" H_0 "in favour" of H_a

Hypothesis Testing: steps

5. Compute the observed value of T from the observed data
6. Compare the observed value of the T to the Rejection Region
7. Decide and conclude
    - "Reject H_0 in favour of H_a", or
    - "Do not reject H_0"

Hypothesis Testing: steps

Note

We

- compute the sampling distribution of T,
- form the Rejection Region

assuming that H_0 is true.

This sampling distribution is called the null distribution; we will write P_0 to denote probabilities computed under the null distribution.

Type 1 and Type 2 Error

When we make a decision on the hypotheses, we have the potential to make one of two possible errors:

- A Type 1 error occurs when we reject H_0 when, in fact, H_0 is true.
- A Type 2 error occurs when we fail to reject H_0 when, in fact, H_0 is not true (H_a is true).

Type 1 and Type 2 Error

In creating hypothesis testing procedures,

- choosing T
- computing the Rejection Region

it is mathematically easier to control Type 1 error than it is to control Type 2 error

Thus, we set up our hypotheses such that the more “serious” error will be a Type 1 error, and try to ensure that the Type 2 error is not too large.

Type 1 and Type 2 Error

- The probability of a Type 1 error is denoted α, so

$$\alpha = P_0(T \text{ lies in R. R.})$$

- The probability of a Type 2 error is denoted β.

Both quantities will depend on how we choose H_0 and H_a.

Type 1 and Type 2 Error

We will study three types of sets of hypotheses regarding a parameter θ (i.e. μ or p):

$$\begin{aligned} \text{One-sided:} \quad & H_0 : \theta = \theta_0 \\ & H_a : \theta > \theta_0 \end{aligned}$$

$$\begin{aligned} \text{One-sided:} \quad & H_0 : \theta = \theta_0 \\ & H_a : \theta < \theta_0 \end{aligned}$$

$$\begin{aligned} \text{Two-sided:} \quad & H_0 : \theta = \theta_0 \\ & H_a : \theta \neq \theta_0 \end{aligned}$$

Which one we use is based on the wording of the question (and prior belief about θ).

Test Statistics and rejection regions

We will see different experiment scenarios.

- Pre-experiment: the test statistic is a random variable T

$$T = T(X_1, \ldots, X_n)$$

- Post-experiment, once we substitute our observed data, the test statistic is simply a number:
  - we will denote this number T_{obs}

$$T_{obs} = T(x_1, \ldots, x_n)$$

Test Statistics and rejection regions

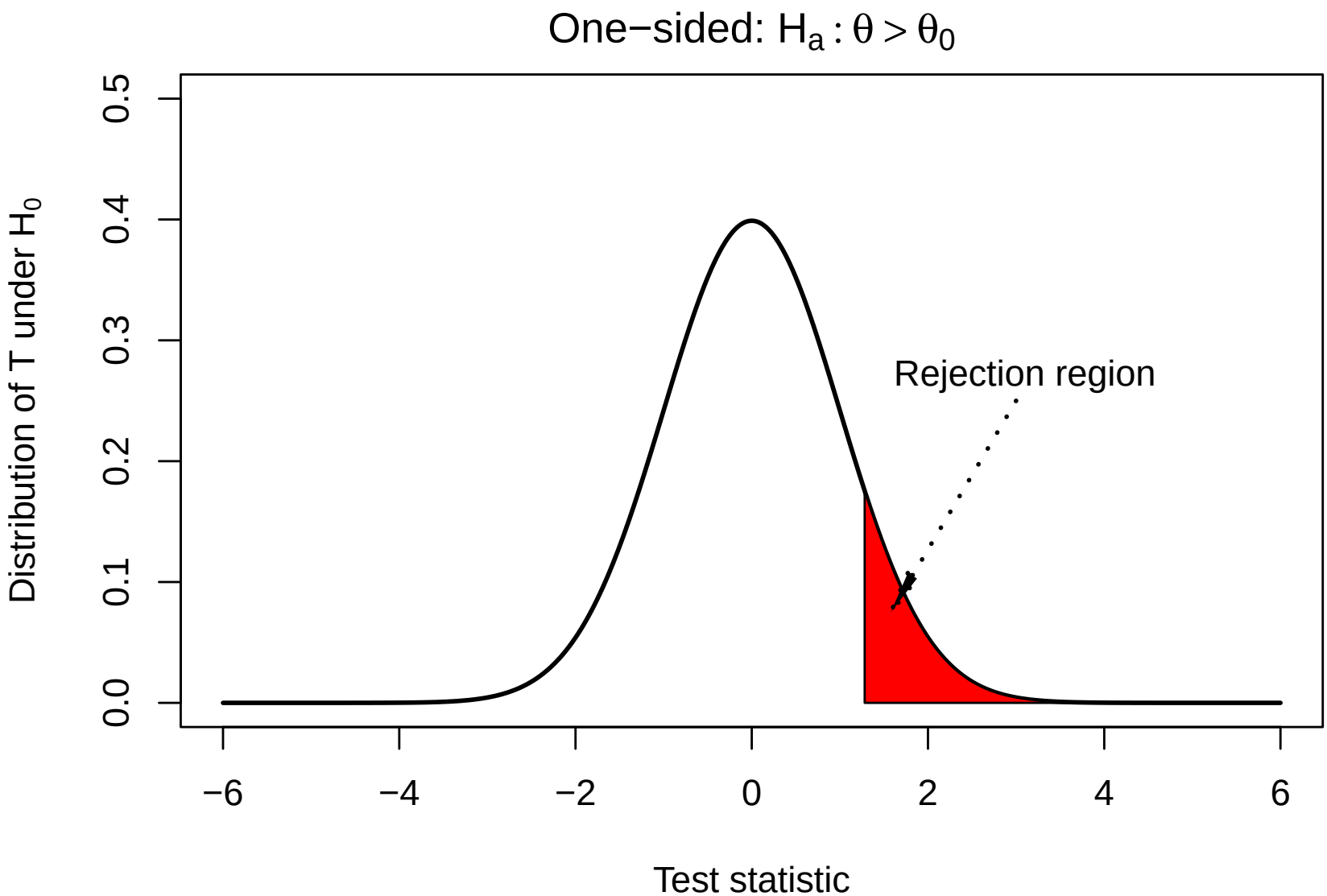
The Rejection region is in the form of a set R, which is determined by H_0, H_a, α, and the scenario, and is not random.

- If T_{obs} falls within R, we will reject H_0 in favour of H_a;
- If it does not, we will not reject H_0

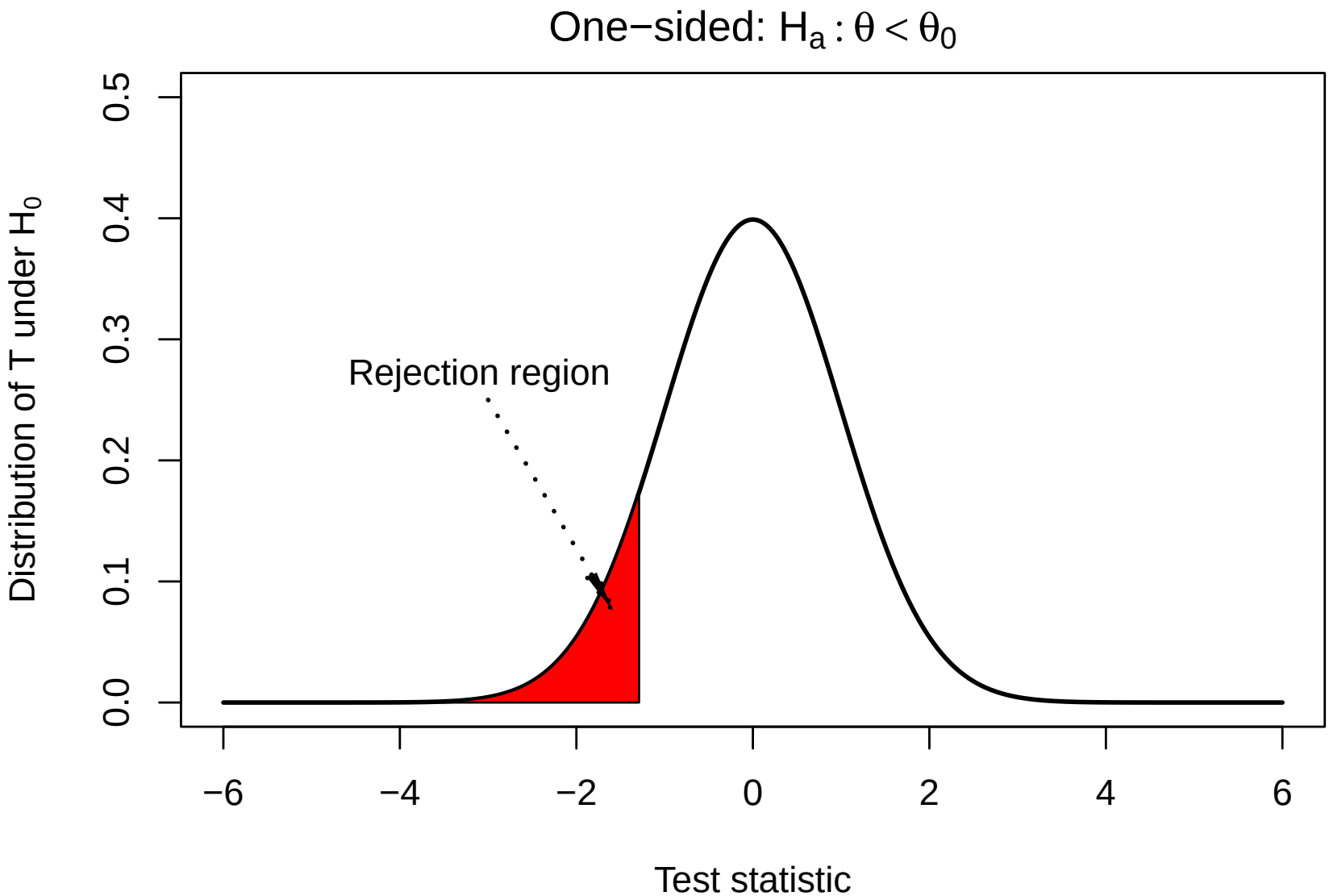
Test Statistics and rejection regions

- Pre-experiment: we may assess the probability that T falls within R
- Post-experiment: either T_{obs} fell within R or not, and as a result we either rejected or did not reject H_0.

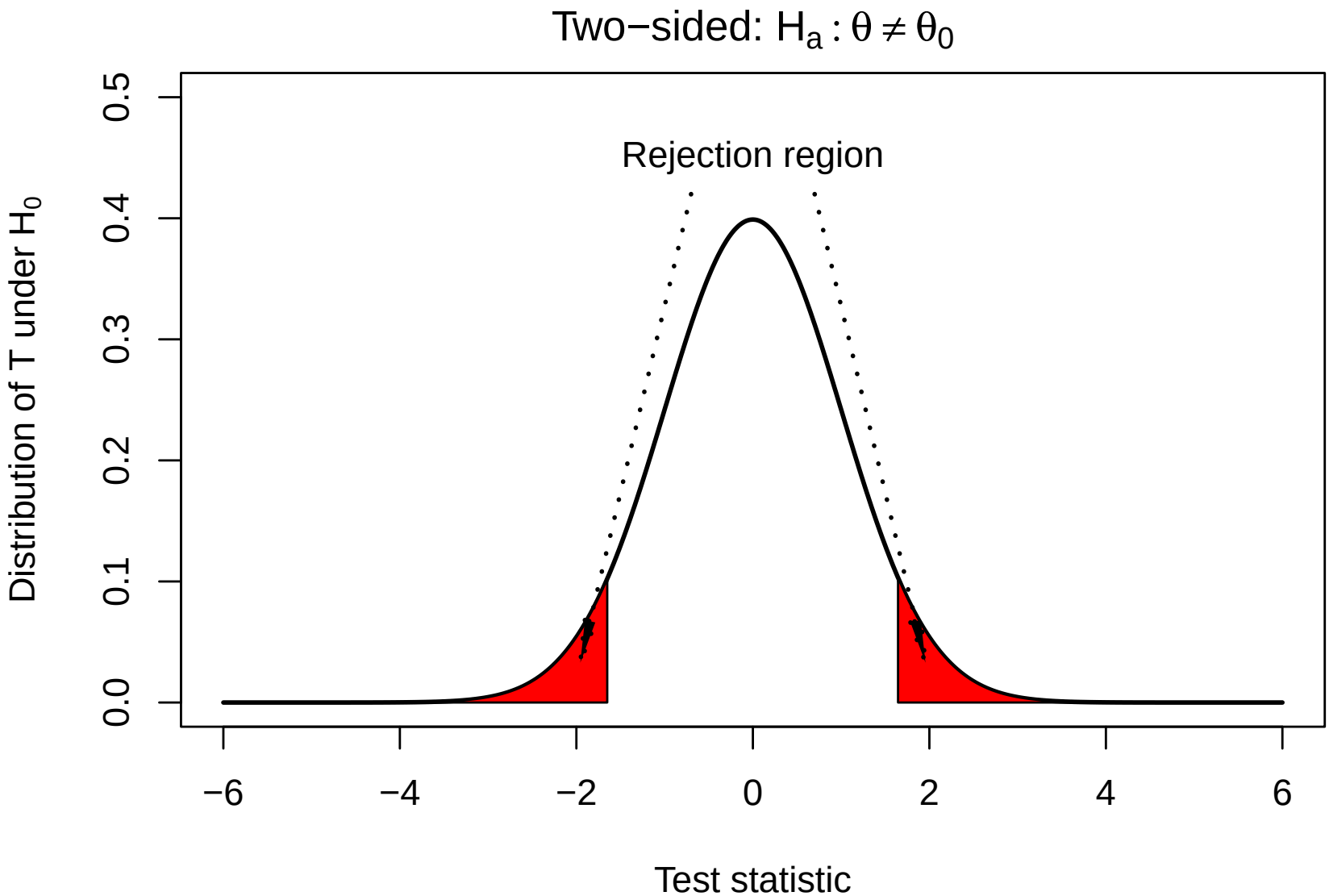
Test Statistics and rejection regions



Test Statistics and rejection regions



Test Statistics and rejection regions



Test Statistics and rejection regions

- All conclusions are made "at the α level", as the chosen value of α affects your decision.
- If the test statistic falls within the rejection region: you reject H_0 in favour of H_a.
- If the test statistic does not fall within the rejection region: you fail to reject H_0;
  - you do not accept H_0 as true, you merely do not have sufficient evidence to decide that it is false.

Test Statistics and rejection regions

Either H_0 or H_a is true, but we do not know which.

After the decision, we are in one of four different situations:

1. Rejected H_0, H_a true
2. Rejected H_0, H_0 true (Type 1 error)
3. Failed to reject H_0, H_0 true
4. Failed to reject H_0, H_a true (Type 2 error)

		Reject H_0 ?	
		Yes	No
H_0 true ?	Yes	2.	3.
	No	1.	4.

Test Statistics and rejection regions

Since we do not know which of H_0 and H_a is true, we do not know which of these situations has occurred.

We design our hypothesis testing procedure such that, on repeated trials, it will falsely reject H_0 (i.e. make a Type 1 error) $100\alpha\%$ of the time

One sample cases

In this course, we will see three types of sets of hypotheses, regarding the parameters μ or p : for null hypothesis

$$H_0 : \mu = \mu_0 \quad \text{or} \quad H_0 : p = p_0,$$

and for the alternative hypothesis, one of

$$\begin{aligned} H_a : \mu > \mu_0 \quad &\text{or} \quad H_a : p > p_0 \\ H_a : \mu < \mu_0 \quad &\text{or} \quad H_a : p < p_0 \\ H_a : \mu \neq \mu_0 \quad &\text{or} \quad H_a : p \neq p_0 \end{aligned}$$

Which one we use is based on the wording of the question (and the prior belief about θ).

One sample cases

In this specification, μ_0 and p_0 represent specific numerical values, for example

- $H_0 : \mu = 10.0$
- $H_0 : p = 0.5$

that represent the hypothesized null values of the data generating parameters.

One sample cases

- Population Mean μ
    1. Small $n(< 30)$
    2. Large $n(\geqslant 30)$
- Population proportion p

One sample cases

Recap on notation:

- $\overline{X}$: sample mean random variable
- S^2: sample variance random variable

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

- t_ν: t-distribution with ν degrees of freedom
- z_a: upper a point of the standard Normal distribution
- $t_{a,\nu}$: upper a point of the t-distribution with ν degrees of freedom

One sample cases

Case: 1. Parameter μ, small n, σ unknown

Suppose that $X_1, \ldots, X_n$ are a random sample from a Normal distribution with unknown μ and σ, and $n < 30$.

$$H_0 : \mu = \mu_0$$

and

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$$

H_a	$H_a : \mu > \mu_0$	$H_a : \mu < \mu_0$	$H_a : \mu \neq \mu_0$
R. R.	$\{T \geqslant t_{\alpha,n-1}\}$	$\{T \leqslant -t_{\alpha,n-1}\}$	$\{\lvert T\rvert \geqslant t_{\alpha/2,n-1}\}$

One sample cases

Case: 2. Parameter μ, σ known

Suppose that $X_1, \ldots, X_n$ are a random sample from a Normal distribution with unknown μ, with σ known.

$$H_0 : \mu = \mu_0$$

and

$$T = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim Normal(0, 1)$$

H_a	$H_a : \mu > \mu_0$	$H_a : \mu < \mu_0$	$H_a : \mu \neq \mu_0$
R. R.	$\{T \geqslant z_\alpha\}$	$\{T \leqslant -z_\alpha\}$	$\{\lvert T\rvert \geqslant z_{\alpha/2}\}$

Note: Here T is often denoted Z, as it is a z-score.

One sample cases

Case: 3. Parameter μ, large n, σ unknown

Suppose that $X_1, \ldots X_n$ are a random sample from a distribution with unknown μ and σ, where $n \geqslant 30$.

$$H_0 : \mu = \mu_0$$

and

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \stackrel{\cdot}{\sim} Normal(0, 1)$$

H_a	$H_a : \mu > \mu_0$	$H_a : \mu < \mu_0$	$H_a : \mu \neq \mu_0$
R. R.	$\{T \geqslant z_\alpha\}$	$\{T \leqslant -z_\alpha\}$	$\{\lvert T\rvert \geqslant z_{\alpha/2}\}$

Normality of the Xs is not required.

One sample cases

Case: 4. Parameter p, n large

Let $X_1, \ldots, X_n$ be a random sample of Bernoulli random variables with unknown probability of success p

$$H_0 : p = p_0$$

and

$$T = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0\,(1 - p_0)}{n}}} \stackrel{\cdot}{\sim} Normal(0, 1)$$

H_a	$H_a : p > p_0$	$H_a : p < p_0$	$H_a : p \neq p_0$
R. R.	$\{T \geqslant z_\alpha\}$	$\{T \leqslant -z_\alpha\}$	$\{\lvert T\rvert \geqslant z_{\alpha/2}\}$

One sample cases

Problem: Painkiller duration

A drug manufacturer is testing the duration of their new over-the-counter painkiller.

Duration is measured on 18 subjects, with sample mean $\overline{x} = 5.2$ hours and sample standard deviation $s = 1.1$ hours.

Previous research has shown that ibuprofen has a mean duration of 4.7 hours.

Conduct an appropriate test of hypotheses at the 5% level to determine whether the new drug lasts longer than ibuprofen. State any assumptions you make.

One sample cases

Solution: Painkiller duration (one-sided)

We are in Case 1. We must assume that the observations X_i are Normally distributed. We are testing

$$\begin{aligned} H_0 : &\quad \mu = 4.7 \\ H_a : &\quad \mu > 4.7 \end{aligned}$$

that is, we are carrying out a one-sided test. We have

- $n = 18$
- $\overline{x} = 5.2$
- $s = 1.1$
- $\mu_0 = 4.7$

so n is small, and σ is unknown.

One sample cases

Solution: Painkiller duration (one-sided)

Here

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$$

Thus, from the data, the observed value is

$$T_{obs} = \frac{5.2 - 4.7}{1.1/\sqrt{18}} = 1.928.$$

One sample cases

Solution: Painkiller duration (one-sided)

$$\text{R. R. :} \quad \{T \geqslant t_{0.05,17} = 1.740\}$$

Since

$$T_{obs} = 1.928 > 1.740$$

we reject H_0 in favour of H_a at the 5% level.

One sample cases

Problem: Painkiller duration (two-sided)

Suppose that the manufacturer claims that the drug has a mean duration of 6 hours.

The drug is tested on 65 subjects, with a sample mean of $\overline{x} = 5.8$ hours and a sample standard deviation of $s = 0.8$ hours.

Conduct an appropriate test of hypotheses at the 1% level

One sample cases

Solution: Painkiller duration (two-sided)

We are in Case 3. We are testing

$$H_0: \quad \mu = 6.0$$
$$H_a: \quad \mu \neq 6.0$$

that is, we are carrying out a two-sided test. We have

- $n = 65$
- $\overline{x} = 5.8$
- $s = 0.8$
- $\mu_0 = 6.0$

so n is large.

One sample cases

Solution: Painkiller duration (two-sided)

Here

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}.$$

We have

$$T_{obs} = \frac{5.8 - 6.0}{0.8/\sqrt{65}} = -2.016.$$

and

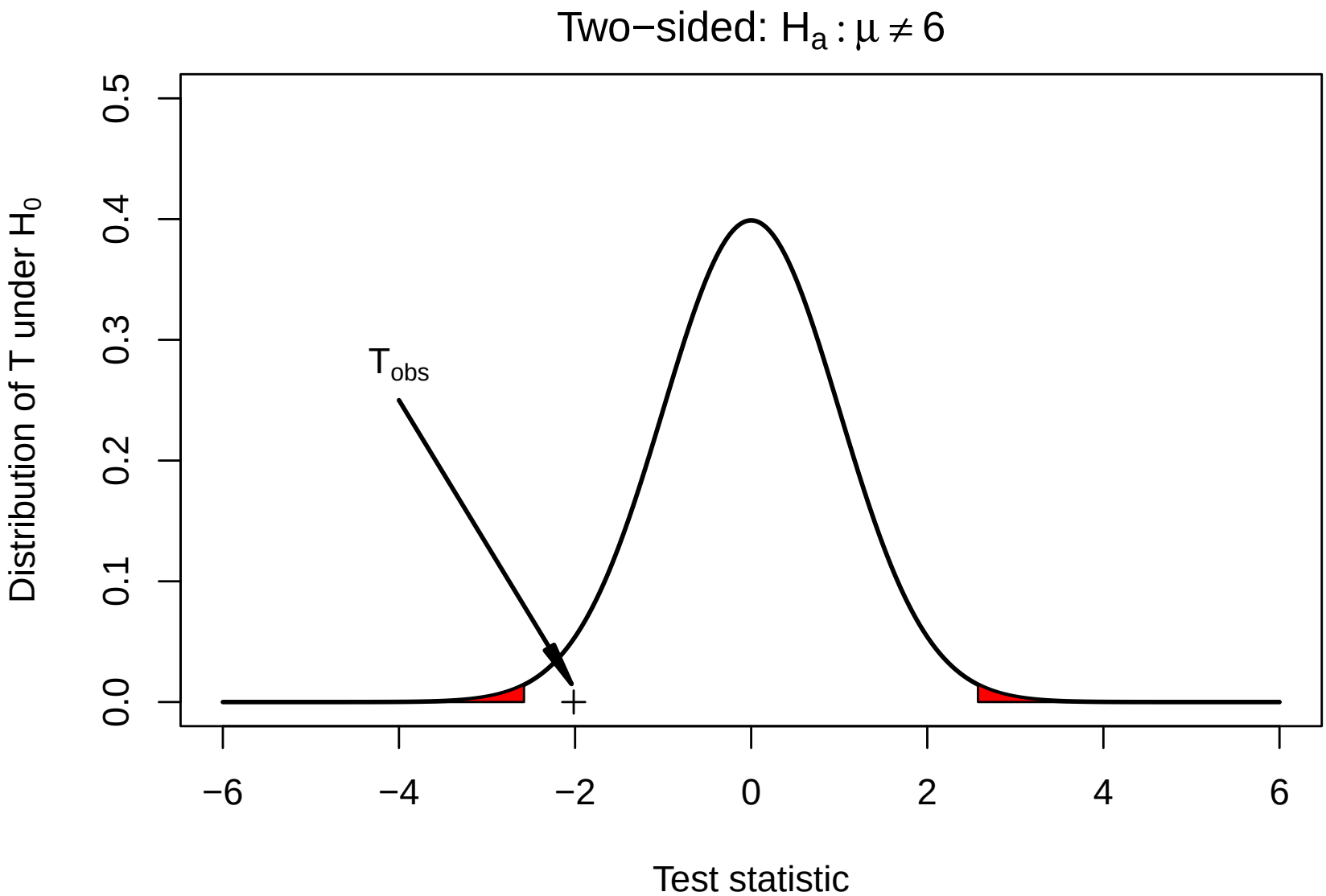
$$\text{R. R. :} \quad \{|T| \geqslant z_{0.005} = 2.576\}$$

Since

$$|T_{obs}| = |-2.016| = 2.016 < 2.576$$

we do not reject H_0 at the 1% level.

One sample cases



One sample cases

Problem: Student stress (one-sided)

McGill is conducting a survey on undergrad student stress levels at the end of the semester, to test whether stress levels are lower than usual. Previous studies have found that 85% of students report high stress.

112 students are recruited, and asked to report their stress levels as either high or low.

87 of these students report that their stress level is high.

Conduct an appropriate test of hypotheses at the 5% level

One sample cases

Solution: Student stress (one-sided)

We are in Case 4. We are testing

$$\begin{aligned} H_0: &\quad p = 0.85 \\ H_a: &\quad p < 0.85 \end{aligned}$$

that is, we are carrying out a one-sided test. We have

- $n = 112$
- $x = 87$
- $\hat{p} = 0.777$
- $p_0 = 0.85$

so n is large.

One sample cases

Solution: Student stress (one-sided)

The test statistic is

$$T = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0\left(1 - p_0\right)}{n}}}$$

Thus

$$T_{obs} = \frac{0.777 - 0.850}{\sqrt{\frac{0.850(1 - 0.850)}{112}}} = -2.170$$

One sample cases

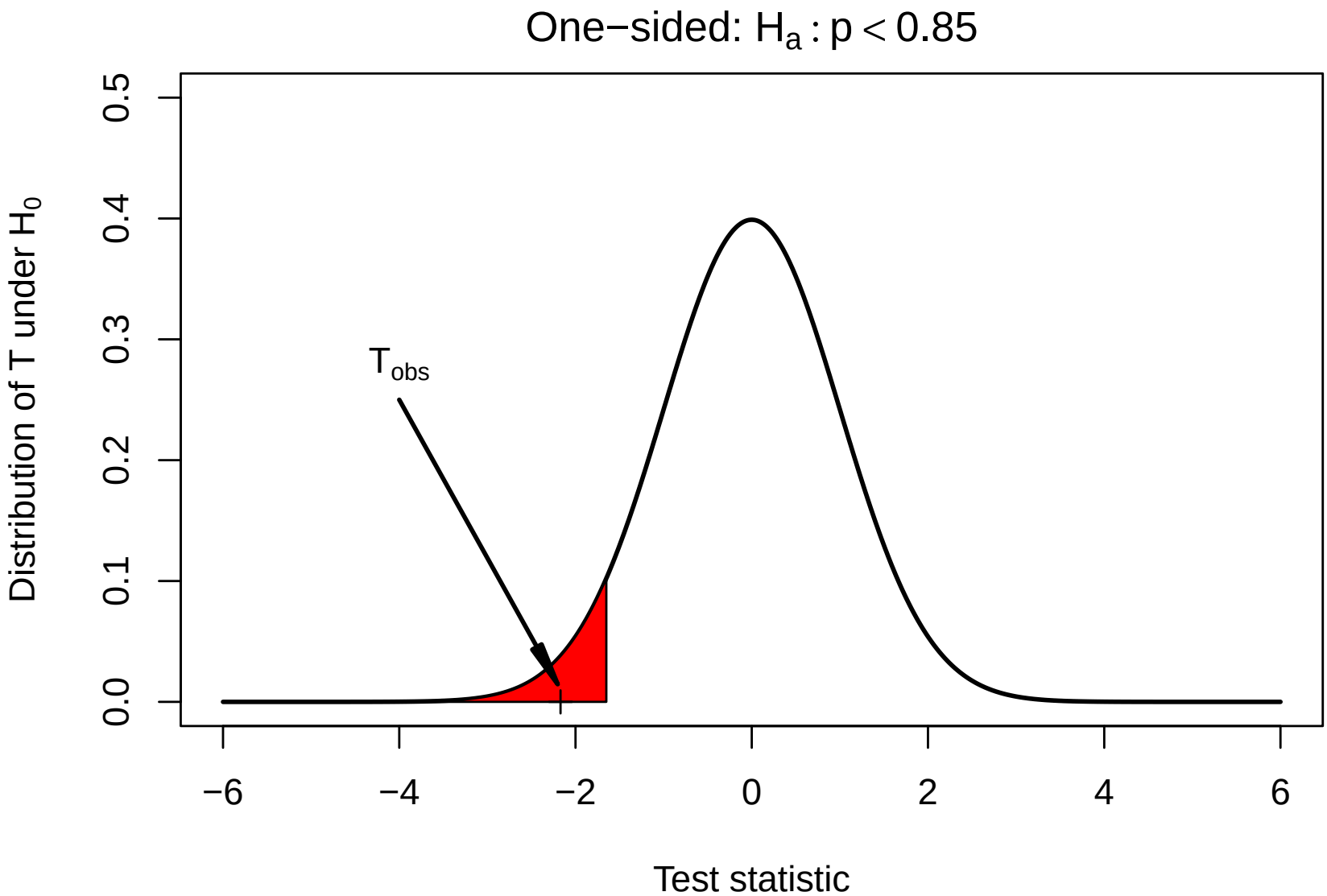
Solution: Student stress (one-sided)

Here

$$\text{R. R. :} \quad \{T \leqslant -z_{0.05} = -1.645\}$$

Since $-2.170 < -1.645$, we reject H_0 in favour of H_a at the 5% level

One sample cases



One sample cases

Note: computation in R

```
n<-112
x<-87
phat<-x/n
p0<-0.85
se.hat<-sqrt(p0*(1-p0)/n)
alph<-0.05
t.obs<-(phat-p0)/se.hat
z.a<-qnorm(alph)
c(phat,se.hat,t.obs,z.a)

## [1]  0.77678571  0.03374008 -2.16995011 -1.64485363
```

One sample cases

Note: computation in R with rounding to 2 d.p.

```
n<-112
x<-87
phat<-round(x/n,2)
p0<-0.85
se.hat<-round(sqrt(p0*(1-p0)/n),2)
alph<-0.05
t.obs<-round((phat-p0)/se.hat,2)
z.a<-round(qnorm(alph),2)
c(phat,se.hat,t.obs,z.a)

## [1]  0.78  0.03 -2.33 -1.64
```

So rounding does change the value of the T_{obs} statistic.

p-values

Many practitioners are uncomfortable with selecting α before seeing their data

- The choice of α is somewhat arbitrary
- Once you look at your data, you can no longer change α

p-values

Hypothesis testing does not take into account where T_{obs} lies in (or out of) the rejection region.

- If you fail to reject H_0, is T_{obs} very close to, or farther away from, the edge of the R. R. ?
- If you reject H_0 in favour of H_a, is T_{obs} close to, or far away from, the edge of the R. R.?

p-values

It is common to avoid a decision-based approach and instead just report a p-value.

The p-value is a probability that may be thought of as "evidence for H_a", and reporting it does not force you to make a decision.

p-values: definition

For a test of hypotheses H_0 and H_a with test statistic T and its observed value T_{obs}, the p-value is defined as follows:

Assuming H_0 is true

- For $H_a : \theta > \theta_0$: p-value $= P_0(T \geqslant T_{obs})$
- For $H_a : \theta < \theta_0$: p-value $= P_0(T \leqslant T_{obs})$
- For $H_a : \theta \neq \theta_0$: p-value $= 2 \times P_0(T \geqslant |T_{obs}|)$

p-values: interpretation

In essence: the p-value is the probability of obtaining a value of T at least as "extreme" as T_{obs}, assuming that H_0 is true,

- "extreme" is determined by the "direction" of H_a.

p-values: interpretation

If p-value is "small": this could be due to two different reasons:

- H_0 is true and the data we observed are merely an unusual outcome (by chance)
- H_0 is not true; H_a is a preferable explanation of the observed data

The smaller the p-value, the less we believe the first case, and the more we believe the second.

p-values: interpretation

If the p-value is large, then the evidence against H_0 is weak.

This does not mean that you believe H_0 is true, just that your current data does not give enough evidence to reject it in favour of H_a.

p-values: implementation

1. Determine H_0 and H_a
2. Determine T and its distribution assuming H_0 is true
3. Compute T_{obs}
4. Calculate p-value based on H_a and the distribution of T
5. Conclude, in terms of "strength of evidence"

p-values: implementation

We may consider "evidence against H_0" based on p-values on the following scale:

p-value	Evidence against H_0
< 0.001	Extremely strong
$0.001 \leqslant p < 0.01$	Very strong
$0.01 \leqslant p < 0.05$	Strong
$0.05 \leqslant p < 0.10$	Modest
> 0.10	Not very strong

However, this is simply a guide, and these interpretations can vary by field.

p-values: implementation

z-test: If $T \sim Normal(0,1)$:

- $H_a : \theta > \theta_0$

$$p\text{-value} = P_0\left(T \geqslant T_{obs}\right)$$

- $H_a : \theta < \theta_0$

$$p\text{-value} = P_0\left(T \leqslant T_{obs}\right)$$

- $H_a : \theta \neq \theta_0$

$$\begin{aligned} p\text{-value} &= P_0\left(T \geqslant \left|T_{obs}\right|\right) + P_0\left(T \leqslant -\left|T_{obs}\right|\right) \\ &= 2P_0\left(T \geqslant \left|T_{obs}\right|\right) \end{aligned}$$

These p-values can be calculated using tables or software.

p-values: implementation

t-test: If $T \sim t_\nu$

- $H_a : \theta > \theta_0$

$$p\text{-value} = P_0\left(T \geqslant T_{obs}\right)$$

- $H_a : \theta < \theta_0$

$$p\text{-value} = P_0\left(T \leqslant T_{obs}\right)$$

- $H_a : \theta \neq \theta_0$

$$\begin{aligned} p\text{-value} &= P_0\left(T \geqslant \left|T_{obs}\right|\right) + P_0\left(T \leqslant -\left|T_{obs}\right|\right) \\ &= 2P_0\left(T \geqslant \left|T_{obs}\right|\right) \end{aligned}$$

These can be calculated using tables or software.

p-values: Examples

Problem: Balls of yarn

A yarn manufacturer is concerned that a particular plant is producing balls that contain either too little or too much yarn; each ball should contain 100 grams.

The manufacturer randomly samples 45 balls, which have a sample mean weight of 99.2 grams, with a sample standard deviation of 1.7 grams.

State a set of appropriate hypotheses and calculate the corresponding p-value.

p-values: Examples

Solution: Balls of yarn

We are in Case 3. We are testing

$$\begin{aligned} H_0 : &\quad \mu = 100 \\ H_a : &\quad \mu \neq 100 \end{aligned}$$

that is, we are carrying out a two-sided test. We have

- $n = 45$
- $\overline{x} = 99.2$
- $s = 1.7$
- $\mu_0 = 100$

so n is large.

p-values: Examples

Solution: Balls of yarn

Here

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \stackrel{\cdot}{\sim} Normal(0,1).$$

and we have

$$T_{obs} = \frac{99.2 - 100.0}{1.7/\sqrt{45}} = -3.157.$$

Since T has a standard Normal distribution, the p-value is:

$$\begin{aligned} p\text{-value} &= 2 \times P_0\left(T \geqslant |T_{obs}|\right) \\ &= 2 \times P_0(T \geqslant |-3.157|) = 2 \times P_0(T \geqslant 3.157) \\ &= 0.002 \end{aligned}$$

p-values: Examples

```
n<-45
xbar<-99.2
mu0<-100
sv<-1.7
t.obs<-(xbar-mu0)/(sv/sqrt(n))
p.val<-2*pnorm(-abs(t.obs))
c(t.obs,p.val)

## [1] -3.156801851  0.001595097
```

p-values: Examples

Problem: Ice cream flavours

An ice cream company is creating a new ice cream flavour and wants to find out whether people prefer a base with chocolate or caramel swirls; they anticipate that caramel will be preferable

The company gives the ice cream to 45 taste testers, 21 of whom prefer caramel.

State a set of appropriate hypotheses and calculate the corresponding p-value.

p-values: Examples

Solution: Ice cream flavours

We are in Case 4. We consider testing

$$H_0: \quad p = 0.5$$
$$H_a: \quad p > 0.5$$

that is, we are carrying out a one-sided test. We have

- $n = 45$
- $x = 21$
- $\hat{p} = 0.467$
- $p_0 = 0.50$

so n is large.

p-values: Examples

Solution: Ice cream flavours

Note that $\hat{p}$ is on the wrong side of p_0 to support H_a.

This means we need to take care to apply the formula

$$p\text{-value} = P_0\left(T \geqslant T_{obs}\right)$$

carefully. Under H_0, T approximately has a standard Normal distribution.

p-values: Examples

Solution: Ice cream flavours

The test statistic is

$$T = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0\,(1 - p_0)}{n}}}$$

Thus

$$T_{obs} = \frac{0.467 - 0.500}{\sqrt{\frac{0.500(1 - 0.500)}{45}}} = -0.447$$

p-values: Examples

Solution: Ice cream flavours

Since T has a standard Normal distribution assuming H_0 is true, the p-value is

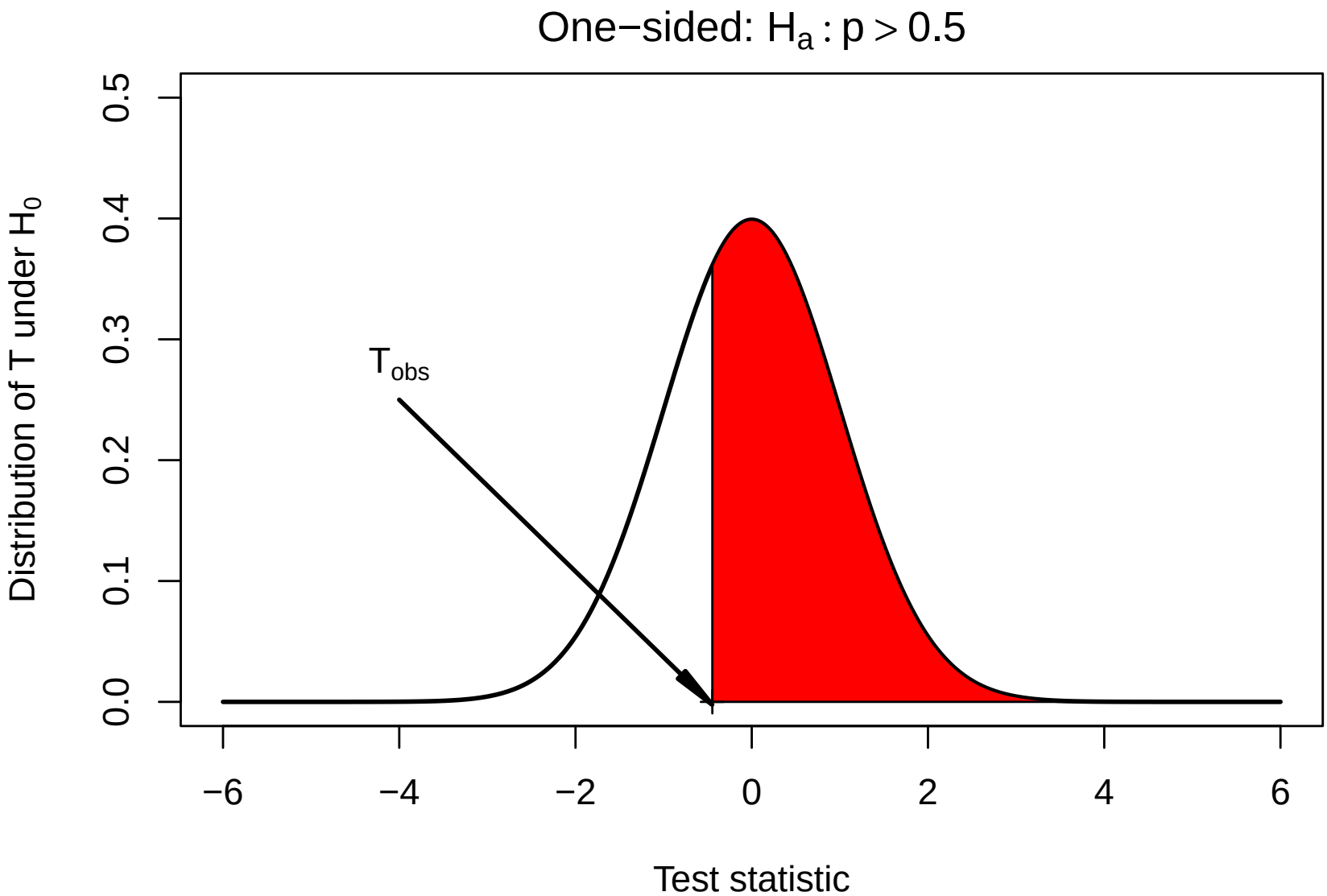
$$p\text{-value} = P_0\left(T \geqslant T_{obs}\right) = P_0(T \geqslant -0.447) = 0.673$$

p-values: Examples

```
n<-45
x<-21
phat<-x/n
p0<-0.5
se.hat<-sqrt(p0*(1-p0)/n)
t.obs<-(phat-p0)/se.hat
p.val<-1-pnorm(t.obs)
c(t.obs,p.val)

## [1] -0.4472136  0.6726396
```

p-values: Examples



Two sample comparisons

We may extend the ideas of

- confidence intervals
- hypothesis tests

to allow the comparison of two samples.

- $X_1, \ldots, X_m$ are a random sample from one population
- $Y_1, \ldots, Y_n$ are a random sample from a second population

Two sample comparisons

We can compute the sample mean and sample variance for each sample

$$\overline{X} = \frac{1}{m}\sum_{i=1}^{m} X_i \qquad S_1^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_i - \overline{X})^2$$

and

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{m} Y_i \qquad S_2^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

Two sample comparisons: paired data

In the two sample case, we can also consider making two observations on each of n experimental units:

- X_i: pre-treatment, or Treatment 1
- Y_i: post-treatment, or Treatment 2

and then looking at the difference

$$D_i = X_i - Y_i$$

to assess whether there is any difference. We call this the case of paired data.

Two sample case: confidence intervals

To determine the correct strategy, we need to identify

(A) the assumed probability model/parameter of interest
- Normal, $\mu_1 - \mu_2$
- Binomial, $p_1 - p_2$

(B) the sample size assumption
- m, n small ($m, n < 30$)
- m, n large ($m, n \geqslant 30$)

(C) paired observations
- No
- Yes

Two sample case: confidence intervals

Theorem: 1. Parameter $\mu_1 - \mu_2$, unpaired, small m, n

Suppose that

- $X_1, \ldots, X_m$ are a random sample from a Normal distribution with unknown μ_1 and σ_1
- $Y_1, \ldots, Y_n$ are a random sample from a Normal distribution with unknown μ_2 and σ_2
- the X_is are independent of the Y_is
- $m < 30$ or $n < 30$
- $\sigma_1 = \sigma_2$

Two sample case: confidence intervals

Theorem: 1. Parameter $\mu_1 - \mu_2$, unpaired, small m, n

A $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is (L, R), where

$$L = (\overline{X} - \overline{Y}) - t_{\alpha/2,m+n-2} S_P \sqrt{\frac{1}{m} + \frac{1}{n}}$$

$$R = (\overline{X} - \overline{Y}) + t_{\alpha/2,m+n-2} S_P \sqrt{\frac{1}{m} + \frac{1}{n}}$$

and

$$S_P^2 = \frac{(m-1)S_1^2 + (n-1)S_2^2}{m+n-2}$$

Two sample case: confidence intervals

Theorem: 2. Parameter $\mu_1 - \mu_2$, unpaired, large m, n

Suppose that

- $X_1, \ldots, X_m$ are a random sample from a distribution with unknown μ_1 and σ_1
- $Y_1, \ldots, Y_n$ are a random sample from a distribution with unknown μ_2 and σ_2
- the X_is are independent of the Y_is
- $m \geqslant 30$ and $n \geqslant 30$

Normality of the Xs and Ys is not required.

Two sample case: confidence intervals

Theorem: 2. Parameter $\mu_1 - \mu_2$, unpaired, large m, n

A $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is (L, R), where

$$L = (\overline{X} - \overline{Y}) - z_{\alpha/2}\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}$$

$$R = (\overline{X} - \overline{Y}) + z_{\alpha/2}\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}$$

Two sample case: confidence intervals

Theorem: 3. Parameter $\mu_1 - \mu_2$, paired, small n

Suppose

- $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ be paired observations from distributions with unknown means μ_1 and μ_2, respectively,
- $n < 30$

Let

$$D_i = X_i - Y_i$$

and

$$\overline{D} = \frac{1}{n}\sum_{i=1}^{n} D_i \qquad S_d^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \overline{D})^2.$$

Two sample case: confidence intervals

Theorem: 3. Parameter $\mu_1 - \mu_2$, paired, small n

Assume that the D_i are a random sample from a Normal distribution with unknown mean

$$\mu_d = \mu_1 - \mu_2$$

and standard deviation σ_d.

Two sample case: confidence intervals

Theorem: 3. Parameter $\mu_1 - \mu_2$, paired, small n

A $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is (L, R), where

$$L = \overline{D} - t_{\alpha/2,n-1}\frac{S_d}{\sqrt{n}}$$

$$R = \overline{D} + t_{\alpha/2,n-1}\frac{S_d}{\sqrt{n}}$$

Two sample case: confidence intervals

Theorem: 4. Parameter $\mu_1 - \mu_2$, paired, large n

Suppose

- $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ be paired observations from distributions with unknown means μ_1 and μ_2, respectively,
- $D_i = X_i - Y_i$, $i = 1, \ldots, n$ are a random sample from a distribution with unknown mean

$$\mu_d = \mu_1 - \mu_2$$

  and standard deviation σ_d.
- $n \geqslant 30$

Normality is **not required**.

Two sample case: confidence intervals

Theorem: 4. Parameter $\mu_1 - \mu_2$, paired, large n

A $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is (L, R), where:

$$L = \overline{D} - z_{\alpha/2}\frac{S_d}{\sqrt{n}}$$

$$R = \overline{D} + z_{\alpha/2}\frac{S_d}{\sqrt{n}}$$

Two sample case: confidence intervals

Theorem: 5. Parameter $p_1 - p_2$, large m, n

Suppose

- $X_1, \ldots, X_m$ be a random sample of Bernoulli random variables with unknown probability of success p_1
- $Y_1, \ldots, Y_n$ be a random sample of Bernoulli random variables with unknown probability of success p_2
- the X_is are independent of the Y_is,
- m and n are large enough

Two sample case: confidence intervals

Theorem: 5. Parameter $p_1 - p_2$, large m, n)

Let

- X be the number of "successes" in the sample $X_1, \ldots, X_m$
- Y be the number of "successes" in the sample $Y_1, \ldots, Y_n$

and

$$\hat{p}_1 = \frac{X}{m} \qquad \text{and} \qquad \hat{p}_2 = \frac{Y}{n}$$

Two sample case: confidence intervals

Theorem: 5. Parameter $p_1 - p_2$, large m, n

A $100(1-\alpha)\%$ confidence interval for $p_1 - p_2$ is (L, R), where

$$L = (\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{m} + \frac{\hat{p}_2(1-\hat{p}_2)}{n}}$$

$$R = (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{m} + \frac{\hat{p}_2(1-\hat{p}_2)}{n}}$$

Examples

Problem: Painkiller duration

A drug manufacturer is testing the duration of pain-relief of their new over-the-counter painkiller, and wishes to quantify the difference in duration compared to ibuprofen.

Examples

Problem: Painkiller duration

The following data are available

- New drug
  - $m = 11$ observations
  - sample mean $\overline{x} = 5.4$ hours
  - sample standard deviation $s_1 = 0.9$ hours
- Ibuprofen
  - $n = 9$ observations
  - sample mean $\overline{y} = 4.8$ hours
  - sample standard deviation $s_2 = 1.1$ hours

Calculate a 90% confidence interval for the difference in mean durations between the two drugs.

State any assumptions you make.

Examples

Solution: Painkiller duration

We are in Scenario 1: $m = 11, n = 9$ and we are finding a 90%Cl for $\mu_1 - \mu_2$, with $\alpha = 0.1$.

We must assume that the Xs and Ys are Normally distributed and independent, and that the population variances are equal, i.e. $\sigma_1 = \sigma_2$. Then

$$\begin{aligned} s_p^2 &= \frac{(m-1)s_1^2 + (n-1)s_2^2}{m+n-2} \\ &= \frac{10 \times 0.9^2 + 8 \times 1.1^2}{11+9-2} \\ &= 0.988 \end{aligned}$$

so that $s_p = \sqrt{0.988} = 0.994$.

Examples

Solution: Painkiller duration

Let $\nu = n + m - 2$: we have that

$$(\overline{x} - \overline{y}) = 5.4 - 4.8 = 0.6,$$

and that

$$t_{\alpha/2,\nu} = t_{0.05,\nu} = 1.734$$

and

$$\sqrt{\frac{1}{m} + \frac{1}{n}} = \sqrt{\frac{1}{11} + \frac{1}{9}} = 0.449.$$

Examples

Solution: Painkiller duration

The 90% CI for $\mu_1 - \mu_2$ is

$$\begin{aligned}(\overline{x} - \overline{y}) \pm t_{0.05,\nu} s_p \sqrt{\frac{1}{m} + \frac{1}{n}} \\ = 0.6 \pm 1.734 \times 0.994 \times 0.449 \\ = (-0.175, 1.375)\end{aligned}$$

Examples

Problem: Painkiller duration (large sample)

Now suppose that 55 patients are given the new drug and 62 are given ibuprofen.

- new drug:
  - the sample mean is $\overline{x} = 5.7$ hours
  - sample standard deviation is $s_1 = 1.1$ hours
- ibuprofen:
  - the sample mean is $\overline{y} = 4.6$ hours
  - the sample standard deviation is $s_2 = 0.4$ hours

Calculate a 95% confidence interval for the difference in mean durations between the two drugs

Examples

Solution: Painkiller duration (large sample)

We are in Scenario 2: $m = 55, n = 62$ and we are finding a 95% interval for $\mu_1 - \mu_2$.

Here

$$\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}} = \sqrt{\frac{1.1^2}{55} + \frac{0.4^2}{62}} = 0.157$$

Examples

Solution: Painkiller duration (large sample)

Let $\nu = n + m - 2$: we have that

$$(\overline{x} - \overline{y}) = 5.7 - 4.6 = 1.1,$$

and that

$$z_{\alpha/2} = z_{0.025} = 1.960$$

Examples

Solution: Painkiller duration (large sample)

The 95% CI for $\mu_1 - \mu_2$ is

$$
\begin{aligned}
(\overline{x} - \overline{y}) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}} \\
&= 1.1 \pm 1.960 \times 0.157 \\
&= (0.793, 1.407)
\end{aligned}
$$

Examples

Problem: Painkiller duration (paired samples)

Now suppose that 15 patients are given ibuprofen, then given the new drug at a later date.

The sample mean difference in drug duration is $\overline{d} = 1.1$ hours, with sample standard deviation $s_d = 0.3$ hours.

Calculate a 99% confidence interval for the difference in mean durations between the two drugs.

Examples

Solution: Painkiller duration (paired samples)

We are in Scenario 3: $n = 15$ paired observations, and we are finding a 99%CI for $\mu_d = \mu_1 - \mu_2$, with $\alpha = 0.01$. We have that

$$t_{\alpha/2,\nu} = t_{0.005,14} = 2.977.$$

Examples

Solution: Painkiller duration (paired samples)

Then

$$\left(\overline{d} - t_{\alpha/2,\nu}\frac{s_d}{\sqrt{n}}, \overline{d} + t_{\alpha/2,\nu}\frac{s_d}{\sqrt{n}}\right)$$
$$= \left(1.1 - 2.977\frac{0.3}{\sqrt{15}}, 1.1 + 2.977\frac{0.3}{\sqrt{15}}\right)$$
$$= (0.869, 1.331)$$

We must make the assumption that the differences in drug duration are Normally distributed.

Examples

Problem: Painkiller duration (paired samples, large n)

Suppose that 72 patients are given both drugs;

- the sample mean difference is $\overline{d} = 1.2$ hours,
- the standard deviation is $s_d = 0.4$ hours

Calculate a 95% confidence interval for the difference in mean durations between the two drugs

Examples

Solution: Painkiller duration (paired samples, large n)

We are in Scenario 4: $n = 72$ paired observations, and we are finding a 95%CI for $\mu_d = \mu_1 - \mu_2$, with $\alpha = 0.05$.

$$\begin{aligned}
&\left(\overline{d} - z_{\alpha/2}\ \frac{s_d}{\sqrt{n}}, \overline{d} + z_{\alpha/2}\frac{s_d}{\sqrt{n}}\right) \\
&\qquad = \left(1.2 - 1.960\frac{0.4}{\sqrt{72}}, 1.2 + 1.960\frac{0.4}{\sqrt{72}}\right) \\
&\qquad = (1.108, 1.292)
\end{aligned}$$

Examples

Problem: Student stress

Suppose that McGill students are being compared to Concordia students with respect to reporting stress.

- McGill
  - 112 students surveyed
  - 87 report high stress
- Concordia
  - 131 students surveyed
  - 92 report high stress

Calculate a 95% confidence interval for the difference in the proportions of students reporting high stress levels between the two schools

Examples

Solution: Student stress

We are in Scenario 5: Two Binomial setups, and we are finding a 95%Cl for $p_1 - p_2$, with $\alpha = 0.05$

$$\hat{p}_1 = \frac{87}{112} = 0.777 \qquad \hat{p}_2 = \frac{92}{131} = 0.702$$

and

$$\sqrt{\frac{\hat{p}_1\left(1-\hat{p}_1\right)}{m} + \frac{\hat{p}_2\left(1-\hat{p}_2\right)}{n}} = \sqrt{0.003} = 0.056$$

Examples

Solution: Student stress

The 95% CI is

$$
\begin{aligned}
(\hat{p}_1 - \hat{p}_2) \pm z_{0.025}\sqrt{\frac{\hat{p}_1\,(1-\hat{p}_1)}{m} + \frac{\hat{p}_2\,(1-\hat{p}_2)}{n}} \\
&= (0.777 - 0.702) \pm 1.960 \times 0.056 \\
&= (-0.035, 0.184)
\end{aligned}
$$

The limits of the CI can be negative as we are looking at a difference in probabilities.

Two sample hypothesis testing

Case: 1. Parameter $\mu_1 - \mu_2$, unpaired, small n, m

Suppose that

- $X_1, \ldots, X_m$ are a random sample from a Normal distribution with unknown μ_1 and σ_1.
- $Y_1, \ldots, Y_n$ are a random sample from a Normal distribution with unknown μ_2 and σ_2.

Two sample hypothesis testing

Case: 1. Parameter $\mu_1 - \mu_2$, unpaired, small n, m

To test

$$H_0 : \mu_1 = \mu_2$$

at level α, with

$$T = \frac{\overline{X} - \overline{Y}}{S_p \sqrt{\frac{1}{m} + \frac{1}{n}}}$$

and if $\nu = m + n - 2$,

	$H_a : \mu_1 > \mu_2$	$H_a : \mu_1 < \mu_2$	$H_a : \mu_1 \neq \mu_2$
R. R.	$\{T \geqslant t_{\alpha,\nu}\}$	$\{T \leqslant -t_{\alpha,\nu}\}$	$\{\lvert T\rvert \geqslant t_{\alpha/2,\nu}\}$

where S_p^2 is the pooled sample variance.

Two sample hypothesis testing

Case: 2. Parameter $\mu_1 - \mu_2$, unpaired, large n, m

Suppose that

- $X_1, \ldots, X_m$ are a random sample from a distribution with unknown μ_1 and σ_1.
- $Y_1, \ldots, Y_n$ are a random sample from a distribution with unknown μ_2 and σ_2.

Further, assume $m \geqslant 30$ and $n \geqslant 30$ and the X_is are independent of the Y_is.

Normality of the Xs and Ys is not required.

Two sample hypothesis testing

Case: 2. Parameter $\mu_1 - \mu_2$, unpaired, large n, m

To test

$$H_0 : \mu_1 = \mu_2$$

at level α with

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}} \doteqdot Normal(0,1)$$

	$H_a : \mu_1 > \mu_2$	$H_a : \mu_1 < \mu_2$	$H_a : \mu_1 \neq \mu_2$
R. R.	$\{T \geqslant z_\alpha\}$	$\{T \leqslant -z_\alpha\}$	$\{\lvert T\rvert \geqslant z_{\alpha/2}\}$

Two sample hypothesis testing

Case: 3. Parameter $\mu_1 - \mu_2$, paired, small n

Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ be paired observations from distributions with unknown μ_1 and μ_2, respectively, and $n < 30$.

Further, let $D_i = X_i - Y_i$ and assume that the D_i are a random sample from a Normal distribution with unknown $\mu_d = \mu_1 - \mu_2$ and σ_d

Two sample hypothesis testing

Case: 3. Parameter $\mu_1 - \mu_2$, paired, small n

To test

$$H_0 : \mu_1 = \mu_2$$

with

$$T = \frac{\overline{D}}{S_d/\sqrt{n}} \sim t_{n-1}$$

	$H_a : \mu_1 > \mu_2$	$H_a : \mu_1 < \mu_2$	$H_a : \mu_1 \neq \mu_2$
R. R.	$\{T \geqslant t_{\alpha,n-1}\}$	$\{T \leqslant -t_{\alpha,n-1}\}$	$\{\lvert T\rvert \geqslant t_{\alpha/2,n-1}\}$

Two sample hypothesis testing

Case: 4. Parameter $\mu_1 - \mu_2$, paired, large n

Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ be paired observations from distributions with unknown μ_1 and μ_2, respectively, and $n \geqslant 30$.

Further, let $D_i = X_i - Y_i$ and assume that the D_i are a random sample from a distribution with unknown $\mu_d = \mu_1 - \mu_2$ and σ_d.

Normality is not required.

Two sample hypothesis testing

Case: 4. Parameter $\mu_1 - \mu_2$, paired, large n

To test

$$H_0 : \mu_1 = \mu_2$$

with

$$T = \frac{\overline{D}}{S_d/\sqrt{n}} \dot{\sim} \mathit{Normal}(0,1)$$

	$H_a : \mu_1 > \mu_2$	$H_a : \mu_1 < \mu_2$	$H_a : \mu_1 \neq \mu_2$
R. R.	$\{T \geqslant z_\alpha\}$	$\{T \leqslant -z_\alpha\}$	$\{\lvert T\rvert \geqslant z_{\alpha/2}\}$

Two sample hypothesis testing

Case: 5. Parameter $p_1 - p_2$, m, n large

Suppose that

- $X_1, \ldots, X_m$ is a random sample of Bernoulli random variables with unknown probability of success p_1
- $Y_1, \ldots, Y_n$ is a random sample of Bernoulli random variables with unknown probability of success p_2

Suppose the $X_i s$ are independent of the Y_is, and that m and n are large enough. Let

- X be the number if successes in the sample $X_1, \ldots, X_m$,
- Y be the number if successes in the sample $Y_1, \ldots, Y_n$.

Two sample hypothesis testing

Case: 5. Parameter $p_1 - p_2$

To test

$$H_0 : p_1 = p_2$$

with

$$T = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{m} + \frac{1}{n}\right)}} \dot{\sim} \mathit{Normal}(0,1)$$

where

$$\hat{p} = \frac{X + Y}{m + n}$$

	$H_a : p_1 > p_2$	$H_a : p_1 < p_2$	$H_a : p_1 \neq p_2$
R. R.	$\{T \geqslant z_\alpha\}$	$\{T \leqslant -z_\alpha\}$	$\{\lvert T\rvert \geqslant z_{\alpha/2}\}$

Two sample hypothesis testing: Examples

Problem: Painkiller duration

A drug manufacturer is testing the duration of pain-relief of their new over-the-counter painkiller, and wishes to quantify the difference in duration compared to ibuprofen.

- New drug
  - $m = 11$ observations
  - sample mean $\overline{x} = 5.4$ hours
  - sample standard deviation $s_1 = 0.9$ hours
- Ibuprofen
  - $n = 9$ observations
  - sample mean $\overline{y} = 4.8$ hours
  - sample standard deviation $s_2 = 1.1$ hours

Two sample hypothesis testing: Examples

Problem: Painkiller duration

Conduct an appropriate test of hypotheses at the 1% level to determine whether the new drug lasts longer than ibuprofen. State any assumptions you make.

Two sample hypothesis testing: Examples

Solution: Painkiller duration

We are in Case 1. We must assume that the observations X_i and Y_i are Normally distributed, that $\sigma_1 = \sigma_2$, and that the Xs are independent of the Ys. We are testing

$$\begin{aligned} H_0 : &\quad \mu_1 = \mu_2 \\ H_a : &\quad \mu_1 > \mu_2 \end{aligned}$$

$$\begin{aligned} s_p^2 = \frac{(m-1)s_1^2 + (n-1)s_2^2}{m+n-2} &= \frac{10 \times 0.9^2 + 8 \times 1.1^2}{11+9-2} \\ &= 0.988 \end{aligned}$$

so that $s_p = \sqrt{0.988} = 0.994$.

Two sample hypothesis testing: Examples

Solution: Painkiller duration

Now

$$T_{obs} = \frac{\overline{x} - \overline{y}}{s_p\sqrt{\frac{1}{m} + \frac{1}{n}}} = \frac{5.4 - 4.8}{0.994\sqrt{\frac{1}{11} + \frac{1}{9}}} = 1.343.$$

so

$$\text{R. R. :} \quad \{T \geqslant t_{\alpha, n+m-2} = t_{0.01,18} = 2.552\}$$

Since

$$1.343 < 2.552$$

we do not reject H_0 at the 1% level.

Two sample hypothesis testing: Examples

Solution: Painkiller duration

Assuming H_0 is true, T has a t-distribution with 18 degrees of freedom. We observed

$$T_{obs} = 1.343$$

Because the test is one-sided, the p-value is given by:

$$p\text{-value} = P_0(T \geqslant T_{obs}) = P_0(T \geqslant 1.343) = 0.09796$$

Two sample hypothesis testing: Examples

Problem: Painkiller duration

Now suppose that 55 patients are given the new drug and 62 are given ibuprofen.

- New drug
  - sample mean $\overline{x} = 5.7$
  - sample standard deviation $s_1 = 1.1$
- ibuprofen
  - sample mean $\overline{y} = 4.6$
  - sample standard deviation $s_2 = 0.4$

Conduct an appropriate test at the 5% level to determine whether the new drug lasts longer than ibuprofen.

Two sample hypothesis testing: Examples

Solution: Painkiller duration

We are in Case 2. We are testing

$$H_0: \quad \mu_1 = \mu_2$$
$$H_a: \quad \mu_1 > \mu_2$$

with $\alpha = 0.05$. We have

$$T_{obs} = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}} = \frac{5.7 - 4.6}{\sqrt{\frac{1.1^2}{55} + \frac{0.4^2}{62}}} = 7.016.$$

$$\text{R. R.:} \quad \{T \geqslant z_{0.05} = 1.645\}$$

Since $7.016 \geqslant 1.645$, we reject H_0 in favour of H_a at the 5% level

Two sample hypothesis testing: Examples

Solution: Painkiller duration

Assuming H_0 is true, T has a Normal $(0, 1)$ distribution approximately. We observed

$$T_{obs} = 7.016$$

Because the test is one-sided, the p-value is given by:

$$p\text{-value} = P_0(T \geqslant T_{obs}) = P_0(T \geqslant 7.016) = 0.00000$$

to five decimal places.

Two sample hypothesis testing: Examples

Problem: Painkiller duration (paired samples)

Now suppose that 15 patients are given ibuprofen, then given the new drug at a later date:

- the sample mean difference is $\overline{d} = 1.1$ hours,
- the standard deviation is $s_d = 0.3$ hours

Conduct an appropriate test of hypotheses at the 10% level to determine whether the new drug lasts longer than ibuprofen. State any assumptions you make

Two sample hypothesis testing: Examples

Solution: Painkiller duration (paired samples)

We are in Case 3. We must assume that the differences in drug duration are Normally distributed.

We are testing

$$H_0: \quad \mu_1 = \mu_2$$
$$H_a: \quad \mu_1 > \mu_2$$

Two sample hypothesis testing: Examples

Solution: Painkiller duration (paired samples)

We have

$$T_{obs} = \frac{\overline{d}}{s_d/\sqrt{n}} = \frac{1.1}{0.3/\sqrt{15}} = 14.201.$$

and

$$\text{R. R.:} \quad \{T \geqslant t_{0.1,14} = 1.345\}$$

Since $14.201 > 1.345$ we reject H_0 in favour of H_a at the 10% level

Two sample hypothesis testing: Examples

Problem: Painkiller duration (paired samples, large n)

Suppose that 72 patients are given both drugs;

- the sample mean difference is $\overline{d} = 1.2$ hours,
- the standard deviation is $s_d = 0.4$ hours

Conduct an appropriate test of hypotheses at the 1% level to determine whether the new drug lasts longer than ibuprofen

Two sample hypothesis testing: Examples

Solution: Painkiller duration (paired samples, large n)

We are in Case 4. We must assume that the differences in drug duration are Normally distributed.

We are testing

$$H_0: \quad \mu_1 = \mu_2$$
$$H_a: \quad \mu_1 > \mu_2$$

Two sample hypothesis testing: Examples

Solution: Painkiller duration (paired samples, large n)

We have

$$T_{obs} = \frac{\overline{d}}{s_d/\sqrt{n}} = \frac{1.2}{0.4/\sqrt{72}} = 25.456$$

and

$$\text{R. R. :} \qquad \{T \geqslant z_{0.01} = 2.326\}$$

Since $25.456 > 2.326$ we reject H_0 in favour of H_a at the 1% level.

Two sample hypothesis testing: Examples

Problem: Student stress

Suppose that McGill students are being compared to Concordia students with respect to reporting stress.

- McGill
  - 112 students surveyed
  - 87 report high stress
- Concordia
  - 131 students surveyed
  - 92 report high stress

Conduct an appropriate test of hypotheses at the 10% level.

Two sample hypothesis testing: Examples

Solution: Student stress

We are in Case 5. We are testing

$$H_0: \quad p_1 = p_2$$
$$H_a: \quad p_1 \neq p_2$$

at the $\alpha = 0.1$ level. We have

$$\hat{p}_1 = 87/112 = 0.777$$
$$\hat{p}_2 = 92/131 = 0.702$$
$$\hat{p} = (87 + 92)/(112 + 131) = 179/243 = 0.737.$$

Two sample hypothesis testing: Examples

Solution: Student stress

We have

$$T_{obs} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{m}+\frac{1}{n}\right)}}$$

$$= \frac{0.777 - 0.702}{\sqrt{0.737(1-0.737)\left(\frac{1}{112}+\frac{1}{131}\right)}} = 1.314$$

and

$$\text{R. R.} \quad \left\{|T| \geqslant z_{\alpha/2} = z_{0.05} = 1.645\right\}$$

Since $|1.314| < 1.645$ we do not reject H_0 at the 10% level.

Two sample hypothesis testing: Examples

Problem: Jellyfish

Researchers are cataloguing two previously undiscovered species of deep-sea jellyfish, and are interested in assessing whether the lengths of the two species are different.

Two sample hypothesis testing: Examples

Problem: Jellyfish

- Species 1
  - sample size $m = 38$
  - sample mean length $\overline{x} = 21.4$ cm
  - sample standard deviation $s_1 = 3.9$ cm
- Species 2
  - sample size $n = 41$
  - sample mean $\overline{y} = 22.8$ cm
  - sample standard deviation $s_2 = 2.3$ cm

State a set of appropriate hypotheses, carry out the test at appropriate α levels, and calculate the corresponding p-value.

Two sample hypothesis testing: Examples

Solution: Jellyfish

We are in Case 2. We should test

$$H_0: \quad \mu_1 = \mu_2$$
$$H_a: \quad \mu_1 \neq \mu_2$$

We have

$$T_{obs} = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}} = \frac{21.4 - 22.8}{\sqrt{\frac{3.9^2}{38} + \frac{2.3^2}{41}}} = -1.924.$$

and note that $m, n > 30$.

Two sample hypothesis testing: Examples

Solution: Jellyfish

For $\alpha = 0.05$, in this two-sided test

$$\text{R. R.:} \quad \{|T| \geqslant z_{0.025} = 1.960\}$$

Since

$$|-1.924| < 1.960,$$

we do not reject H_0 in favour of H_a at the 5% level.

Two sample hypothesis testing: Examples

Solution: Jellyfish

For $\alpha = 0.1$, in this two-sided test

$$\text{R. R.:} \quad \{|T| \geqslant z_{0.05} = 1.645\}$$

Since

$$|-1.924| > 1.645,$$

we do reject H_0 in favour of H_a at the 10% level.

Two sample hypothesis testing: Examples

Solution: Jellyfish

Assuming H_0 is true, T has a Normal $(0, 1)$ distribution approximately. We observed

$$T_{obs} = -1.924$$

Because the test is two-sided, the p-value is given by:

$$\begin{aligned} p\text{-value} &= 2 \times P_0(T \geqslant |T_{obs}|) \\ &= 2 \times P_0(T \geqslant |-1.924|) \\ &= 2 \times 0.02716 \\ &= 0.05431 \end{aligned}$$

rounding to five decimal places.