Unit 3: Histograms



PREREQUISITES

Histograms require the ability to group numbers by size into categories. Students will need to be able to compute proportions (relative frequencies) and percentages. This unit continues the discussion of describing distributions that began in Unit 2, Stemplots.

ADDITIONAL TOPIC COVERAGE

Additional information on histograms and other graphic displays can be found in *The Basic Practice of Statistics*, Chapter 1, Picturing Distributions with Graphs.

ACTIVITY DESCRIPTION

The Unit 3 activity focuses on quality control in the production of polished wafers used in the manufacture of microchips. The Wafer Thickness tool found in the Interactive Tools menu is required for this activity. Using this interactive, students can set three controls at three different levels. These controls affect the thickness distribution of polished wafers. The final task asks students to make a recommendation for the control settings so that the product is consistently close to the target thickness of 0.5 mm.

MATERIALS

Students will need access to the Wafer Thickness interactive from the online Interactive Tools menu.

The activity introduces students to histograms and the concept of variability. Students learn how histograms are constructed by watching a histogram being made in real time as the data are generated by the Wafer Thickness interactive. In addition, students should discover (at an informal level) that there are different sources of variability – this understanding will be useful preparation for future units. Here are some sources of variability that students should observe:

- Under the same control settings, thickness varies from wafer to wafer.
- Under the same control settings, the histograms from two samples of wafers will differ (variability due to sampling).
- Changing the control settings changes the distribution of wafer thickness (variability due to control settings).

Questions 4 and 5 are ideal for group work. In question 4a students must design a strategy for determining the effect that changes in the control settings have on the sample data. There are three control settings, each having three levels. Hence, there are $3 \times 3 \times 3 = 27$ distinct sets of possible control levels. A carefully designed plan may reduce the number of settings used in the investigation. The variability due to sampling makes 4b somewhat difficult to answer.

Students may have to view more than one sample from a given set of control settings. If students get frustrated, have them start with Control 3. Control 3's effect on the spread of the data is probably the easiest to spot. Control 1's shift in location is also not difficult to observe, particularly if Control 3 is set at level 3. Control 2's effect on shape as well as location is the most difficult to ascertain and students may not be able to figure out how Control 2 affects the distribution of wafer thickness. (That's OK – this happens in the real world.)

In question 5, students need to make a recommendation on the best choice of settings for the three controls and to support that recommendation based on the histograms they have constructed. There is not a single correct answer to this question. Some settings clearly give better results than others – but a "best" choice of control settings is a point open to argument. Students should make a decision and defend it against other possibilities.

It should be noted that data from a single sample can be saved in a CSV file. Data from CSV files can be imported into statistics packages or worked with in Excel.

Students may be interested in seeing how real data on microchip thickness are gathered. The video clip at the following site shows a technician taking measurements from wafers on which microchips have been embedded:

http://www.youtube.com/watch?v=jG84UjCZboo

THE VIDEO SOLUTIONS

1. Time of first lightning flash.

2. Horizontal scale: Time of day in hours. Vertical Scale: Percent of days with first lightning flash within that hour.

3. Roughly symmetric.

4. These were values that were separated from the overall pattern by a gap in the data.

5. The classes need to have equal width.

6. Using too many classes can make it difficult to summarize patterns connected with specific values on the horizontal axis. (In other words, you can't see the forest for the trees.) Too few classes can mask important patterns.

UNIT ACTIVITY SOLUTIONS

1. a. Sample answer based on the following sample data (in mm): 0.591, 0.483, 0.489, 0.452, 0.639, 0.523, 0.601, 0.511, 0.498, 0.467.

After the first wafer was measured, a rectangle was drawn above the interval 0.550 to 0.600 since the thickness 0.591 fell between these values. The second, third, and fourth rectangles were stacked on top of each other over the interval 0.450 to 0.500, since 0.483, 0.489, and 0.452 all fell in that interval. The process continued until a rectangle was drawn for each of the 10 measurements.



b. Sample answer: The histogram for the sample data from 1a appears below.

The histogram does not appear to be symmetric.

The interval 0.450 mm to 0.500 mm has the tallest bar and hence more wafers had thicknesses that fell in this interval than any other interval.

There are no gaps between the bars. However, none of the wafers had thicknesses that fell in the intervals 0.300 to 0.450 and 0.650 to 0.900.

The smallest data value fell in the interval from 0.450 to 0.500 and the largest data value fell in the interval from 0.600 to 0.650.

The thickness 0.5 mm does not appear to be a good choice for summarizing the location of these data. One bar falls to the left of 0.500 mm (the tallest bar) and three bars fall to the right of this value; perhaps 0.525 mm would be a better choice. The controls do not appear to be properly set to produce wafers of consistent 0.5 mm thickness.

2. a. Sample answer data for second sample: 0.389, 0.541, 0.525, 0.621, 0.543, 0.500, 0.638, 0.392, 0.382, 0.602.

b. Sample answer:



The histogram does not appear to be symmetric. (It would be symmetric if the second bar had been closer to the first bar.)

The interval 0.500 mm to 0.550 mm has the tallest bar and hence, more wafers had thicknesses that fell in this interval than any other interval.

There are gaps between each of the bars. None of the wafers had thicknesses that fell in the intervals 0.300 mm to 0.350 mm, 0.400 mm to 0.500 mm, 0.550 mm to 0.600 mm and 0.650 mm to 0.900 mm.

The smallest data value fell in the interval from 0.350 to 0.400 and the largest data value fell in the interval from 0.600 to 0.650.

The thickness 0.5 mm might be a good choice for summarizing the location of these data. It's the lower endpoint of the interval corresponding to the tallest bar. The outside bars are the same height. Given there is a larger gap between the first and the second bar than there is

between the second and the third bar, using the lower value of the middle bar's interval seems reasonable. So, it is somewhat reasonable to assume that the controls are properly set to produce wafers that are fairly consistently close to 0.5 mm in thickness.

3. Sample answer from sample data shown in histogram that follows descriptions of common features and differences.

Common features: Neither histogram is symmetric. In both samples, the data values are spread from 0.35 to 0.65. The highest bar occurs over the interval 0.400 to 0.450 in both histograms.

Differences: The two histograms appear different in shape. In the left histogram, the heights of the bars are irregular – down, up, down, up, down. However, in the right histogram, from the second bar to the last bar, the heights of the bars decrease.



4. a. Sample: Change one control setting at a time and compare histograms to see what has changed. For example, start with the following settings: Control 1 = 1, Control 2 = 1 and Control 3 = 1. Change Control 3 from 1 to 2 to 3 and describe the change. Then choose different settings for Controls 1 and 2 and repeat the process described above. See if the observed pattern remains the same. If so, describe how the settings of Control 3 affect wafer thickness in sample data.

Adapt the strategy described above to determine how Controls 1 and 2 affect the thickness of wafers.

b. Sample answer: Samples were collected with Control 1 = 1 and Control 2 = 1 and then changing Control 3 from 1 to 2 to 3. The notation (1, 1, 1), (1, 1, 2), and (1, 1, 3) is used to identify the three control settings. In the histograms below, the most apparent change appears to be to in the spread of the data. The data are least spread out (thicknesses are most consistent) when Control 3 = 3. (It is almost as if the right tail shrinks as the level of Control 3 is increased.)



Next, we kept Control 1 = 1, set Control 2 = 2, and then changed Control 3 from 1 to 2 to 3. The histograms appear below. The same pattern of reduced spread occurred. The data are more consistent (less spread out) when Control 3 = 3.



Next, we focus on the effect of Control 2. The histograms below compare settings (1,1,1), (1,2,1) and (1,3,1). When Control 2 = 1, the data appear more concentrated to the left. When Control 2 = 2, the data appear more symmetrical, and when Control 2 = 3, the data appear more concentrated to the right. So, we conclude that Control 2 affects the shape of the data.



Last, we change the settings for Control 1, leaving settings for Control 2 and Control 3 fixed. Below are histograms for samples from settings (1, 1, 1), (2, 1, 1) and (3, 1, 1). Changing the settings on Control 1 from 1 to 2 to 3 appeared to shift the bars in the histogram to the right – hence, increasing the thicknesses.



5. Sample answer (student answers will vary): We recommend settings (3,2,3). We chose Control 3 = 3 to reduce variability. We chose Control 2 = 2 so that we had balance between high and low values. Finally, we chose Control 1 = 3 to increase the thickness. We compare this choice of settings with (2,2,3) and (2,3,3) in the histograms below.



EXERCISE SOLUTIONS





b. Sample answer (assuming the student's home state is Massachusetts): For Massachusetts, there were 903 thousand people 65 or older in 2010. Massachusetts' population of 65 and over appears to be fairly typical.



c. Sample answer: The distribution is skewed to the right. There are two gaps – one between 2,000 thousand and 2,500 thousand and the other between 3,500 thousand and 4,000 thousand. California with 4,247 thousand people 65 or over could be an outlier. Florida with 3,260 thousand, New York with 2,618 thousand, and Texas with 2,602 thousand might also be outliers (or they could simply be the tail of the overall pattern in the distribution).

d. In the histogram below, the gaps in the data are hidden. However, you still can observe an overall pattern that is skewed to the right.



2. a. Sample answer (this time assuming the student's home state is Florida): For Florida, 17.3% of the people were 65 or older. Florida has a higher percentage of people 65 or older than all other states and the District of Columbia.



b. Sample answer: The overall pattern is roughly symmetric. There is a small gap – there are no percentages between 8% and 9%. South Carolina (7.8%) and Alaska (7.7%) might be outliers. However, they really don't appear to be unusual values – the gap is small and these values are at the upper end of the class interval from 7% to 8%.



3. a. Sample answer (students could have made other choices for the class sizes):

b. Sample answer: The overall pattern of the distribution of states' population sizes is skewed to the right. There are two gaps in the data, one between 20,000 thousand and 24,000 thousand and the other between 28,000 thousand and 36,000 thousand. California (37,254 thousand) is definitely an outlier. In addition, Texas (25,146 thousand) is a potential outlier.

4. a. Yes.

Sample explanation: If you look at the breaking strengths recorded in the first column, all the entries are different. In fact, all but four of the breaking strengths are distinct. So, breaking strength varied from stake to stake even though the stakes were nearly identical.



c. The interval from 160 to 165 contained the most data.

d. The histogram below looks exactly the same as the histogram in (b) – it has the same shape, the same gaps, and the same potential outliers. The only thing that changed was the scaling on the vertical axis.



e. Sample answer. The overall pattern in the data is skewed to the left. The three data values between 115 and 125 represent a departure from the overall pattern and are sufficiently far from the rest of the data that they may be considered outliers. There are three class intervals containing no data that separate these potential outliers and the rest of the data.

REVIEW QUESTIONS SOLUTIONS

1. Sample answer: The overall pattern in the first histogram is skewed to the right. There is a gap between 600 and 700 and one outlier between 700 and 800. The outlier is Babe Ruth's record 714 career home runs. Although the pattern in the second histogram could still be described as skewed to the right (because the tail of the data on the right is stretched out), the pattern is more jagged compared to the first histogram. There are a few secondary peaks and valleys apparent in the second histogram, which are not visible in the first histogram. Also interesting is the fact that the data values in the second class interval (100 to 200) of the first histogram are not evenly distributed when that class interval is divided in half, 100 to 150 and 150 to 200. There are 24 data values in the class interval 100 to 150 but only one-quarter as many from 150 to 200. A similar pattern holds when the class interval from 200 to 300 is divided into two class intervals.





Histogram 1

Histogram 2

b. Sample answer: Yes, for example Rod Carew had 19 career years. In the first histogram, his data value was classified in the class 19 - 24 and in the second histogram it was classified in the class 19 - 24 and in the second histogram it was classified in the class 19 - 21.

c. Histogram 1: The shape appears unimodal and skewed right. Histogram 2: The shape appears bimodal and roughly symmetric. Changing the class intervals had a big effect on the overall shape.

Duration (minutes)	Frequency	Percent
0-6	13	32.5
6 – 12	9	22.5
12 – 18	10	25
18 – 24	3	7.5
24 – 30	1	2.5
30 – 36	1	2.5
36 – 42	0	0
42 – 48	1	2.5
48 – 54	2	5

3. a.

b. 55%

c. 10%

d. (See histogram on next page...)



e. Sample answer: The distribution is skewed to the right. There is a gap between 30 and 36 minutes. There are two distinct groups of phone calls, those lasting under 30 minutes and a few lasting 36 or more minutes.