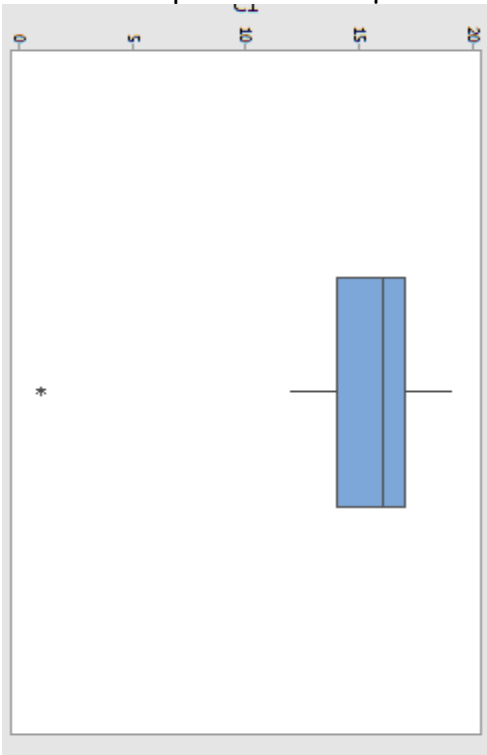


- Find the **mean** and **median** of the following data set: 2, 4.5, 8, 1, 2, 6.8, 12. Round answers to the nearest hundredth. **Mean=5.19, median=4.5**
- Find the ADM = $\frac{\sum|x-\bar{x}|}{n}$ of the following data set: 9, 7, 11, 4, 14, 2, 16. Round answer to the nearest hundredth. **4**
- The following data set shows the quiz raw scores (out of 20 points) for 11 students in a biology class.

Scores	1	12	14	15	15	16	16	16	17	18	19
--------	---	----	----	----	----	----	----	----	----	----	----

- Find the quartiles Q_1 , Q_2 (median) and Q_3 . **$Q_1=14$, $Q_2=16$ and $Q_3=17$**
- Find the Interquartile Range (IQR). **3**
- Interpret the meaning of the interquartile range in **context** to this problem.
The spread of the middle 50% of the data is 3 quiz points. OR
A typical quiz score ranged from 14 to 17.
- Check the quiz data for outliers. Are there any? **Outliers are quiz scores below 9.5 and above 21.5. Therefore 1 is an outlier.**
- Write the five number summary. **1 14 16 17 19**
- Create a box plot from the quiz data using the five number summary.

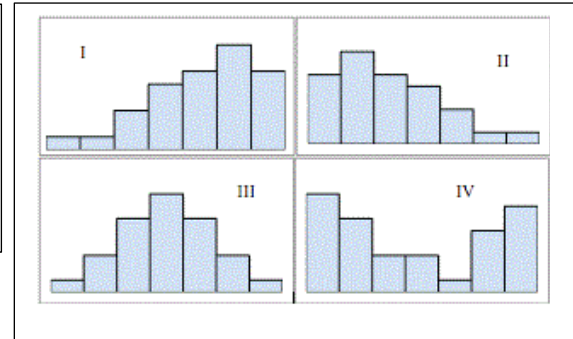


- Label each as categorical or quantitative:
 - Temperatures in SCV for the past year **quantitative**

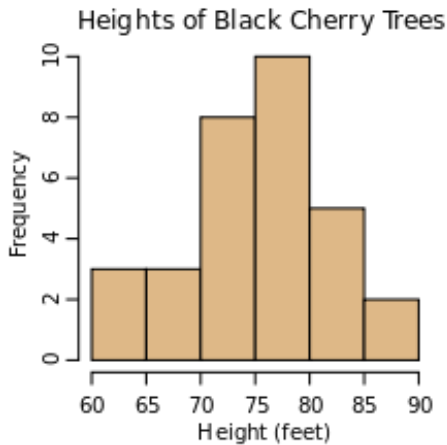
b. Weather conditions in SCV in past year categorical

5. Consider the following histograms and distributions. Choose an appropriate histogram for each of the following distributions.

- a. The distribution of length measurements at birth for 10,000 babies? **III**
 b. The distribution of quiz scores on an easy quiz? **I**
 c. The distribution of annual income for school employees where a high percentage of employees are entry-level teachers and only a few are high-paid administrators? **II**

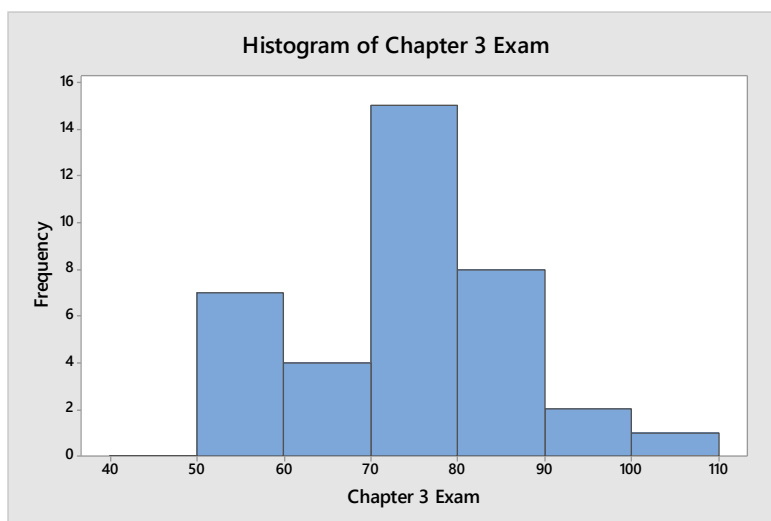


6. Given the following distribution of the heights of black cherry trees, answer the following:



- a) What is the x-axis measuring? **Heights of cherry trees in feet**
 b) What is the y-axis measuring? **Number of cherry trees that fell in each height group**
 c) How many cherry trees are measured in this sample? **31**
 d) How many trees are less than 70 feet tall? **6**
 e) Approximately how many trees are between 65 to 75 feet in height? **11**
 f) How many trees are at least 80 feet tall? **7**
 g) What percentage of trees are at least 80 feet tall? **$7/31 = 22.58\%$**

7. Answer the following questions given the distribution of results from Chapter 3 exam for a class. The exam was worth 100 points.



- How many students took the chapter 3 exam? **37 students**
- What is the shape of the distribution of exam scores? **Symmetric or slight skew right**
- What was a typical score for this class (center)? Around **75 points**
- What was the typical spread for this class? **Around 65 to 85 points**
- How many students got at least 80 points on the exam? **11 students**
- What percentage of students got at least an 80 on the exam? $11/37 = 29.7\%$
- How many students scored less than 80 points on the exam? **26**
- What percentage of students scored less than 80 points on the exam? $26/37 = 70.3\%$

8. The body temperature of students is taken each time a student goes to the nurse's office. The five-number summary for the temperatures (in degrees Fahrenheit) of students on a particular day is:

Min	Q1	Median	Q3	Max
96.6°	97.85°	98.25°	98.6°	101.8°

One can expect that a typical temperature for a student would fall between **97.85** and **98.6** degrees.

9. All students in the physical education class completed a basketball free-throw shooting event and the highest number of shots made was 32. The next day, the PE teacher realized that he had made a mistake. The best student had actually made 52 shots (instead of 32). Indicate whether changing the student's score made each of these summary statistics increase, decrease, or stay about the same:

- Mean **increase**
- Median **about the same**

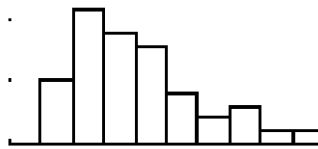
- c) overall **range increase**
- d) IQR **about the same**

10) We collect these data from 50 male students. Which variable is categorical and which is quantitative?

- A) eye color **categorical**
- B) head circumference **quantitative**
- C) marital status **categorical**
- D) number of cigarettes smoked daily **quantitative**
- E) number of TV sets at home **quantitative**

11) Which one of the quantitative variables in problem 10 is most likely to be symmetric? Why? **Head circumference because most people will be near the average head size with very few extreme values. Specie measurements are symmetric.**

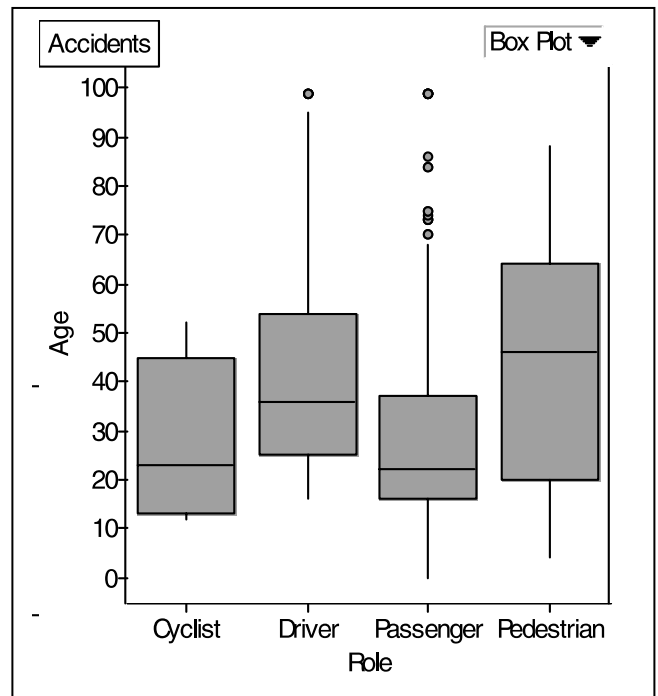
12) Which is true of the data whose distribution is shown?



- I. The distribution is skewed to the right. **True**
- II. The mean is smaller than the median. **False, the mean is larger than the median.**
- III. We should summarize with mean and standard deviation. **False. The graph is skewed therefore the median and IQR would be a more appropriate measure of center and spread.**

13) The boxplots show the ages of people involved in accidents according to their role in the accident.

- a) Which role involved the youngest person, and what is the age? **Passenger, 0 years (infant)**
- b) Which role involved the person with the lowest median age, and what is the age? **Passenger, 23**
- c) Which role involved the smallest range of age, and what is it? **Cyclist, 52-10=42**
- d) Which role involved the largest IQR of age, and what is it? **Pedestrian, 64-20=44**
- e) Which role has the most symmetric distribution? Explain. **Pedestrian because the distance from the median to extremes are about the same**
- f) Which role has the most skewed distribution? Explain. **Passenger because the distance from the median to the max is much longer than the distance from the median to the min**
- g) 50% of cyclists involved in accidents were above what age? **Near 22 years old**
- h) What percent of pedestrians involved in



14) A class of fourth graders takes a diagnostic reading test, and scores are reported by reading grade level. The five number summaries for the boys and girls are shown below.

Boys: 2.8 4.1 4.8 5.5 5.6
 Girls: 2.1 4.5 4.9 5.6 5.8

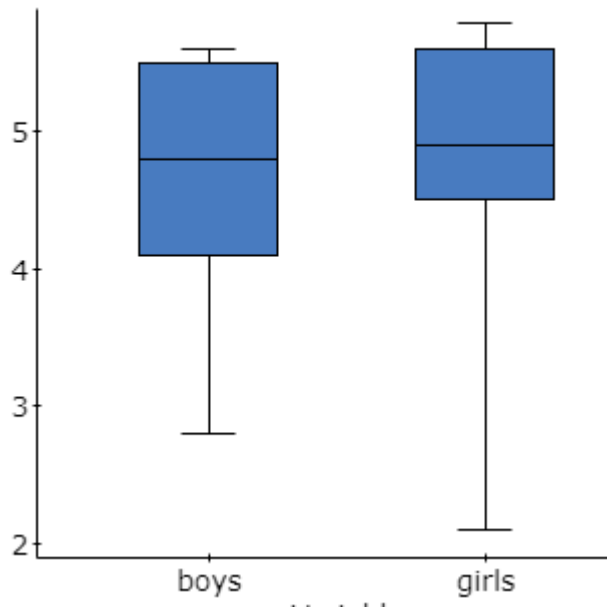
- a) Which group has the highest score? Circle one: Boys / **Girls (with a max of 5.8)**
- b) Which group has the greatest range? **Range can be interpreted in two ways, either typical range or overall range. Usually we are more interested in typical range. Here are the two possible answers:**

Circle one: **Boys** (with a typical range of $5.5-4.1=1.4$) / Girls

Circle one: Boys / **Girls** (with an overall range of $5.8-2.1=3.7$)

c) Which group has the highest IQR?

Circle one: **Boys** ($5.5-4.1=1.4$) / Girls



Which group's scores appear to be more skewed? Explain.

Girls:

$$\text{Median} - \text{Min} = 4.9 - 2.1 = 2.8$$

$$\text{Max} - \text{Median} = 5.8 - 4.9 = .9$$

Boys:

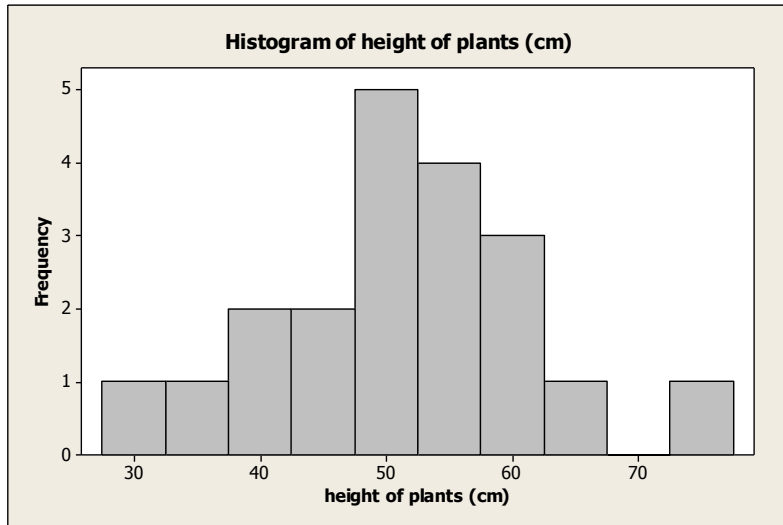
$$\text{Median} - \text{Min} = 4.8 - 2.8 = 2$$

$$\text{Max} - \text{Median} = 5.6 - 4.8 = .8$$

By visual inspection, one can see that both graphs are left skewed with the girl's graph more skewed to the left. The girls score is more skewed because the distance from the Median to the Min (2.8) is three times bigger than the distance from the Max to the Median (.9). The boys distances (2 and .8) are closer to each other than the girls distances (.9 and 2.8).

d) Which group generally did better on the test? Explain. **Girls did better because they have a higher Q1, median, Q3 and Max. They had a low min, but the distribution was skewed to the left meaning that there weren't very many low scores**

15) The students in a biology class kept a record of the height (in centimeters) of plants for a class experiment. The following is a list of the data, a histogram, and the descriptive statistics.

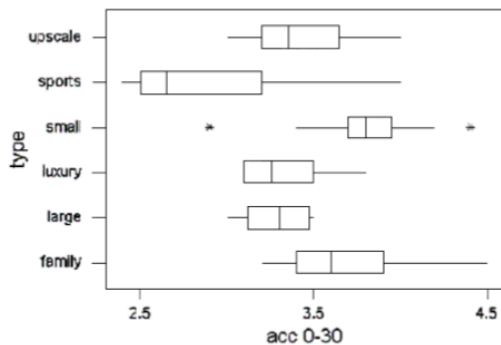


Variable	N	N*	Mean	StDev	Minimum	Q1	Median
height of plants (cm)	20	0	51.05	10.63	32.00	43.25	50.50
Variable	Q3	Maximum					
height of plants (cm)	58.25	75.00					

- a) Is it appropriate to use the mean to summarize these data? Explain. **Yes, the graph is roughly symmetric (with a possible small skew to the right). The center is approximately equidistant to both ends of the histogram.**
- b) Describe the distribution of plant heights. **The distribution is unimodal (one peak) and symmetric with no outliers. Most plants have a height of about 40-60cm. The smallest plant is 32cm. The largest plant is 75cm, which gave us a slight skew and higher mean. The range is $75 - 32 = 43$ cm. There are some plants that are twice as big as others.**
- c) When would you use SD instead of IQR? **Standard deviation is used for symmetric graphs and the interquartile range is used for all other distributions (shapes).**
- d) Interpret the meaning of the SD in context for this example. **68% (or roughly 70%) of the plants heights' will fall between 40.42cm and 61.68 cm. (mean + SD = $51.05 + 10.63 = 61.68$ and mean - SD = $51.05 - 10.63 = 40.42$)**
- e) Using the SD, between which two heights was the growth typical? **40.42 cm and 61.68 cm**

16) The 1999 Consumer Reports new Car Buying Guide reported the number of seconds required for a variety of cars to accelerate form 0 to 30 mph. The cars were also

classified into six categories by type. The following boxplots display the distributions of acceleration times for each type of car. (Note: the astericks on the boxplot for the small type of cars, these denote outliers.)



- a) If we compare a *typical* car in each category, which type accelerates the fastest? What part of the boxplots did you compare to make your choice?

The sports car accelerates the fastest. I used the median to make my choice.

- b) If we compare the *typical* range of acceleration times for each car type, which type performs the most consistently? What part of the boxplots did you compare to make your choice?

The large car has the most consistent acceleration time. I looked at the IQR (distance from Q1 to Q3 or the length of the entire box). I also looked at the overall range (distance from min to max or the length of the entire graph.) I noticed that the small car has the smallest IQR (but had two outliers a larger overall range). The large car has the smallest overall range. I decided to pick the large car as the most consistent, because it had the smallest overall range and no unpredictable outliers.

- c) Now, let us only focus on the Small cars. If the outliers were removed from the dataset of Small cars, which of the following measures of spread would be least affected: *Overall range*, *interquartile range* (the distance between the 1st and 3rd quartile marks), or *standard deviation*?

The IQR would be the least affected because the middle 50% values would remain about the same after the outliers are removed. The SD and overall range would definitely decrease.

17. The math department at a particular college wants to investigate the use of the newly developed math tutorial program. They decide to sample students to find out about their participation. Several plans for choosing the sample are proposed.

i) Students are divided into groups according to their math level (below average, average, and above average). Then twenty students are selected from each group and interviewed to determine whether they participated in the school's tutorial program.

ii) Every hundredth student who registers is asked whether they participated in the school's tutorial program.

iii) Students are divided into groups according to their math level (below average, average, and above average). Then all students in the average and above average groups are chosen and interviewed to determine whether they participated in the school's tutorial program.

iv) Students are selected to be interviewed to determine whether they participated in the school's tutorial program. The researcher goes to the tutoring room and interviews students as they come in.

v) 100 students are chosen according to student ID numbers generated by a computer program.

vi) Students are mailed a questionnaire to determine whether they participated in the school's tutorial program.

a. Which of the above would be a good method to randomly select students? **Why?** a, b, e subjects are randomly selected controlling possible bias.

b. Which of the above methods might result in being biased? **Why?** c, d, f.

c: not all groups have the same opportunity to participate or be chosen

d: Going to the tutoring room excludes students who do not go to tutoring.

f: voluntary response questionnaires are biased because the person chooses to participate

c. Name the type of sampling used for each of the above.

a. stratified

b. systematic

c. cluster

d. convenience

e. random

f. voluntary

18. Researchers reported that a newly discovered herb helps lower cholesterol. To test this claim, a study is conducted among 1000 high cholesterol patients. Doctors are told to give their patients either the herb in capsule form or a placebo. The doctors are given the information as to who has been randomly assigned the herb or the placebo. The patients are not aware whether they are being given the herb or the placebo.

Identify the following:

- i. the sample 1000 high cholesterol patients
- ii. the explanatory and response variables explanatory: herb or placebo capsule
response: cholesterol level

Follow up question: why was it important to assign two treatments?

(herb/placebo)

iii. whether or not the experiment is blind (or double-blind) blind: patients not aware but doctors are

19. Identify the following research studies as observational or experimental. Explain why.

a. Data from the Census was studied to investigate the claim that the average number of children per household has decreased throughout the years. **Observational no treatment assigned**

b. A college is proposing to provide in class tutors to improve the success rate for the lower level math courses. One instructor is given two lower level classes, one with a tutor and one without. The average grades for the class are then compared. The students were not aware that they had been chosen to participate in this study. **Experiment, treatment assigned (tutoring)**

c. A recent company report stated that Costco customers taste at least 2 samples during one shopping visit. This is of interest to the company because studies have shown an increase sale of the product when customers are allowed to sample first. The afternoon manager at one store does not believe this is accurate. He believes his customers sample at least 4 to 5 products during a single visit. He has an employee track selected customers by viewing store video cameras and documenting the sampling for fifty customers at 5:00 pm every day for two weeks. **Observational no treatment assigned**

20. In problem above, identify a possible lurking variable for each study. **Many possible answers: for example:**

a. Not everyone participates in the census.

b. Time of day: one class is a morning class, another might be a night class.

c. Time of day or people going after work and being hungrier.

21. In problem 19b above, why is it important that the students did not know they were part of the study? What is the name of this technique? **Blinding is used to avoid biases when subjects change behavior when they are aware they are participating in a study.**