

# STT 231: August 28, 2019

- Four ways to the semester on the right foot!
  1. Purchase course pack by next class meeting
  2. Enroll in course Top Hat: JOIN CODE **830383**
  3. Bring your laptop to 1<sup>st</sup> recitation, next Tuesday
  4. Ensure D2L notifications are turned on for course
  
- Agenda:
  - Introductions, course overview, begin Chapter 1

# Introductions

■ Instructor: John Keane, [keanejoh@msu.edu](mailto:keanejoh@msu.edu)

1. Where are you from?
2. What do you do on-campus? What are you studying / working on?
3. What was in your high school locker?
4. Would you rather fight 100 duck-sized horses or one horse-sized duck?
5. What is the average class size at MSU?

# Why study statistics?

Consider a simple example:

- 150 students at a university are taking a statistics course.
- There are five sections of the course, with class sizes shown below:

Section	A	B	C	D	E
Class size	15	15	15	15	90

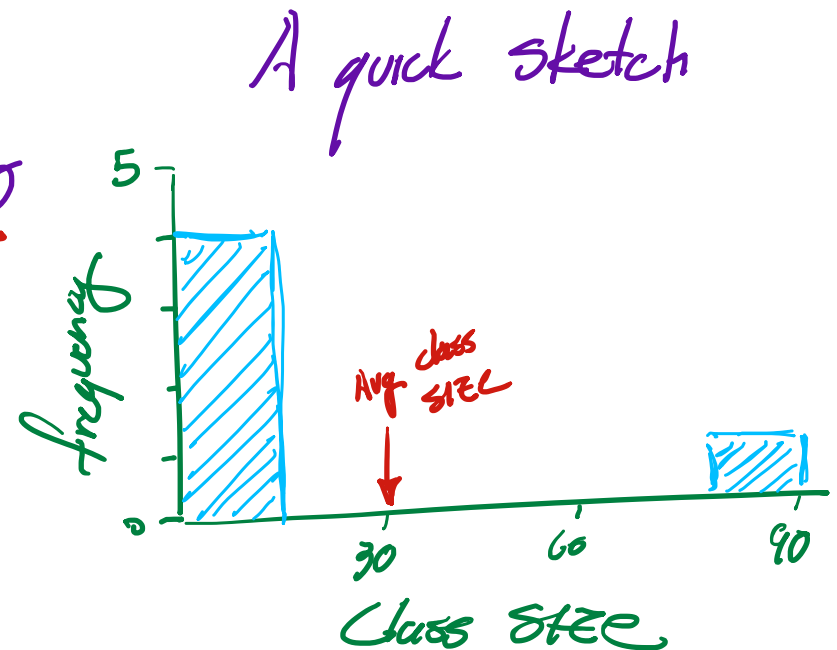
- What is the average class size, per instructor, across the five sections?

# Why study statistics?

Section	A	B	C	D	E
Class size	15	15	15	15	90

- What is the average class size, per instructor, across the five sections?

$$\text{Avg class size} = \frac{15 + 15 + 15 + 15 + 90}{5} = \underline{30}$$



# Why study statistics?

Section	A	B	C	D	E
Class size	15	15	15	15	90

= 150  
students  
in total

- What is the average class size, per student, across the five sections?

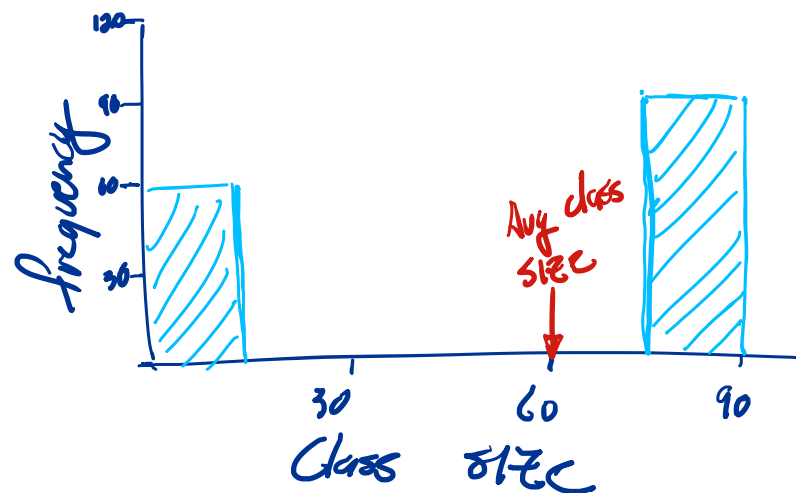
A quick sketch

15 + 15 + 15 + 15 = 60 students

have class size 15

90 students have  
class size 90!

$$\text{Avg class size} = \frac{60(15) + 90(90)}{150} = 60!$$



# Why study statistics?

- Comment on how these two average class sizes compare and explain why they differ as they do.
- Which average would you say is more relevant?

*Aside: many colleges report avg class size per section, not per student.*

# Why study statistics?

- So again... why study statistics?
- Too frequently, we engage only in discussions involving typical values...

...when we should also consider *variability* and *distribution* of cases.

# Course overview

- Instructor: John Keane, [keanejoh@msu.edu](mailto:keanejoh@msu.edu)
- Lecture: MW 9:10AM-10:00AM, G008 Golden Hall
- Recitation: Thursday afternoons (times vary)
- Office hours: Mondays 11:00AM – 12:00PM or by appointment, Wells Hall C427
- **Description and prerequisites**
- Calculus-based course in probability and statistics. Probability models and random variables. Estimation, confidence intervals, tests of hypotheses, and simple linear regression with applications in sciences.
- ***Prerequisites:*** MTH 124 or MTH 132 or MTH 152H or LB 118

# Required materials

- **Required – WebWork subscription:** Used for homework assignments. Students will be required to purchase a subscription via CashNet at a cost of \$50.00.
- **Required – Course pack:** STT 231 uses an interactive course pack in lieu of a hardcover textbook that includes reading assignments and in-class activities.
- **Required – Graphing-Calculator:** Any basic graphing calculator (has basic STATS functions) is fine, although student support is offered for Texas Instrument models (for example: 83, 84, 89, Inspire). No cell phones (or other devices with access to data plans) allowed during exams.

*Pay window:  
9/18 - 9/25*

# Required materials

- **Required – Top Hat account:** The Top Hat Response System will be used regularly during and between lectures as a method of gauging how students are understanding content.
  - **JOIN CODE: 830383**
- **Required – R and RStudio:** This course uses a free statistical programming language ('R') and a convenient interface ('RStudio') as an aid in exploring statistical concepts and practicing various procedures.

# Grading

	WebWork	In-class participation	Recitation activities	At-home Readings	Exams
%	20	5	10	10	55

Grade	Range
4.0	100% - 90%
3.5	85% - 89.99%
3.0	80% - 84.99%
2.5	75% - 79.99%

Grade	Range
2.0	70% - 74.99%
1.5	65% - 69.99%
1.0	60% - 64.99%
0	0% - 59.99%

# Student evaluation

## ■ In-class participation: 5%

- Top Hat questions are frequently posed in-lecture. Your answers to these questions are evaluated evenly on completion and accuracy and are worth a total of 5% of your grade.
- Make-up questions offered weekly.

## ■ Recitation: 10%

- Thursday recitations are required. Submitted activities worth 10% of your grade.
- Lowest recitation grade dropped at end of semester.

# Student evaluation

## ■ WebWork: 20%

- WebWork exercises most weeks (total of 11). Extensions / make-ups are not offered, so make sure you start them early!
- Lowest HW score dropped at the end of semester.

## ■ At-home readings: 10%

- In between lectures, short readings are assigned to prepare you for the following lecture. Readings often require short responses via Top Hat.

## ■ Exams: 55%

- Three, non-cumulative exams, each worth ~18.33% of your final grade.

# Exam Dates and Times

*\* Heavily emphasizes  
final third of course*

- There will be two midterm exams and one ~~\*~~ comprehensive ~~\*~~ final exam on the following dates and times:
  - Exam 1: Wednesday, October 9, 2019
    - Time: 7:00 – 8:20pm
  - Exam 2: Wednesday, November 20, 2019
    - Time: 7:00 – 8:20pm
  - Exam 3: Thursday, December 12, 2019
    - Time: 7:45 – 9:45am
- Locations will be announced in the week before each exam and posted on D2L

# Exam policies

- Closed book
- Formula sheet will be provided
- Calculators are allowed/necessary
- Make-up Exam
  - Available the day after the exam, time and place TBD.
  - Must provide documentation to the instructor no later than 1 week before the exam date.
  - Valid reasons to take the make-up include:
    - medical emergency
    - university sanctioned event (including class at that time)
    - religious holidays
    - military obligation

# Where to get help

- STT help room
  - A102 Wells Hall
  - Hours are posted on D2L
- Email/visit your TA
  - Contact info and office hours posted on D2L
- Email/visit the instructor
  - Contact info and office hours posted on D2L



# STT 231

# STATISTICAL METHODS

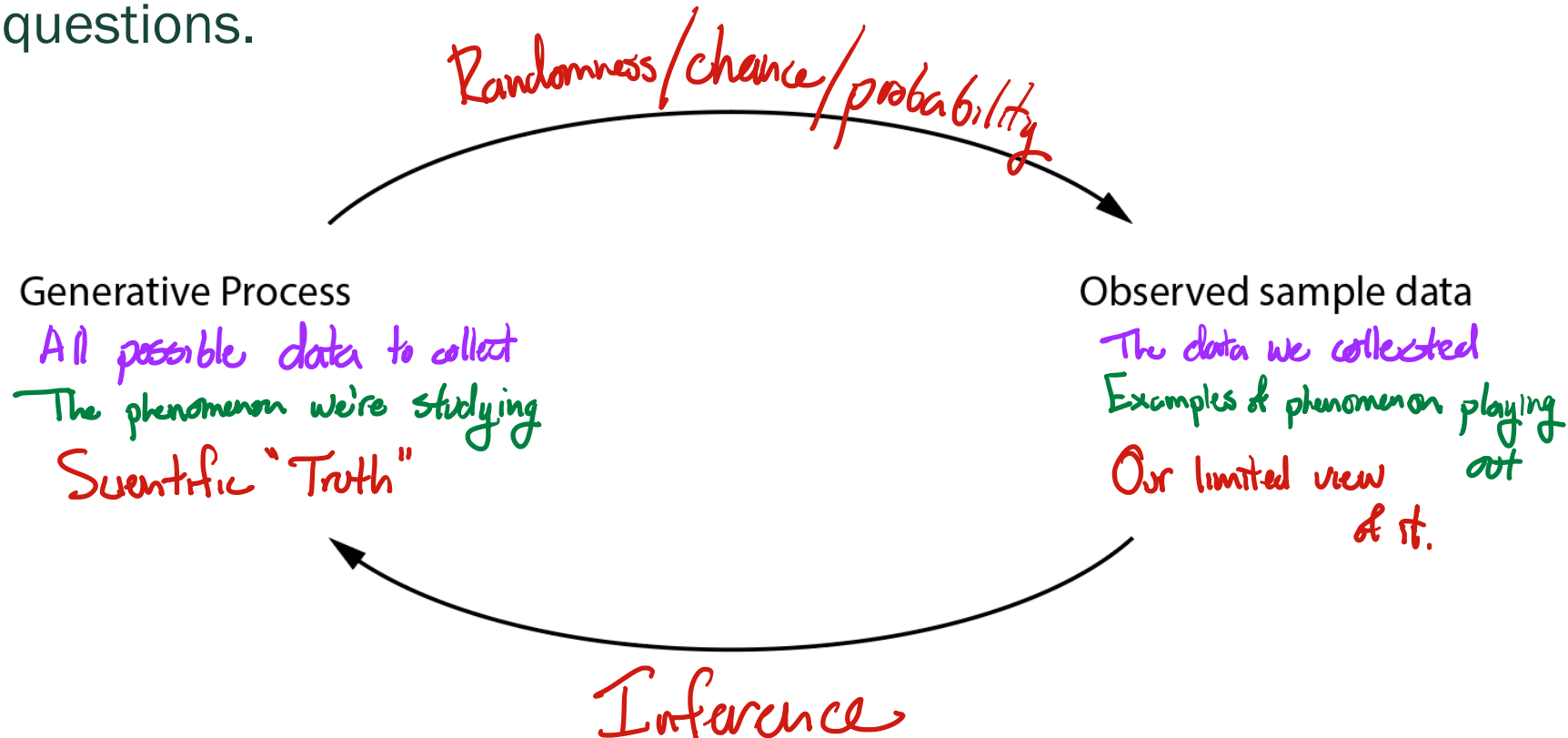
Chapter 1: Introduction to Data

# Lecture 1-1: Variable Types

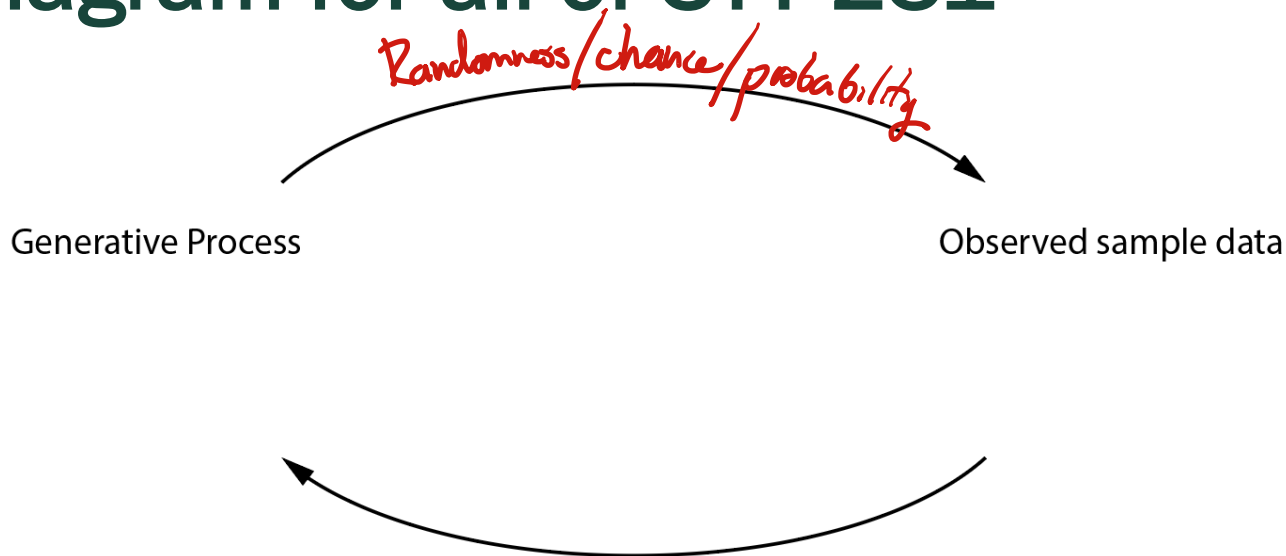
- Consider each of the following research questions:
  1. Do students who self-identify as 'night owls' have higher GPAs than those who consider themselves 'morning larks'?
  2. Is swimming with dolphins therapeutic for patients suffering from clinical depression?
  3. Is gestation length associated with life expectancy among mammalian animal species?
  4. Does exposure to light pollution at night influence weight gain?

# A diagram for all of STT 231

We might conduct studies to attempt to answer these questions.



# A diagram for all of STT 231



Scientists are often looking at

*observed samples* of data in order to reach conclusions that extend to some larger or generative process. We'll call that process the

*population of interest*.

# Populations & samples

Definition of a population:

The **generative process** is sometimes concrete and other times only an abstraction, but it is object of study when we are collecting and analyzing data. The **sample data** is our way of finding out, with some level of (un)certainty, about the facts of the process that produced the sample data.

# Cases and variables

Sample data is made up of Cases ;

a Variable is any characteristic recorded for each case.

All variables are either *categorical* or *quantitative*.

- Categorical variables classify cases into groups/categories, placing each case into exactly one of two or more categories.
- Quantitative variables measure/record a numerical quantity for each case. Unlike categorical variables, numerical operations like adding and averaging make sense for quantitative variables.

# Quantitative variables

Quantitative variables can be divided further into two groups:

discrete and continuous

- Quantitative, **discrete** variables can only take on certain values within a given interval.
- On the other hand, quantitative, **continuous** variables can take on any value within an interval.

# Categorical variables

Categorical variables can be divided further into two groups:

ordinal and non-ordinal

## Example 1.1: Owls vs. Larks

A recent study examines the relationship between class start times, sleep, circadian preference, alcohol use, academic performance, and other variables in college students.

The data were obtained from a sample of  $n = 253$  students who completed skills tests to measure cognitive function, completed a survey that asked many questions about attitudes and habits, and kept a sleep diary to record time and quality of sleep over a two-week period.

# Example 1.1: Owls vs. Larks

Below are some of the recorded variables.

Variable	Coding	Type
<i>Gender</i>	1 = male, 0 = female	categorical, non-ordinal
<i>ClassYear</i>	Year in school, 1 = first year, ..., 4 = senior	categorical, ordinal
<i>LarkOwl</i>	Early riser or night owl? Lark, Neither, or Owl	categorical, non-ordinal
<i>ClassesMissed</i>	Number of classes missed in a semester	quantitative, discrete
<i>AnxietyScore</i>	Measure of amount of anxiety	Depends on how variable was encoded
<i>AnxietyStatus</i>	Coded anxiety score: normal, moderate or severe	categorical, ordinal

## Example 1.2: Identifying variable types

- Consider each of the following variables and classify them by their type.

1. The zip code in which you were born. *categorical, non-ordinal*
2. The score you earn on a test, as a letter grade. *ordinal*
3. The score you earn on a test, as a percentage of total available points. *quantitative, continuous (why not discrete?)*
4. Your handedness, i.e., your tendency to use either the right or left hand more naturally than the other.

*Depends on how you encode it!*

## Example 1.3: Looking more closely at handedness

- a. Before observing any sample data, what proportion of students do you think are right-handed?
- b. Submit via Top Hat whether you identify as right- or left-handed. How many cases are there in our observed sample? What type of variable have we recorded?

→ Categorical, non-ordinal



## Example 1.3: Looking more closely at handedness

c. Please indicate which hand you use for each of the following activities by putting a (+) in the appropriate column...

...or (++) if you would never use the other hand for that activity.

If in any case you are really indifferent, put (+) in both columns.

Task	Left	Right
Writing	++	
Drawing	+	
Throwing	+	
Scissors	+	+
Toothbrush	+	
Knife (without fork)	+	+
Spoon	+	+
Broom (upper hand)	+	+
Striking match (hand that holds the match)	++	
Opening box (hand that holds the lid)	+	+
Total	12	5

## Example 1.3: Looking more closely at handedness

d. Create a Left and a Right score by counting the total number of (+) signs in each column.

Your handedness score is

$$\frac{Right - Left}{Right + Left}$$

Submit your Handedness score via Top Hat.

What do you expect the distribution of responses to look like?

$$\frac{5 - 12}{5 + 12} = -0.4118$$

Task	Left	Right
Writing	++	
Drawing	+	
Throwing	+	
Scissors	+	+
Toothbrush	+	
Knife (without fork)	+	+
Spoon	+	+
Broom (upper hand)	+	+
Striking match (hand that holds the match)	++	
Opening box (hand that holds the lid)	+	+
Total	12	5



## Example 1.3: Looking more closely at handedness

e. What proportion of students were right-handed? What was the average 'handedness' score?



# Lecture 1-2: Summaries of observed data

Study of post-herpetic neuralgia, the most common complication of shingles where nerve endings cause an ongoing burning sensation.

Each patient was given placebo (treatment = 0) or vincristine (treatment = 1).

Six weeks later, patients were interviewed to see whether an improvement had occurred (1 = Improvement).

Case	Outcome	Treatment	Age	Sex	Pre.Dur
1	1	1	76	M	36
2	1	1	52	M	22
3	0	0	80	F	33
4	0	1	77	M	33
5	0	1	73	F	17
6	0	0	82	F	84
7	0	1	71	M	24
8	0	0	78	F	96
9	1	1	83	F	61
10	1	1	75	F	60
11	0	0	62	M	8
12	0	0	74	F	35
13	1	1	78	F	3
14	1	1	70	F	27
15	0	0	72	M	60
16	1	1	71	F	8
17	0	0	74	F	5
18	0	0	81	F	26

# Post-herpetic neuralgia

a. Classify each of the variables recorded in this study by their type.

Case	Outcome	Treatment	Age	Sex	Pre.Dur
1	1	1	76	M	36
2	1	1	52	M	22
3	0	0	80	F	33
4	0	1	77	M	33
5	0	1	73	F	17
6	0	0	82	F	84
7	0	1	71	M	24
8	0	0	78	F	96
9	1	1	83	F	61
10	1	1	75	F	60
11	0	0	62	M	8
12	0	0	74	F	35
13	1	1	78	F	3
14	1	1	70	F	27
15	0	0	72	M	60
16	1	1	71	F	8
17	0	0	74	F	5
18	0	0	81	F	26



# Post-herpetic neuralgia

a. Classify each of the variables recorded in this study by their type.

Variable	Coding	Type
Outcome	1 = Improvement, 0 = No Improvement	<i>non-ordinal</i>
Treatment	1 = vincristine, 0 = placebo	<i>non-ordinal</i>
Age	Years	<i>continuous</i>
Sex	Male / Female	<i>non-ordinal</i>
Pre.Dur	Pretreatment duration of symptoms (months)	<i>continuous</i>

# Summarizing with proportions

Because *Outcome* and *Treatment* are

Categorical, we'll summarize the collected data on these variables using proportions.

The proportion of a categorical variable that takes on a particular outcome is found by:

$$\text{Proportion} = \frac{\text{Cases in a category}}{\text{Total number of cases}}$$

## Notation:

The proportion for an observed sample is denoted  $\hat{p}$  and read “p-hat.”

The proportion for the population that is the focus of study is denoted  $p$ .

# Summarizing with proportions

b. Compute the sample proportion  $\hat{p}$  of patients who reported improvement six weeks after treatment.

$$\hat{p} = \frac{7}{18} = 0.3889$$

Case	Outcome	Treatment	Age	Sex	Pre.Dur
1	1	1	76	M	36
2	1	1	52	M	22
3	0	0	80	F	33
4	0	1	77	M	33
5	0	1	73	F	17
6	0	0	82	F	84
7	0	1	71	M	24
8	0	0	78	F	96
9	1	1	83	F	61
10	1	1	75	F	60
11	0	0	62	M	8
12	0	0	74	F	35
13	1	1	78	F	3
14	1	1	70	F	27
15	0	0	72	M	60
16	1	1	71	F	8
17	0	0	74	F	5
18	0	0	81	F	26

\*\*\*NOTE: We do not use the notation  $p$  here because

$p$  represents true improvement rate.\*\*\*  
of all patients with post herpetic neuralgia.

# Summarizing with proportions

c. Compute the sample proportion  $\hat{p}$  of patients who received the *vincristine*.

$$\hat{p} = \frac{10}{18} = 0.5555$$

Case	Outcome	Treatment	Age	Sex	Pre.Dur
1	1	1	76	M	36
2	1	1	52	M	22
3	0	0	80	F	33
4	0	1	77	M	33
5	0	1	73	F	17
6	0	0	82	F	84
7	0	1	71	M	24
8	0	0	78	F	96
9	1	1	83	F	61
10	1	1	75	F	60
11	0	0	62	M	8
12	0	0	74	F	35
13	1	1	78	F	3
14	1	1	70	F	27
15	0	0	72	M	60
16	1	1	71	F	8
17	0	0	74	F	5
18	0	0	81	F	26

# Summarizing with proportions

d. What notation would be used to represent the difference in observed improvement rates between control and treatment groups?

$$\hat{p}_0 - \hat{p}_1$$

Case	Outcome	Treatment	Age	Sex	Pre.Dur
1	1	1	76	M	36
2	1	1	52	M	22
3	0	0	80	F	33
4	0	1	77	M	33
5	0	1	73	F	17
6	0	0	82	F	84
7	0	1	71	M	24
8	0	0	78	F	96
9	1	1	83	F	61
10	1	1	75	F	60
11	0	0	62	M	8
12	0	0	74	F	35
13	1	1	78	F	3
14	1	1	70	F	27
15	0	0	72	M	60
16	1	1	71	F	8
17	0	0	74	F	5
18	0	0	81	F	26



# Summarizing quantitative variables

Because variables *Age* and *Pre.Dur* are quantitative, a smart way to summarize their recorded values is to report their mean, median, standard deviation.

Additionally, it is often valuable to report percentile rankings for each variable, often referred to as the five-number summary.

- **Mean** -- the numerical average value  
We represent the mean of a sample (called a statistic) by ...

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} = \frac{1}{n} \sum_n x_i$$

(NOTE: The mean of a population is denoted ' $\mu$ ')

- **Median** -- the 50<sup>th</sup> percentile of a dataset, i.e., the point that splits the observed data in half.

# Central tendency

a. Compute the mean and median age of the study patients. Both of these measures offer an estimate of a typical value for the observed Age of the 18 patients.

76 52 80 77 73 82 71 78 83

75 62 74 78 70 72 71 74 81

$$\bar{X} = \frac{1}{n} \sum x_i = \frac{76+52+\dots+81}{18} = 73.83$$

$$\text{Location of median} = \frac{n+1}{2} = \frac{19}{2} = 9.5 \text{ (halfway between 9th, 10th sorted observations)}$$

52 62 70 71 71 72 73 74 74 75 76 77 78 78 80 81 82 83

9th 10th

$$M = \frac{74+75}{2} = 74.5$$

# Central tendency

b. What if the youngest patient's age was mis-recorded as '5' instead of '52'?

New mean  $\bar{x} = 71.22$

New median  $M = \text{still } 74.5$

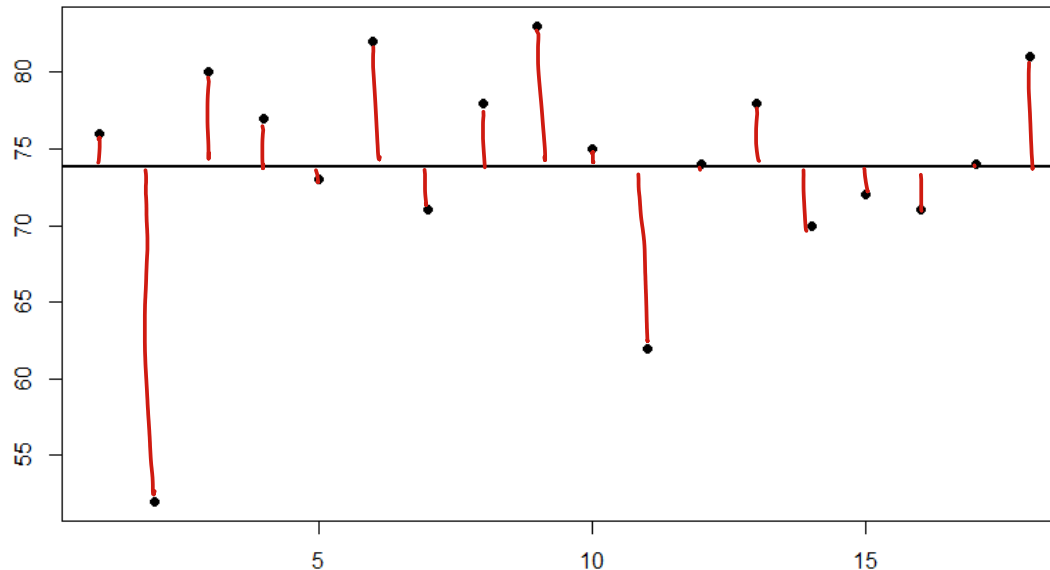
KEY IDEA:

The mean is sensitive to extreme observations.

The median is robust / resistant to extreme observations.

# Standard deviation

In addition to summarizing the typical age of the patients, we also want to describe how much (or little) their ages vary. One of the most common measures of variation is the **standard deviation**.



By definition, the standard deviation first computes the distance (or **deviation**) of every observation in a dataset from their mean,  $x_i - \bar{x}$ . It squares these deviations, finds an approximate average of them, and then takes the square root of that average to undo the earlier squaring. Let's walk through these step-by-step on the next page.

# Standard deviation

The standard deviation gives the

approximate average distance from  
the average .

The larger the standard deviation, the  
more variability there is in

the data and the more spread out the individual observations are.

# Example 1.4: SD practice

Choose one of the four statements below to describe the relationship between the data sets compared.

- i. The quantity in column A is greater.
- ii. The quantity in column B is greater.
- iii. The two quantities are equal.
- iv. The relationship cannot be determined from the given information.

Statement	Column A	Column B
<u>ii</u>	The standard deviation of {0.2, 0.4, 0.6, 0.8}	The standard deviation of {2, 4, 6, 8}

$$\bar{x} = 0.5$$

Deviations

$$x_i - \bar{x} = \begin{array}{l} -0.3 \\ -0.1 \\ 0.1 \\ 0.3 \end{array}$$

$$\bar{x} = 5$$

Deviations:

$$x_i - \bar{x} = \begin{array}{l} -3 \\ -1 \\ 1 \\ 3 \end{array}$$

# Example 1.4: SD practice

Choose one of the four statements below to describe the relationship between the data sets compared.

- The quantity in column A is greater.
- The quantity in column B is greater.
- The two quantities are equal.
- The relationship cannot be determined from the given information.

Statement	Column A	Column B
<u>iii</u>	The standard deviation of {1,3,5,7,9}	The standard deviation of {3,5,7,9,11}

$$\bar{x} = 5$$

deviations:

-4

-2

0

2

4

$$\bar{x} = 7$$

Deviations

-4

-2

0

2

4

## Example 1.4: SD practice

Choose one of the four statements below to describe the relationship between the data sets compared.

- i. The quantity in column A is greater.
- ii. The quantity in column B is greater.
- iii. The two quantities are equal.
- iv. The relationship cannot be determined from the given information.


Statement	Column A	Column B
<u>ii</u>	The standard deviation of {1,3,5,7,9}	The standard deviation of {1,3,5,7,9,9}



## Example 1.4: SD practice

Choose one of the four statements below to describe the relationship between the data sets compared.

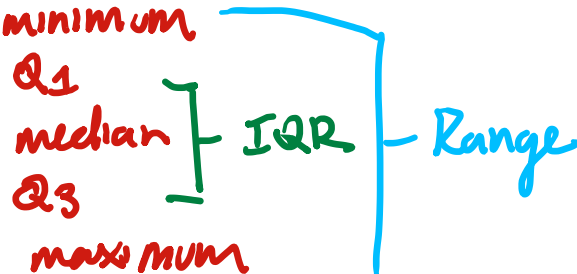
- i. The quantity in column A is greater.
- ii. The quantity in column B is greater.
- iii. The two quantities are equal.
- iv. The relationship cannot be determined from the given information.

Statement	Column A	Column B
	The standard deviation of {1,3,5,7,9}	The standard deviation of {1,3,5,5,7,9}



# Percentiles

**Percentiles, Range, & IQR:** The  $P^{\text{th}}$  percentile is the value of a quantitative variable which is greater than  $P$  percent of the data. The most frequently-reported percentiles constitute a **5-number summary** of a distribution:

- 0<sup>th</sup> percentile *minimum*
  - 25<sup>th</sup> percentile *Q<sub>1</sub>*
  - 50<sup>th</sup> percentile *median*
  - 75<sup>th</sup> percentile *Q<sub>3</sub>*
  - 100<sup>th</sup> percentile *maximum*
- 

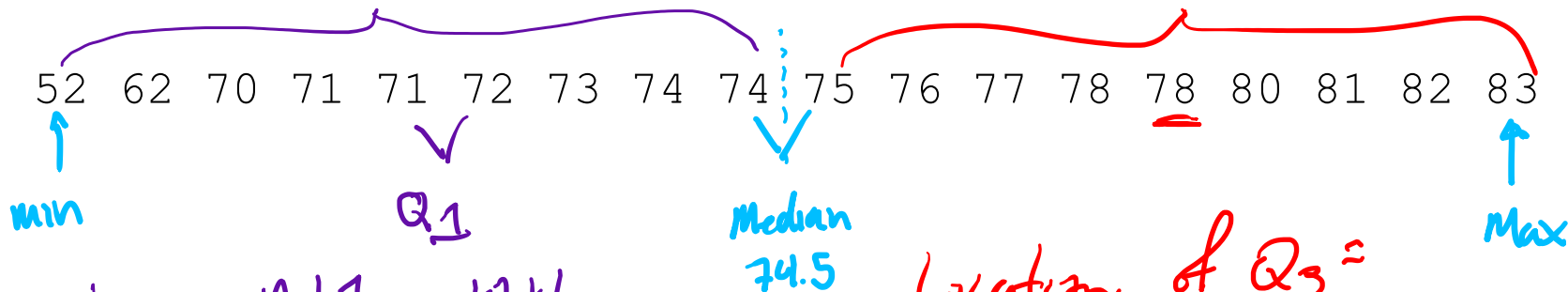
**Range** = 100<sup>th</sup> percentile – 0<sup>th</sup> percentile

**Interquartile Range (IQR)** = 75<sup>th</sup> percentile – 25<sup>th</sup> percentile

\*\*\*Note: There are a variety of ways to compute percentiles. TI-calculators do not necessarily compute them in the same manner as RStudio or other software applications.

# Percentiles

d. Compute the 5-number summary for the Age data of the patients.



$$\text{Location of } Q_1 = \frac{n+1}{2} = \frac{10+1}{2} = 5.5$$

$$Q_1 = 71 + 0.25(72 - 71) \\ = 71.25$$

5-number summary:

(52, 71.25, 74.5, 78, 83)

$$\text{Location of } Q_3 = \frac{n+1}{2} = \frac{9+1}{2} = 5$$

$$Q_3 = 78$$

Note: TI-calculators, Rstudio, excel, etc., often compute percentiles differently.

# Recap

Summary measures of the variable Age for our 18 cases:

Measures of central tendency	Measure of variation	Additional percentile rankings
$\bar{x} = \underline{73.83}$	$s = \underline{7.4459}$	$Q_0 = \underline{52}$
$M = \underline{74.5}$	$IQR = \underline{6.75}$	$Q_1 = \underline{71.25}$
	$Range = \underline{31}$	$Q_3 = \underline{78}$
		$Q_4 = \underline{83}$

# Using calculators

Although there was value in exploring these computations in detail, we usually use calculators or software to compute all the statistical summaries listed above. Provide a summary of the patients' pretreatment neuralgia duration *Pre.Dur* by first entering the data into your calculator and then having it do all the heavy computational lifting.

36	22	33	33	17	84	24	96	61
60	08	35	03	27	60	08	05	26

1. Navigate to 2<sup>ND</sup>  
> STAT > EDIT

L1	L2	L3	1
██████	-----	-----	

L1(1) =

2. Navigate to STAT > CALC > 1-VAR STATS

```

EDIT CH10 TESTS
1:1-Var Stats
2:2-Var Stats
3:Med-Med
4:LinReg(ax+b)
5:QuadReg
6:CubicReg
7:QuartReg

```

# Lecture 1.3 – Graphical summaries

If we'd like to graphically represent a single quantitative variable, we might choose between a

histogram

or boxplot

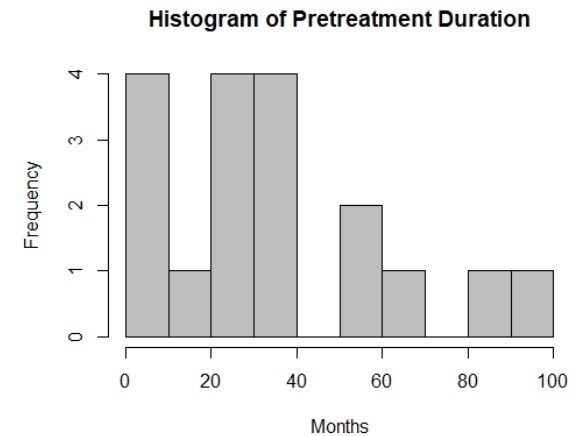
(or, better, both!).

Case	Outcome	Treatment	Age	Sex	Pre.Dur
1	1	1	76	M	36
2	1	1	52	M	22
3	0	0	80	F	33
4	0	1	77	M	33
5	0	1	73	F	17
6	0	0	82	F	84
7	0	1	71	M	24
8	0	0	78	F	96
9	1	1	83	F	61
10	1	1	75	F	60
11	0	0	62	M	8
12	0	0	74	F	35
13	1	1	78	F	3
14	1	1	70	F	27
15	0	0	72	M	60
16	1	1	71	F	8
17	0	0	74	F	5
18	0	0	81	F	26

# Histograms

Histograms take observations of a numeric variable and classify them into *bins* of equal width, spanning the interval over which the variable is observed.

\* i.e., cannot exactly determine mean, std. dev., etc.



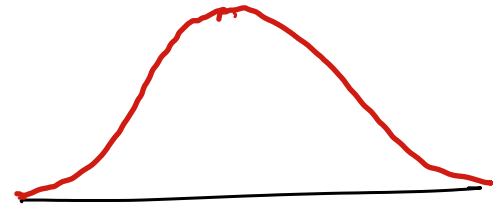
Note that a histogram does not show the exact

Values of summary statistics <sup>\*</sup> . Rather, it describes the shape and density of the data. Higher bars represent where data are denser.

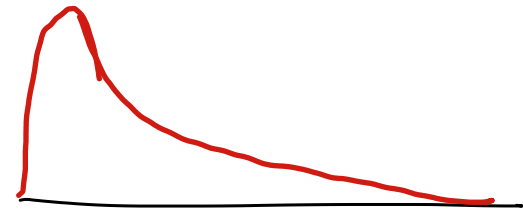
# Interpreting histograms

We often draw smooth curves to illustrate the general shape of a distribution. This allows us to see the shape of a distribution with fewer jagged edges at the corners. There are *five* common shapes we might use to describe the histogram of observed data:

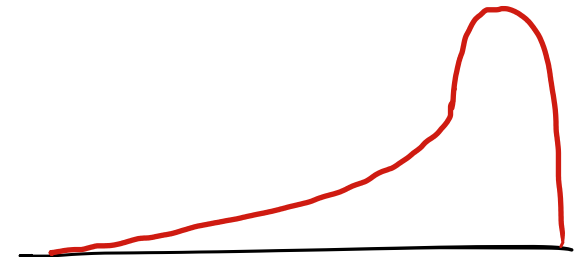
1. *Symmetric & bell shaped*



2. *Right- or positive-skewed*

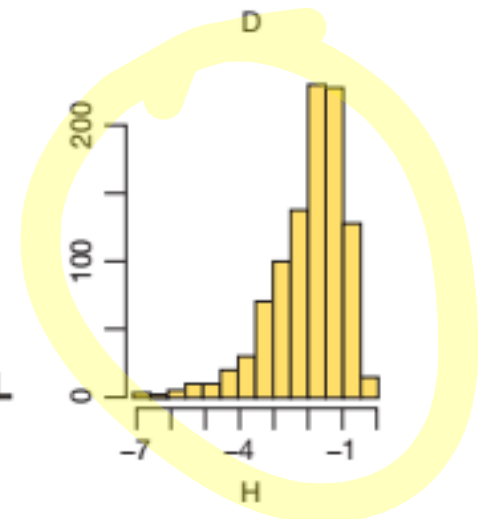
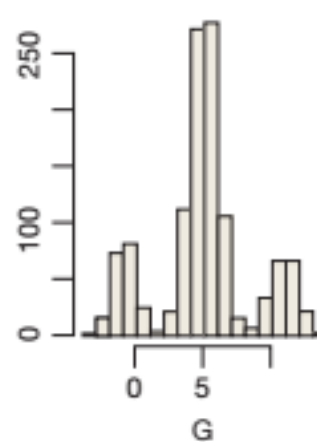
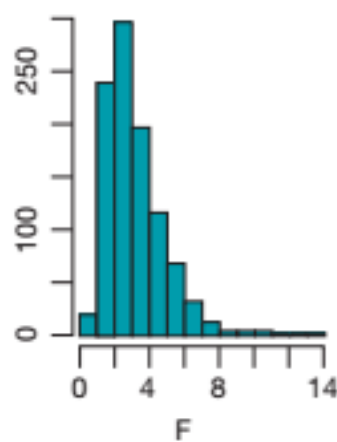
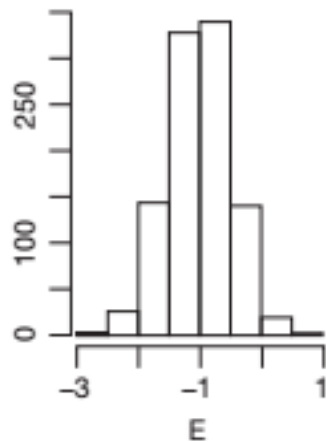
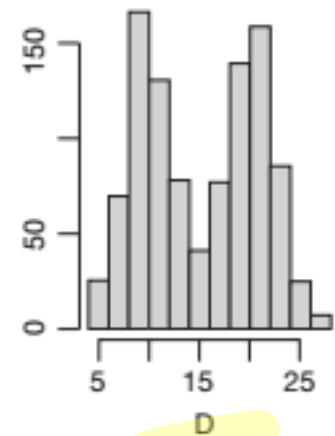
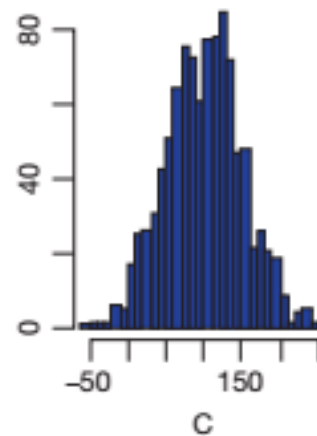
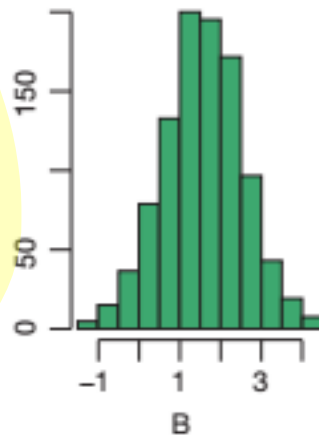
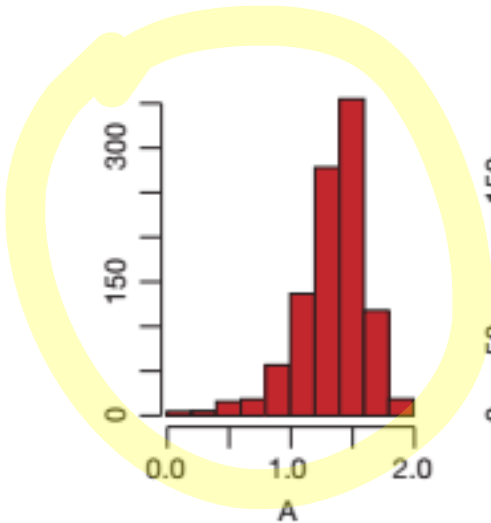


3. *Left- or negative-skewed*



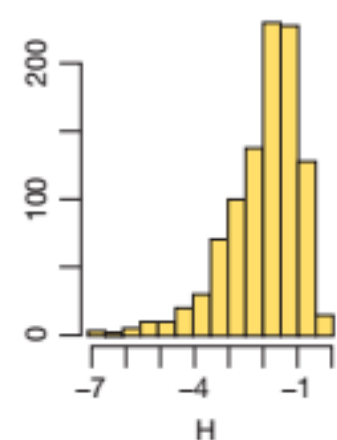
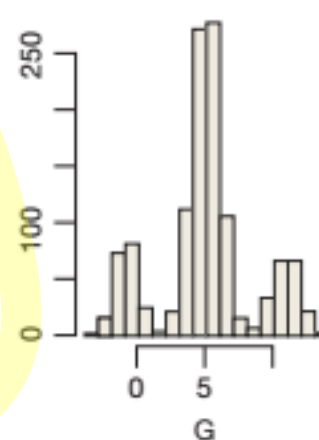
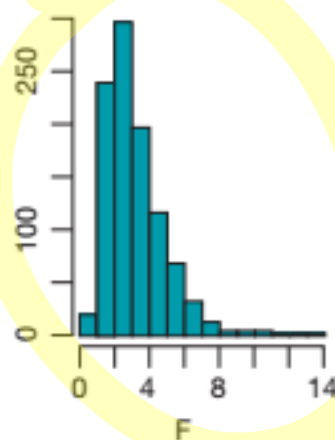
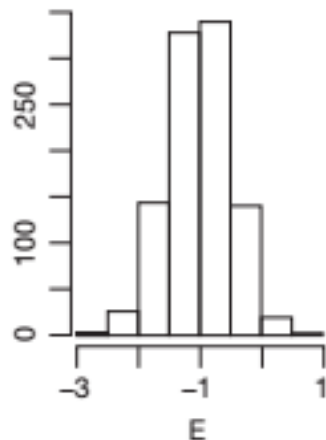
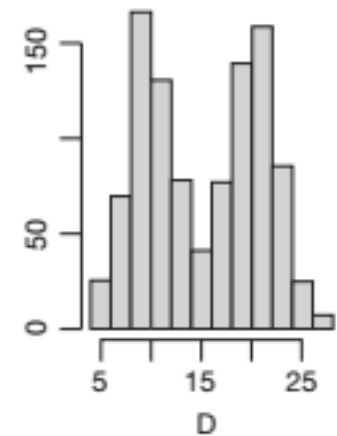
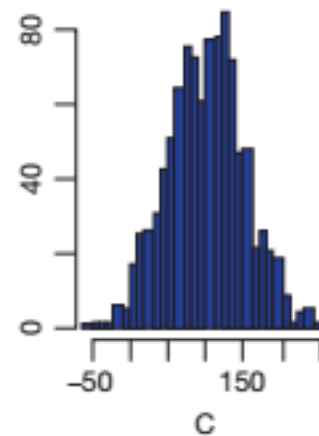
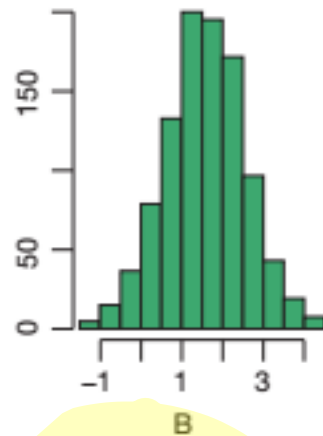
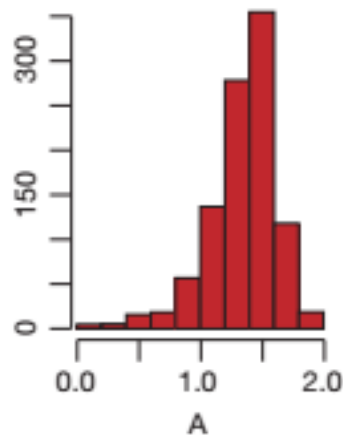
# Example 1.6 Histogram practice

a. Which histograms are skewed to the left?



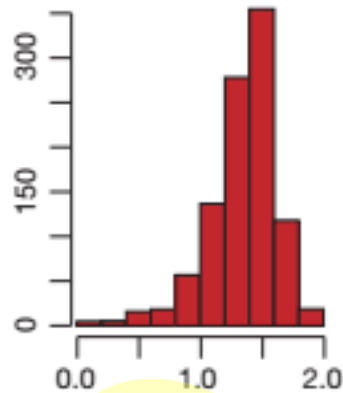
# Example 1.6 Histogram practice

b. Which histograms are skewed to the right?

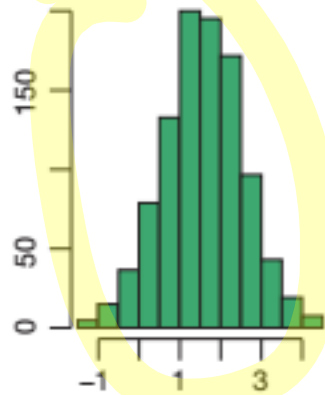


# Example 1.6 Histogram practice

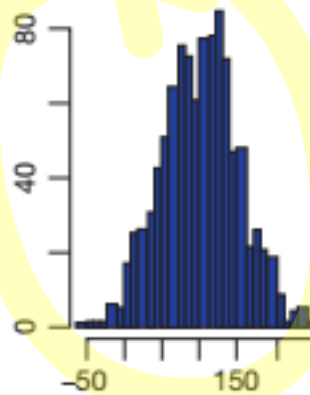
c. Which histograms are skewed to the symmetric?



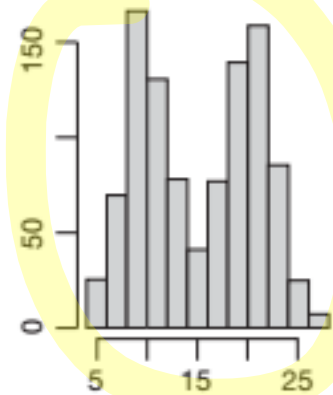
A



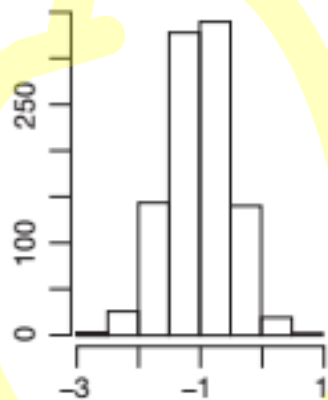
B



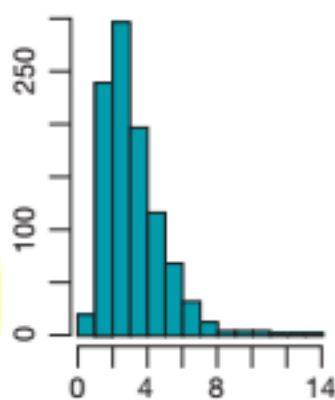
C



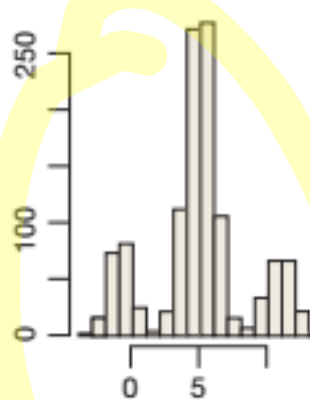
D



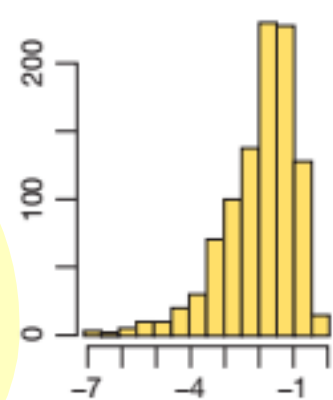
E



F



G

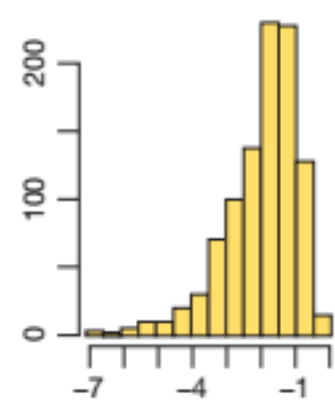
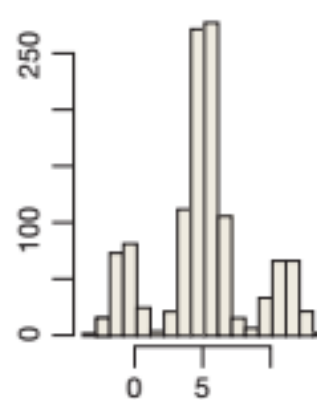
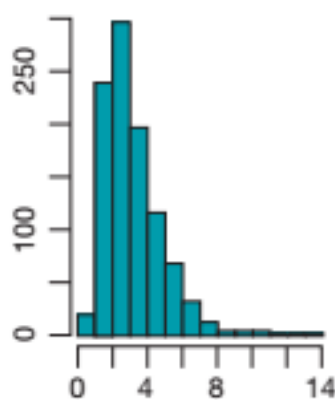
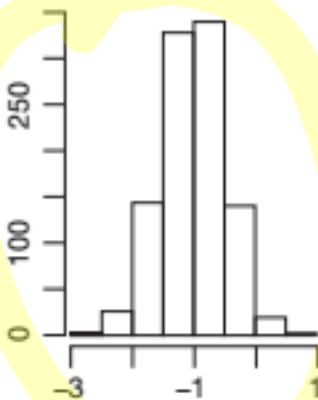
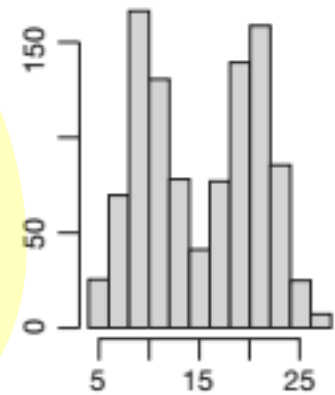
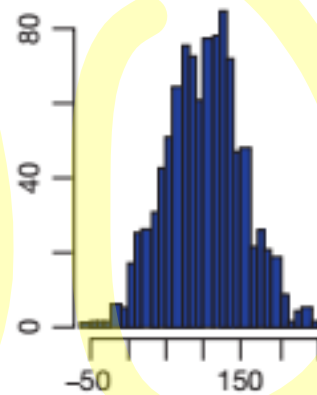
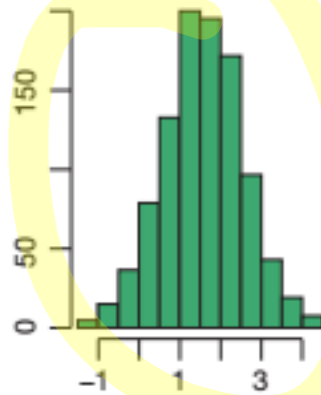
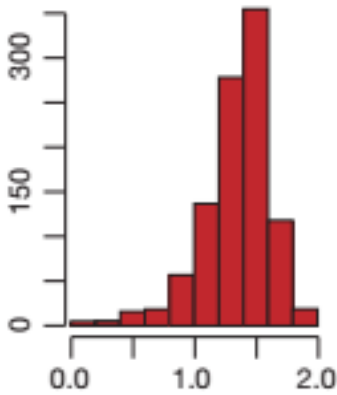


H



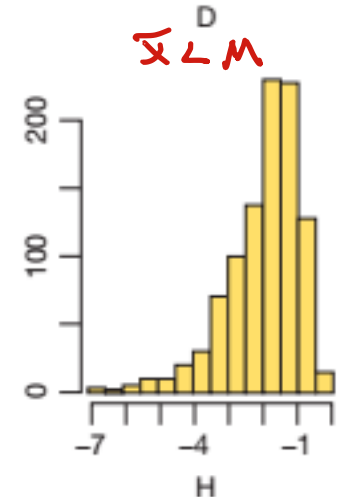
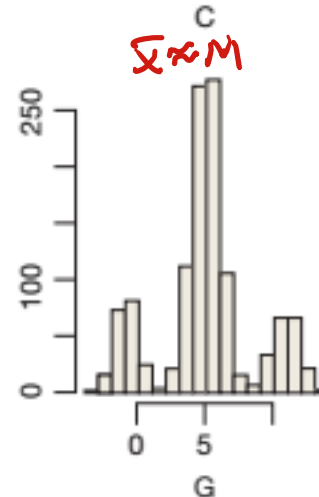
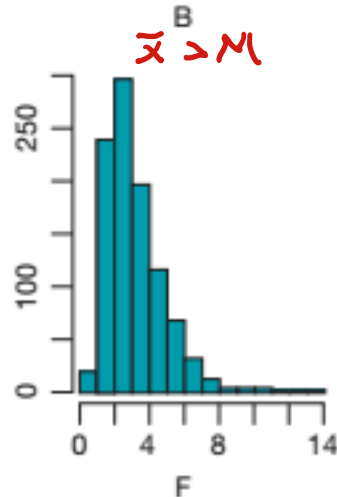
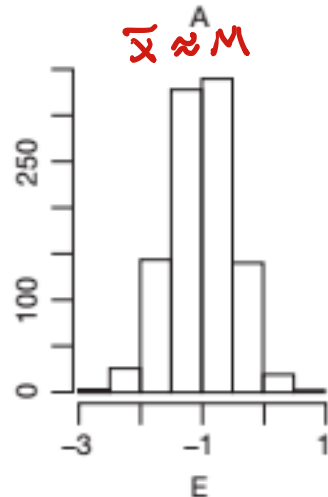
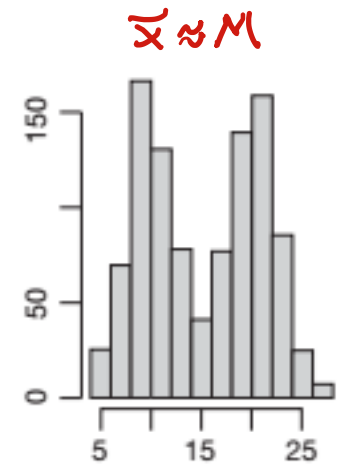
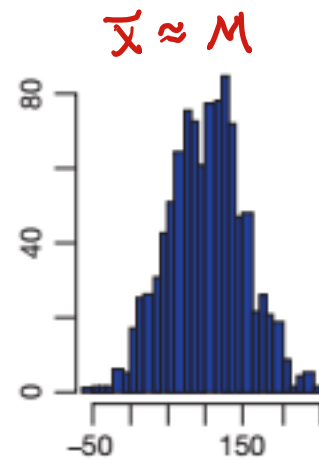
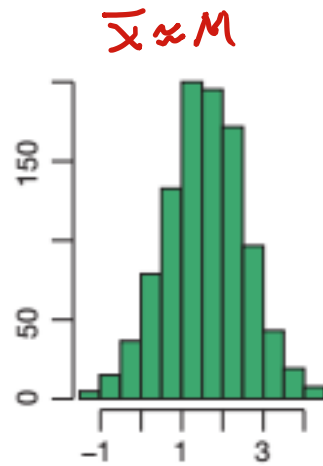
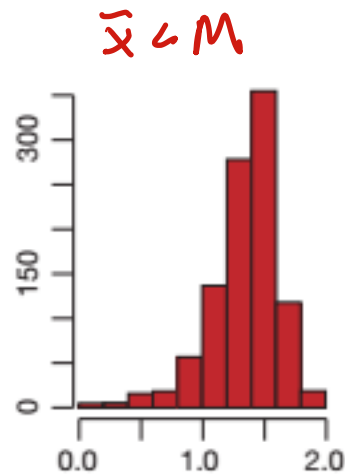
# Example 1.6 Histogram practice

d. Which histograms are skewed to the symmetric & bell-shaped?



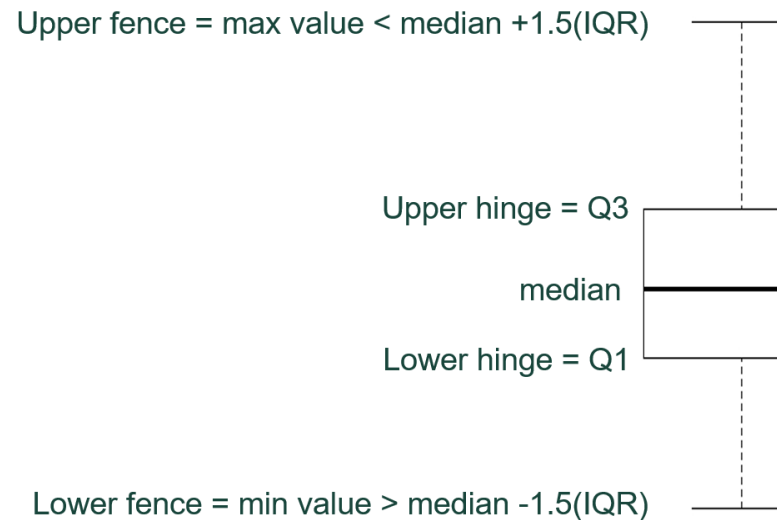
# Example 1.6 Histogram practice

e. For each of the histograms, state whether the mean is likely to be larger than the median, smaller than the median, or approximately equal to the median.



# Boxplots

A boxplot is a graphical display of the five-number summary for a single quantitative variable. It shows the general shape of the distribution, identifies the middle 50% of the data, and highlights any outliers



# Boxplots

We first draw a heavy line at the median, with lighter lines at Q1 (the lower hinge) and Q3 (the upper hinge), and connect Q1 and Q3 to form a box. This helps to visualize the IQR of the data. The points at distances  $1.5(IQR)$  from each hinge define the fences of the data set. Lines (sometimes called whiskers) are drawn from each hinge to the most extreme measurements inside the inner fence. Points that are beyond the fences are plotted individually and are often considered **outliers**.

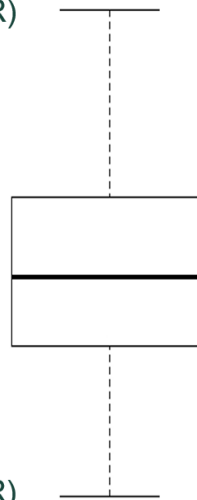
Upper fence = max value  $<$  median  $+1.5(IQR)$

Upper hinge = Q3

median

Lower hinge = Q1

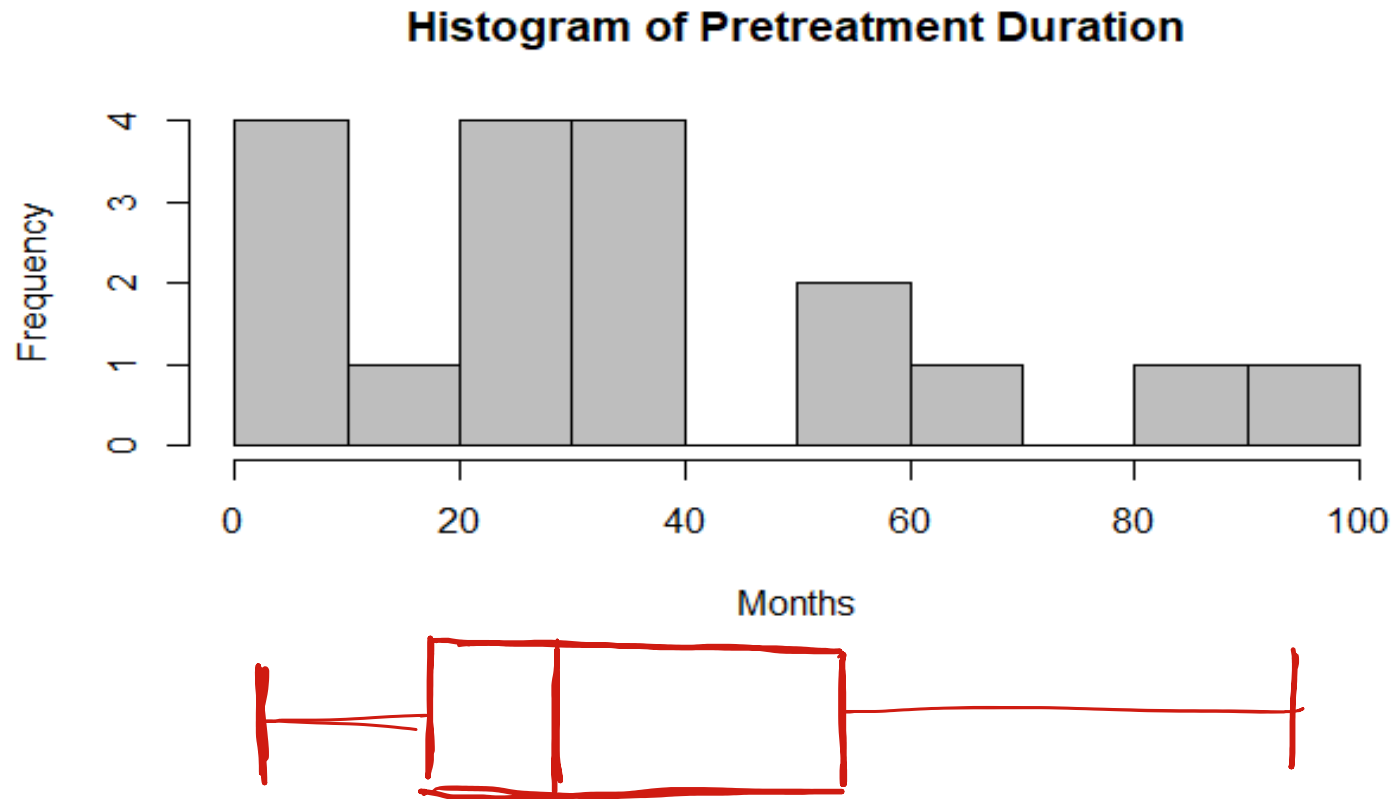
Lower fence = min value  $>$  median  $-1.5(IQR)$



# Post-herpetic neuralgia

a. Draw a boxplot for the data on Pretreatment Duration from our Neuralgia study with the five number summary (3, 18.25, 30, 54, 96).

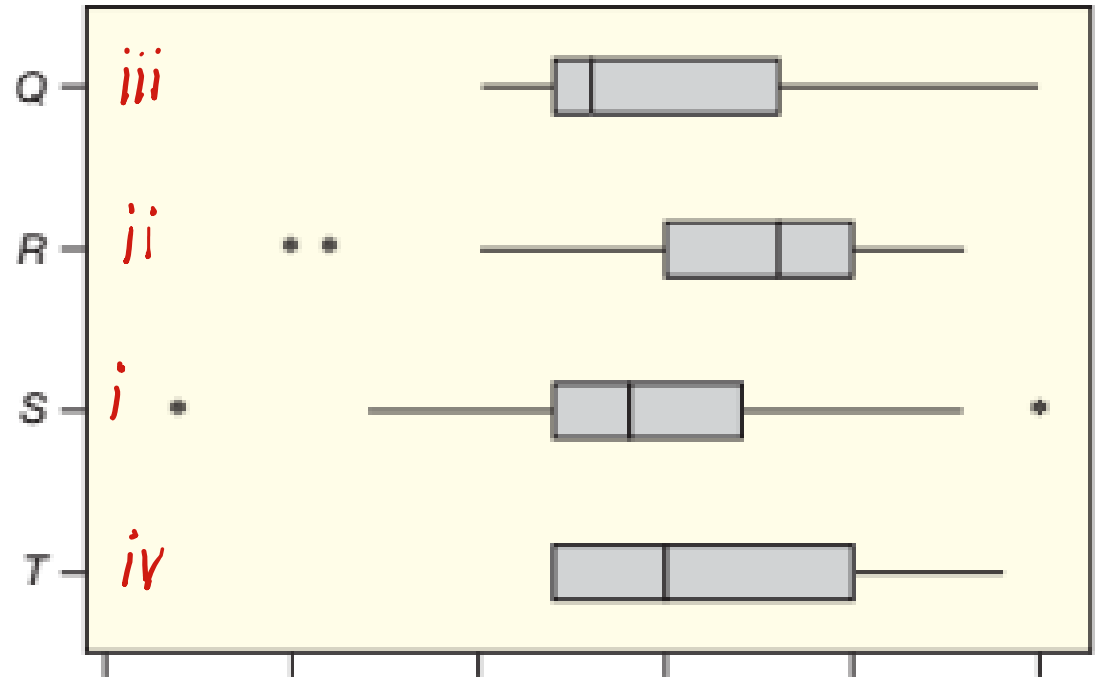
Data: 3 5 8 8 17 22 24 26 27 33 33 35 36 60 60 61 84 96



## Example 1.7: Matching graphical and quantitative summaries.

a. Match each five-number summary to one of the available boxplots

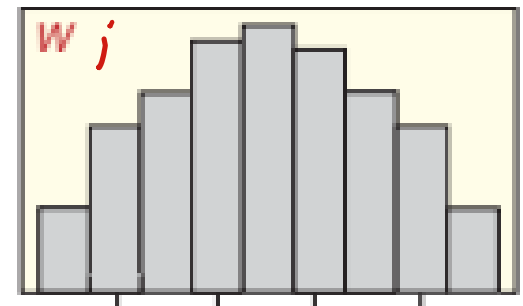
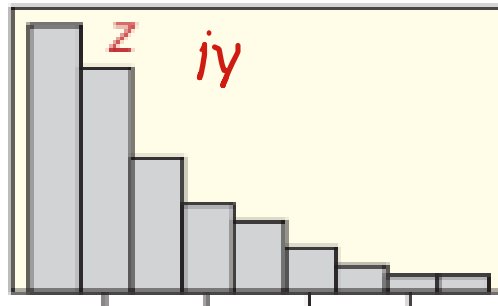
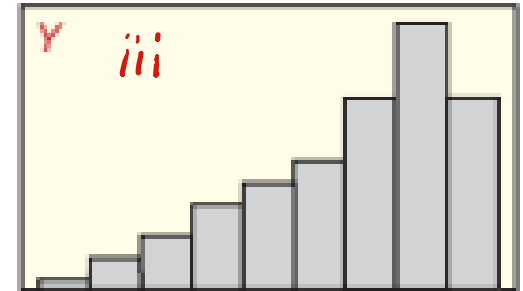
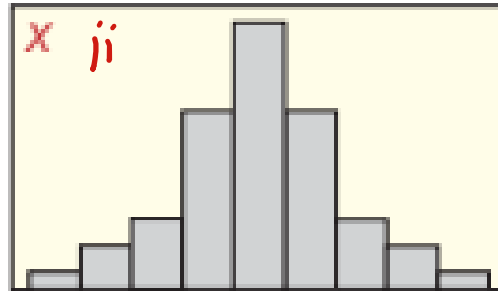
- i.* (2,12,14,17,25)
- ii.* (5,15,18,20,23)
- iii.* (10,12,13,18,25)
- iv.* (12,12,15,20,24)



## Example 1.7: Matching graphical and quantitative summaries.

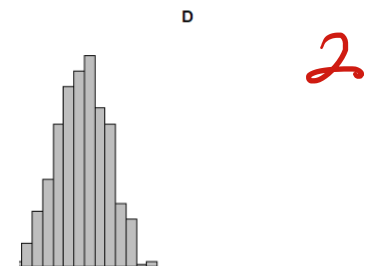
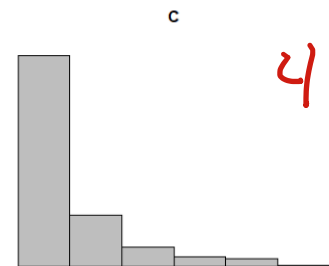
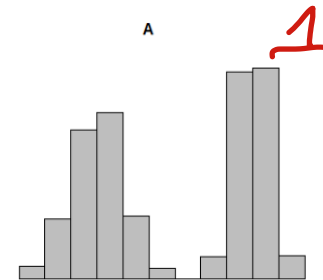
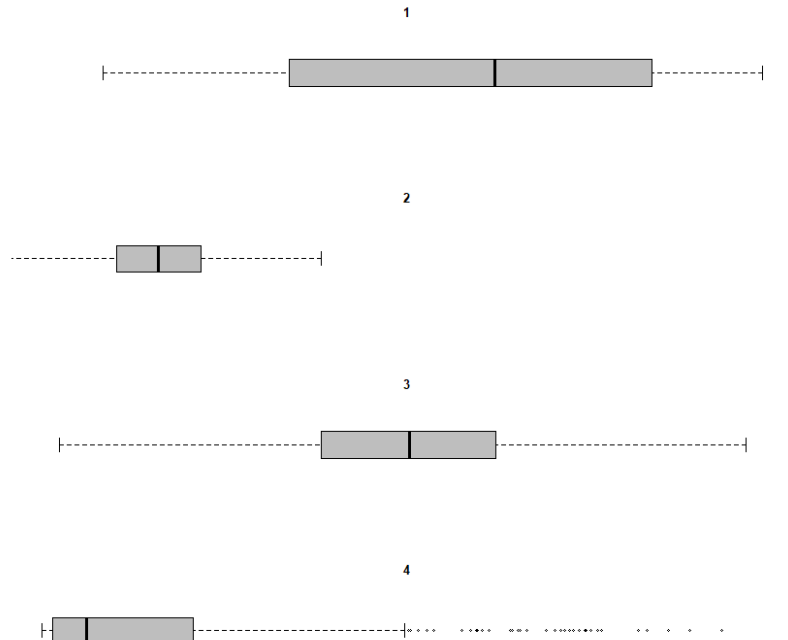
b. Match each five-number summary to one of the available histograms.

- i.* (1,3,5,7,9)
- ii.* (1,4,5,6,9)
- iii.* (1,5,7,8,9)
- iv.* (1,1,2,4,9)



# Example 1.7: Matching graphical and quantitative summaries.

c. Match each boxplot to one of the available histograms..



## Z-scores

Often, a single data value is not easily interpreted without knowing how it relates to other observations of the same variable.

A common way of determining how usual or unusual a single observation is to count how many standard deviations it is away from the mean. This quantity is

known as a z-score.

The number of standard deviations from the mean: z-Score

The standardized score for a data value is defined to be:

$$z = \frac{X - \bar{X}}{S} = \frac{\text{"observed" - "expected"}}{\text{"measure of variability"}}$$

For a population, the sample mean  $\bar{x}$  is replaced with the population mean  $\mu$  and the sample standard deviation  $s$  is replaced with the population standard deviation  $\sigma$ .

Definition: The z-score tells how many standard deviations the value is from the mean and is independent of the unit of measurement.

## Example 1.8: Working with z-scores

A. Consider each of the following cases, which gives the value of an observed data point and the mean and standard deviation of the sample from which is came. Rank these cases from the smallest to largest z-score.

- i. The value 243 in a dataset with mean 200 and standard deviation 25.
- ii. The value 88 in a dataset with mean 96 and standard deviation 10.
- iii. The value 5.2 in a dataset with mean 12 and standard deviation 2.3.
- iv. The value 8.1 in a dataset with mean 5 and standard deviation 2.

$z_{iii}$   
(smallest)

$z_{ii}$

$z_{iv}$

$z_i$   
(largest)

$$z_i = \frac{243 - 200}{25} = 1.72$$

$$z_{iii} = \frac{5.2 - 12}{2.3} = -2.9565$$

$$z_{ii} = \frac{88 - 96}{10} = -0.8$$

$$z_{iv} = \frac{8.1 - 5}{2} = 1.55$$



## Example 1.8: Working with z-scores

B. If the mean of the dataset below is 6.8 and the standard deviation is 3.6, how many observations will have a z-score between -1 and 1? (Hint: this is the same as asking how many fall within 1 standard deviation of the mean.)

Set A: {2, 9, 2, 6, 9, 10, 7, 4, 5, 14}

$$6.8 \pm 3.6 = (3.2, 10.4)$$

7 observations fall within one standard deviation of the mean.



## Bar & Mosaic plots

Recall the sleep study from Lecture 1-1 that examined the relationship between class start times, sleep, circadian preference, alcohol use, academic performance, and other variables in college students.

*Does the proportion of students who have an early class differ across the class year of our  $n = 253$  students?*

This question asks about the relationship between two categorical variables.

## Bar & Mosaic plots

To answer it, we'll need a **contingency** table, where the categories for one variable are listed across the rows and the categories for the second variable are listed across the columns.

	Freshman	Sophomore	Junior	Senior	Total
0 = No Early Class	8	31	21	25	85
1 = Early Class	39	64	33	32	168
Total	47	95	54	57	253

a. Complete the table on the previous page by filling in the row and column **margins**.

## Bar & Mosaic plots

	Freshman	Sophomore	Junior	Senior	Total
0 = No Early Class	8	31	21	25	85
1 = Early Class	39	64	33	32	168
Total	47	95	54	57	253

b. What proportion of the students surveyed have an early class?

$$\hat{p} = \frac{168}{253} = 0.6640$$

## Bar & Mosaic plots

	Freshman	Sophomore	Junior	Senior	Total
0 = No Early Class	8	31	21	25	85
1 = Early Class	39	64	33	32	168
Total	47	95	54	57	253

c. What proportion of freshman have an early class?  $\hat{p} = \frac{39}{47} = 0.8298$

d. What proportion of sophomores have an early class?  $\hat{p} = \frac{64}{95} = 0.6737$

e. What proportion of students with an early class are sophomores?

$$\hat{p} = \frac{64}{168} = 0.3810$$

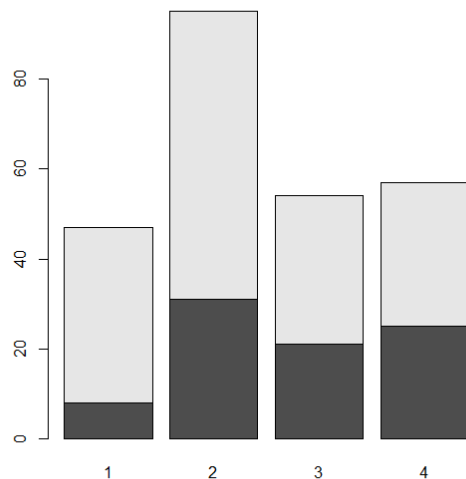
Note that the solutions to (d) and (e) are

not the same !



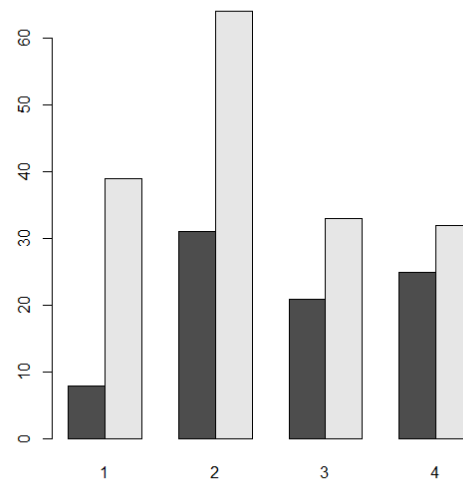
# Bar & Mosaic plots

Bar Chart of EarlyClass by ClassYear Variables



Dark Grey = Early Class

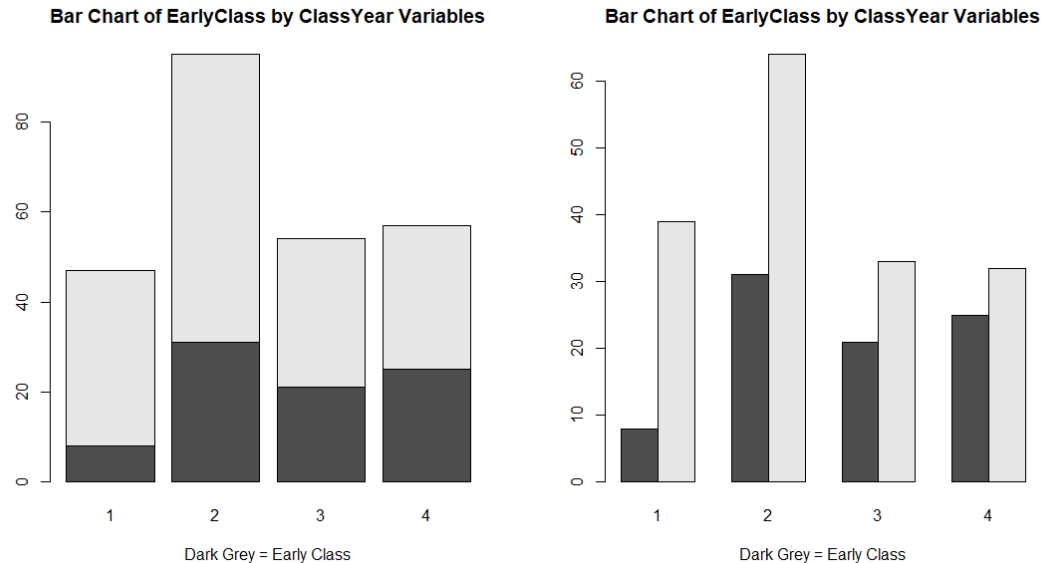
Bar Chart of EarlyClass by ClassYear Variables



Dark Grey = Early Class

	Freshman	Sophomore	Junior	Senior	Total
0 = No Early Class	8	31	21	25	85
1 = Early Class	39	64	33	32	168
Total	47	95	54	57	253

## Bar & Mosaic plots

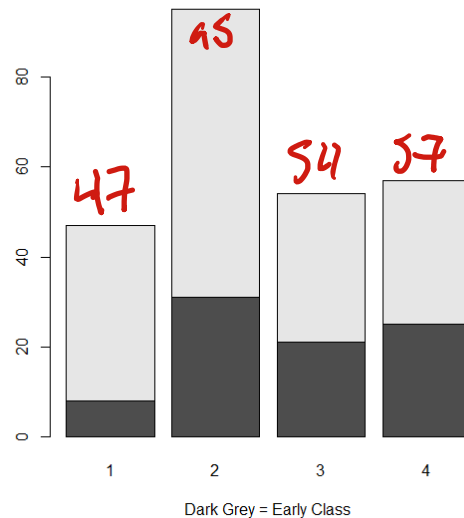


- a. Were more juniors or more seniors present in the sample of  $n = 253$  students? Which plot does a better job of showing this?

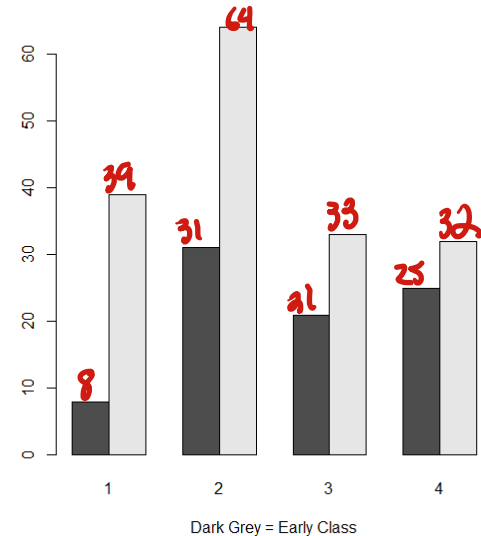
*Seniors ; left plot*

# Bar & Mosaic plots

Bar Chart of EarlyClass by ClassYear Variables

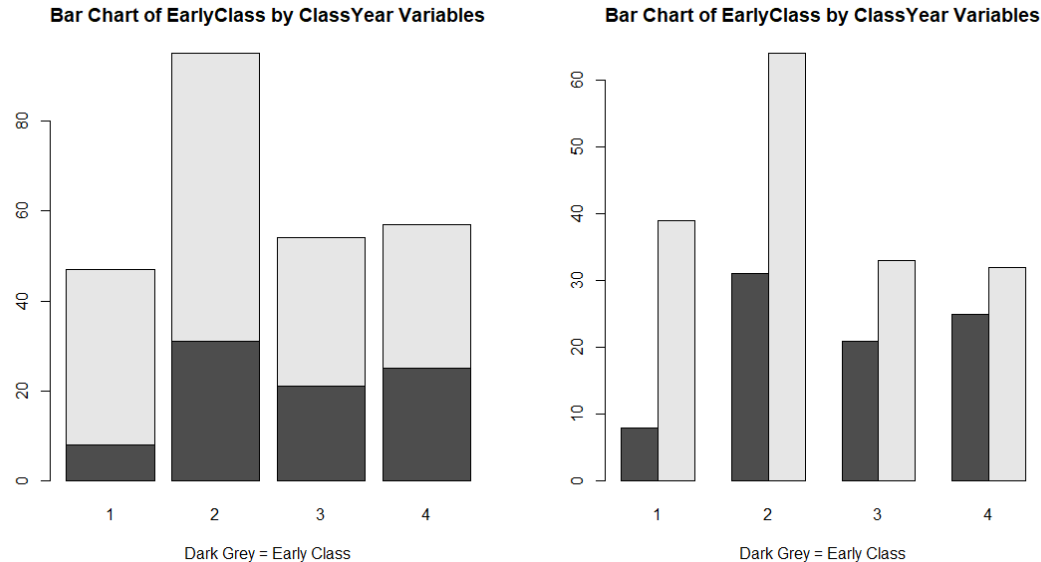


Bar Chart of EarlyClass by ClassYear Variables



	Freshman	Sophomore	Junior	Senior	Total
0 = No Early Class	8	31	21	25	85
1 = Early Class	39	64	33	32	168
Total	47	95	54	57	253

# Bar & Mosaic plots

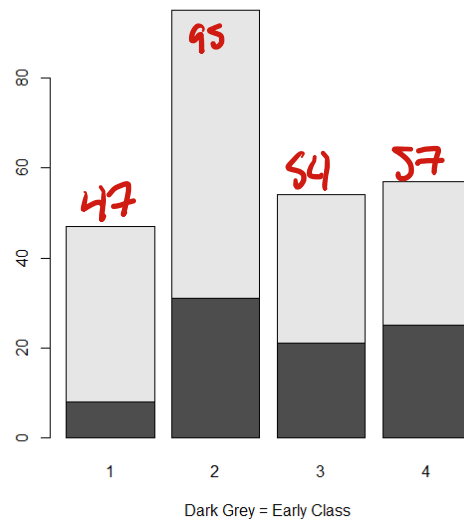


b. Do a greater *number* of freshman or sophomore students have early classes at least once per week? Which graph is more helpful to answer this question?

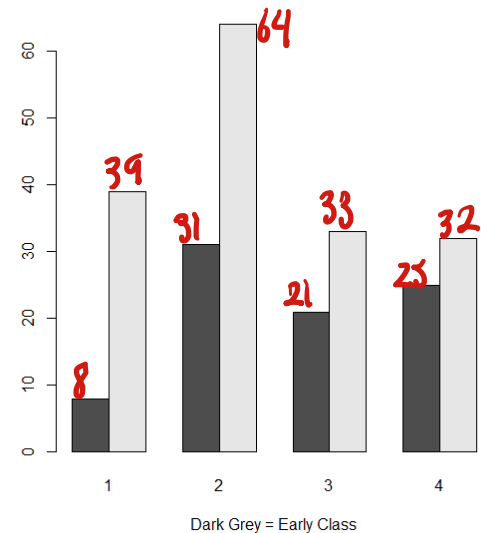
*Sophomores; right plot*

# Bar & Mosaic plots

Bar Chart of EarlyClass by ClassYear Variables

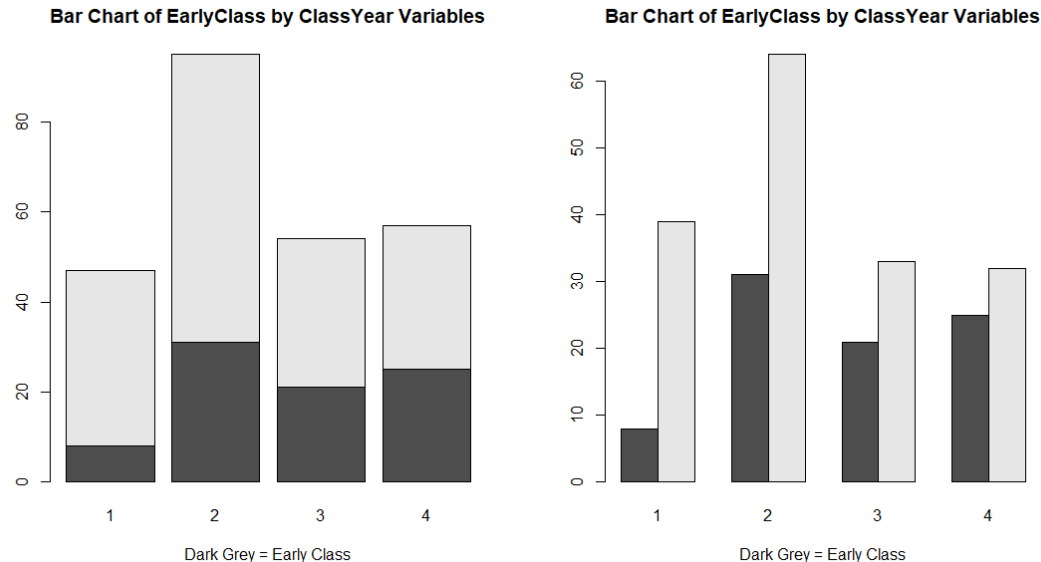


Bar Chart of EarlyClass by ClassYear Variables



	Freshman	Sophomore	Junior	Senior	Total
0 = No Early Class	8	31	21	25	85
1 = Early Class	39	64	33	32	168
Total	47	95	54	57	253

# Bar & Mosaic plots



c. Do a greater *percentage* of freshman or sophomore students have early classes at least once per week? Which graph is more helpful to answer this question?

*freshmen ; left plot.*

d. Which of these two plots provides a better way of analyzing the relationship between *EarlyClass* and *ClassYear*? Explain.

*Depends on what question you're trying to answer.*

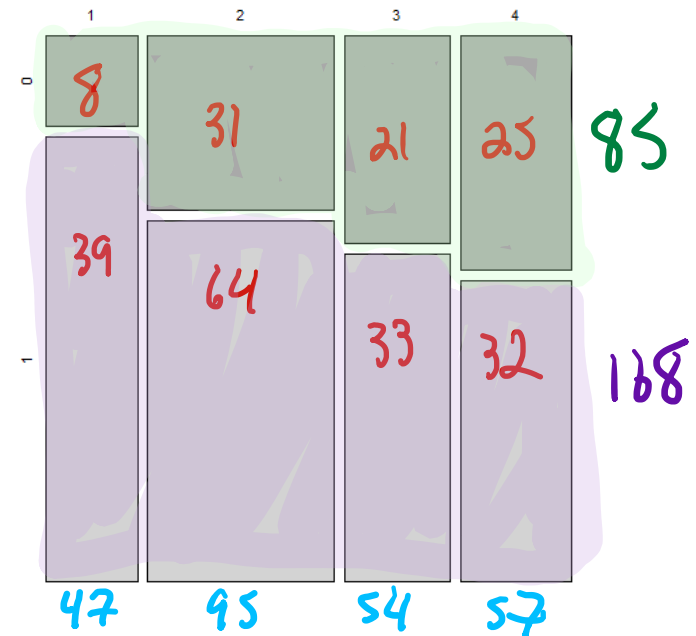


## Bar & Mosaic plots

Mosaic plots use area to represent the number of observations in a given cell of a two-way table. No form of bar-chart does this, making mosaic plots somewhat unique.

Mosaic plots are particularly useful at describing the relationship (or lack of one) between two categorical variables.

Mosaic Plot of EarlyClass by ClassYear



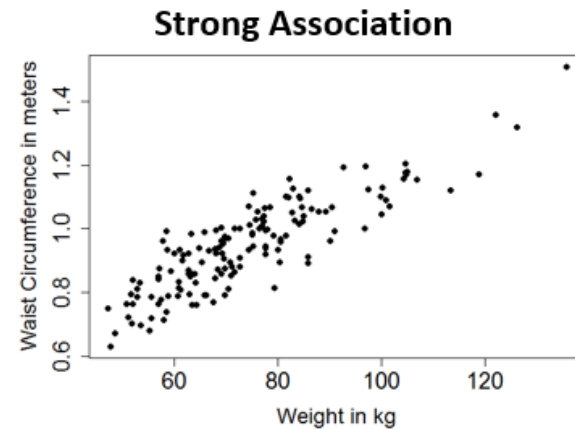
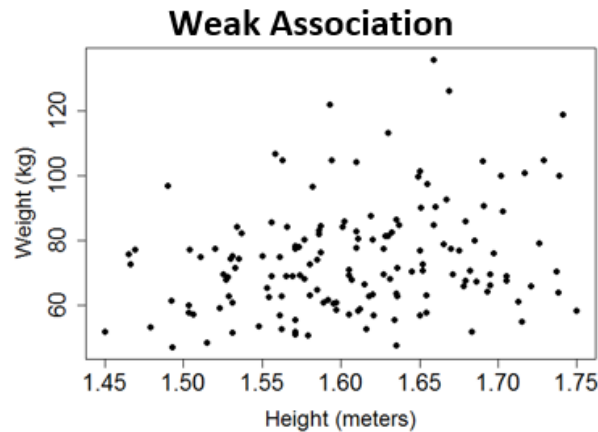
	Freshman	Sophomore	Junior	Senior	Total
0 = No Early Class	8	31	21	25	85
1 = Early Class	39	64	33	32	168
Total	47	95	54	57	253

# Scatterplots

A scatterplot provides a case-by-case view of data for two **quantitative** variables. Each point represents a single case measured on the two variables. Typically, we want to assess three characteristics of the visualized relationship:

## Strength of association

---



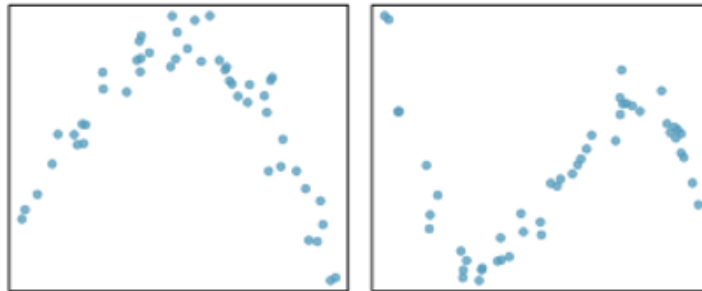
# Scatterplots

A scatterplot provides a case-by-case view of data for two **quantitative** variables. Each point represents a single case measured on the two variables. Typically, we want to assess three characteristics of the visualized relationship:

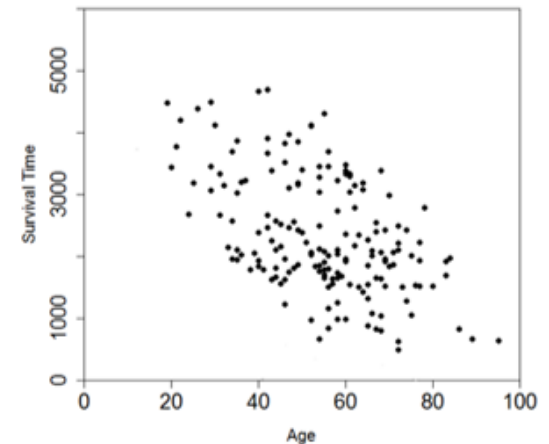
## Linearity of association



**Non-linear Association**



**Linear Association**



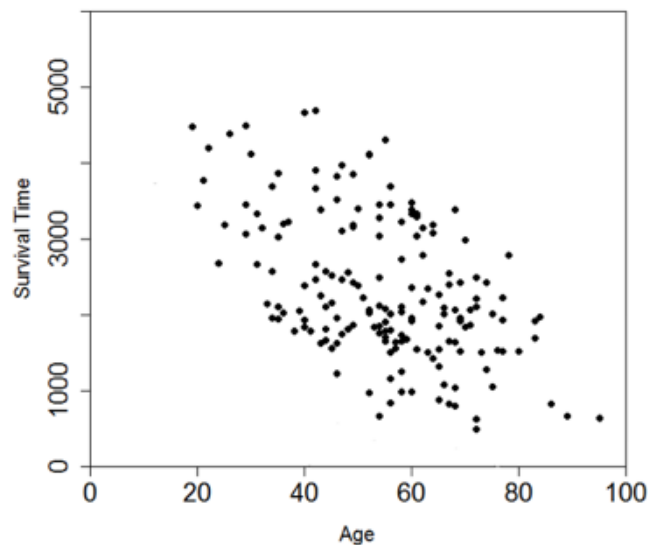
# Scatterplots

A scatterplot provides a case-by-case view of data for two **quantitative** variables. Each point represents a single case measured on the two variables. Typically, we want to assess three characteristics of the visualized relationship:

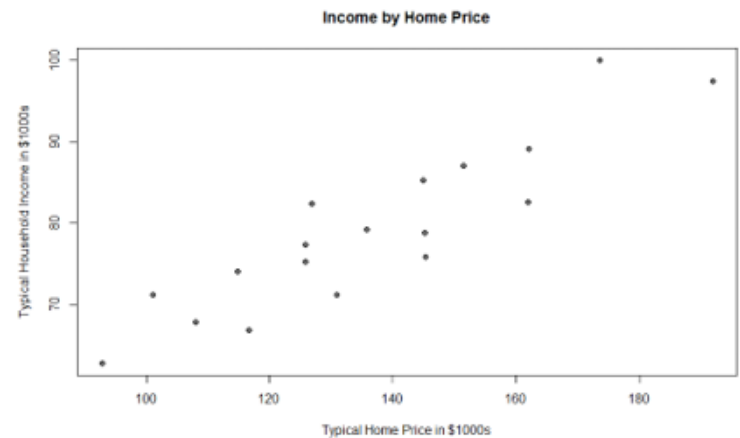
## Direction of association



**Negative Association**



**Positive Association**



## Correlation

The correlation coefficient, denoted by the letter  $r$ , quantifies the strength of the linear association (or clustering about a line) between two variables  $x$  and  $y$ .

Where the standard deviation describes the variability of a single set of data, the correlation describes the joint variability of two sets of data, together.

The value of  $r$  is always between  $-1$  and  $1$ .  
Values closer to  $-1$  correspond to a...

strong, negative association.

Values closer to  $1$  correspond to a...

strong, positive association.

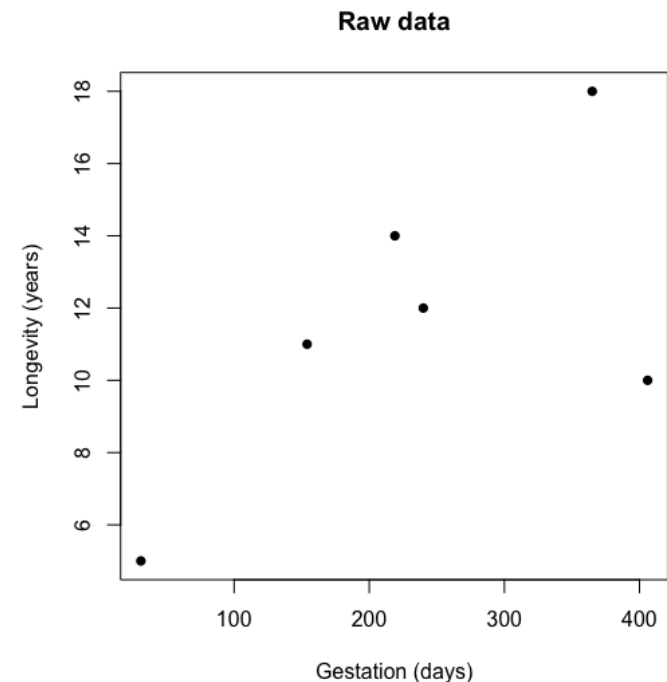
Values closer to  $0$  correspond to...

a lack of a linear association.

## Computing correlation

Earlier in Chapter 1-1, we asked the question, “Is gestation length associated with life expectancy among mammalian species?” Consider the following data set, which gives both values for just six mammalian species.

Species	Gestation (days)	Lifespan (years)
donkey	365	14
bear,black	219	18
moose	240	12
rabbit	31	5
sheep	154	11
camel	406	10
$\bar{x}$	235.8333	11.6667
$s$	137.5462	4.3205



What do you estimate is the correlation of the variables *Gestation* and *Lifespan* for these six observed cases?



## Computing correlation

You'll notice that species with longer gestations often have higher life expectancies. But how strong is this relationship? To compute the **correlation** between the variables *Gestation* and *Lifespan*, which quantifies the *linear* relationship between these two variables, take the following steps:

1. Transform each coordinate  $(x, y)$  into z-score  $(z_x, z_y)$
2. Find the product  $z_x \cdot z_y$  for each coordinate.

Consequence 1: Pairs that are  $(+, +)$  or  $(-, -)$  produce positive products.

Consequence 2: Pairs that are  $(-, +)$  or  $(+, -)$  produce negative products.

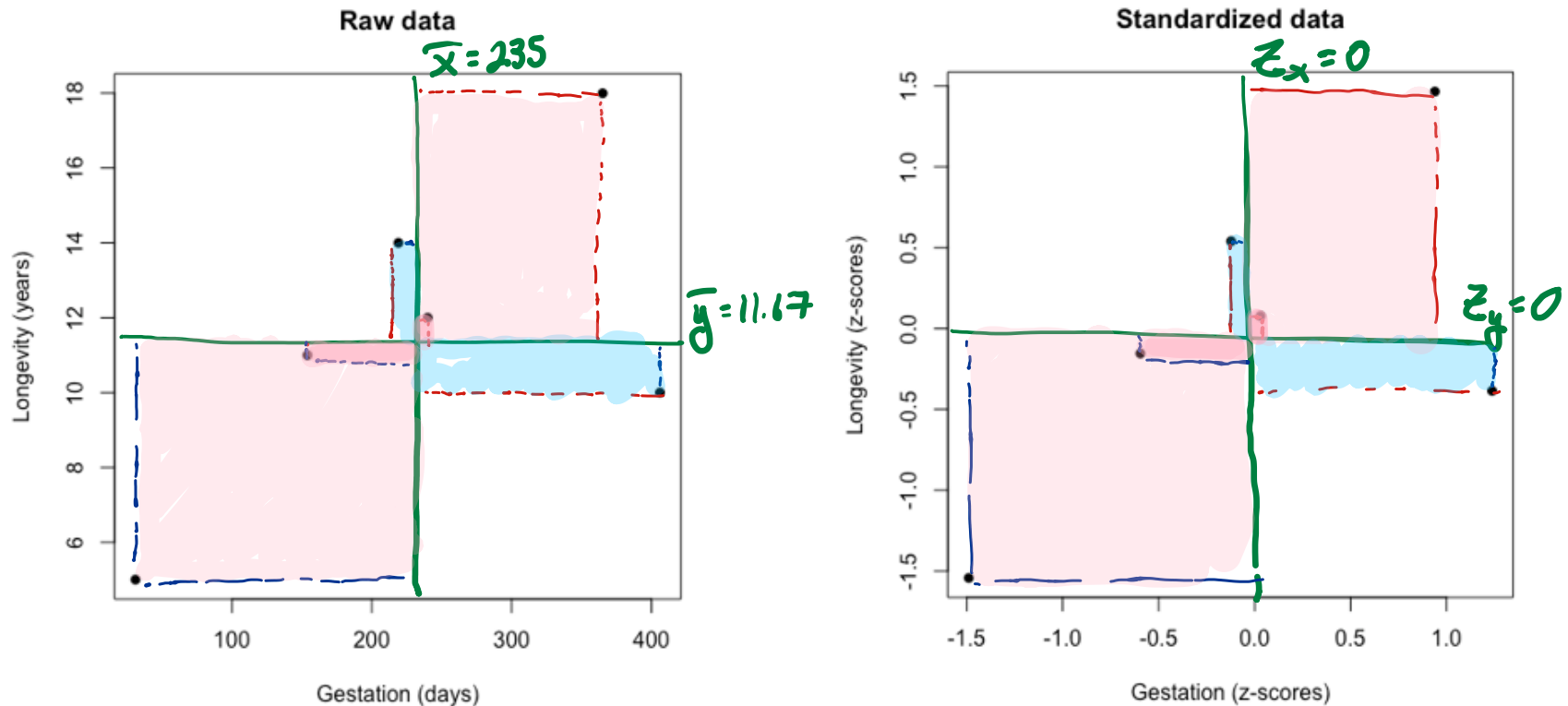
3. Find the approximate average size of products. This is Pearson's product moment correlation.

$$Z_x = \frac{154 - 235.83}{137.5462} = -0.5950 \quad \text{Rabbits' } Z_y = \frac{5 - 11.6667}{4.3205} = -1.5430$$

Species	Gestation (days)	Lifespan (years)	$Z_{\text{gestation}}$	$Z_{\text{lifespan}}$
donkey	365	14	0.9391	0.5401
bear, black	219	18	-0.1224	1.4659
moose	240	12	0.0303	0.0772
rabbit	31	5	-1.4892	-1.5430
sheep	154	11	-0.5950	-0.1543
camel	406	10	1.2373	-0.3878
$\bar{x}$	235.8333	11.6667	0	0
$s$	137.5462	4.3205	1	1

## Computing correlation

Consider two scatterplots of the raw data vs. their standardized z-scores.



The correlation multiplies the both values in the standardized  $(x, y)$  coordinates and then finds an approximate average size of these products. That average is the correlation between  $x$  and  $y$ .

## Computing correlation

$$r = \frac{\sum Z_x * Z_y}{n - 1}$$

$$r = \frac{0.5072 + (-0.1794) + 0.0023 + 2.2979 + 0.0918 + (-0.4773)}{6 - 1}$$

$$r = \frac{2.2425}{5} = 0.4485$$

Species	$Z_{\text{gestation}}$	$Z_{\text{lifespan}}$	$Z_g * Z_l$
donkey	0.9391	0.5401	0.5072
bear,black	-0.1224	1.4659	-0.1794
moose	0.0303	0.0772	0.0023
rabbit	-1.4892	-1.5430	2.2979
sheep	-0.5950	-0.1543	0.0918
camel	1.2372	-0.3858	-0.4773

## Correlation $\neq$ linearity!

*Do NOT necessarily*

A value of  $r$  close to 1 or  $-1$  \_\_\_\_\_ tell you that a relationship is linear! You **must**

*check a scatter plot* ; otherwise, your interpretation of the correlation coefficient might be wrong.

