2.1: Descriptive Statistics - Dotplots and Histograms

During the statistical analysis process, we ask a question, collect data, summarize and analyze the data, and finally, draw a conclusion. Descriptive statistics help us to summarize and analyze data. We will learn about numerical and graphical ways to describe and present data.

In this section, we will summarize and analyze **frequency distributions** of quantitative variables to investigate a question about ages of students at various types of academic institutions. A frequency distribution of a variable provides two important facts about the variable: all values the variable takes on, and how often (or how frequently) the variable takes on each given value.

A **quantitative variable** can be measured or counted and data values are expressed as numbers. **Categorical** or **qualitative variables** cannot be measured or counted and rather, can be expressed as membership of a group called a category.

Distributions of Age

A professor at Chaffey College is curious about the typical age of students who enroll at public two-year institutions compared to public four-year institutions and for-profit institutions.

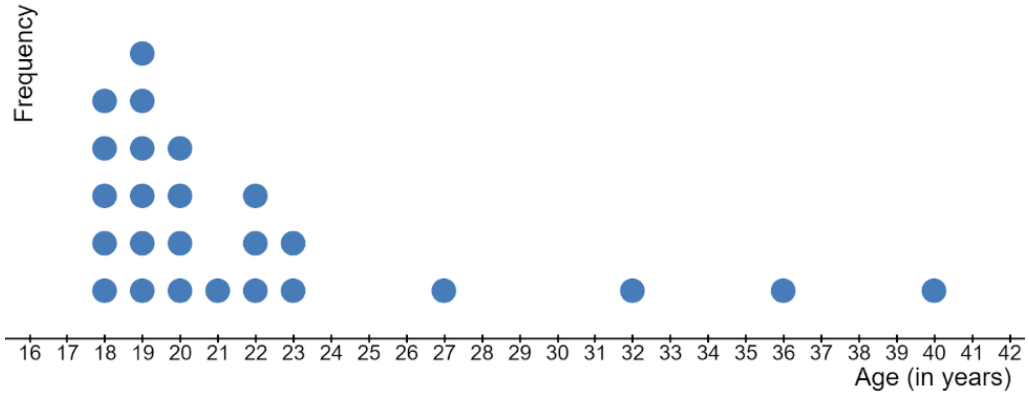
1. Make a prediction: what is the typical age of students at each type of institution? Why do you think this?

2. The variable we are discussing today is age. Is this variable quantitative or qualitative? Justify your answer.

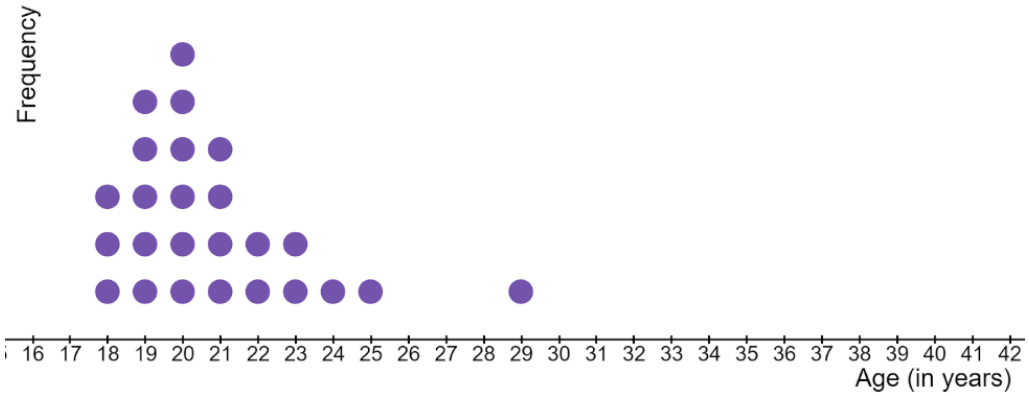
Dotplots

She randomly surveys 25 students in her general education classes at Chaffey College (a public two-year institution). She then asks her colleague at a nearby public four-year institution to randomly survey 25 students. Below are the resulting dotplots.

Ages of Students at a Public Two-Year Institution



Ages of Students at a Public Four-Year Institution



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

1. What does a dot represent in these dotplots?

2. How many students from the two-year sample are older than 25?

3. What proportion of students from the two-year sample are older than 25?

4. How many students from the four-year sample are older than 25?

5. What proportion of students from the four-year sample are older than 25?

6. What is the most frequent age in the two-year sample?

7. What is the most frequent age in the four-year sample?

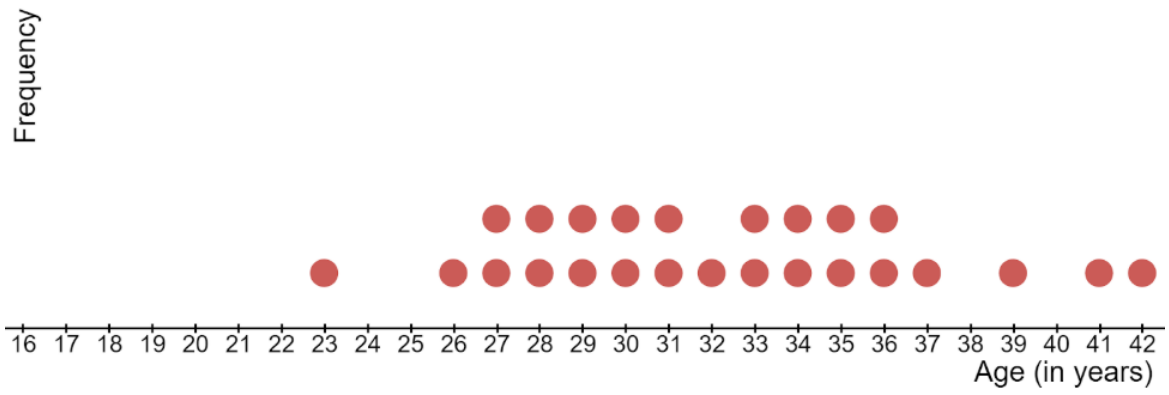
8. What is the typical age of a student in the two-year sample? What is the typical age of a student in the four-year sample? Compare these using the dotplots.

9. Which sample has more variation? Explain using the dotplots.

The shape of these two distributions are **right-skewed** because they have a long right tail. In other words, the higher ages are less likely to occur.

The professor from Chaffey College asks a colleague at a for-profit institution to randomly survey 25 students. The dotplot is given below. This distribution is closer to a **rough bell-shape** (in which the graph is symmetric and has one peak in the middle and two equal tails on each side; values in the tails are less likely to occur) and we could say it might have a slight right skew. It appears as though the typical age of a student at the for-profit institution is higher than at the public institutions (around 32 years).

Ages of Students at a For-Profit Institution

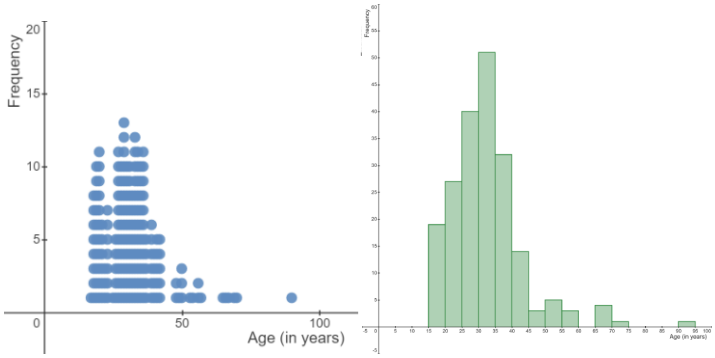


Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

Histograms

Sometimes, the type of data we collect can influence what type of graph we use to summarize the data. Often, with a variable like age, we want to group students into different ranges of ages, especially if we have a large sample of data. In this case, we use a **histogram** to summarize the data graphically.

Here is a dotplot and histogram of ag›es of 200 public four-year institutions. The histogram is more easily readable and we can more easily use it to analyze the data than using the dotplot.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

Let's return to our example of ages at a for-profit institution. Here is the sample of 25 ages:

$$[33, 33, 30, 31, 39, 27, 35, 36, 37, 23, 35, 41, 42, 36, 34, 28, 28, 29, 26, 27, 29, 30, 31, 32, 34]$$

To help us find patterns within the for-profit data set, we will group the data into ranges of ages called **bins**. For this example, we will use intervals of size 5 so each bin will contain 5 ages (15-19, 20-24, etc.). The first bin starts with a value slightly lower than the lowest age in the set. We will create the **frequency distribution table** below prior to graphing the histogram.

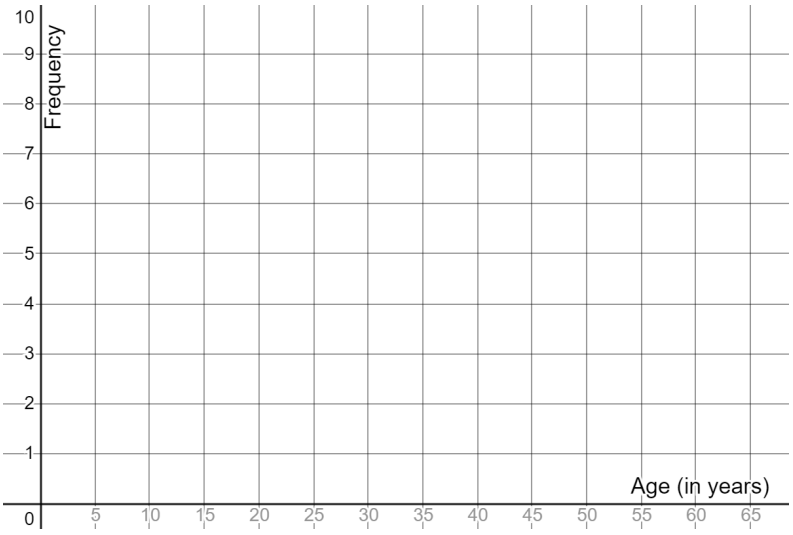
Bin	Tally	Frequency	Relative Frequency (as a fraction)	Relative Frequency (as a decimal)	Relative Frequency (as a percent)
15-19		0	$\frac{0}{25}$	0	0%
20-24	|	1	$\frac{1}{25}$	0.04	4%
25-29					
30-34					
35-39					
40-45					
Total:					

For each data value in the set, determine the bin it falls into. For example, the lowest age in the set is 23 years old, so it belongs in the bin with a range 20 to 24. A tally mark (|) has been written in the tally column next to the row led by 20-24. Continue to make

tally marks in the tally column until you have selected a bin for all data values. Each time a tally reaches the fifth mark, represent it as a horizontal tally mark (5 is the same as 𝍸).

The frequency is the number of data values in each bin, or the number of tally marks for a given bin. Write the frequency as a number in the frequency column. We compute the relative frequency by dividing the frequency by the total number of data values (sample size). We can write the relative frequency as a fraction, decimal, and percent.

12. Now, use the table to create a **frequency histogram**. Each bin in the distribution is represented by a vertical bar. The height of that bar is the frequency of the bin. Draw the bars so that each adjacent bar is touching (there are no gaps between adjacent bars).



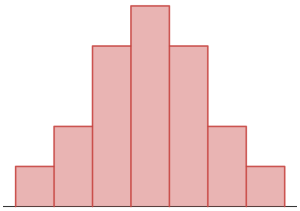
Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

13. What is the sum of all heights of bars in the histogram? What does this sum represent?

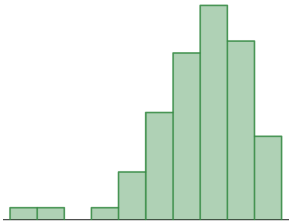
14. What does the height of the second bar represent?

Summary: Center, Shape, and Spread

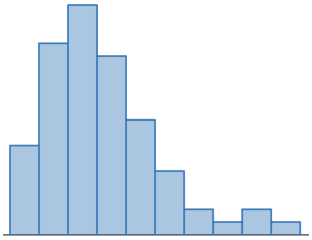
The **center** of a distribution is the typical value in the data set, or the single value that best represents the distribution. The **shape** of a distribution is the overall pattern of the distribution. There are four common shapes we might use to describe a distribution.



Bell-Shaped Distribution



Skewed-Left Distribution



Skewed-Right Distribution

Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

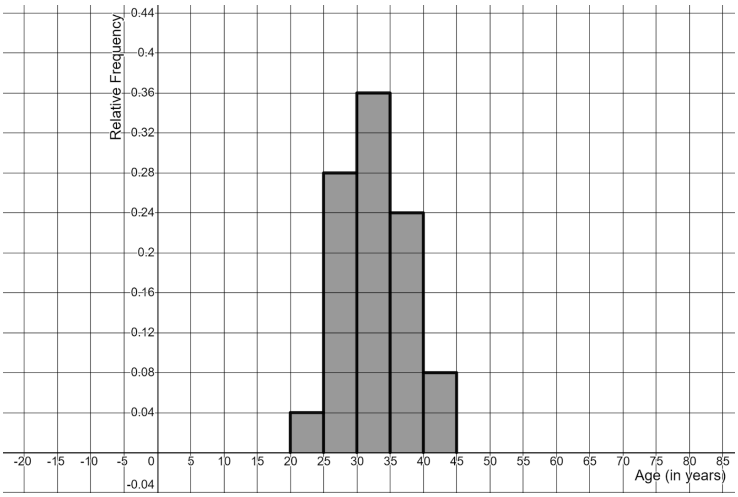
Uniform Distribution

A **uniform** distribution is one in which every data value is equally likely to occur. We can use these graphs to help us identify potential **outliers**. An outlier is a data value that is much higher or lower than most other values. The **spread** of a distribution describes the variation within a data set. It is how far apart the data values are. We often consider the **range** of values in the set, which is found by subtracting the lowest data value from the highest data value.

15. What is the center, shape, and spread of the frequency histogram?

16. How many students are older than 30 in the for-profit sample?

17. A **relative frequency histogram** displays the relative frequencies for the bins instead of the frequencies.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

18. Compare the relative frequency histogram to the frequency histogram you graphed in question 12. Are there any similarities or differences between the two histograms?

19. What proportion of students are older than 30 in the for-profit sample?