Chapter 1: Introduction to data

Yingbo Li

Southern Methodist University

STAT 2331

Outline

Observations and variables

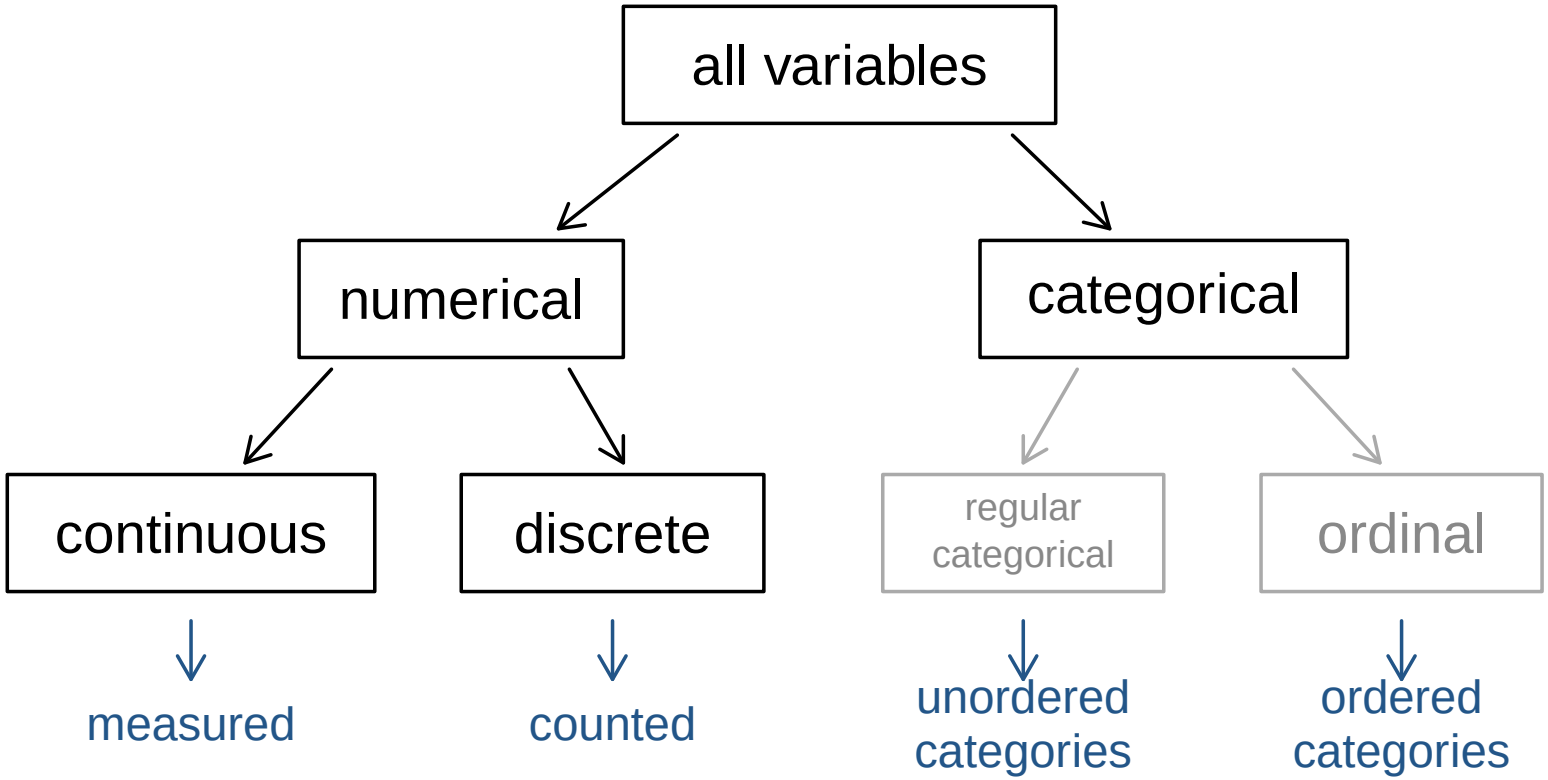
variable ↓

data matrix ⇒

	type	price	···	weight
1	small	15.9	···	2705
2	midsize	33.9	···	3560
⋮	⋮	⋮	⋮	⋮
54	midsize	26.7	···	3245

← *observation/case*

Types of variables



Types of variables (cont.)

	type	price	mpgCity	drivetrain	passengers	weight
1	small	15.9	25	front	5	2705
2	midsize	33.9	18	front	5	3560
⋮	⋮	⋮	⋮	⋮	⋮	⋮
54	midsize	26.7	20	front	5	3245

- `type`: small, midsize or large. *(categorical, ordinal)*
- `price`: average price in \$1000's *(numerical, continuous)*
- `mpgCity`: cite mileage per gallon *(numerical, continuous)*
- `drivetrain`: front, rear, 4WD *(categorical)*
- `passengers`: passenger capacity *(numerical, discrete)*
- `weight`: car weight in pounds *(numerical, continuous)*

Question

1. What type of variable is the number of roses you received on a Valentine's day?

(a) numerical, continuous

(b) *numerical, discrete*

(c) categorical

(d) categorical, ordinal

Question

2. What type of variable is people's age?

(a) *numerical, continuous*

(b) numerical, discrete

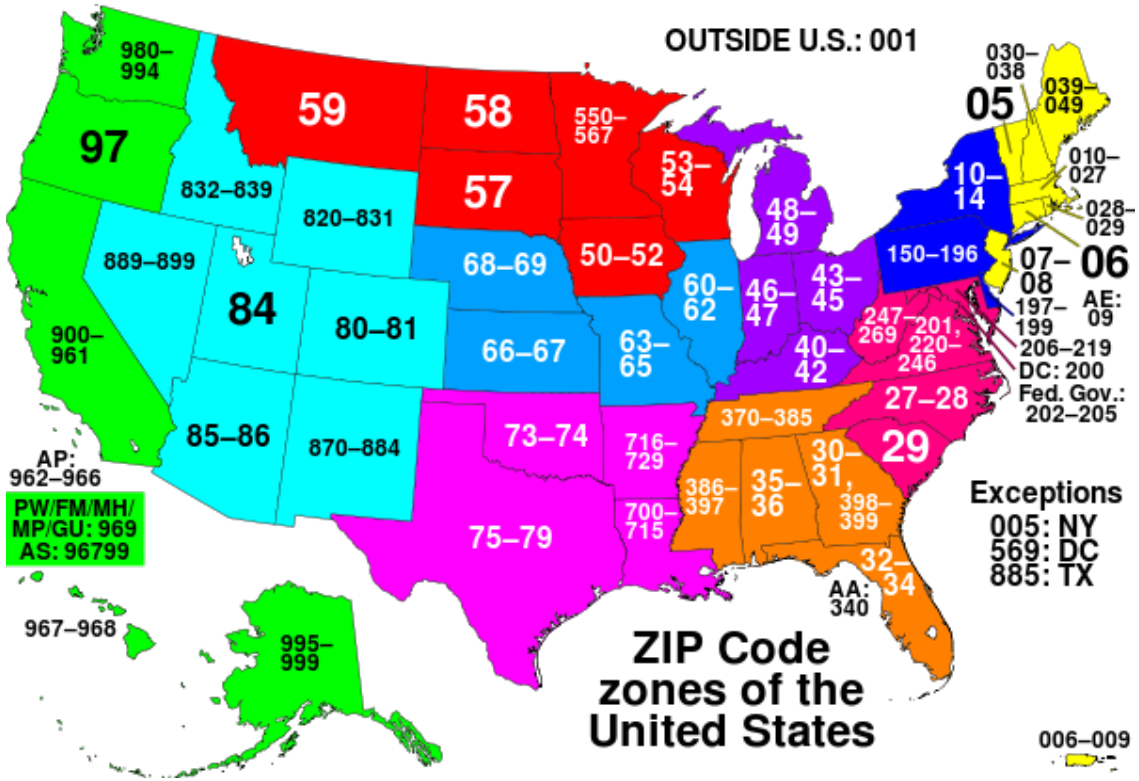
(c) categorical

(d) categorical, ordinal

Question

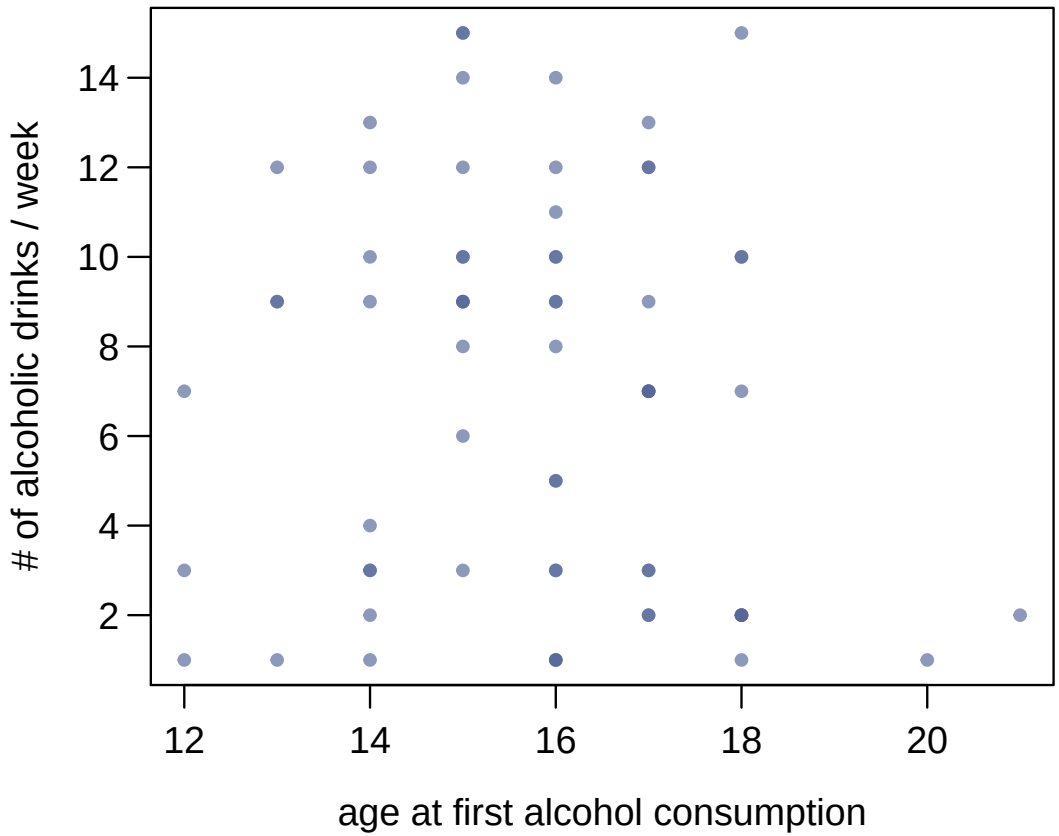
3. What type of variable is a zip code?

(a) numerical, continuous
(b) numerical, discrete
(c) *categorical*
(d) categorical, ordinal



https://en.wikipedia.org/wiki/ZIP_Code

Relationships among variables

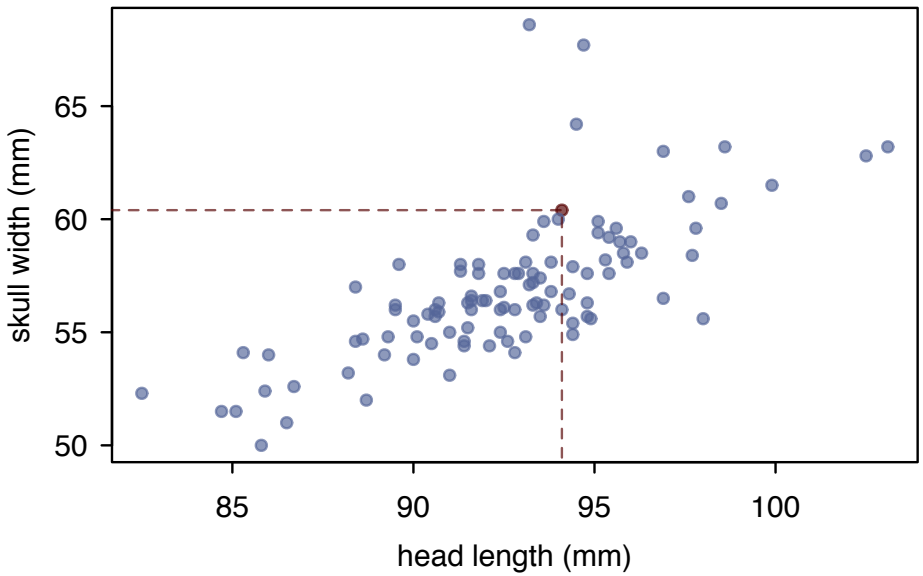


Does there appear to be a relationship between number of alcoholic drinks consumed per week and age at first alcohol consumption?

Associated and independent variables

Question

4. Based on the scatterplot on the right, which of the following statements is correct about the head and skull lengths of possums?



(a) There is no relationship between head length and skull width, i.e. the variables are independent.

(b) *Head length and skull width are positively associated.*

(c) Skull width and head length are negatively associated.

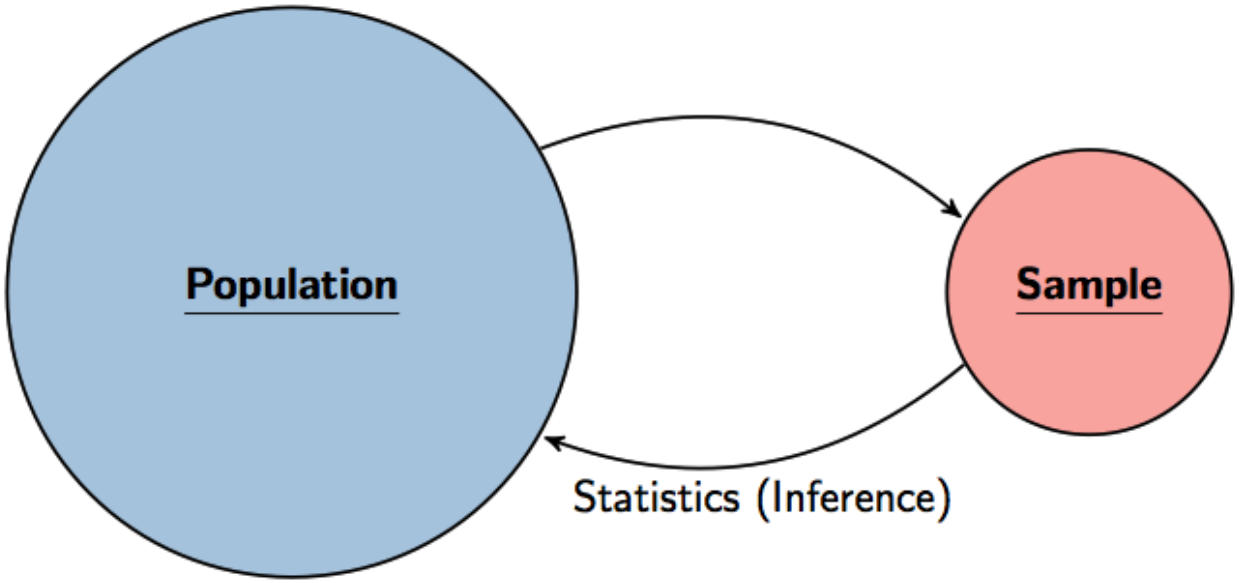
(d) A longer head causes the skull to be wider.

(e) A wider skull causes the head to be longer.

Associated vs. independent

- When two variables show some connection with one another, they are called *associated* variables.
  - Associated variables can also be called *dependent* variables and vice-versa.
- If two variables are not associated, i.e. there is no evident connection between the two, then they are said to be *independent*.
- It is also possible for observations to be independent as well.

Populations and samples



- A *population* is a set of units (usually people, objects, transactions, or events) that we are interested in studying.
- A *sample* is a subset of the units of a population.
- A *statistical inference* is an estimate or prediction or some other generalization about a population based on information.

Populations and samples

Consider the following research questions:

1. What is the average mercury content in swordfish in the Atlantic Ocean?
   population: all swordfish in the Atlantic Ocean
2. Over the last 5 years, what is the average time to degree for SMU undergraduate students?
   population: SMU students who have graduated in the last five years
3. Does the drug sulphinpyrazone reduce the number of deaths in heart attack patients?
   population: all heart attack patients

Note: For question 3, a rigorously way to show the drug being effective is via experiment, to see if patients randomly assigned to take this drug have significant improvement in their health conditions than those randomly assigned to take placebo. Hence, the population may not be restrict to the patients who take the drug. This being said, the answer "all heart attach patients who have used drug sulphinpyrazone" is also acceptable.

Anecdotal evidence

- Let's consider the following possible responses to our research questions:
  1. A man on the news got mercury poisoning from eating swordfish, so the average mercury concentration in swordfish must be dangerously high.
  2. I met two students who took more than 10 years to graduate from SMU, so it must take longer to graduate at SMU than at many other colleges.
  3. My friend's dad had a heart attack and died after they gave him sulphinpyrazone. The drug must not work.
- These are *anecdotal evidence*, based on a limited sample size that might not be representative of the population.
- Anecdotal evidence is typically composed of unusual cases that we recall based on their striking characteristics.

Census

- Wouldn't it be better to just include everyone and "sample" the entire population?
  - Such a special sample is called a *census*.
- There are problems with taking a census:
  - It can be difficult to complete a census: there always seem to be some individuals who are hard to locate or hard to measure. *And there may be a certain characteristic about those individuals who are hard to locate.*
  - Populations rarely stand still. Even if you could take a census, the population changes while you work, so it's never possible to get a perfect measure.
  - Taking a census may be more complex than sampling.

Sampling is natural

- Sampling is a natural..
- Think about sampling something you are cooking - you taste (examine) a small part of what you're cooking to get an idea about the dish as a whole.

Exploratory analysis to inference

- When you taste a spoonful of soup and decide it doesn't taste salty enough, that's *exploratory analysis*.
- If you generalize and conclude that your soup needs salt, that's an *inference*.
- For your inference to be valid the spoonful you tasted (the sample) needs to be *representative* of the entire pot (the population).
  - If your spoonful comes only from the surface and the salt is collected at the bottom of the pot, what you tasted is probably not representative of the whole pot.
  - If you first stir the soup thoroughly before you taste, your spoonful will more likely be representative of the whole pot.

Random vs. biased sampling

- If we want to examine characteristics of all SMU undergraduates, we should take a *random sample* from the population.
- We could do this by getting a list of all SMU undergraduates from the registrar's office and randomly selecting a number of students from that list.
- If we employ a method where *all students are equally likely to be selected*, this would be a *simple random sampling* that is *representative* of the population.

Can you suggest such a method?

- If instead we asked first-year students to sample a number of SMU students and they sample only from students in their classes and dorm, the sample would be *biased* towards first-year students.

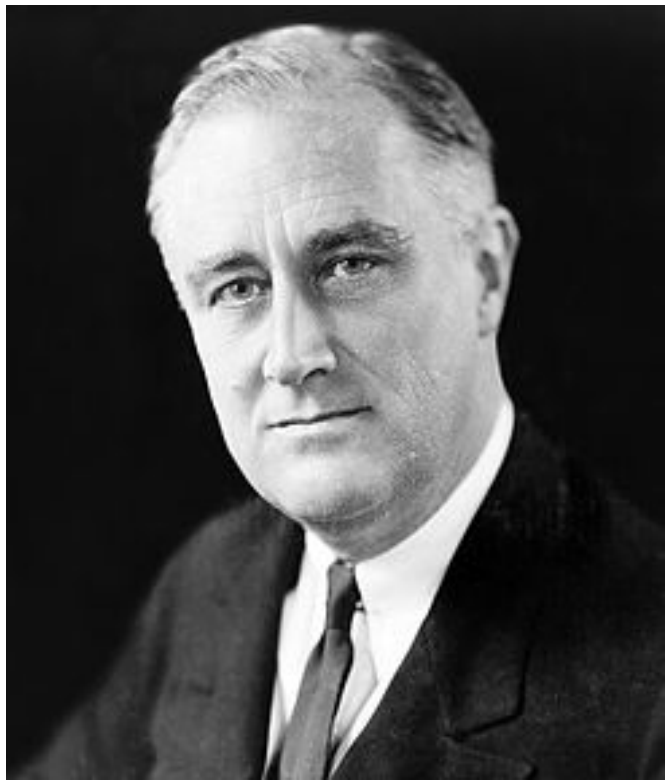
A few sources of bias

- *Non-response bias:* If only a small fraction of the randomly sampled people choose to respond to a survey, the sample may no longer be representative of the population.
- *Voluntary response bias:* Occurs when those who respond have strong opinions on the issue since such a sample will also not be representative of the population.
- *Convenience sample:* Individuals who are easily accessible are more likely to be included in the sample.

Landon vs. FDR

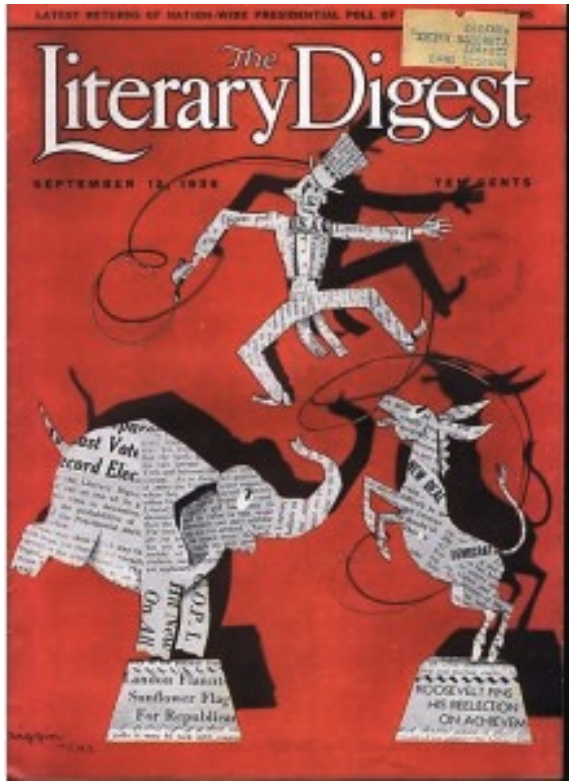
A historical example of a biased sample yielding misleading results:

In 1936, Landon sought the Republican presidential nomination opposing the re-election of FDR.



The Literary Digest Poll

- The Literary Digest polled about 10 million Americans, and got responses from about 2.4 million.
- The poll showed that Landon would likely be the overwhelming winner and FDR would get only 43% of the votes.
- Election result: FDR won, with 62% of the votes.



- The magazine was completely discredited because of the poll, and was soon discontinued.

The Literary Digest Poll - what went wrong?

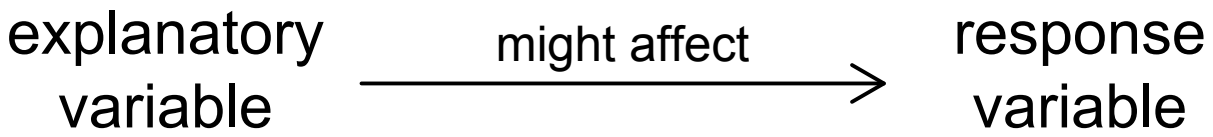
- The magazine had surveyed
  - its own readers,
  - registered automobile owners, and
  - registered telephone users.
- These groups had incomes well above the national average of the day (remember, this is Great Depression era) which resulted in lists of voters far more likely to support Republicans than a truly *typical* voter of the time, i.e. the sample was not representative of the American population at the time.

Large samples are preferable, but...

- The Literary Digest election poll was based on a sample size of 2.4 million, which is huge, but since the sample was *biased*, the sample did not yield an accurate prediction.
- Back to the soup analogy: If the soup is not well stirred, it doesn't matter how large a spoon you have, it will still not taste right. If the soup is well stirred, it doesn't matter whether you have a large or small spoon, it will taste fine either way.

Explanatory and response variables

- To identify the *explanatory variable* and the *response variable* in a pair of variables, identify which of the two is suspected of affecting the other.



- Labeling variables as explanatory and response does not guarantee the relationship between the two is actually causal, even if there is an association identified between the two variables.

Question

5. A study is designed to test the effect of type of light on exam performance of students. 180 students are randomly assigned to three classrooms: one that is dimly lit, another with yellow lighting, and a third with white fluorescent lighting and given the same exam. Which is correct?

(a) explanatory: dimly lit, yellow, white fluorescent
response: exam performance

(b) explanatory: exam performance
response: dimly lit, yellow, white fluorescent

(c) *explanatory: type of light (categorical with 3 levels)*
response: exam performance

(d) explanatory: exam performance
response: type of light (categorical with 3 levels)

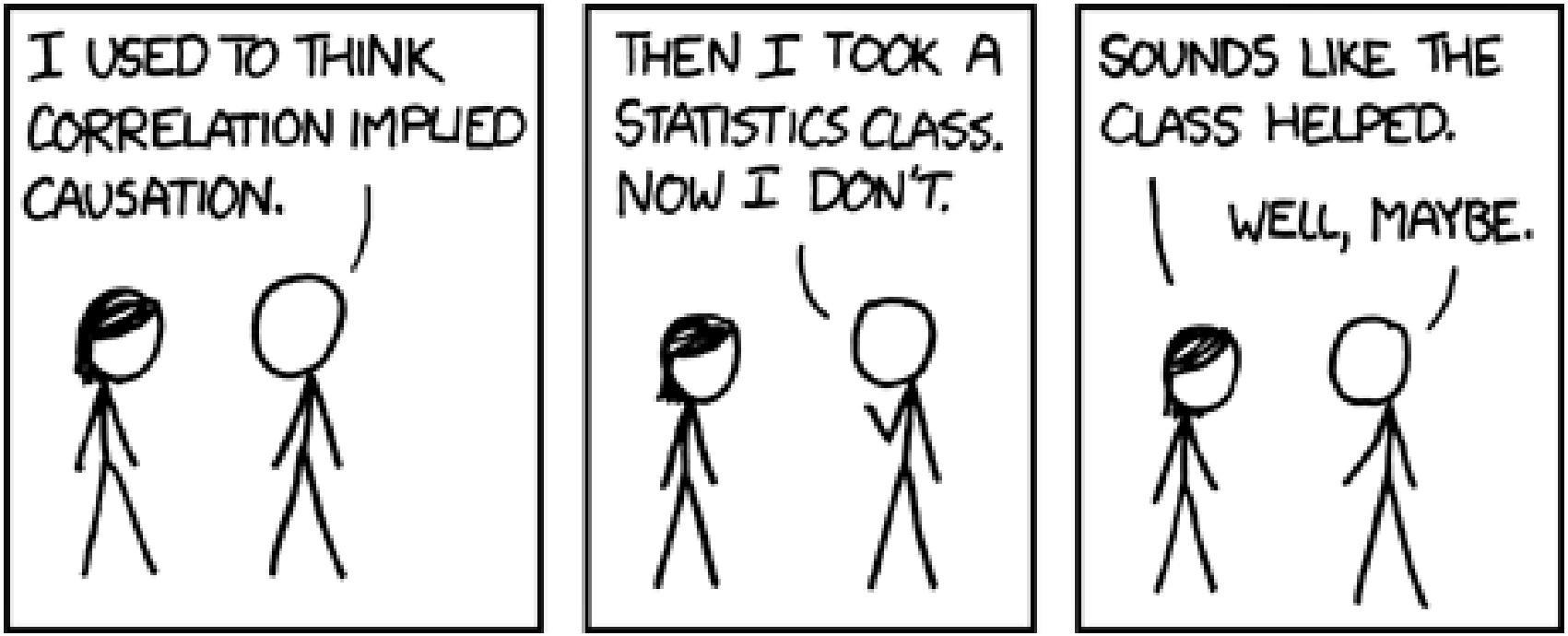
Recap: random sampling

- Random sampling: each case has equal chance of being selected.
- Random sampling yields a sample that is representative of the population.
- Conclusions based on studies using randomly sampled data can be extended to the population at large.
- If a study is based on a non-random (non-representative) sample, conclusions are only true for that particular sample and may not be true for the population at large.

Observational studies and experiments

- *Observational study:* Researchers collect data in a way that *does not directly interfere with how the data arise*, i.e. they merely "observe", and can only establish an association between the explanatory and response variables.
- *Experiment:* Researchers *randomly assign* subjects to various treatments in order to be able *to establish causal connections* between the explanatory and response variables (also called a *randomized experiment*).
- If you're going to walk away with one thing from this class, let it be **"correlation does not imply causation"**.

Correlation does not imply causation



Note: *http://xkcd.com/*

New study sponsored by General Mills says that eating breakfast makes girls thinner

EMAIL THREAD

Study: Breakfast Helps Girls Stay Slim
I love these studies....and finding out who sponsored them!

By ALEX DOMINGUEZ, Associated Press

Girls who regularly ate breakfast, particularly one that includes cereal, were slimmer than those who skipped the morning meal, according to a study that tracked nearly 2,400 girls for 10 years.

Girls who ate breakfast of any type had a lower average body mass index, a common obesity gauge, than those who said they didn't. The index was even lower for girls who said they ate cereal for breakfast, according to findings of the study conducted by the Maryland Medical Research Institute. The study received funding from the National Institutes of Health and cereal-maker General Mills.

"Not eating breakfast is the worst thing you can do, that's really the take-home message for teenage girls," said study author Bruce Barton, the Maryland institute's president and CEO.

The fiber in cereal and healthier foods that normally accompany cereal, such as milk and orange juice, may account for the lower body mass index among cereal eaters, Barton said.

The results were gleaned from a larger NIH survey of 2,379 girls in California, Ohio and Maryland who were tracked between ages 9 and 19. Results of the study appear in the September issue of the Journal of the American Dietetic Association.

Nearly one in three adolescent girls in the United States is overweight, according to the association. The problem is particularly troubling because research shows becoming overweight as a child can lead to a lifetime struggle with obesity.

As part of the survey, the girls were asked once a year what they had eaten during the previous three days. The data were adjusted to compensate for factors such as differences in physical activity among the girls and normal increases in body fat during adolescence.

A girl who reported eating breakfast on all three days had, on average, a body mass index 0.7 units lower than a girl who did not eat breakfast at all. If the breakfast included cereal, the average was 1.65 units lower, the researchers found.

What type of study is this, observational study or an experiment?

"Girls who regularly ate breakfast, particularly one that includes cereal, were slimmer than those who skipped the morning meal, according to a study that tracked nearly 2,400 girls for 10 years. [...] As part of the survey, the girls were asked once a year what they had eaten during the previous three days."

This is an observational study since the researchers merely observed the behavior of the girls (subjects) as opposed to imposing treatments on them.

What is the conclusion of the study?

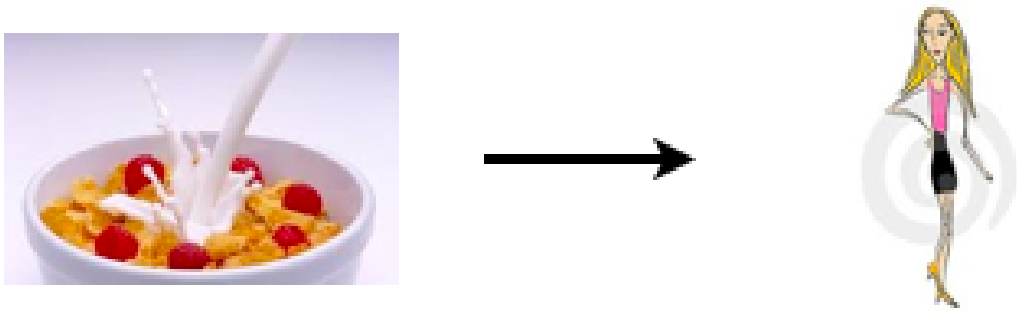
There is an association between girls eating breakfast and being slimmer.

Who sponsored the study?

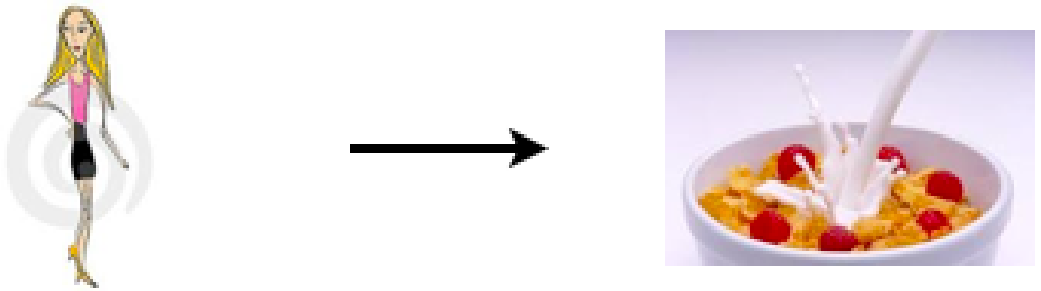
General Mills.

3 possible explanations:

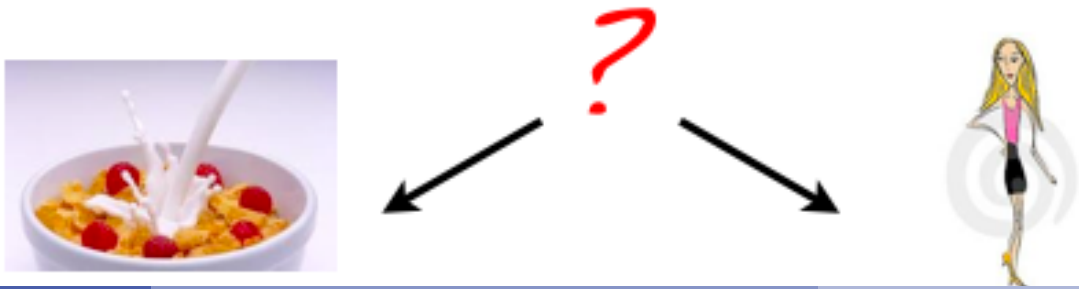
1. Eating breakfast causes girls to be thinner.



2. Being thin causes girls to eat breakfast.



3. A third variable is responsible for both. What could it be?



Lurking variables

- An extraneous variable that affects both the explanatory and the response variable and that make it seem like there is a relationship between the two are called *lurking* or *confounding* variables.
- Observational studies do not control for lurking variables hence they can't be used for establishing causality but they are useful for discovering trends and possible relationships.

Types of observational studies

- A *prospective study* identifies individuals and collects information as events unfold.
  - Ex: Nurses' Health Studies for investigating factors that influence women's health (*http://www.channing.harvard.edu/nhs*).
- *Retrospective studies* collect data after events have taken place.
  - Ex: Use medical records from past patients to estimate the optimum time to start anticoagulant therapy after surgery.

Experiments

- In an *experiment* researchers randomly assign subjects (also called *experimental units*) to *treatment* and *control* groups.
- The objective of *random assignment* (or *randomization*) is to ensure that the treatment and control groups are similar in all characteristics, except for the treatment being investigated, so that any observed difference between the two groups is due to the treatment.
- While observational studies can be used to only infer association, experiments allow researchers to make *causal* statements.

Principles of experimental design

1. *Control:* Compare treatment of interest to a control group.
2. *Randomize:* Randomly assign subjects to treatments.
3. *Replicate:* Within a study, replicate by collecting a sufficiently large sample. Or replicate the entire study.
4. *Block:* If there are variables that are known or suspected to affect the response variable, first group subjects into *blocks* based on these variables, and then randomize cases within each block to treatment groups.

More on blocking

- We would like to design an experiment to investigate if energy gels makes you run faster:
  - Treatment: energy gel
  - Control: no energy gel
- It is suspected that energy gels might affect pro and amateur athletes differently:
  - Block for pro status
  - Divide the sample to pro and amateur
  - Randomly assign pro athletes to treatment and control groups
  - Randomly assign amateur athletes to treatment and control groups

Can you think of any other variables we should block for?

Question

6. A study is designed to test the effect of light level and noise level on exam performance of students. The researcher also believes that light and noise levels might have different effects on males and females, so wants to make sure both genders are represented equally under different conditions. Which of the below is correct?

(a) There are 3 explanatory variables (light, noise, gender) and 1 response variable (exam performance)

(b) *There are 2 explanatory variables (light and noise), 1 blocking variable (gender), and 1 response variable (exam performance)*

(c) There is 1 explanatory variable (gender) and 3 response variables (light, noise, exam performance)

(d) There are 2 blocking variables (light and noise), 1 explanatory variable (gender), and 1 response variable (exam performance)

Reducing bias in human experiments

- Randomized experiments are the gold standard, but they do not ensure an unbiased perspective into the cause and effect relationships in all cases.
- Bias can unintentionally arise especially in human studies.
- Using a control for the treatment is one way to reduce this bias.
- Keeping the patients uninformed about their treatment, *blinding*, can also help reduce bias. This is usually done through the use of a *placebo*.
- If both the patients and the doctors are uninformed about which treatment which patient is getting, this is called a *double-blind* study.

Question

7. What is the main difference between observational studies and experiments?

(a) Experiments take place in a lab while observational studies do not need to.

(b) In an observational study we only look at what happened in the past.

(c) *Experiments use random assignment while observational studies do not.*

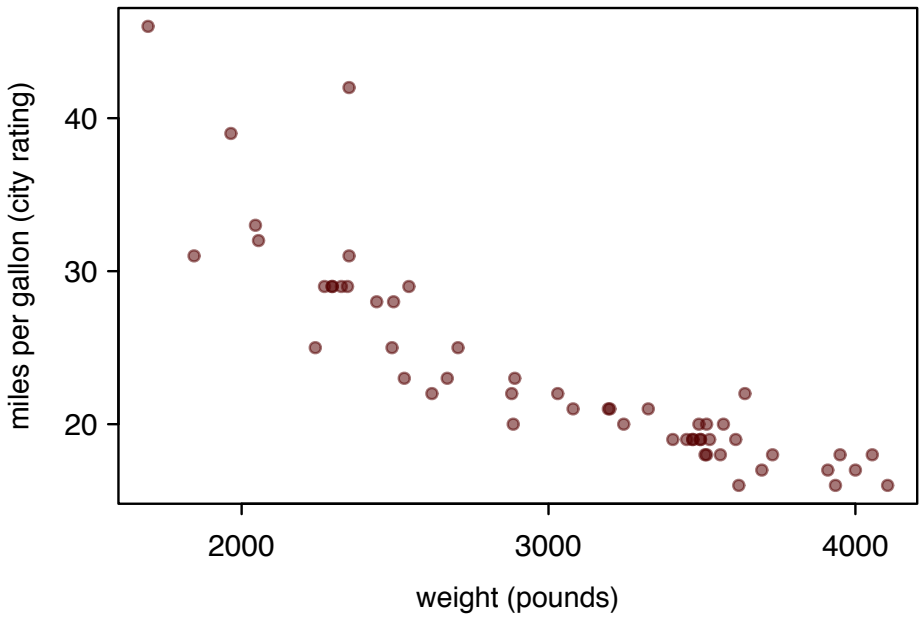
(d) Observational studies are completely useless since no causal inference can be made based on their findings.

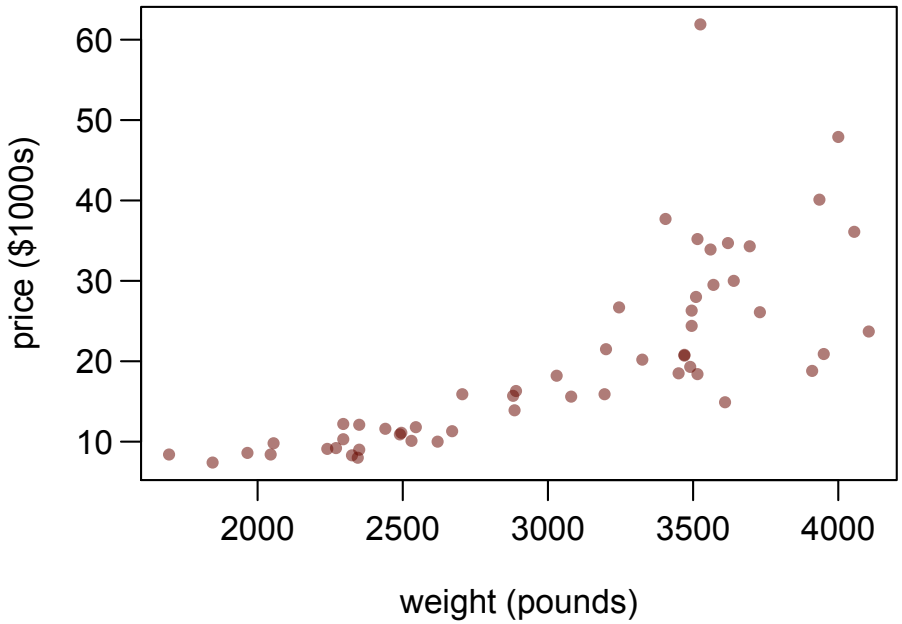
Random assignment vs. random sampling

	Random assignment	No random assignment	
Random sampling	Causal inference, generalized to the whole population.	No causal inference, correlation statement generalized to the whole population.	Conclusions generalized to population.
No random sampling	Causal inference, only for the sample.	No causal inference, correlation statement only for the sample.	Conclusions not generalized to population.
	Causation	Correlation	

Cars: ... vs. weight

From the `cars` data:



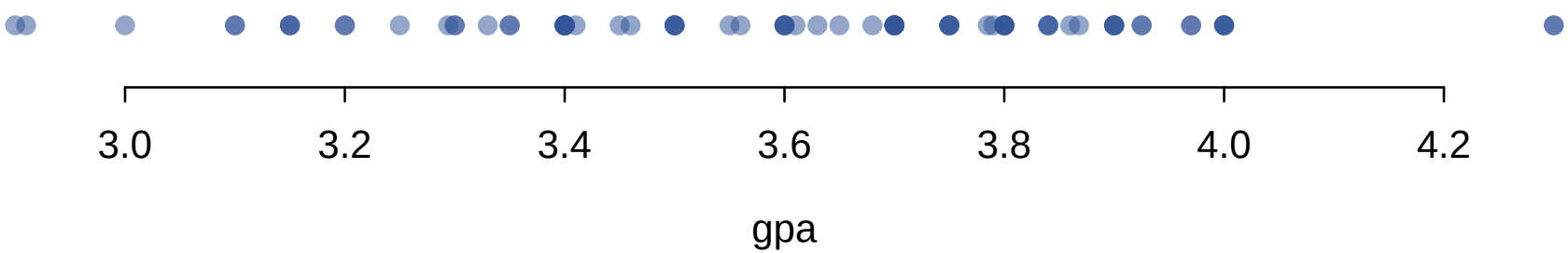


What do these scatterplots reveal about the data? How might they be useful?

- Relationship exists or not?
- If exists, positive or negative?

Dot plots

Useful for visualizing one numerical variable.

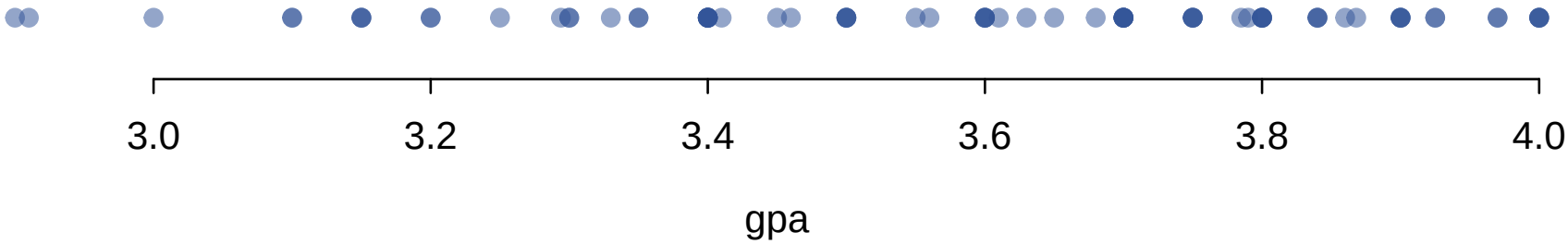


Do you see anything out of the ordinary?

GPA cannot be greater than 4.

Dot plots (cont.)

Useful for visualizing one numerical variable. Darker colors represent areas where there are more observations.

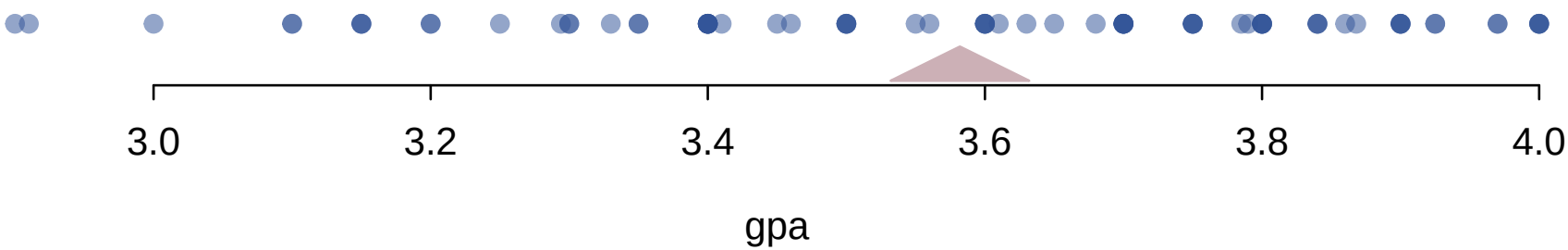


How would you describe the distribution of GPAs in this data set? Make sure to say something about the

- center
- shape
- spread

of this distribution.

Dot plots & mean



- The *mean*, also called the *average* (marked with a triangle in the above plot), is one way to measure the center of a *distribution* of data.
- The mean GPA is 3.58.

Mean

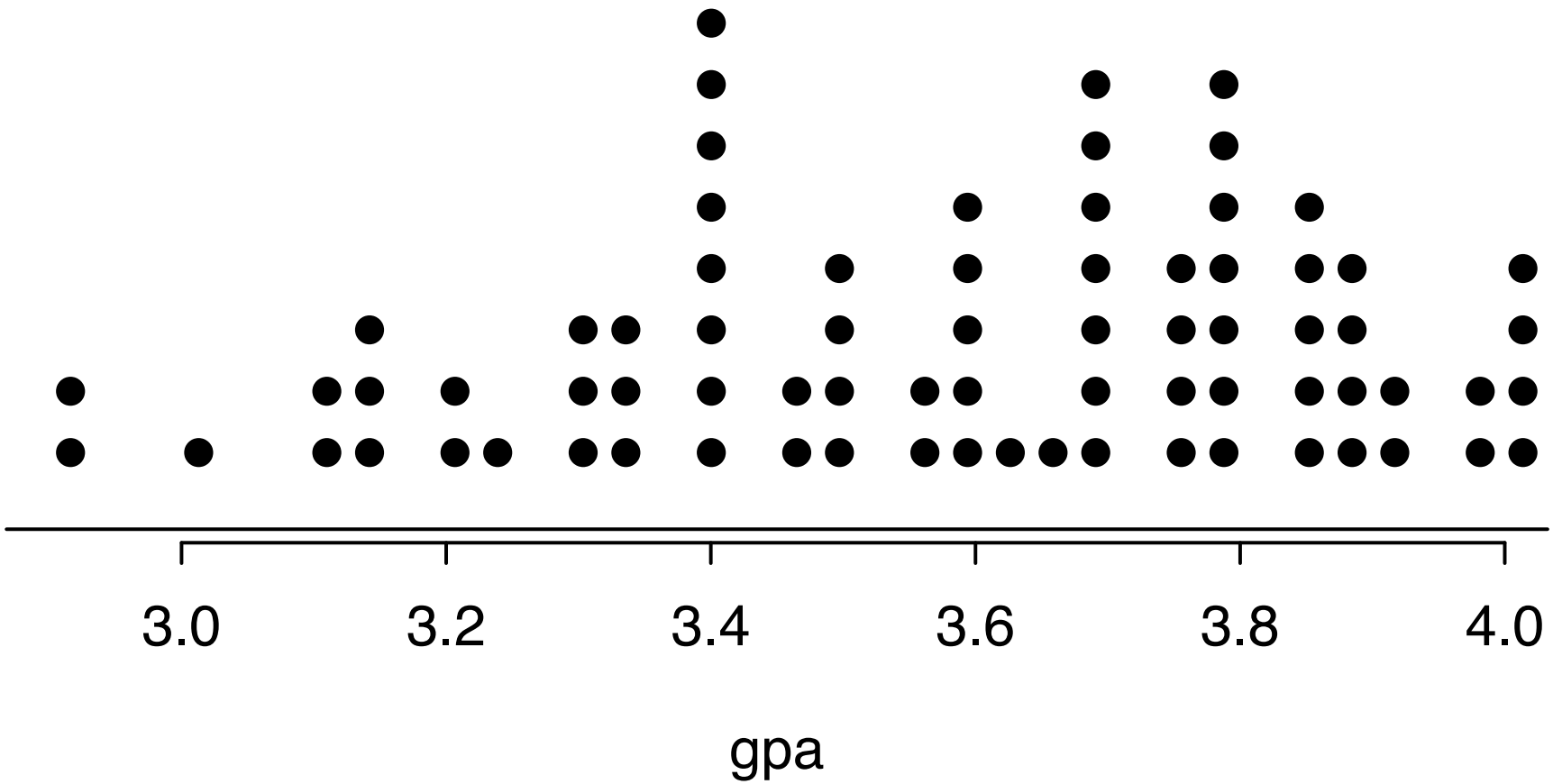
- The *sample mean*, denoted as $\bar{x}$, can be calculated as

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

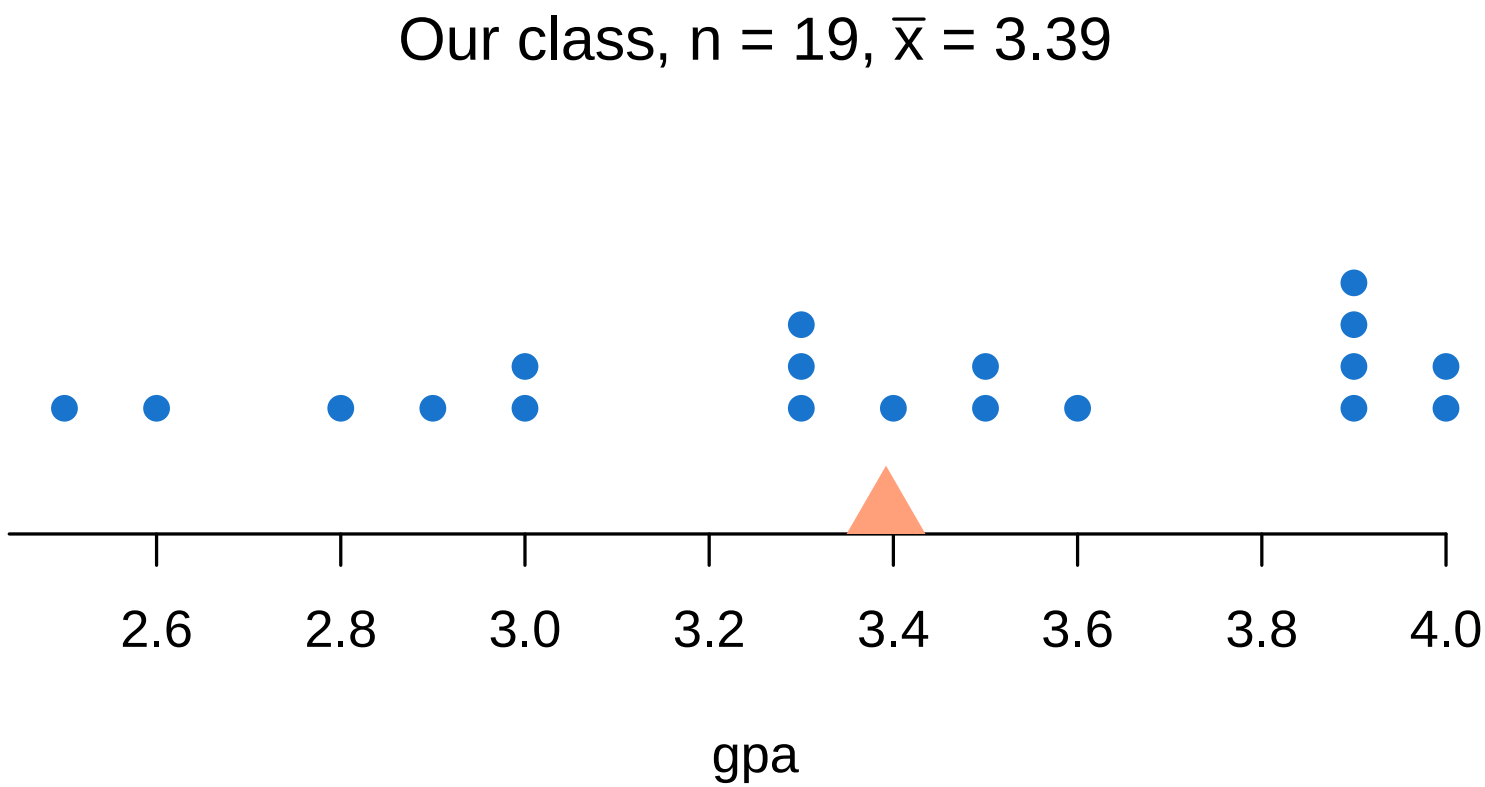
  where $x_1, x_2, \cdots, x_n$ represent the n observed values.
- The *population mean* is also computed the same way but is denoted as μ. It is often not possible to calculate μ since population data is rarely available.
- The sample mean is a *sample statistics*, or a *point estimate* of the population mean. This estimate may not be perfect, but if the sample is good (representative of the population) it is usually a good guess.

Stacked dot plot

Higher bars represent areas where there are more observations, makes it a little easier to judge the center and the shape of the distribution.

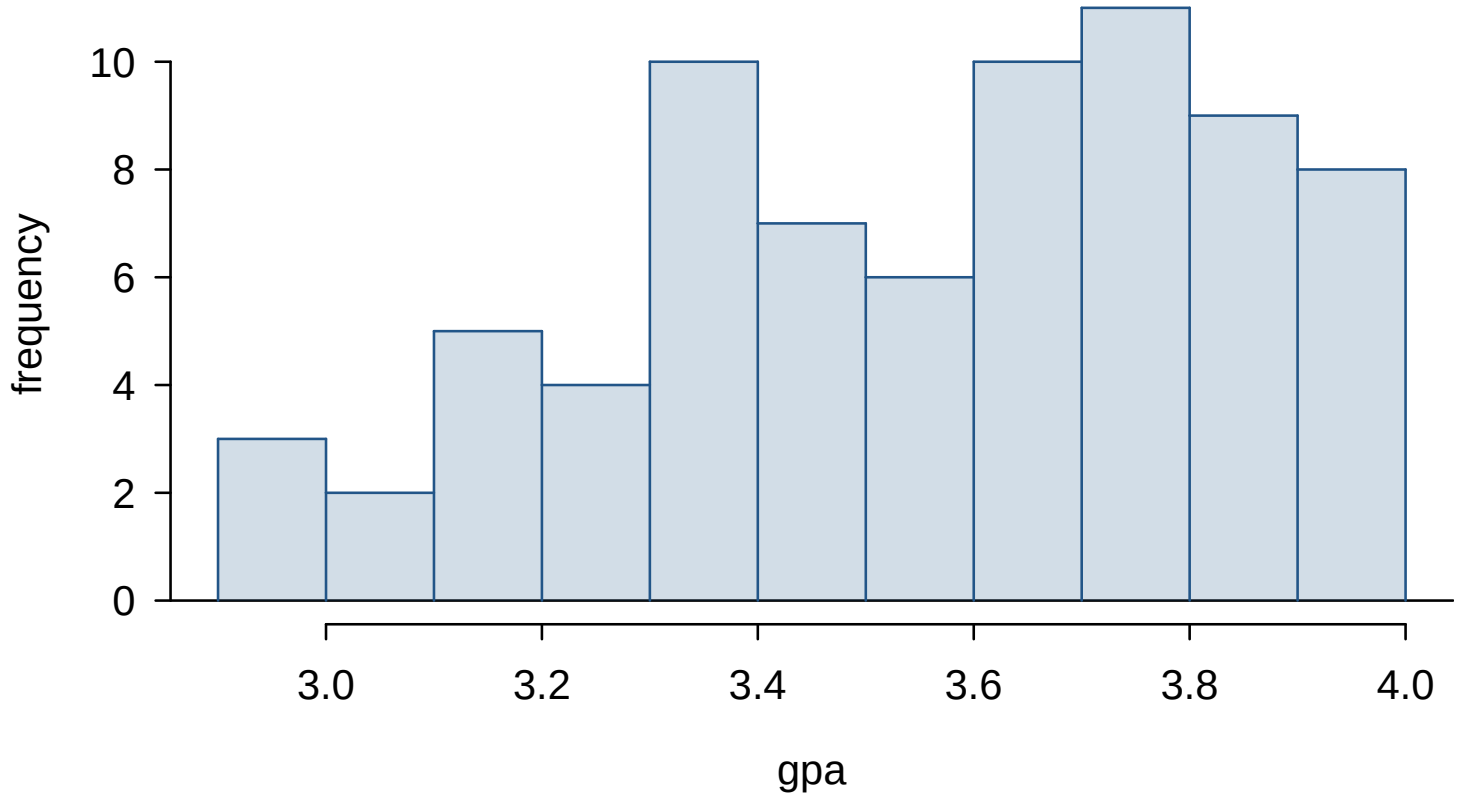


GPA of our class (rounded to 1 decimal place)



Histograms - GPA

Higher bars represent areas where there are more observations, preferable when sample size is large but hides finer details like individual observations.



Note: *Observations that fall on the boundary of a bin (e.g., 3.0) are allocated to the lower bin.*

Anatomy of a histogram

Order the data in ascending order:

```
2.900 2.910 3.000 3.100 3.100 3.150 3.150 3.150 3.200 3.200 3.250 3.294 3.300 3.300 3.330
3.350 3.350 3.400 3.400 3.400 3.400 3.400 3.400 3.400 3.410 3.450 3.460 3.500 3.500 3.500
3.500 3.550 3.560 3.600 3.600 3.600 3.600 3.610 3.630 3.650 3.680 3.700 3.700 3.700 3.700
3.700 3.700 3.750 3.750 3.750 3.750 3.785 3.790 3.800 3.800 3.800 3.800 3.800 3.840 3.840
3.840 3.860 3.868 3.900 3.900 3.900 3.900 3.925 3.925 3.970 3.970 4.000 4.000 4.000 4.000
4.300 4.300
```

Make a *frequency table* where the number of observations that fall in a certain bin are recorded by counting how many observations fall in each bin. Let's use a bin width of 0.1:

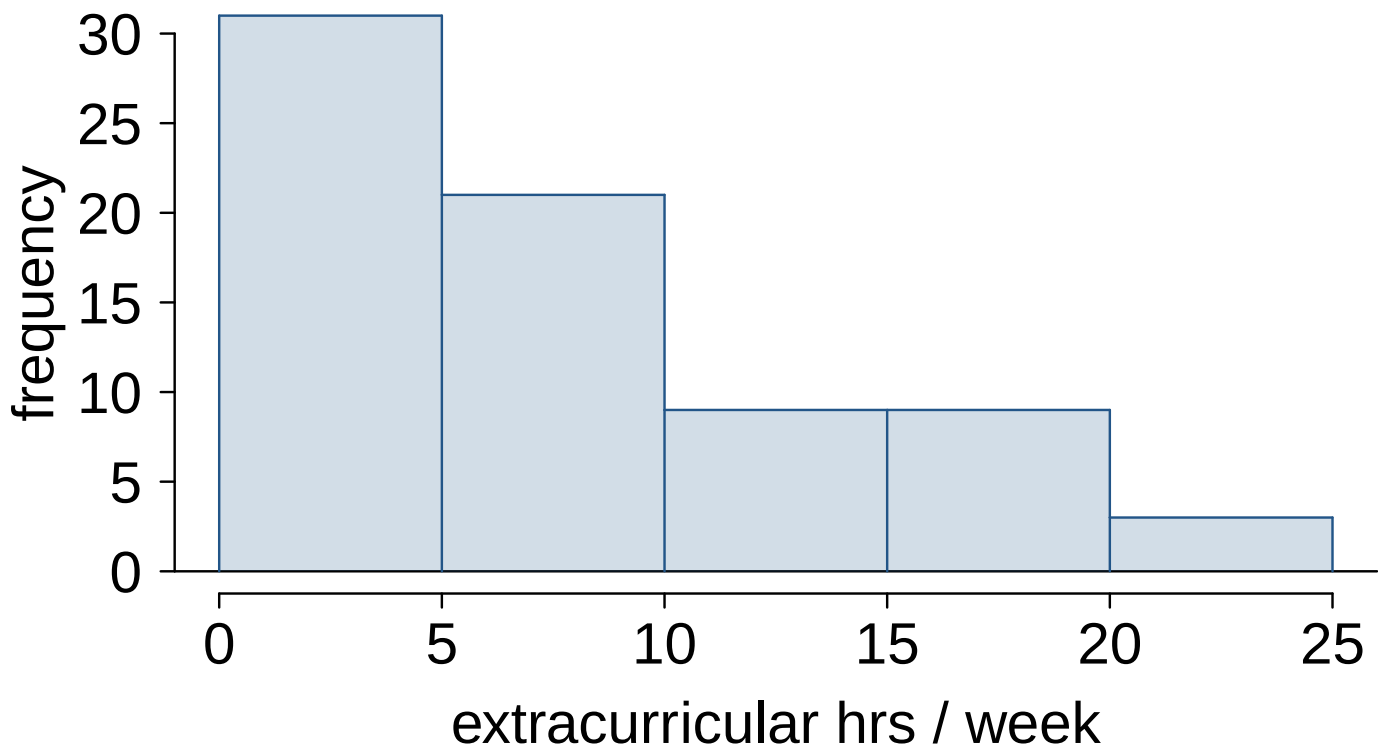
GPA	2.9 to 3	3 to 3.1	3.1 to 3.2	3.2 to 3.3	···	3.8 to 3.9	3.9 to 4
Count					···		

GPA	*2.9 to 3*	*3 to 3.1*	*3.1 to 3.2*	*3.2 to 3.3*	*···*	*3.8 to 3.9*	*3.9 to 4*
Count	*3*	*2*	*5*	*4*	*···*	*9*	*8*

Note: Histogram is shown on the previous slide.

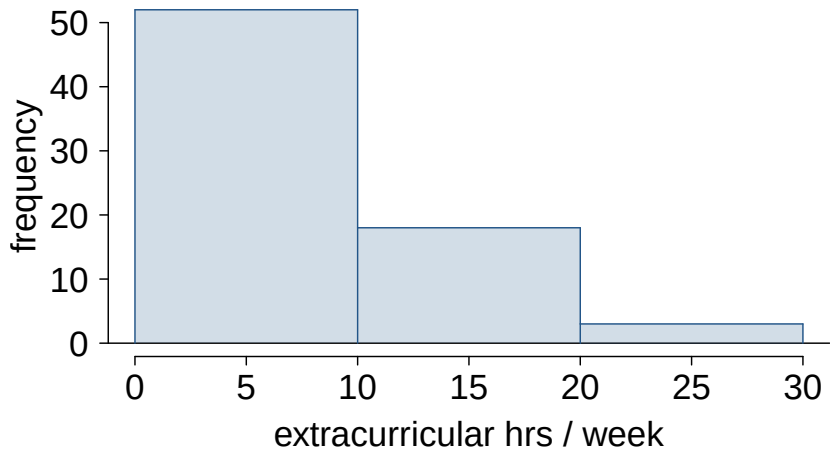
Histograms - Extracurricular hours

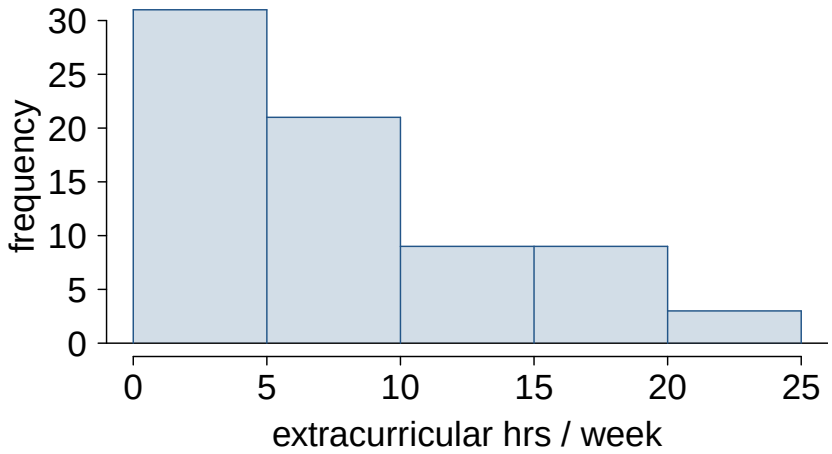
- Histograms provide a view of the *data density*. Higher bars represent where the data are relatively more common.
- Histograms are especially convenient for describing the *shape* of the data distribution.
- The chosen *bin width* can alter the story the histogram is telling.

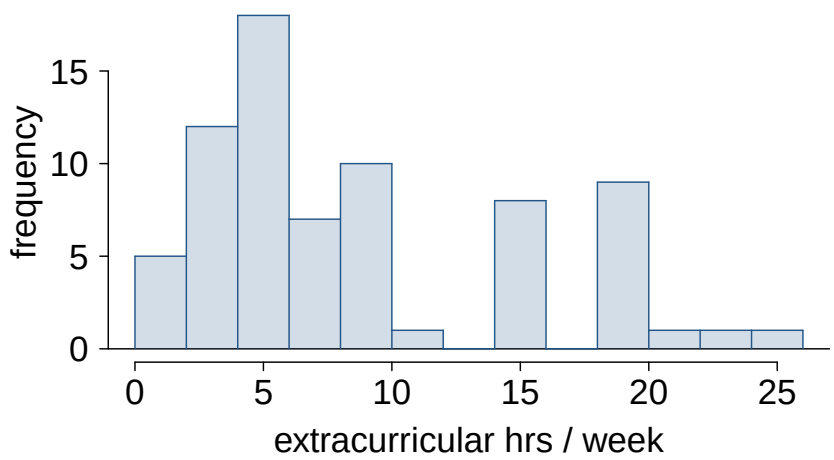


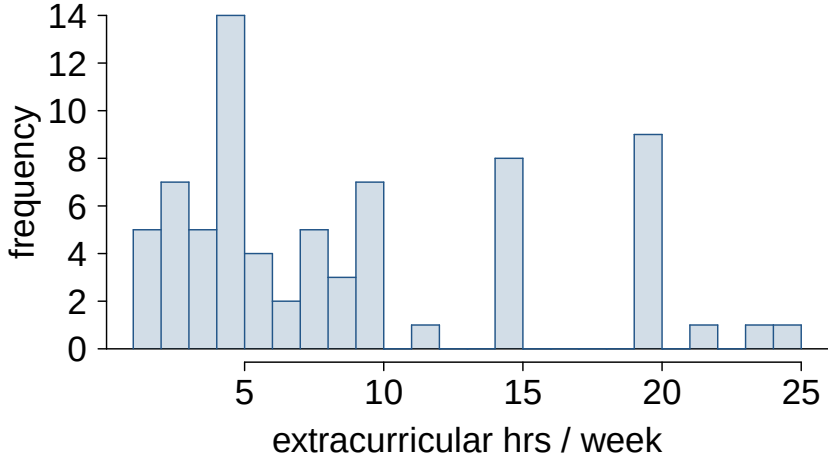
Bin width

Which one(s) of these histograms are useful? Which reveal too much about the data? Which hide too much?



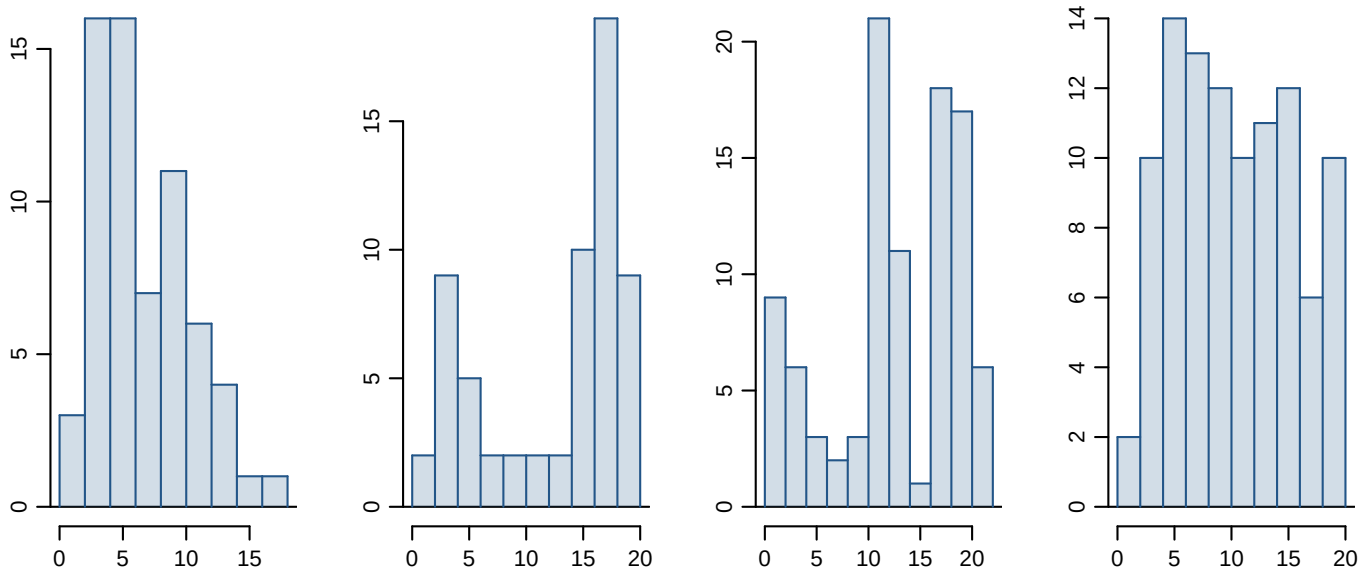






Shape of a distribution: 1. modality

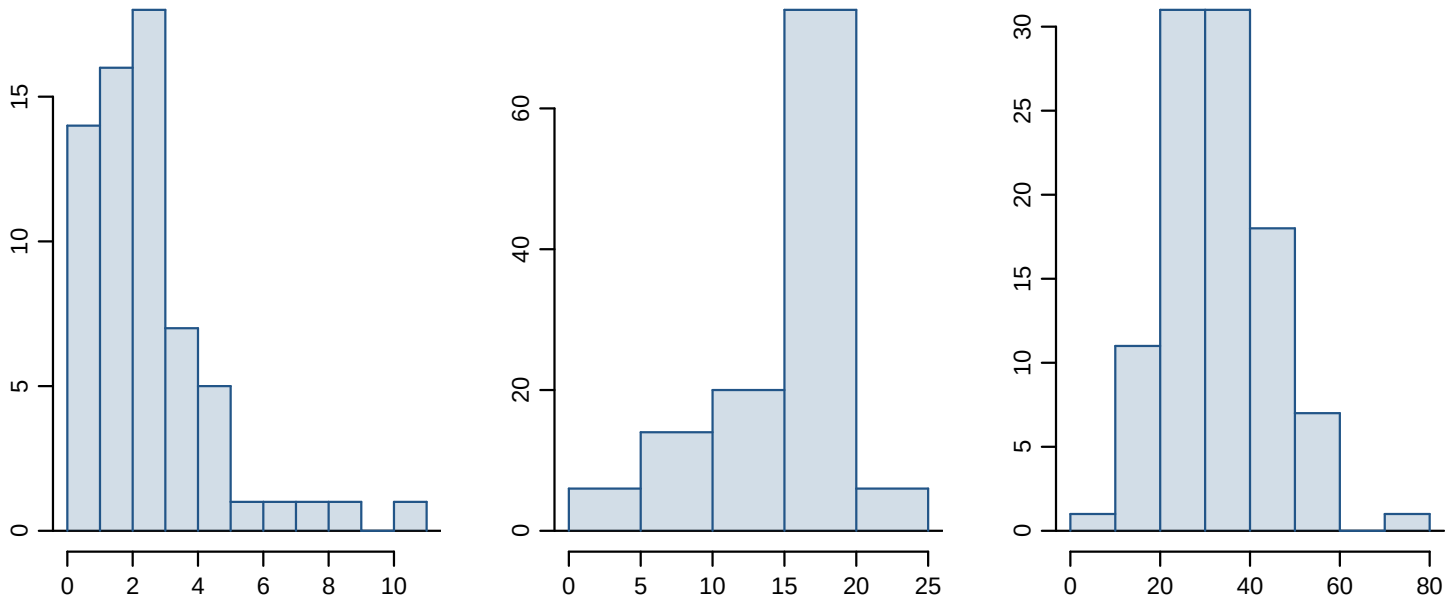
The *mode* is defined as the most frequent observation in the data set. Does the histogram have a single prominent peak (*unimodal*), several prominent peaks (*bimodal/multimodal*), or no apparent peaks (*uniform*)?



Note: In order to determine modality, it's best to step back and imagine a smooth curve over the histogram. Use the limp spaghetti method.

Shape of a distribution: 2. skewness

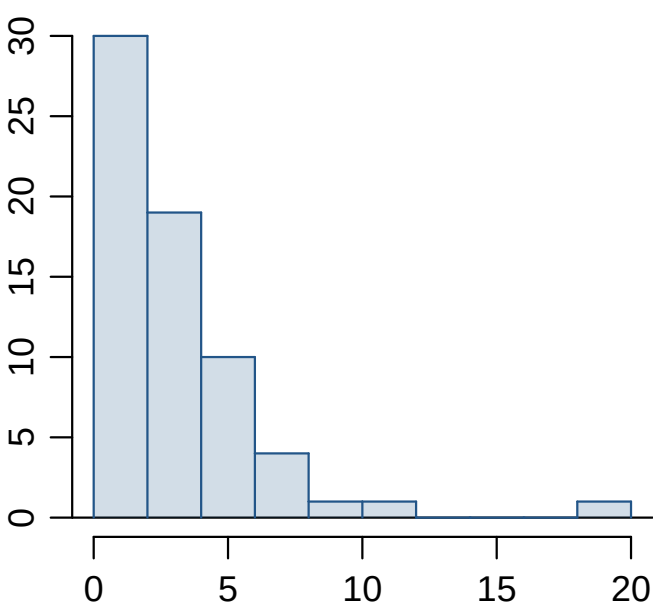
Is the histogram *right skewed*, *left skewed*, or *symmetric*?

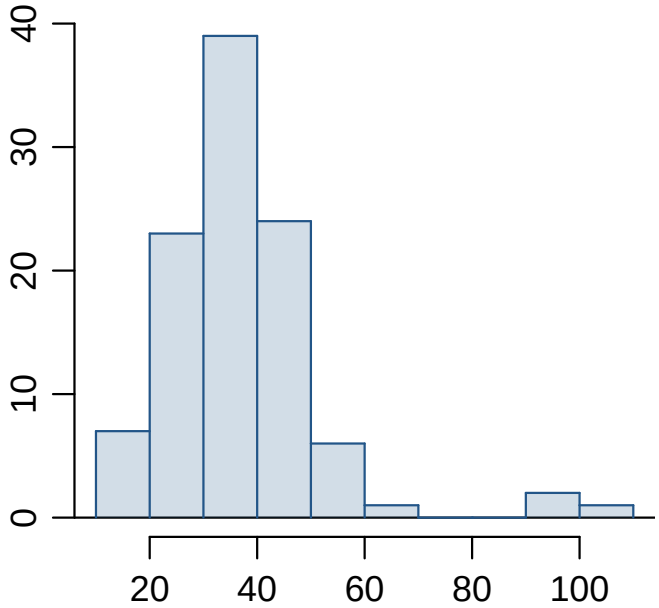


Note: *Histograms are said to be skewed to the side of the long tail.*

Shape of a distribution: 3. unusual observations

Are there any unusual observations or potential *outliers*?





Question

8. Which of these variables do you expect to be uniformly distributed?

(a) weights of adult females

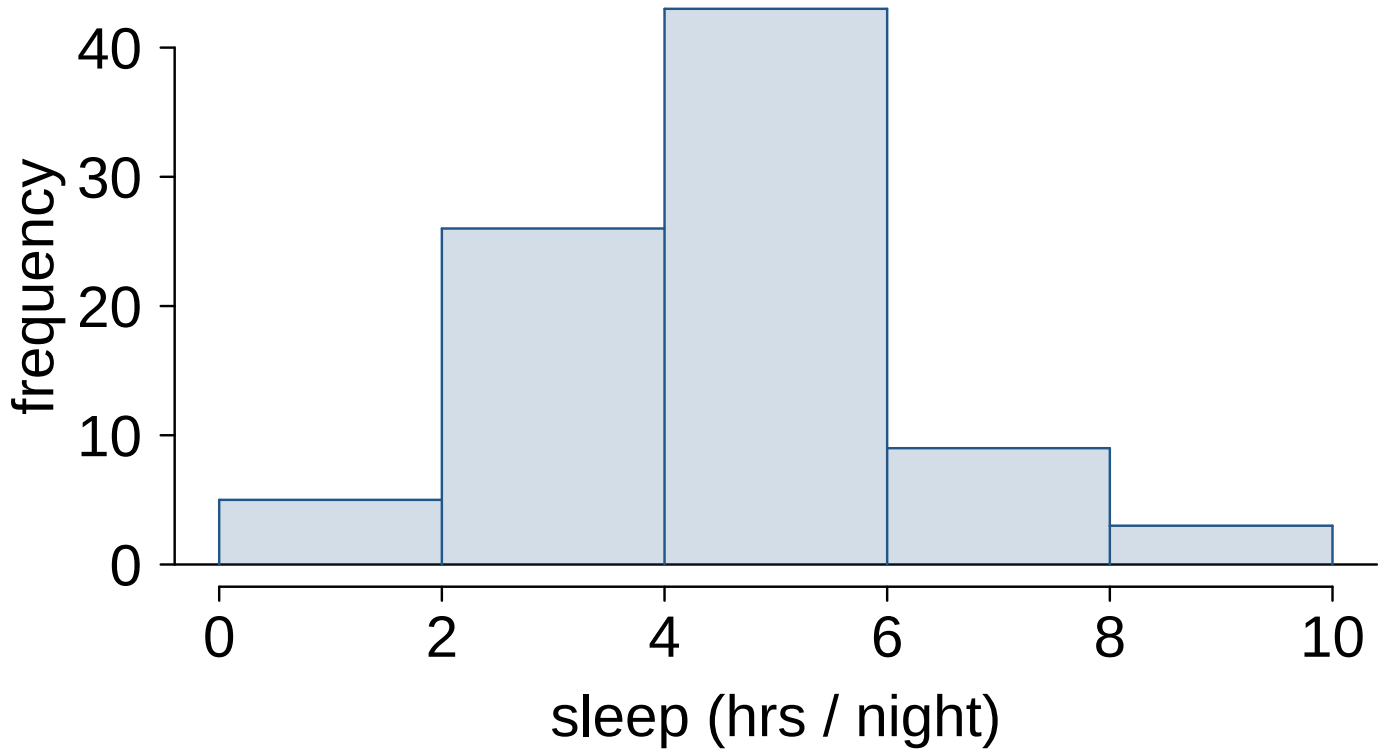
(b) salaries of a random sample of people from North Carolina

(c) exam scores

(d) *birthdays of classmates (day of the month)*

Variability in data

How would you describe the amount of variability in the number of hours of sleep students get per night?



Deviation

The distance of an observation from the mean is its *deviation*: $x_i - \bar{x}$.

1 1 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
4 4 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5
5 5 5 5 5 5 5 5 5 5 5 5 5 5 6 6 7 7 7 7 7 7 7 7 8 9 9 9

$$\bar{x} = 4.6$$

$$x_1 - \bar{x} = 1 - 4.6 = -3.6$$

$$x_2 - \bar{x} = 1 - 4.6 = -3.6$$

$$\vdots$$

$$x_{86} - \bar{x} = 9 - 4.6 = 4.4$$

Variance

Sample variance, s^2, is roughly the average squared deviation from the mean.

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

Note: When calculating the sample variance we divide by $n-1$ instead of n.

What is the variance of amount of sleep students get per night?

$$s^2 = \frac{(-3.6)^2 + (-3.6)^2 + \cdots + (4.4)^2}{86-1} = \frac{12.96 + 12.96 + \cdots + 19.36}{85} = 2.76$$

Variance (cont.)

Why do we use the squared deviation in the calculation of variance?

- To get rid of negatives so that observations equally distant from the mean are weighed equally.
- To weigh larger deviations more heavily

Standard deviation

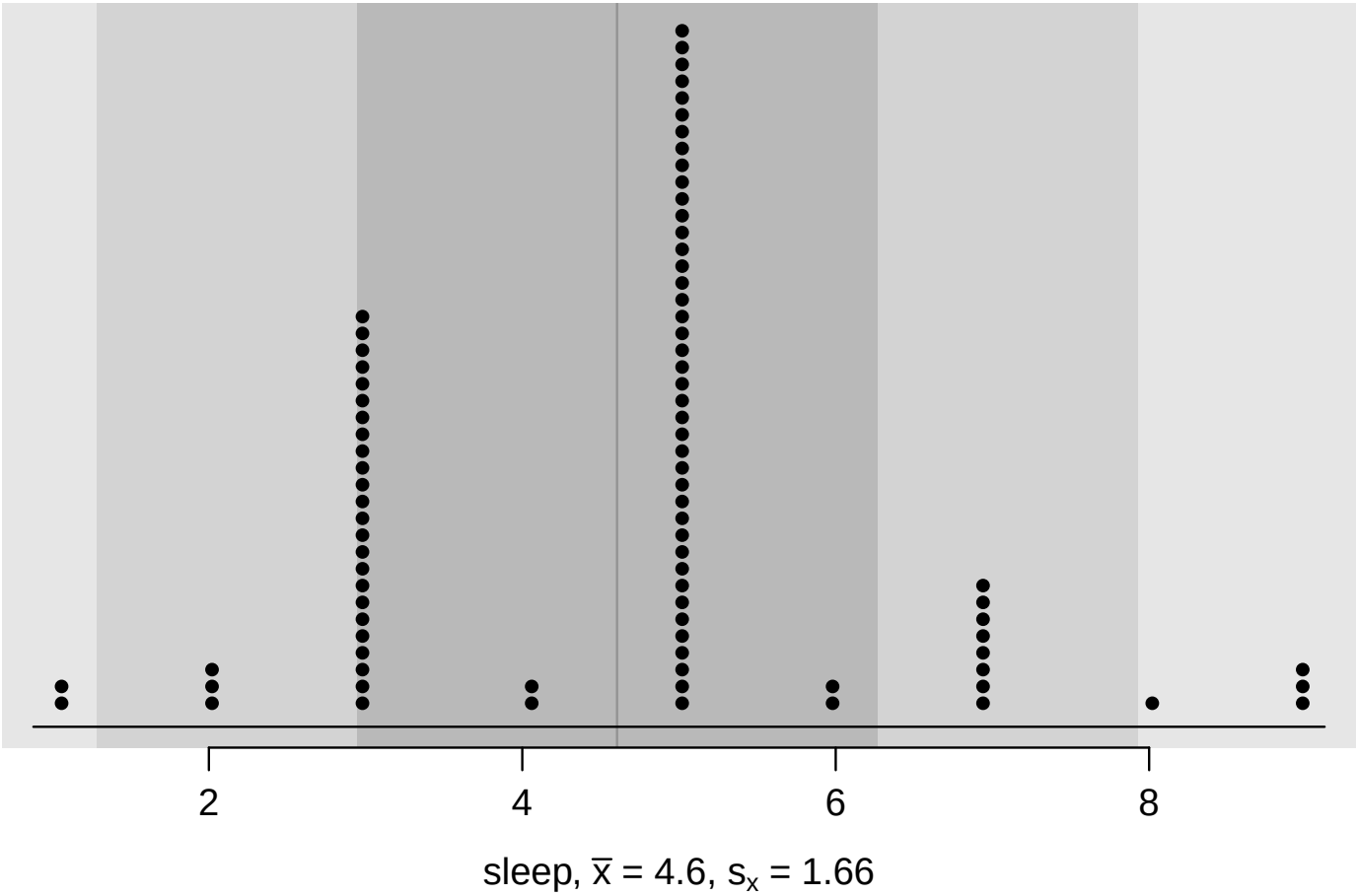
Sample standard deviation, s, is the square root of the variance.

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

The standard deviation of car prices can be calculated as:

$$s = \sqrt{2.76} = 1.66$$

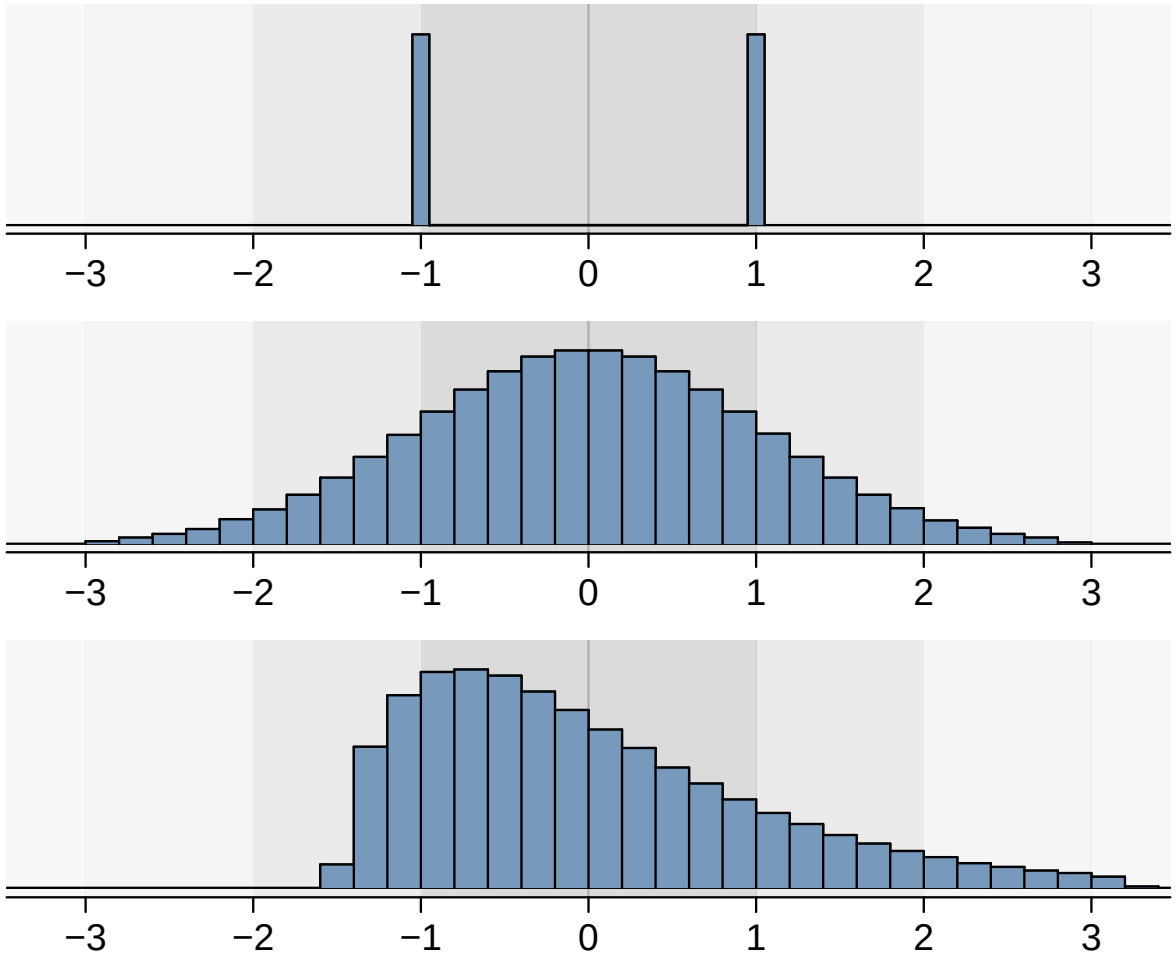
Variability in sleep time

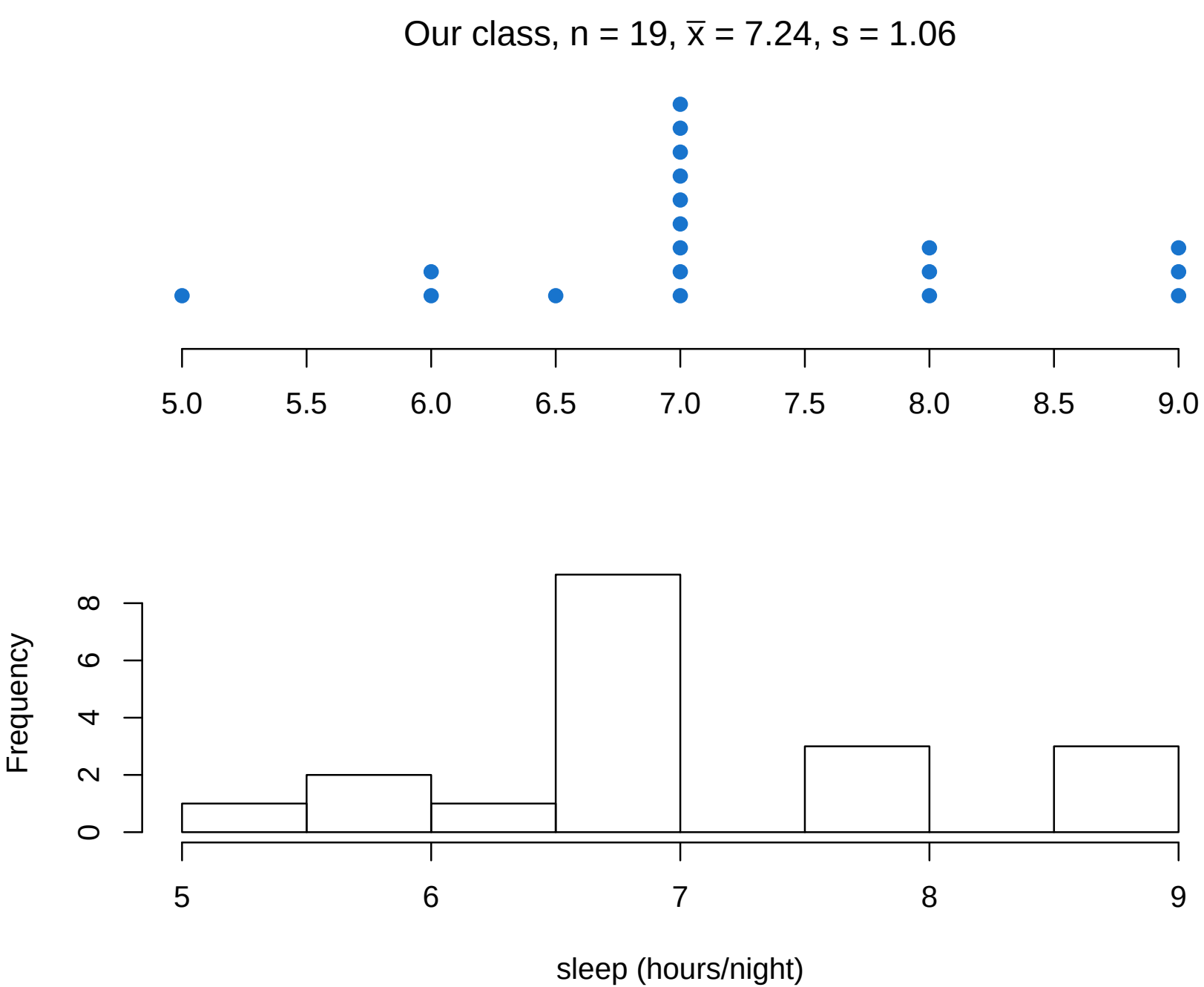


- 69 out of 86 students (80%) are within 1 SD of the mean.
- 80 out of 86 students (93%) are within 2 SDs of the mean.
- 86 out of 86 students (100%) are within 3 SDs of the mean.

Describing distributions

When describing distributions make sure to talk about the shape, center, spread, and if any, unusual observations.




Our class, n = 19, x̄ = 7.24, s = 1.06
5.0
5.5
6.0
6.5
7.0
7.5
8.0
8.5
9.0
Frequency
0
2
4
6
8
5
6
7
8
9
sleep (hours/night)

Notations

	mean	variance	SD
sample	$\bar{x}$	s^2	s
population	μ	σ^2	σ

Do you see a trend in what types of letters are used for sample statistics vs. population parameters?

Latin letters for sample statistics, Greek letters for population parameters.

Median

- The *median* is the value that splits the data in half when ordered in ascending order.

$$0, 1, 2, 3, 4$$

- If there are an even number of observations, then the median is the average of the two values in the middle.

$$0, 1, \underline{2, 3}, 4, 5 \rightarrow \frac{2+3}{2} = 2.5$$

- Since the median is the midpoint of the data, 50% of the values are below it. Hence, it is also the 50^{th} *percentile*.

Q1, Q3, and IQR

- The 25^{th} percentile is also called the first quartile, *Q1*.
- The 50^{th} percentile is also called the median.
- The 75^{th} percentile is also called the third quartile, *Q3*.

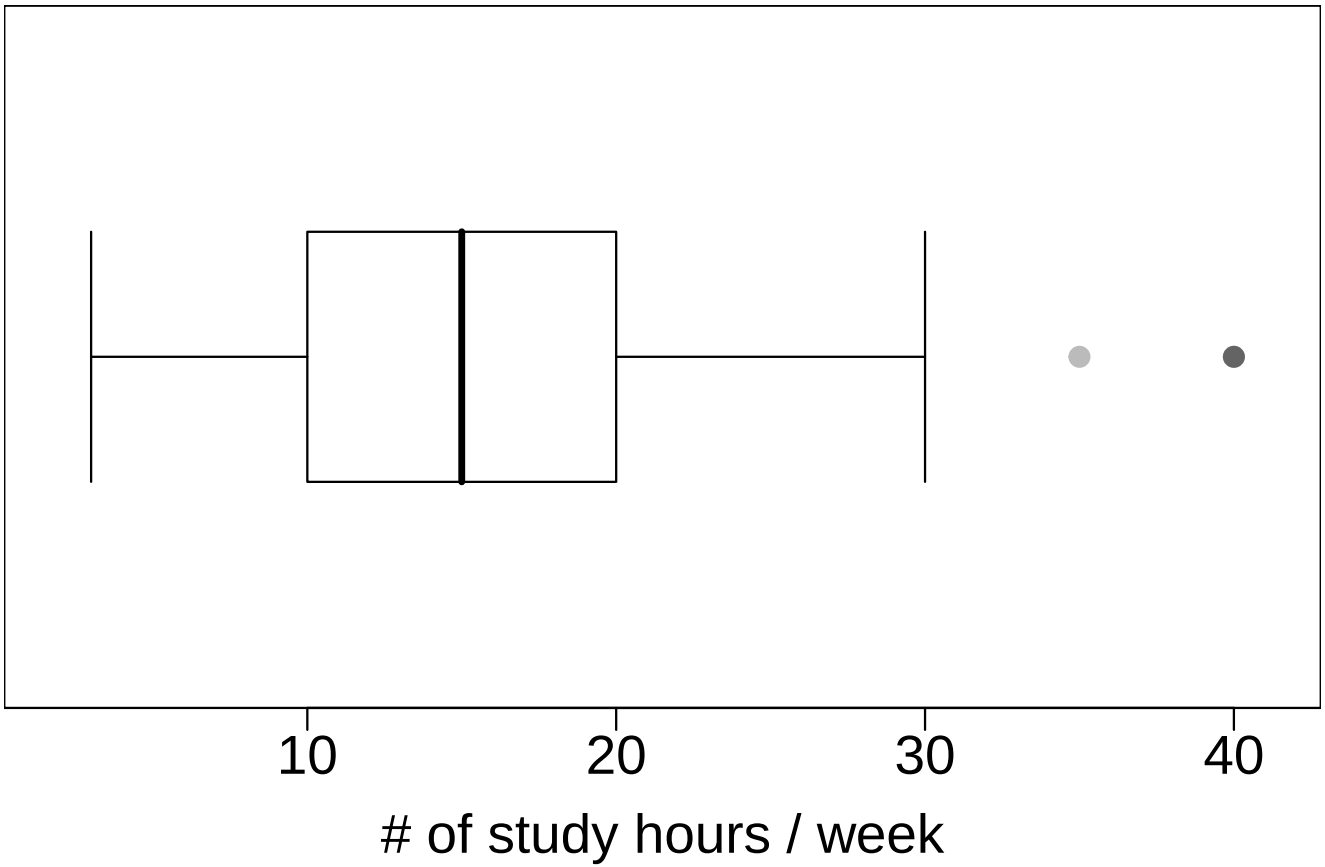
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NAs
3.00	10.00	15.00	17.42	20.00	40.00	13.00

- Between Q1 and Q3 is the middle 50% of the data. The range these data span is called the *interquartile range*, or the *IQR*.

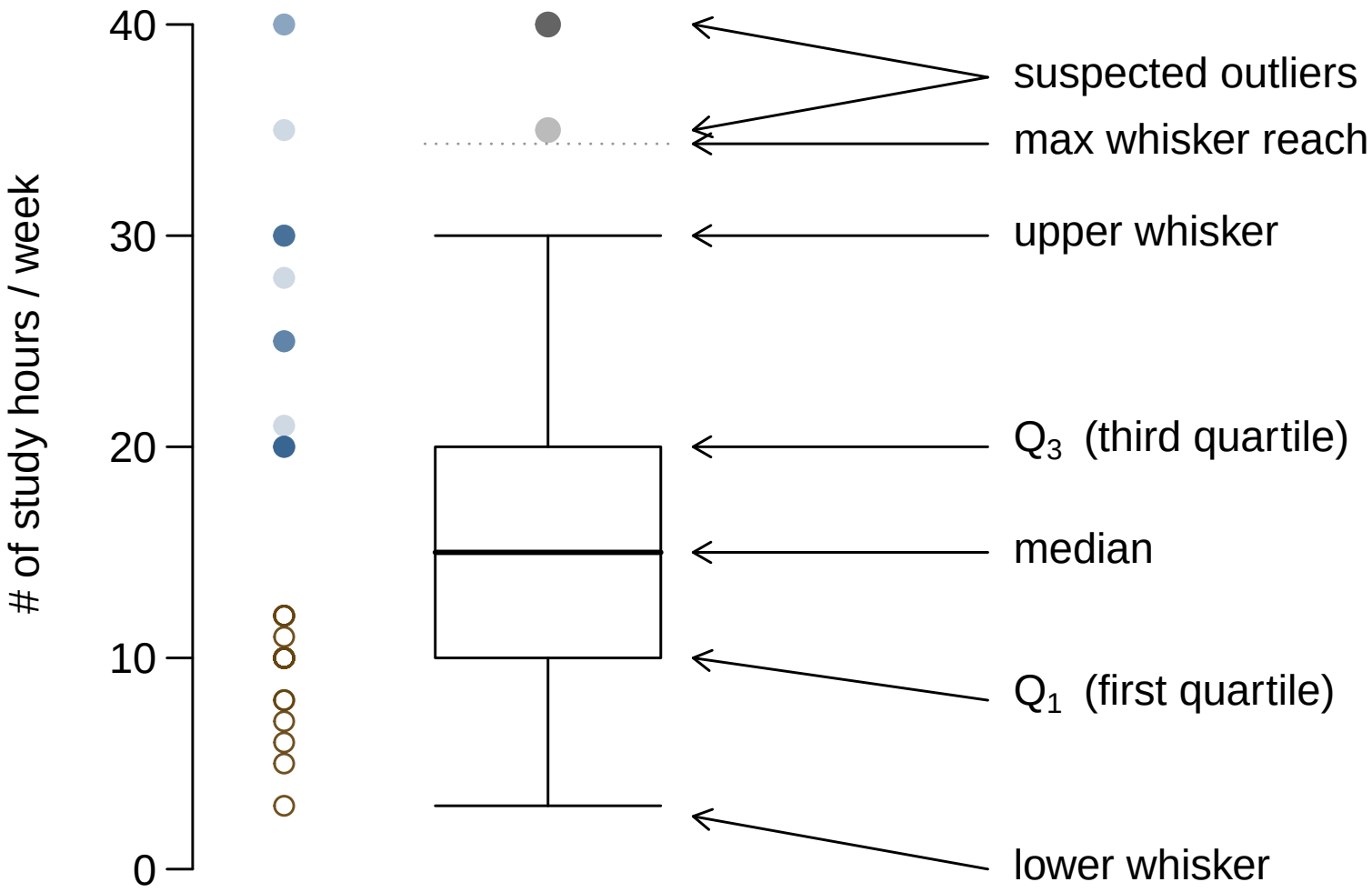
$$IQR = 20 - 10 = 10$$

Box plot

The box in a *box plot* represents the middle 50% of the data, and the thick line in the box is the median.



Anatomy of a box plot



Whiskers and outliers

- *Whiskers* of a box plot can extend up to 1.5 * IQR away from the quartiles.

  max upper whisker reach : $Q3 + 1.5 * IQR = 20 + 1.5 * 10 = 35$

  max lower whisker reach : $Q1 - 1.5 * IQR = 10 - 1.5 * 10 = -5$

- An *outlier* is defined as an observation beyond the maximum reach of the whiskers. It is an observation that appears extreme relative to the rest of the data.

Outliers (cont.)

Why is it important to look for outliers?

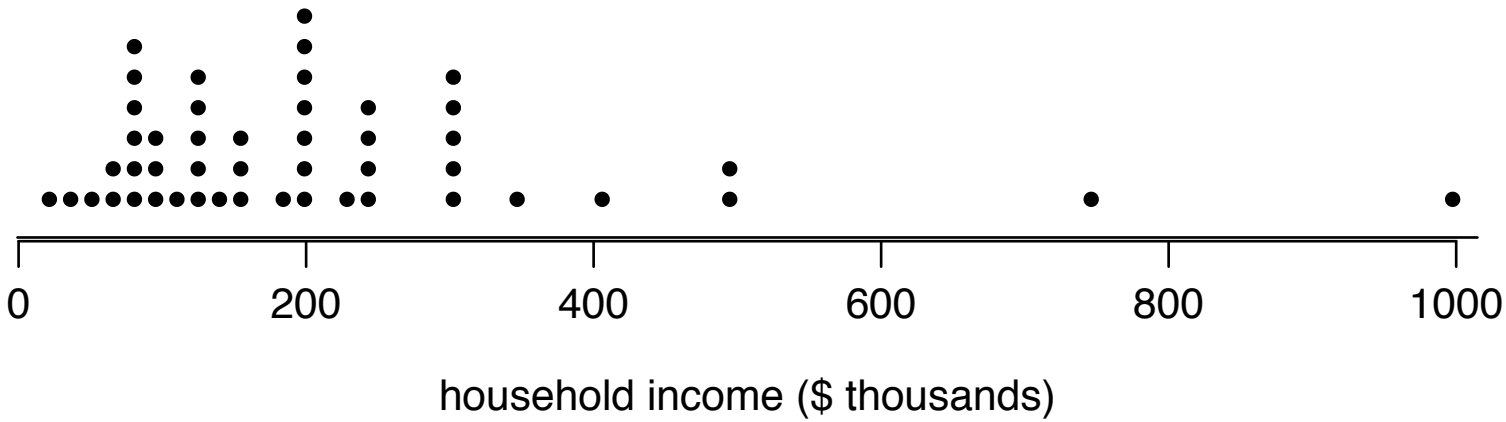
- *Identify extreme skew in the distribution.*
- *Identify data collection and entry errors.*
- *Provide insight into interesting features of the data.*

"I'm not an outlier; I just haven't found my distribution yet!"

Household income

Question

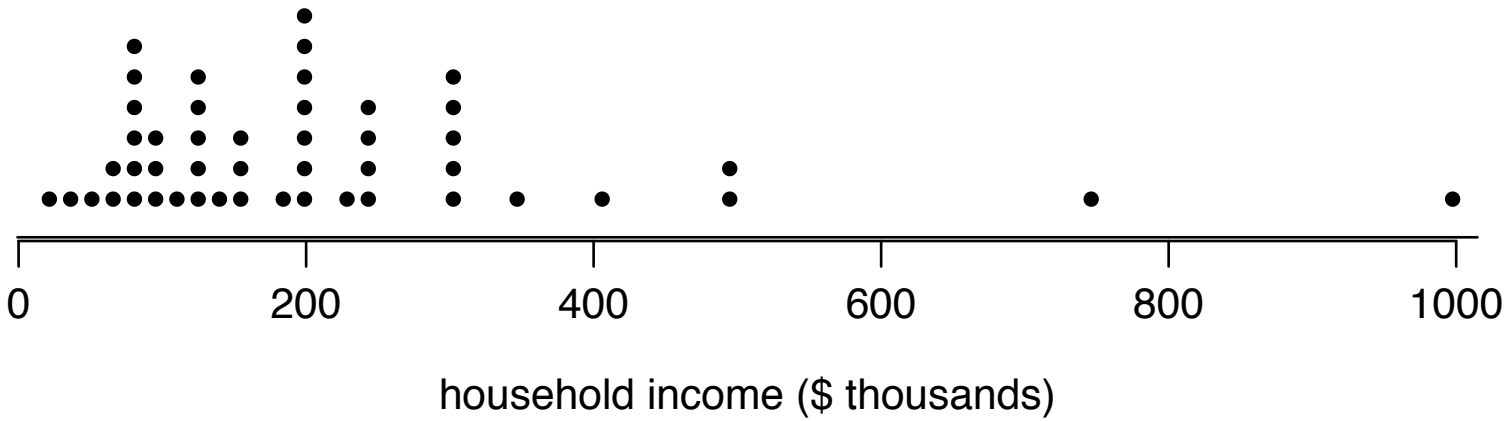
9. Which of the below is the most reasonable estimate for the median household income? (n = 48)



(a) $50K
(b) *$150K*
(c) $300K
(d) $400K
(e) $500K

Robust statistics

How would sample statistics such as mean, median, SD, and IQR of household income be affected if the largest value was replaced with $10 million? What if the smallest value was replaced with $10 million?



	robust		**not robust**	
scenario	median	IQR	$\bar{x}$	s
original data	165K	150K	211K	180K
move largest to $10 million	165K	150K	398K	1,422K
move smallest to $10 million	190K	163K	4,186K	1,424K

Robust statistics

Median and IQR are more robust to skewness and outliers than mean and SD. Therefore, to describe the *center and spread*

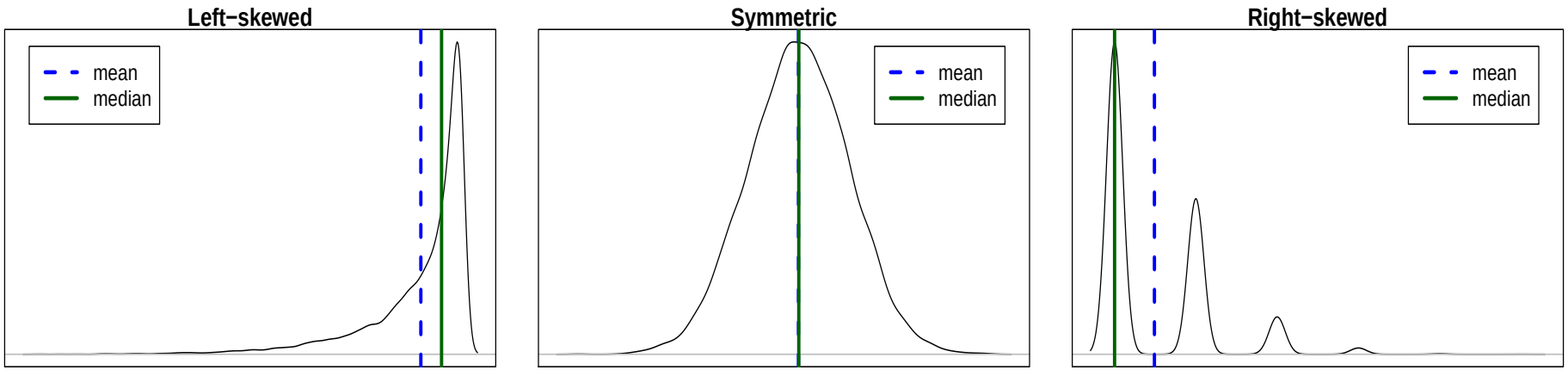
- for skewed distributions it is more appropriate to use *median and IQR*
- for symmetric distributions it is more appropriate to use *mean and SD*

If you would like to estimate the typical household income for a SMU student, would you be more interested in the mean or median income?

Median

Mean vs. median

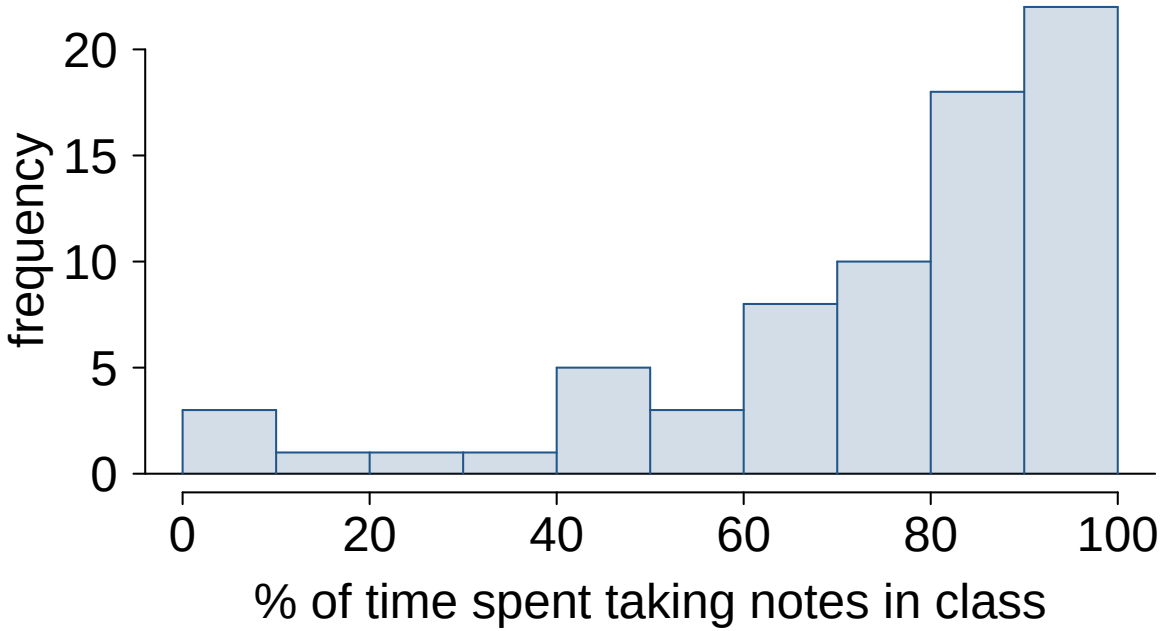
- If the distribution is symmetric, center is the mean
  - Symmetric: mean $\approx$ median
- If the distribution is skewed or has outliers center is the median
  - Left-skewed: mean $<$ median
  - Right-skewed: mean $>$ median



Note: *These rules often hold, but there are exceptions.*

Question

10. Which is true for the distribution of percentage of time actually spent taking notes in class versus on Facebook, Twitter, etc.?



median: 90%
mean: 78%

(a) mean is larger than median

(b) *mean is smaller than median*

(c) mean is roughly equal to the median

(d) impossible to tell

Contingency table

A table that summarizes data for two categorical variables is called a *contingency table*.

		cheat		
		no	yes	Total
major	arts and humanities	5	3	8
	natural sciences	15	12	27
	social sciences	20	21	41
	other	1	0	1
	Total	41	36	77

Note: *The survey question on cheating was worded as "Have you ever cheated on an assignment or exam?"*

Sample proportions

Question

11. Does there appear to be a relationship between major and whether or not a student cheated?

(a) *yes*
(b) no

major	cheat		
	no	yes	Total
arts and humanities	5	3	8
natural sciences	15	12	27
social sciences	20	21	41
other	1	0	1
Total	41	36	77

Proportion of students cheated

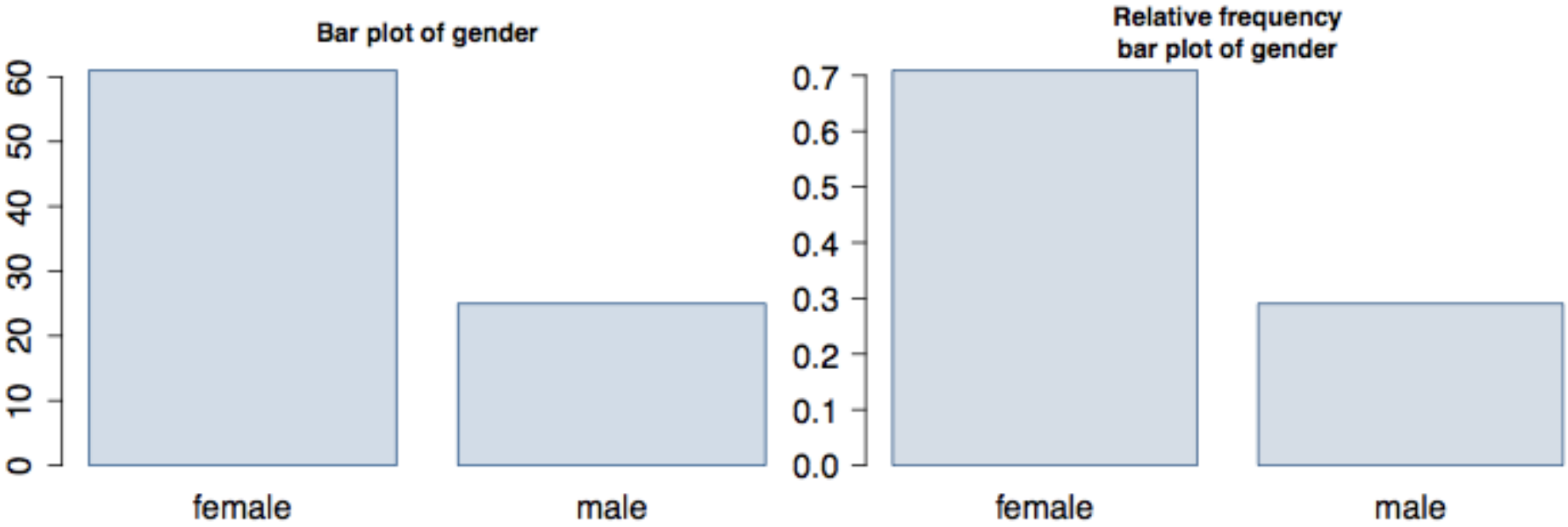
- among arts and humanities majors: $3/8 = 0.375 = 37.5\%$
- among natural sciences majors: $12/27 \approx 0.444 = 44.4\%$
- among social sciences majors: $21/40 \approx 0.512 = 51.2\%$
- among other majors: $0/1 = 0 = 0\%$

Visualization of categorical data

- bar plot
- segmented bar plot
- mosaic plot

Bar plots

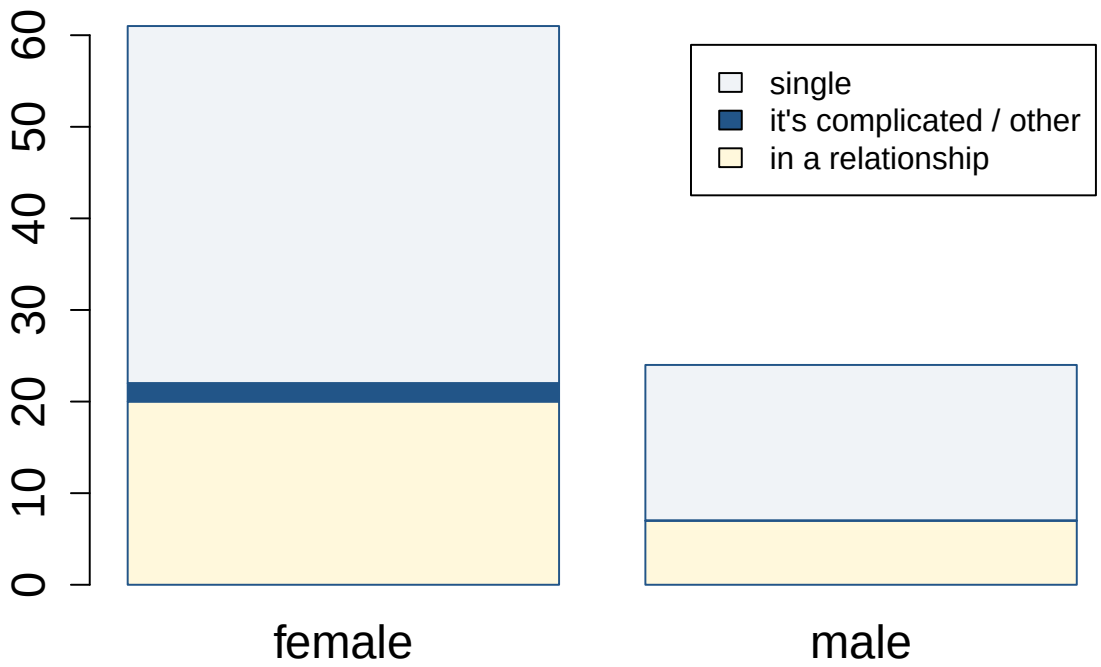
A *bar plot* is a common way to display a single categorical variable. A bar plot where proportions instead of frequencies are shown is called a *relative frequency bar plot*.



How are bar plots different from a histogram?

Segmented bar plot

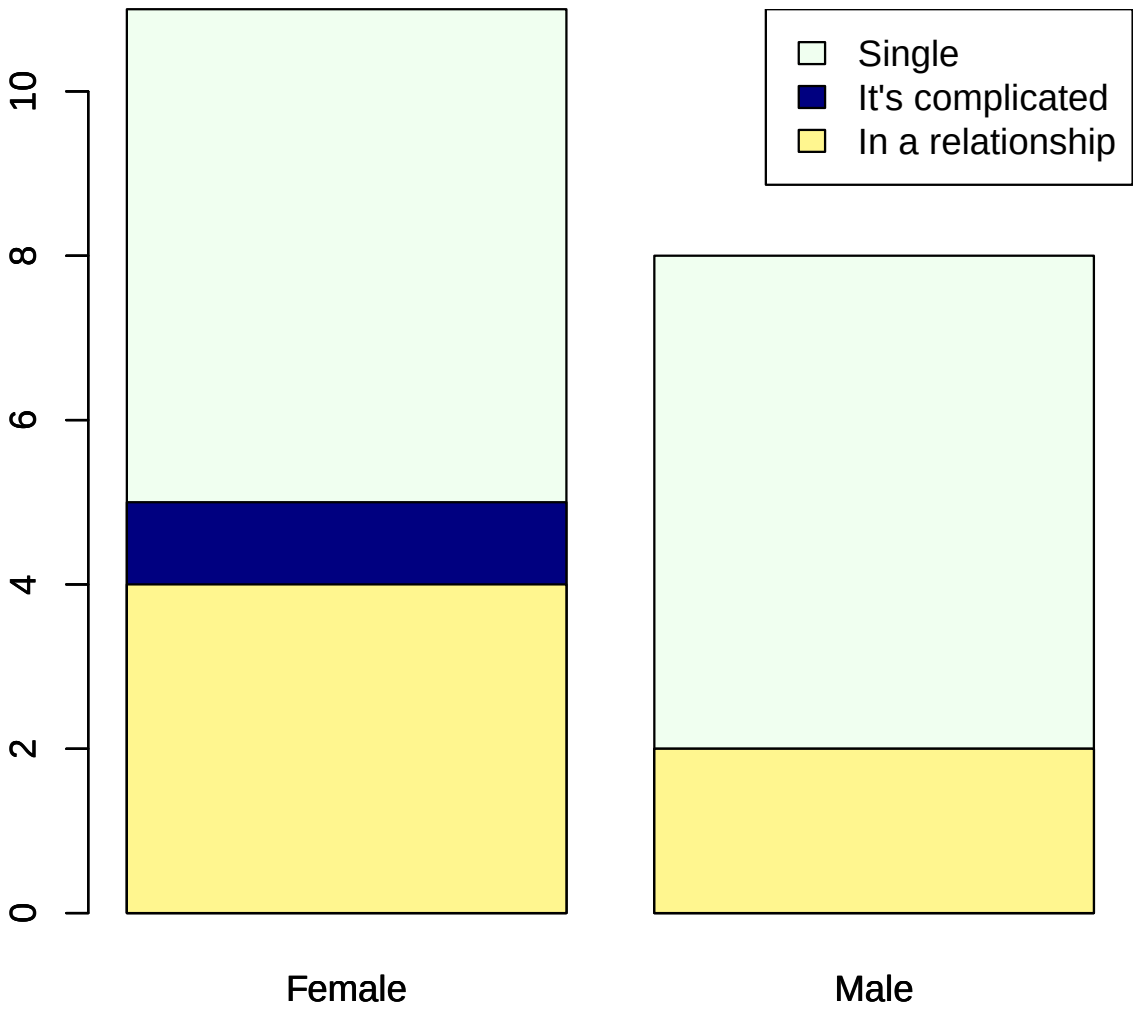
Does there appear to be a relationship between gender and relationship status?



% of females in a relationship = 20 / 61≈ 33%

% males in a relationship = 7 / 25 ≈ 28%

Data from our class



Question

12. A poll conducted by the Pew Research Foundation asked 2,142 people from the general public and 1,055 college presidents if they thought online courses offer an equal educational value compared with courses taken in a classroom. 621 people from the general public and 538 college presidents answered yes. Based on these data, does it appear that the general public and college presidents have differing opinions on this issue?

(a) Yes, the general public is more likely to view online education as comparable to classroom education.

(b) *Yes, college presidents are more likely to view online education as comparable to classroom education.*

(c) No, the two groups do not appear to have differing opinions on this issue.

(d) We cannot tell from the information given.

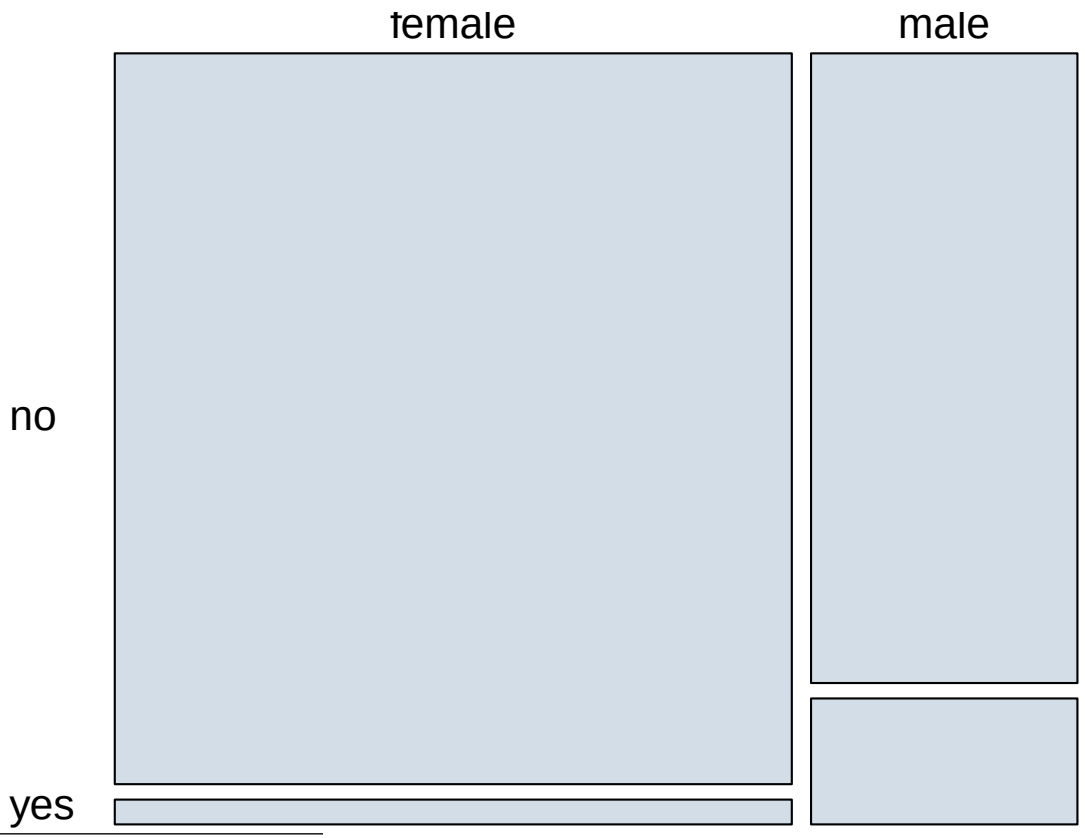
GP:
621 / 2142 = 0.290
Pres:
538 / 1055 = 0.510

Our class:
8 / 19 = 0.421

http://pewresearch.org/pubs/2092/online-courses-students-colleges-universities-technology-laptops-tablets

Mosaic plots

Does there appear to be a relationship between gender and having used Adderall for an exam or to study?

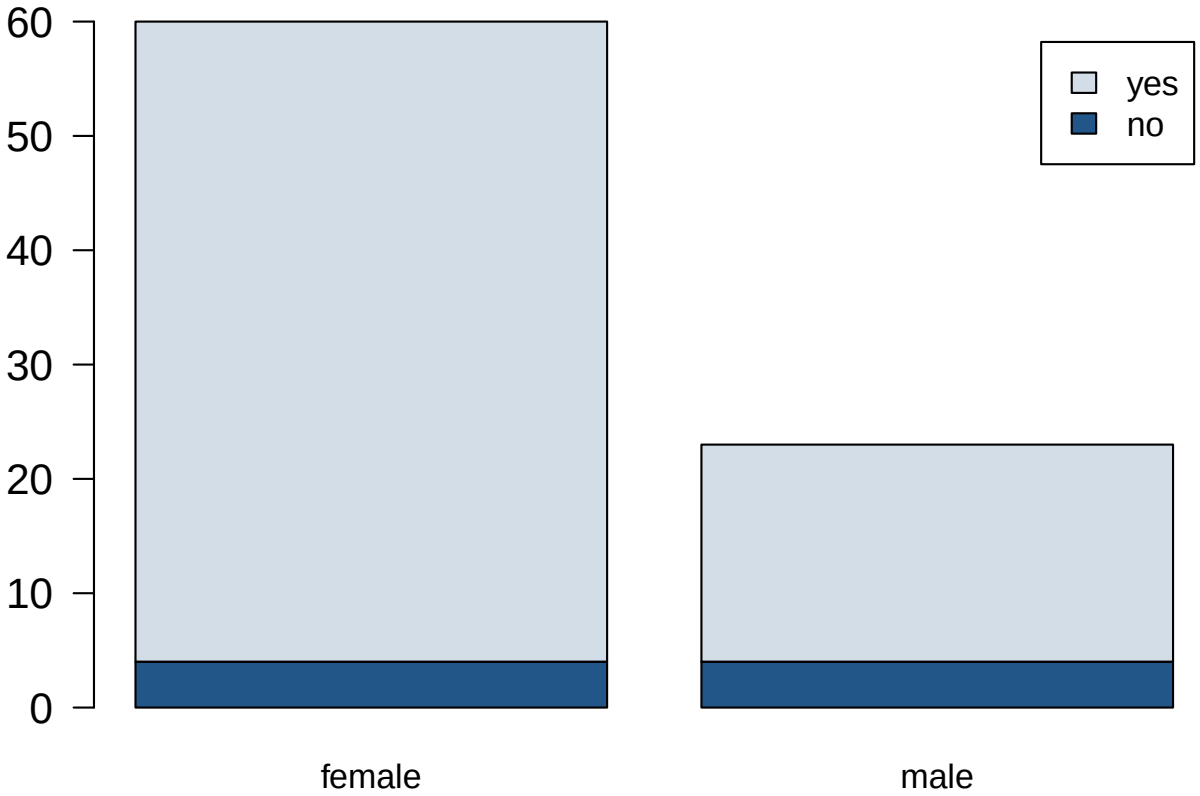


% female who used Adderall < % males

Note: *The survey question on Adderall was worded as "Have you ever used Adderall for an exam or to study?"*

Question

13. Does there appear to be a relationship between gender and opinion on gay marriage?



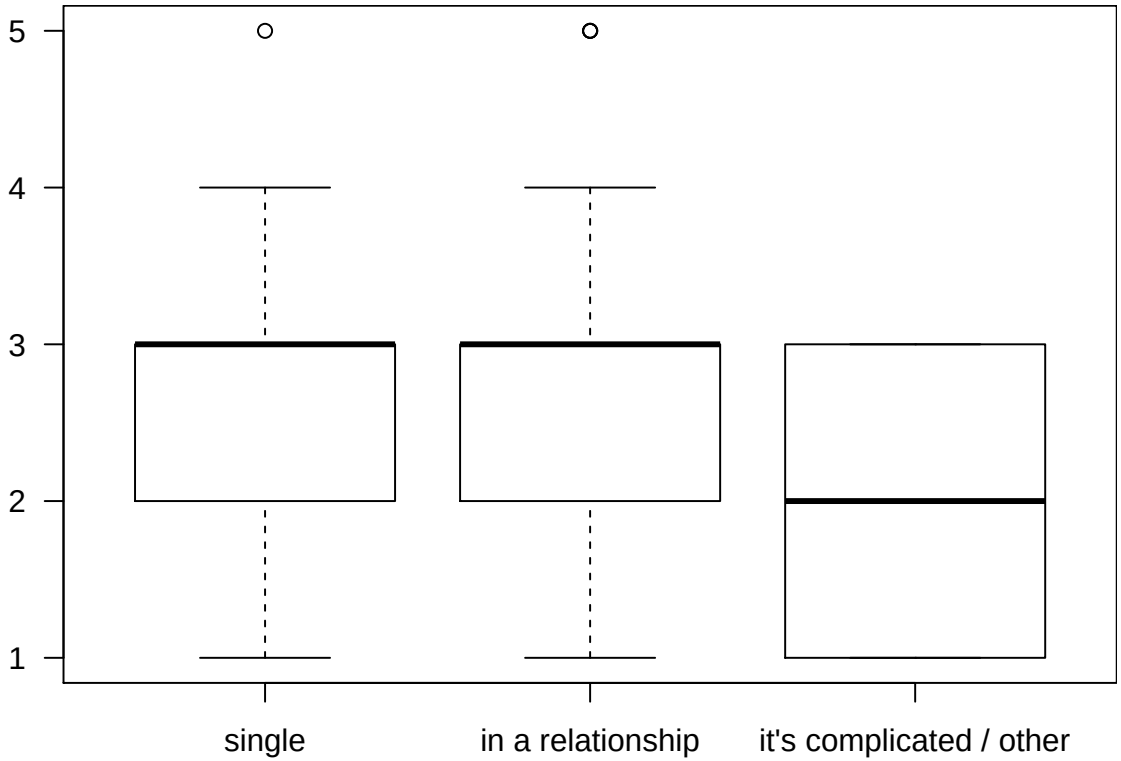
(a) *yes*

(b) no

Note: *The survey question on gay marriage was worded as "Should gay marriage be legal?"*

Side-by-side box plots

Does there appear to be a relationship between how much students dread this semester and their relationship status?



Note: *1 - not at all, 5 - a lot*