

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 same 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

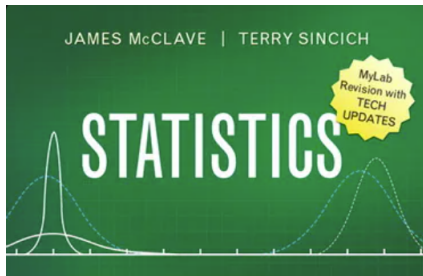
The course is broken down into four parts

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

Textbook

A useful¹ book for the course is:

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



We will be covering Chapters 1 to 9.

¹useful 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

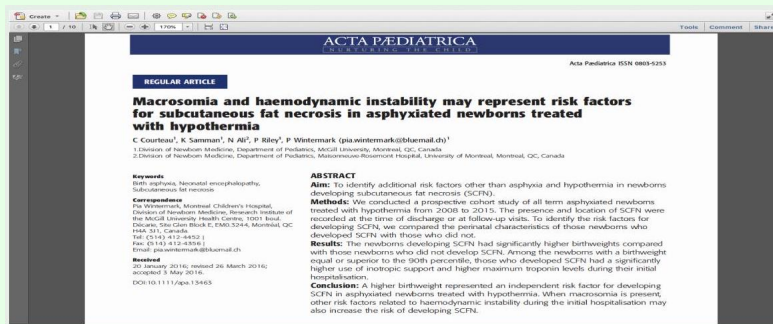
4. Characteristic(s) of interest: variable(s) measured on the experimental units
5. Sample: a subset of a population under study
 - If we could, we would look at the whole population (census); otherwise we need to take a sample

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 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)

Observational studies

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:

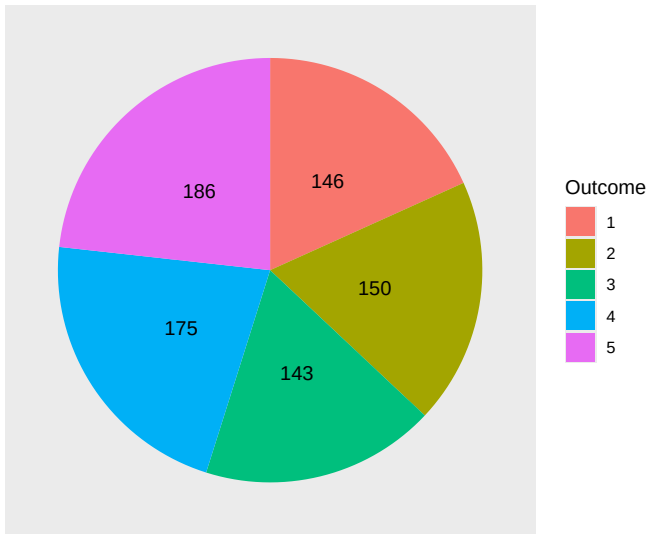
- There many types of graphical displays
- We will discuss the two that are the most commonly used

Methods for describing qualitative data

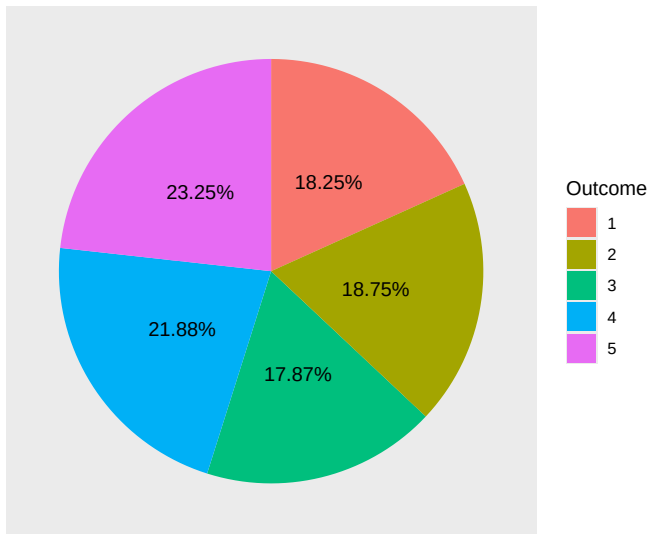
Pie chart

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

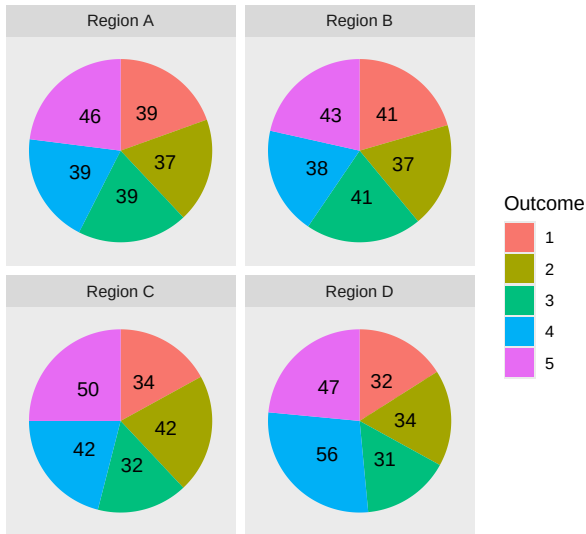
Methods for describing qualitative data



Methods for describing qualitative data



Methods for describing qualitative data

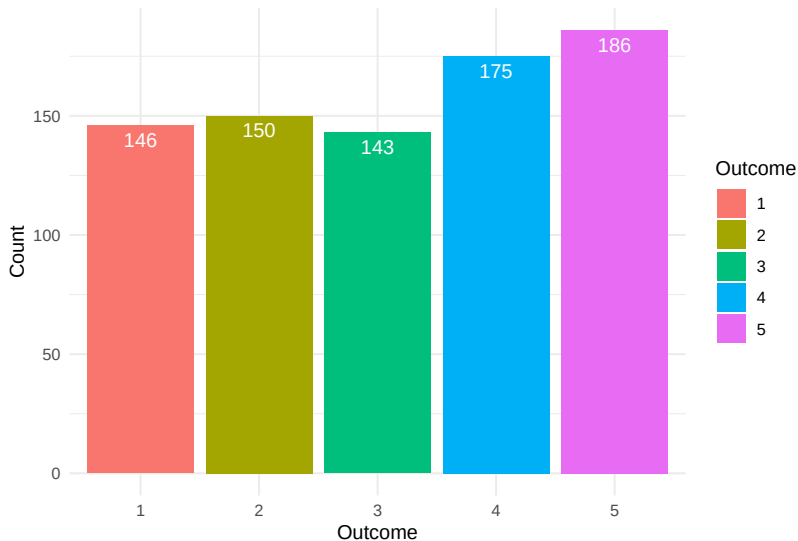


Methods for describing qualitative data

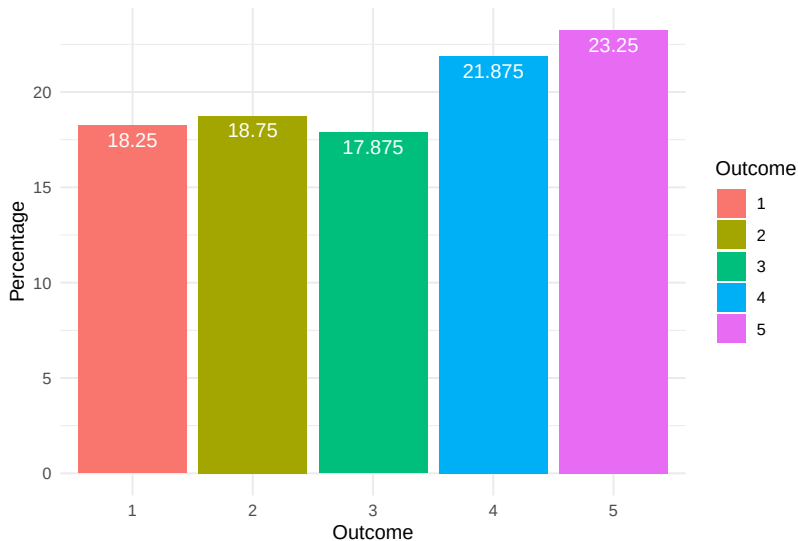
Bar plot

- 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

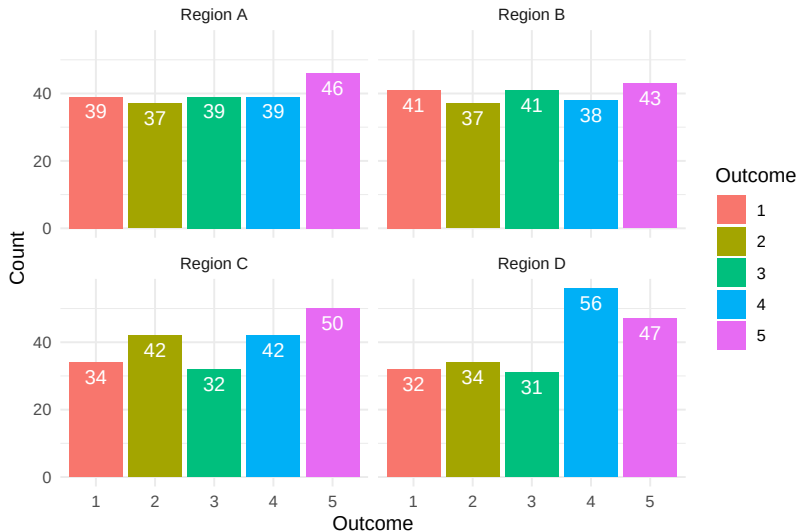
Methods for describing qualitative data



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

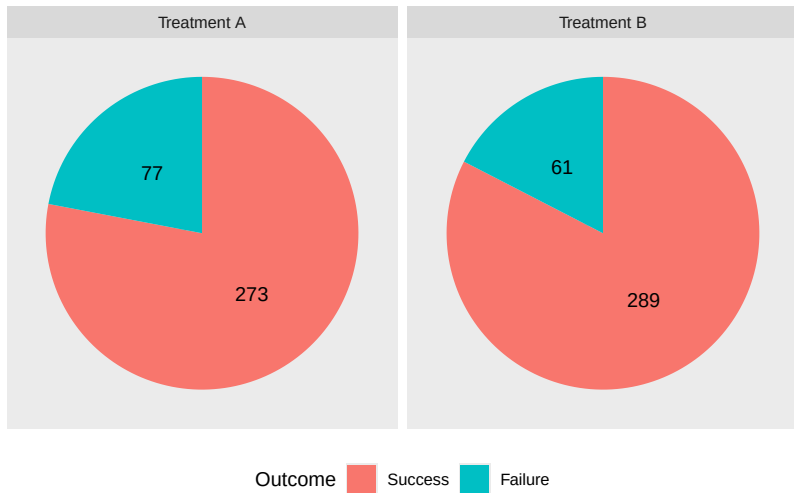
Example: Kidney stone treatment

```
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

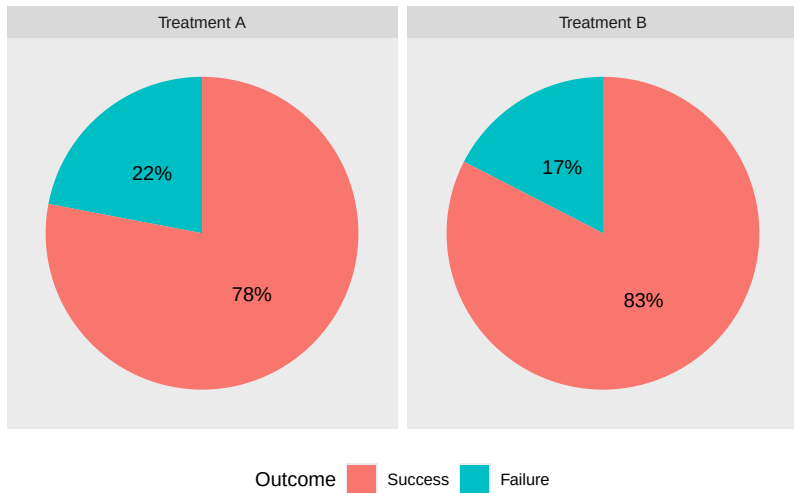
Example: Kidney stone treatment

```
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=labeler(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:

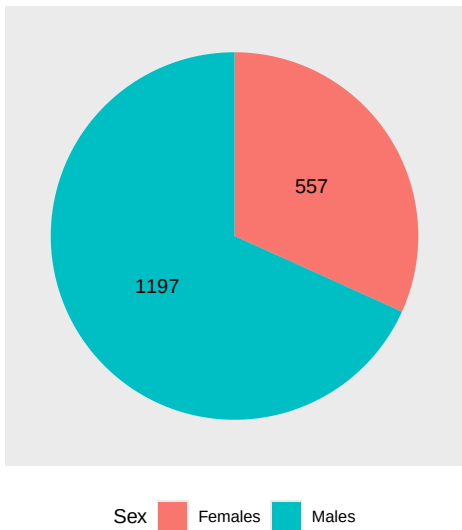
	Admitted
Males	1197
Females	557

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



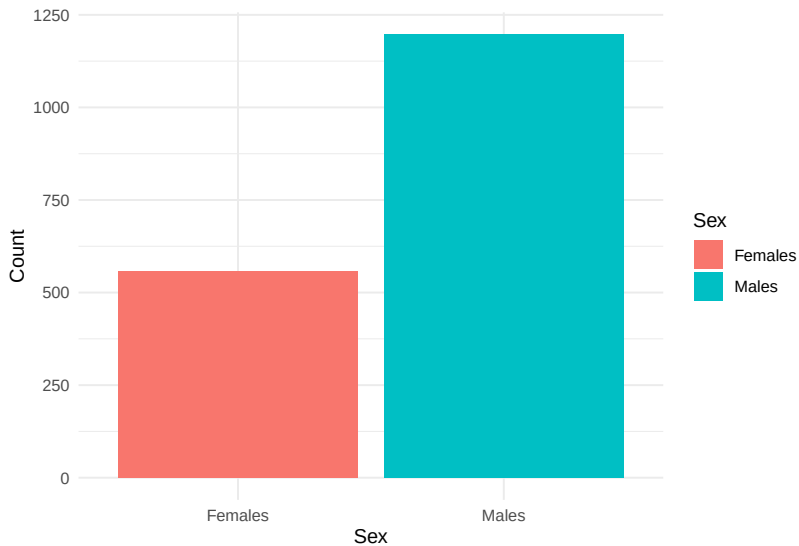
Simpson's Paradox

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:

Major	Males			Females		
	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
```

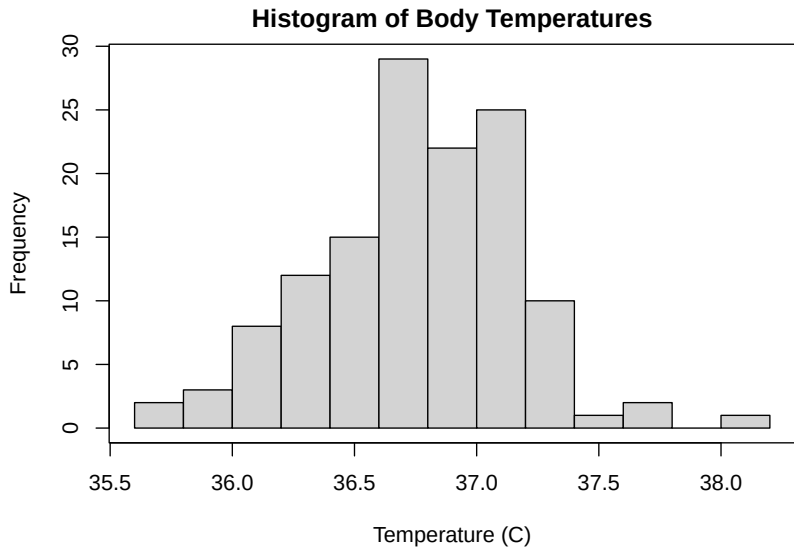
```
+ [1] 35.7 36.4 36.7 36.9 37.1 35.8 36.6 37.1 37.2 36.1 36.4 36.7 36.9 37.1 36.7  
+ [16] 36.8 37.1 37.2 36.2 36.4 36.7 37.2 36.3 36.7 36.9 37.1 37.3 36.2 36.8 37.0  
+ [31] 37.2 36.5 36.7 36.9 37.1 36.3 36.6 36.8 37.0 37.3 36.6 36.8 37.1 37.7 36.3  
+ [46] 36.6 36.8 37.0 37.4 36.8 37.0 37.1 38.2 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()
```

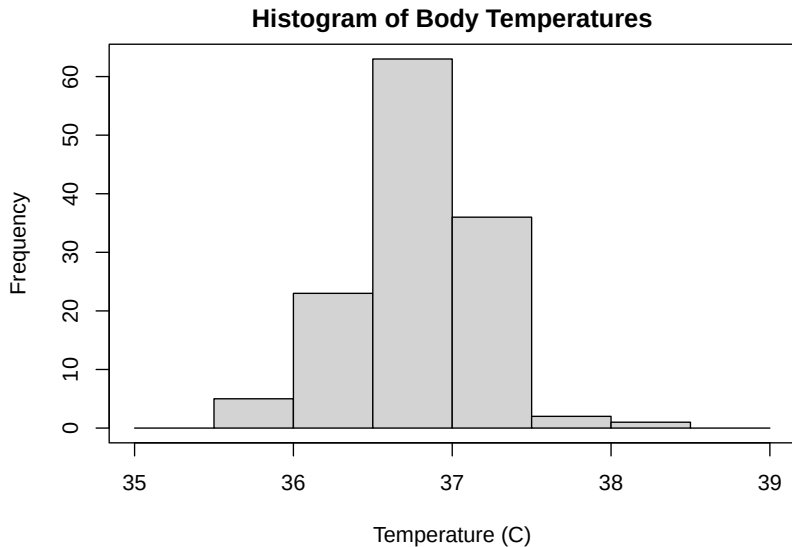
Histograms



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



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 $l = (x_L - x_S)/k$, 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

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

where the notation for intervals is that

$[a, b)$ includes a but **not** b

and

$[a, b]$ includes **both** a 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

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

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=1)
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

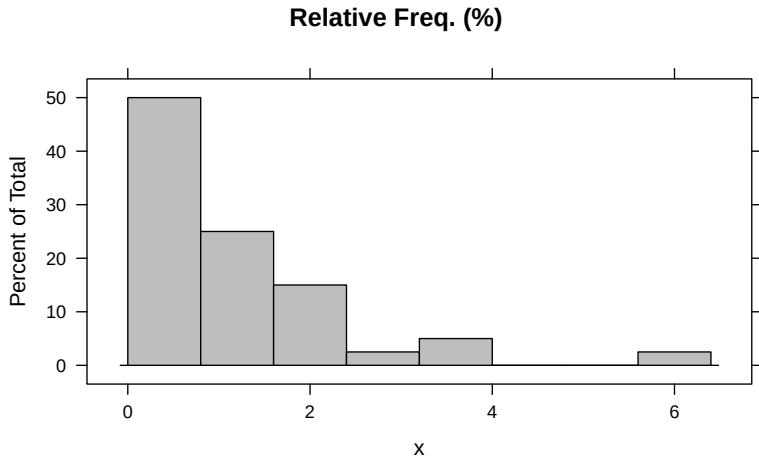
When inspecting or constructing a histogram:

- 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')
```

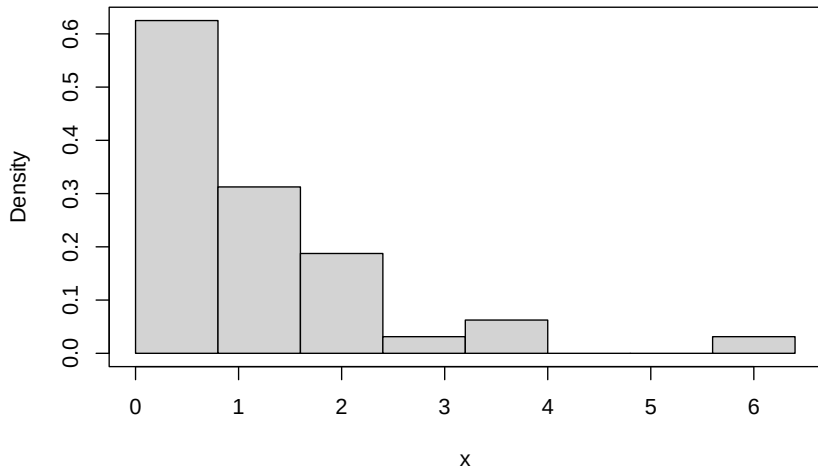
Histograms



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

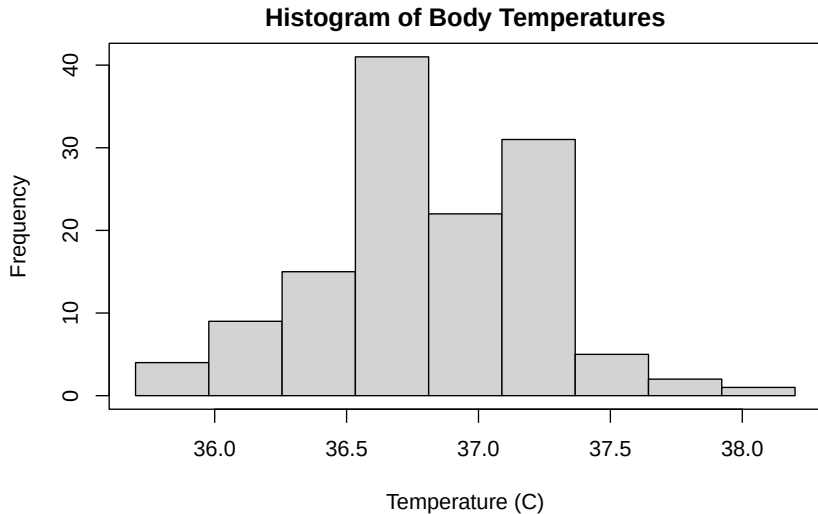
+ [1] 8
```

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



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
```

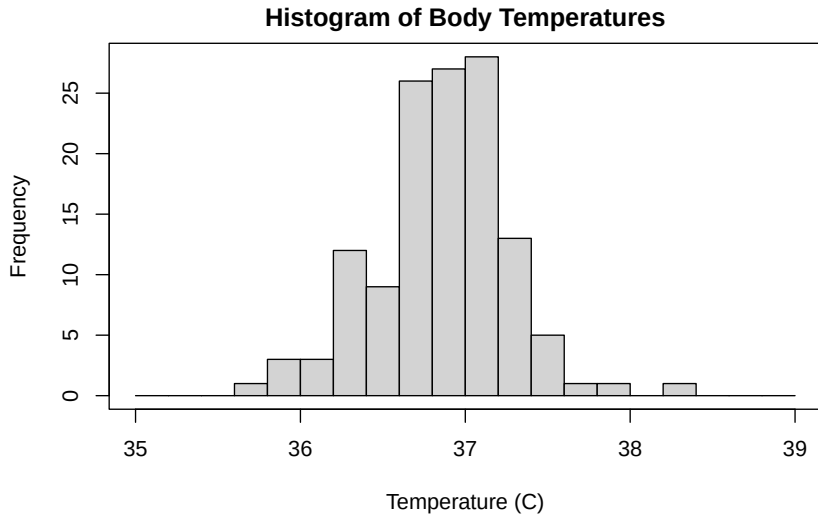
It is sometimes preferable to use more interpretable intervals.

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()
```

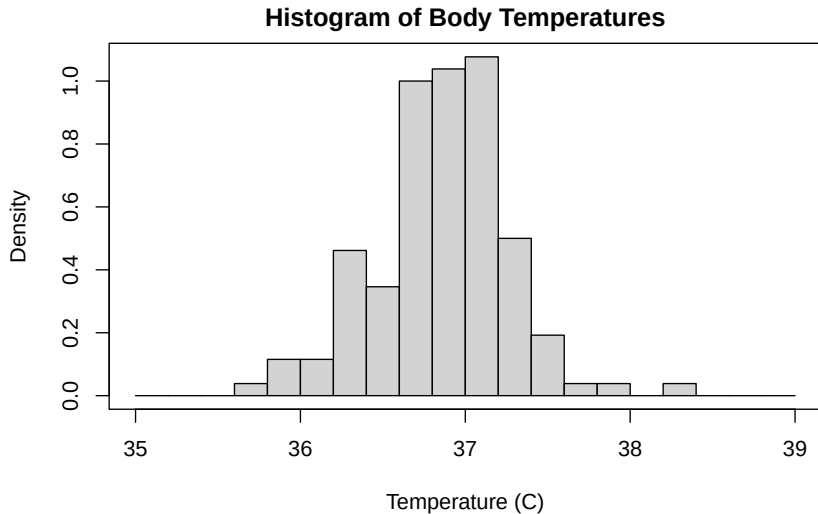
Histograms



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



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
```

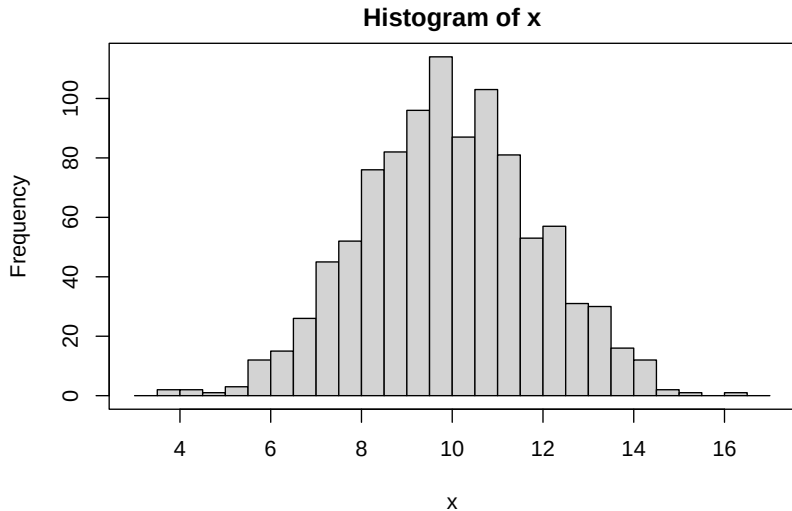
```
+ [1] 3.0 3.5 4.0 4.5 5.0 5.5 6.0 6.5 7.0 7.5 8.0 8.5 9.0 9.5 10.0  
+ [16] 10.5 11.0 11.5 12.0 12.5 13.0 13.5 14.0 14.5 15.0 15.5 16.0 16.5 17.0
```

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



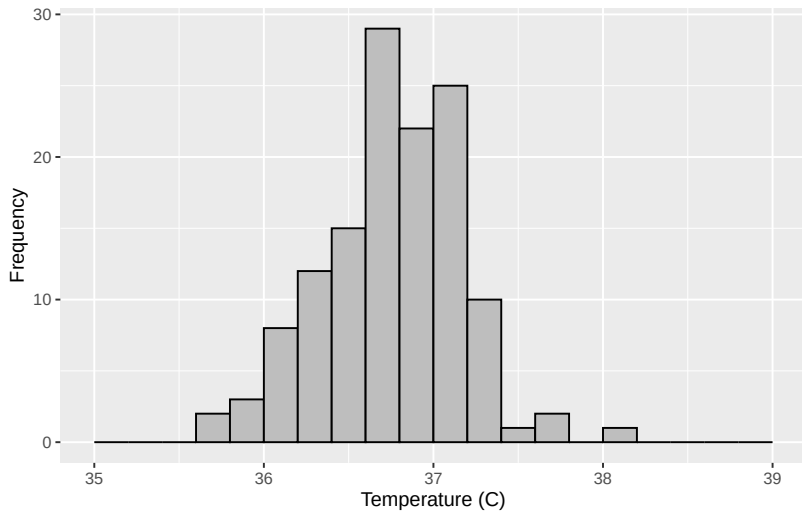
Histograms

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)')
```

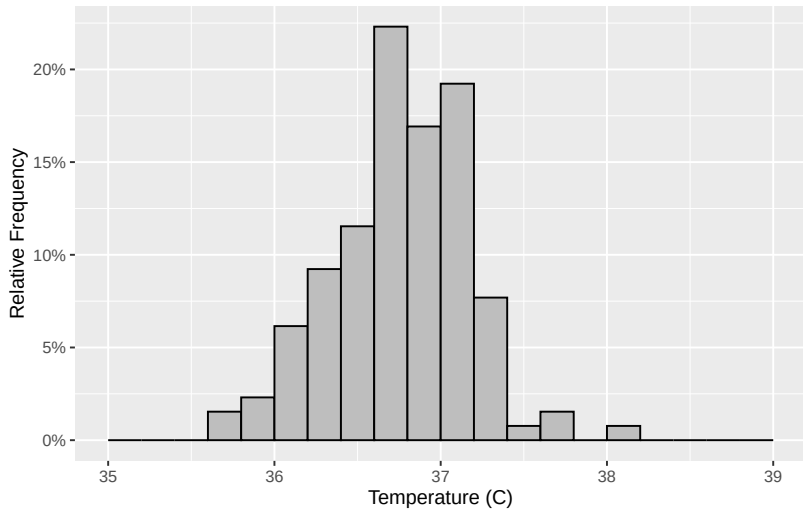
Histograms



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)')
```

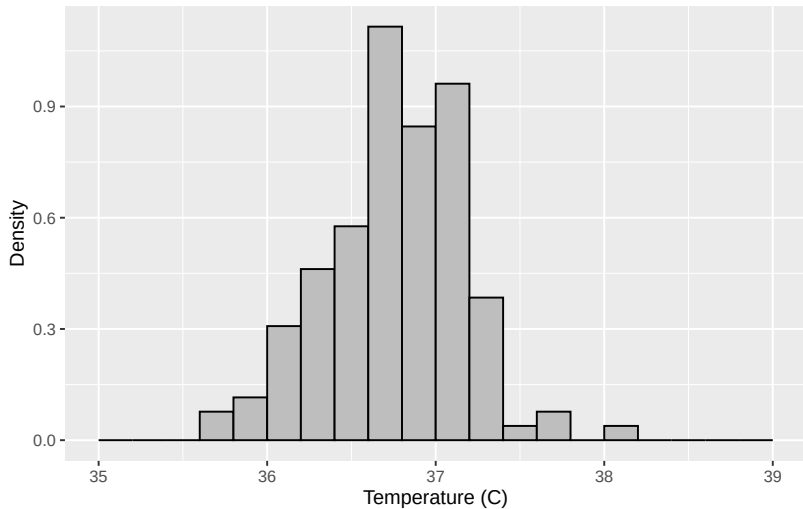
Histograms



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 $x_1, x_2, \dots, x_n$ be n data values.

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

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

- $\bar{x}$ 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

$$\begin{aligned}\bar{x} &= \frac{1}{n} \sum_{i=1}^n x_i = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5} \\ &= \frac{3 + 5 + 5 + 6 + 8}{5} = 5.4\end{aligned}$$

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

$$\begin{aligned}\bar{x} &= \frac{132.3 + 130.7 + 134.1 + 5.3 + 133.9 + 131.3 + 132.0 + 133.3}{8} \\ &= 116.6125\end{aligned}$$

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

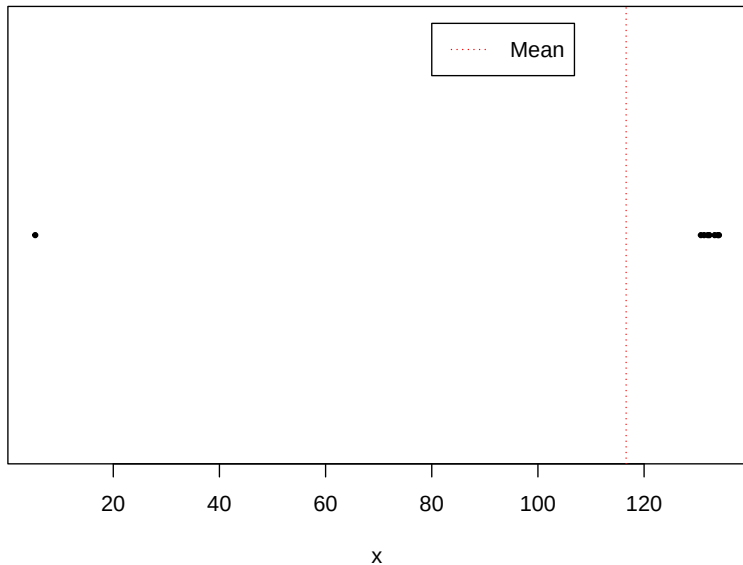
Computing in R:

```
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 $x_1, x_2, \dots, x_n$ 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 2 3 4 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 4 5 6 7 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 2 5 8 9

The median is $m = 5$

Sample median

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

1 2 5 8 900

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)

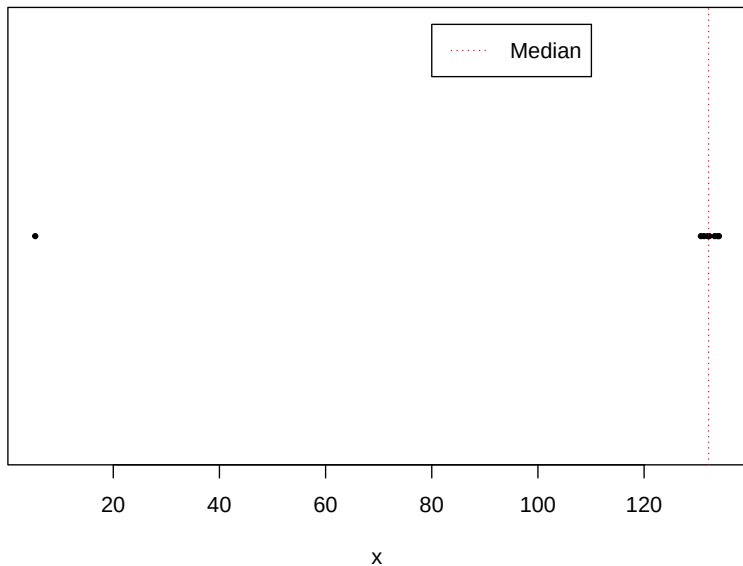
+ [1] 5.3 130.7 131.3 132.0 132.3 133.3 133.9 134.1

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 $x_1, x_2, \dots, x_n$ 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

$$R = x_L - x_S,$$

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 $x_1, x_2, \dots, x_n$ and their sample mean $\bar{x}$

- Distances from the sample mean

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

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:

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

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

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

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

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

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

Sample variance

Note the following: squaring out the bracket

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

Sample variance

we have

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

Sample variance

Therefore

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

Sample variance

Therefore

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

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 $\text{mean} = \text{median}$.

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 $\text{mean} \approx \text{median}$

Symmetry and Skew

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

- $\text{median} < \text{mean}$

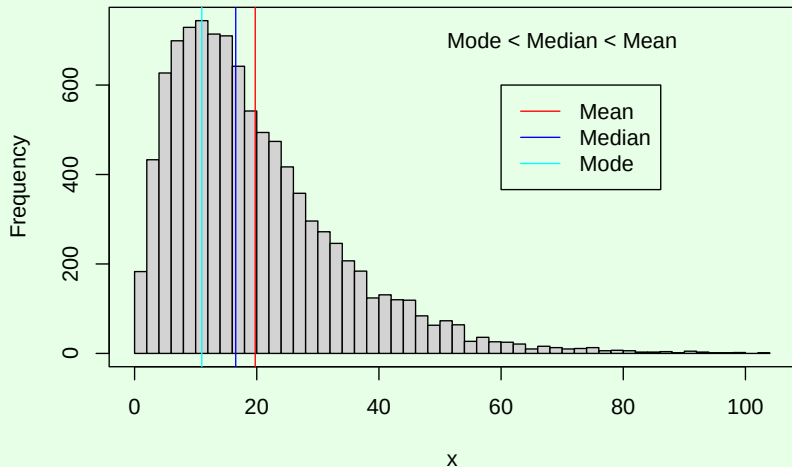
Negatively (left) skewed data: histogram shows a “long tail” on the left.

- $\text{median} > \text{mean}$

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

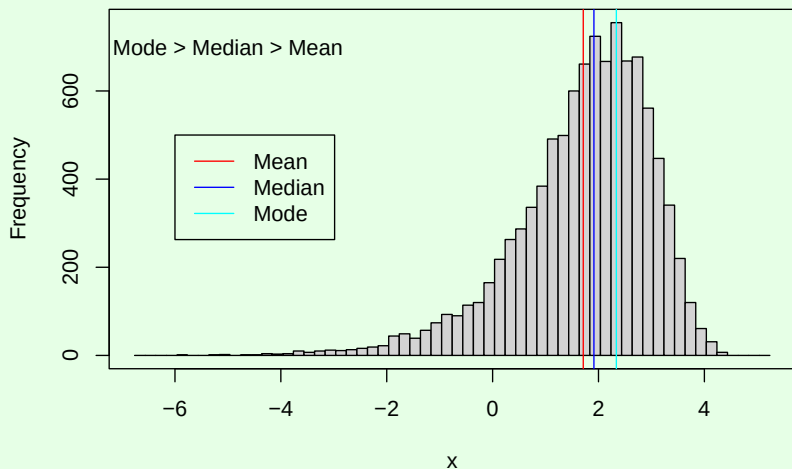
Symmetry and Skew

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

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

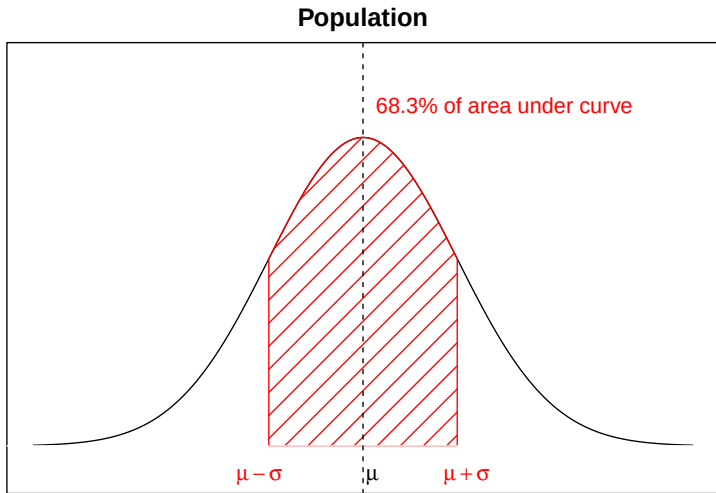
- Roughly 95% of the observations will lie in the range

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

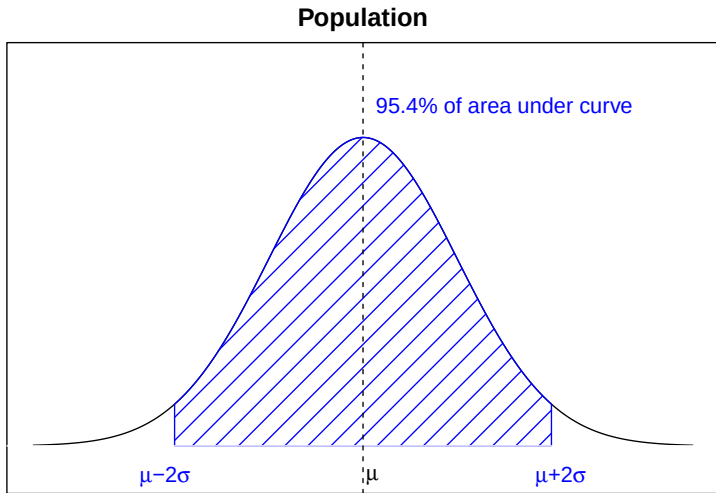
- Roughly 99.7% of the observations will lie in the range

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

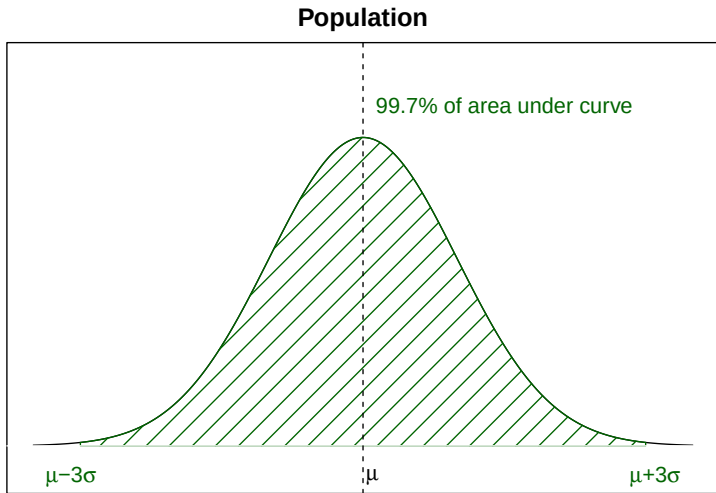
Empirical rule



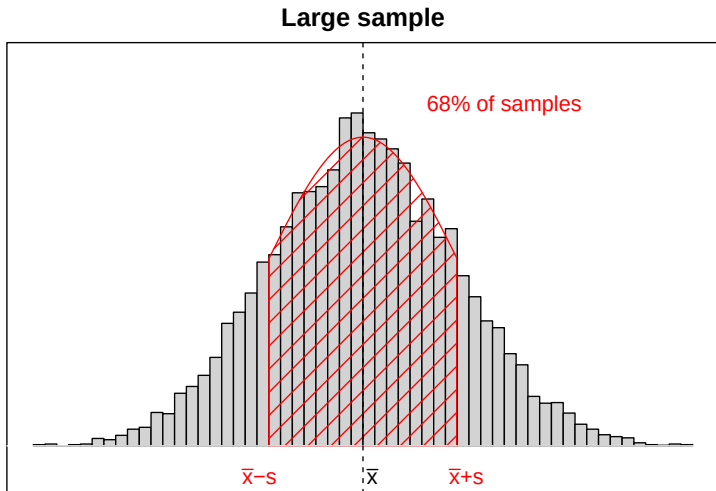
Empirical rule



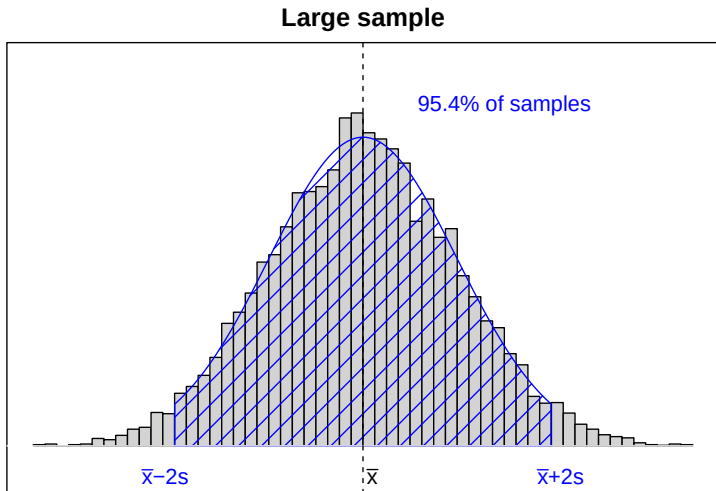
Empirical rule



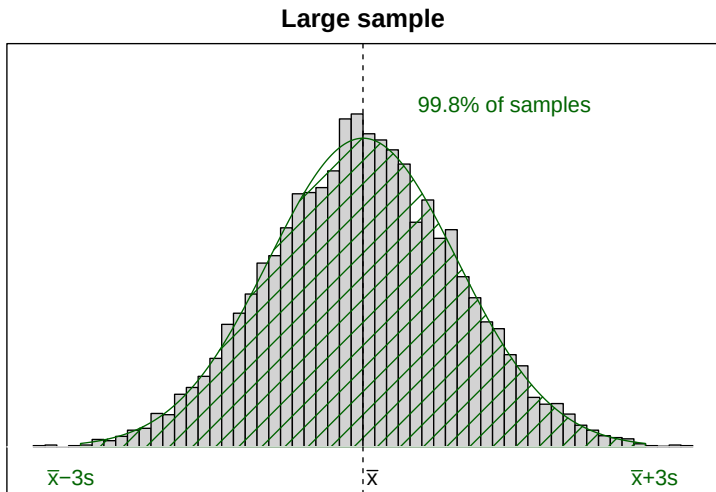
Empirical rule



Empirical rule



Empirical rule



Empirical rule

Example: Body Temperatures

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

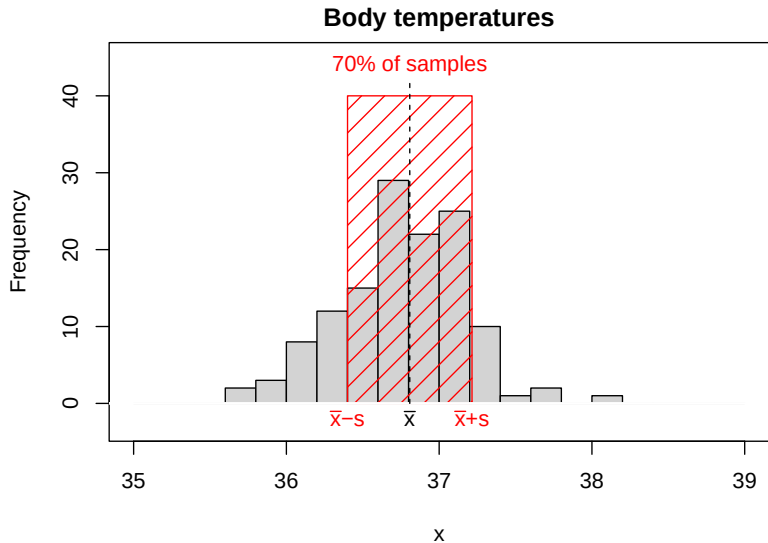
$$(\bar{x} - s, \bar{x} + s) = (36.40, 37.22)$$

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

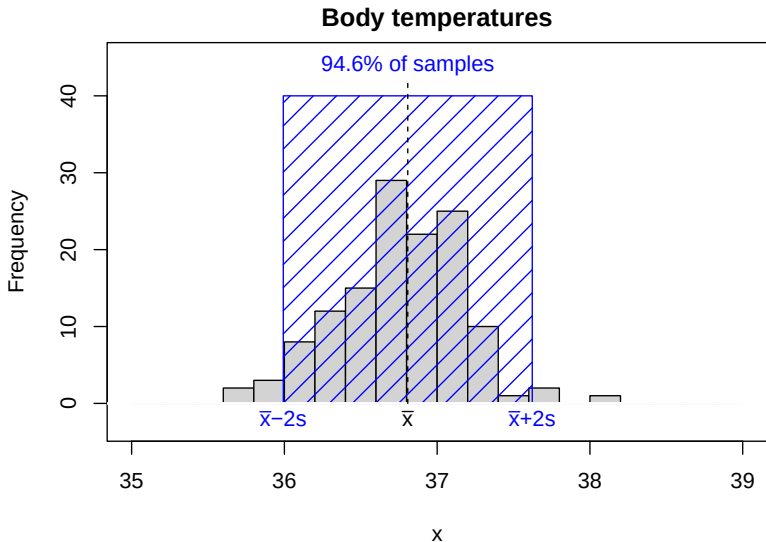
$$(\bar{x} - 3s, \bar{x} + 3s) = (35.59, 38.03)$$

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%

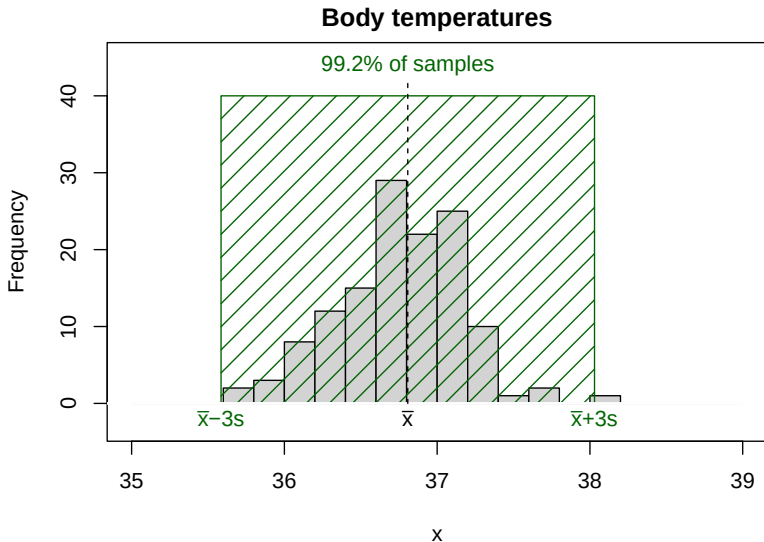
Empirical rule



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

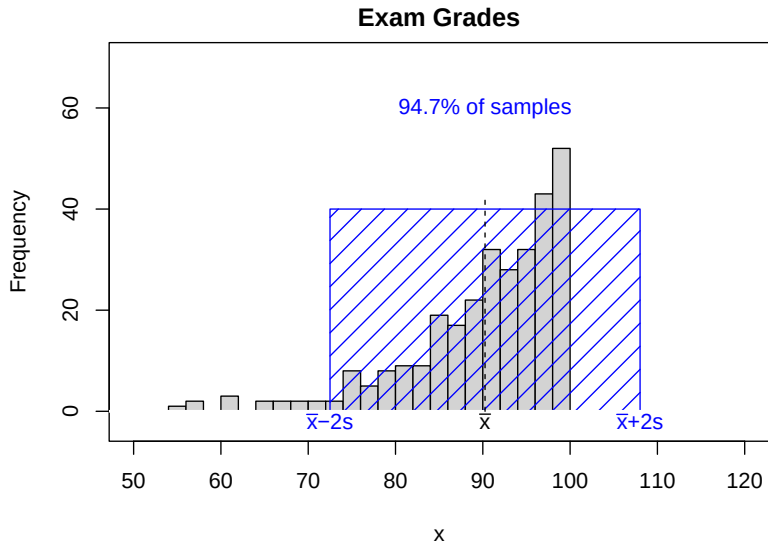
Empirical rule

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



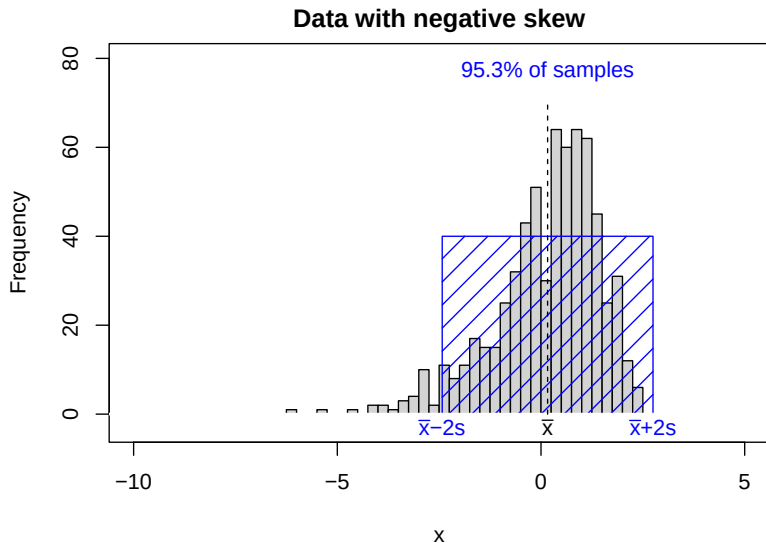
Empirical rule

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



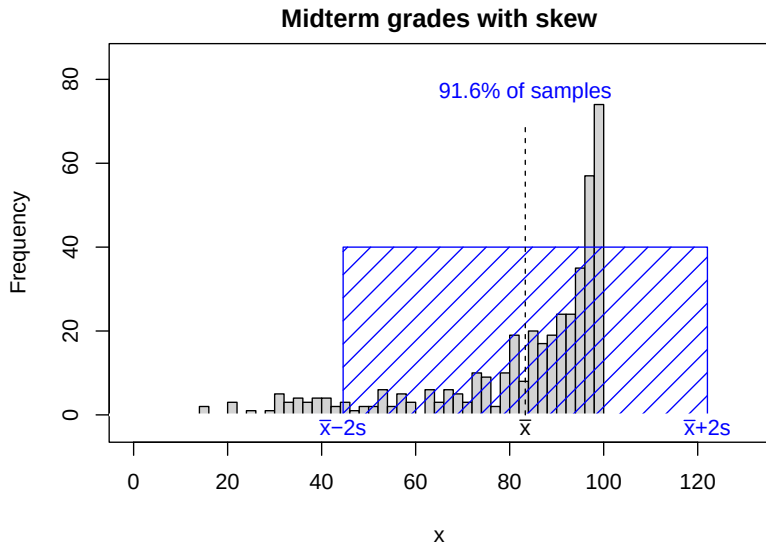
Empirical rule

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

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

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

Chebyshev's rule

Example: Body temperatures

For the sample of $n = 130$ body temperatures

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

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

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

We could also define this quantity for a population

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

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

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,

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

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

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

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

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

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

Graphical methods: Boxplots

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

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

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])
```

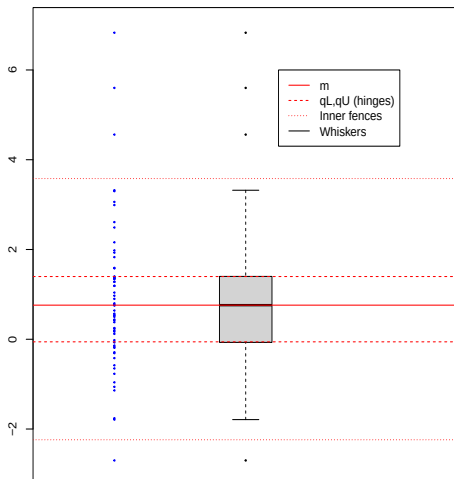
```
c(m,innerL,innerU,outerL,outerU,whiskL,whiskU)
```

```
+      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)

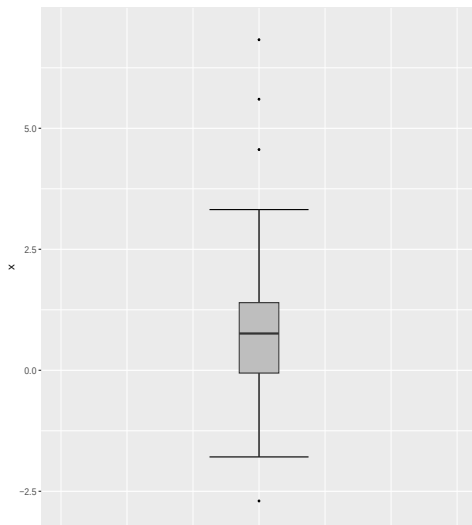
+ $stats
+      [,1]
+ [1,] -1.79
+ [2,] -0.07
+ [3,]  0.76
+ [4,]  1.40
+ [5,]  3.32
+
+ $n
+ [1] 62
+
+ $conf
+      [,1]
+ [1,] 0.4650295
+ [2,] 1.0549705
+
+ $out
+ [1] -2.70  6.83  5.60  4.56
+
+ $group
+ [1] 1 1 1 1
+
+ $names
+ [1] "1"
```

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

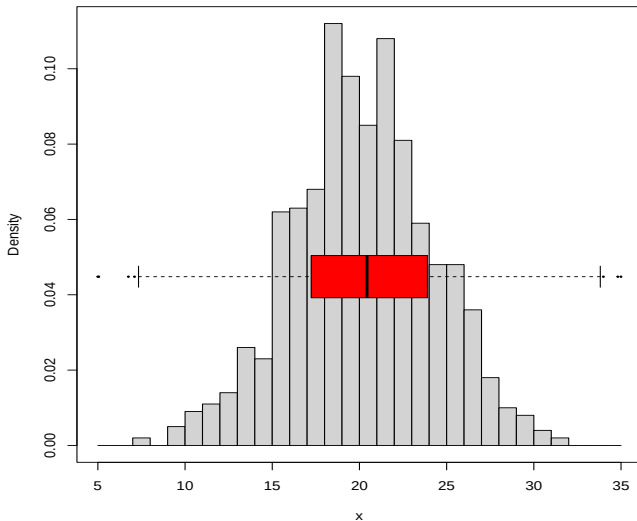
Indication of skewed data

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

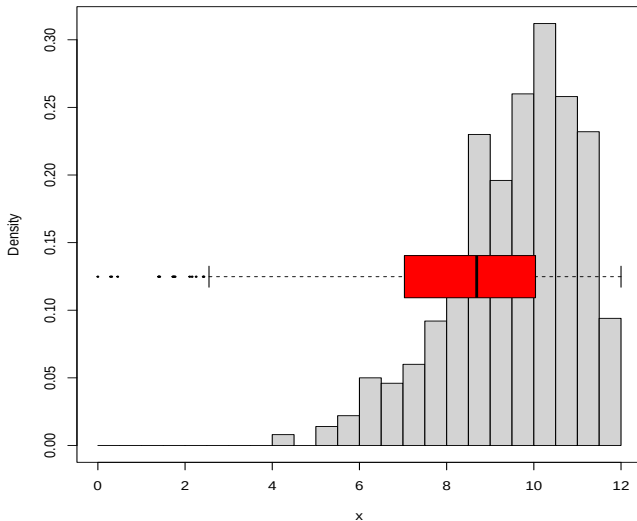
Boxplots vs Histograms

```
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')
```

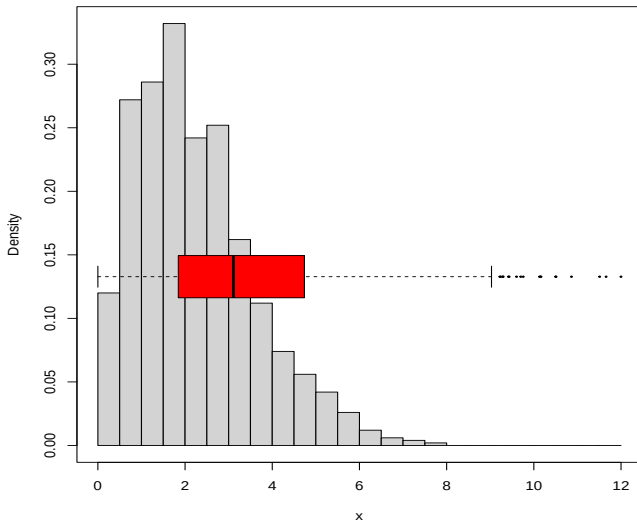
Boxplots vs Histograms



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
```

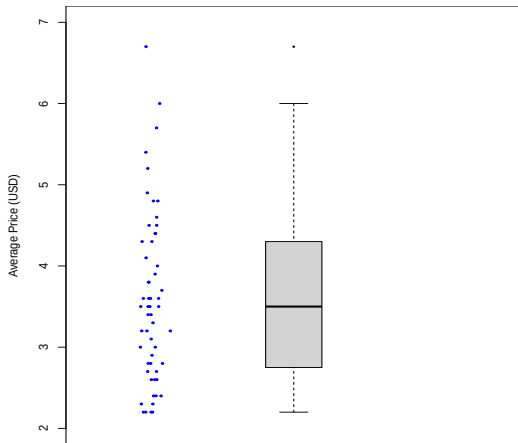
```
+ [1] 56
```

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

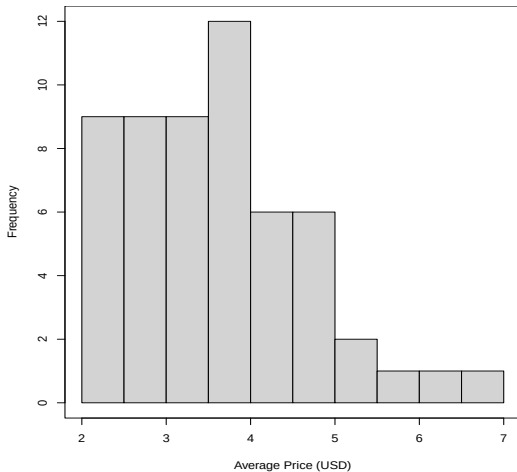
```
+   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
+ 2.200  2.775   3.500   3.573  4.300   6.700
```

More data [here](#)

Boxplots vs Histograms



Boxplots vs Histograms

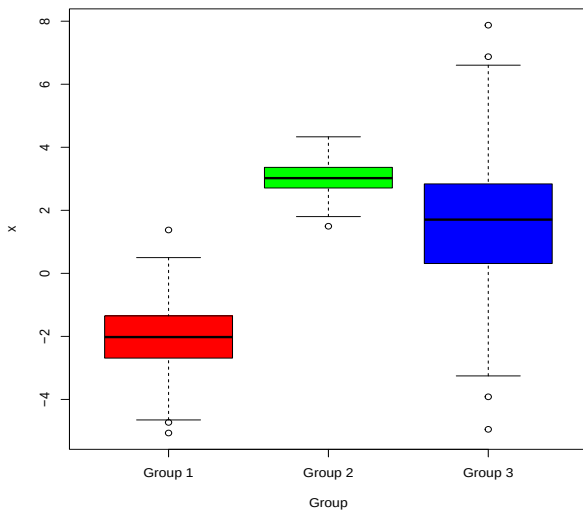


Comparing Distributions

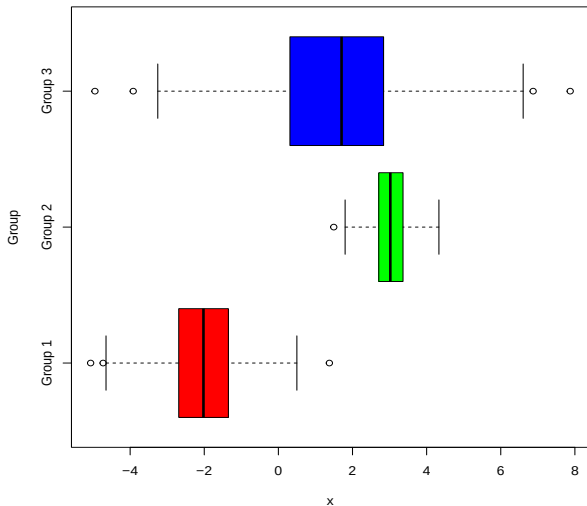
We can plot several boxplots for comparison of groups: we may examine

- 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]))

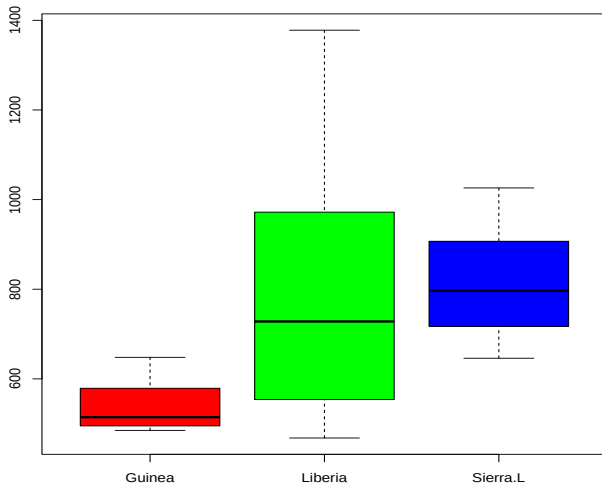
+ [1] 785.90000 288.9162

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

+ [1] 806.80000 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:

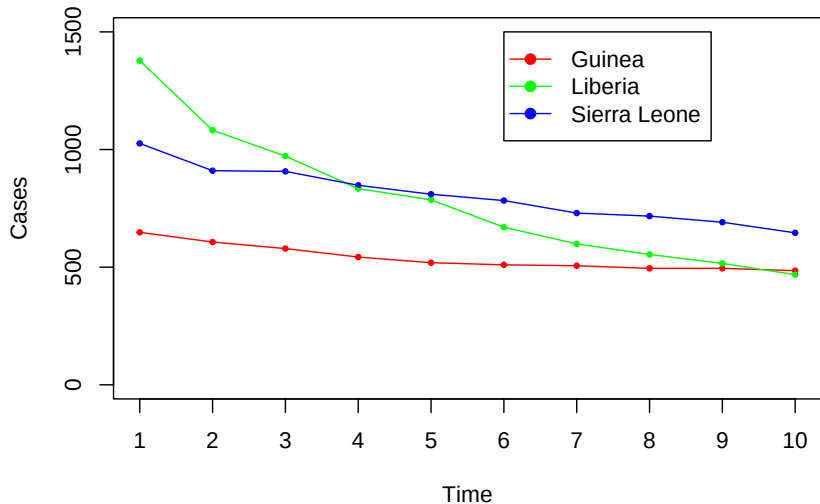
- In this infectious disease setting, we might expect there to be systematic changes across time;
- For each country, the observation at week 1 is not truly comparable with the observation at week 10.

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

