

Chapter 2

Describing, Exploring, and Comparing Data

2-2 Summarizing Data with Frequency Tables

- Subtracting two consecutive lower class limits indicates that the class width is $60 - 55 = 5$. Since there is a gap of 1.0 between the upper class limit of one class and the lower class limit of the next, class boundaries are determined by increasing or decreasing the appropriate class limits by $(1.0)/2 = 0.5$. The class boundaries and class midpoints are given in the table below, and the Excel bin format is shown at the right.

| <u>height</u> | <u>class boundaries</u> | <u>class midpoint</u> | <u>frequency</u> | <u>EXCEL FORMAT</u> | |
|---------------|-------------------------|-----------------------|------------------|---------------------|------------------|
| | | | | <u>bin</u> | <u>frequency</u> |
| 55 - 59 | 54.5 - 59.5 | 57 | 1 | 59 | 1 |
| 60 - 64 | 59.5 - 64.5 | 62 | 3 | 64 | 3 |
| 65 - 69 | 64.5 - 69.5 | 67 | 49 | 69 | 49 |
| 70 - 74 | 69.5 - 74.5 | 72 | 46 | 74 | 46 |
| 75 - 79 | 74.5 - 79.5 | 77 | 1 | 79 | 1 |
| | | | 100 | more | 0 |

NOTE: Although they often contain extra decimal points and may involve consideration of how the data were obtained, class boundaries are the key to tabular and pictorial data summaries. Once the class boundaries are obtained, everything else falls into place. Here the first class width is readily seen to be $59.5 - 54.5 = 5.0$ and the first midpoint is $(54.5 + 59.5)/2 = 57$. In this manual, class boundaries will typically be calculated first and then used to determine other values. In addition, the sum of the frequencies is an informative number used in many subsequent calculations and will be shown as an integral part of each table.

- The bin values are the upper class limits for that class, and the lower class limits of the next class may be obtained by adding the degree of accuracy of .01. The class limits can thus be readily obtained. For the second class, for example, they are .50 to .99. Since the gap between classes as presented is .01, the appropriate class limits are increased or decreased by $(.01)/2 = .005$ to obtain the class boundaries and the following table. The class width is $0.495 - (-0.005) = .50$; the first midpoint is $(-.005 + .495)/2 = 0.245$.

| <u>GPA</u> | <u>class boundaries</u> | <u>class midpoint</u> | <u>frequency</u> |
|-------------|-------------------------|-----------------------|------------------|
| 0.00 - 0.49 | -0.005 - 0.495 | 0.245 | 72 |
| 0.50 - 0.99 | 0.495 - 0.995 | 0.745 | 23 |
| 1.00 - 1.49 | 0.995 - 1.495 | 1.245 | 47 |
| 1.50 - 1.99 | 1.495 - 1.995 | 1.745 | 135 |
| 2.00 - 2.49 | 1.995 - 2.495 | 2.245 | 288 |
| 2.50 - 2.99 | 2.495 - 2.995 | 2.745 | 276 |
| 3.00 - 3.49 | 2.995 - 3.495 | 3.245 | 202 |
| 3.50 - 3.99 | 3.495 - 3.995 | 3.745 | 97 |
| | | | 1140 |

5. The relative frequency for each class is found by dividing its frequency by 100, the sum of the frequencies. NOTE: As before, the sum is included as an integral part of the table. For relative frequencies, this should always be 1.000 (i.e., 100%) and serves as a check for the calculations.

| <u>height</u> | <u>relative frequency</u> |
|---------------|---------------------------|
| 55 - 59 | .01 |
| 60 - 64 | .03 |
| 65 - 69 | .49 |
| 70 - 74 | .46 |
| 74 - 79 | <u>.01</u> |
| | 1.00 |

7. The relative frequency for each class is found by dividing its frequency by 1140, the sum of the frequencies. NOTE: In #5, the relative frequencies were expressed as decimals; here they are expressed as percents. The choice is arbitrary.

| <u>GPA</u> | <u>relative frequency</u> |
|-------------|---------------------------|
| 0.00 - 0.49 | 6.32% |
| 0.50 - 0.99 | 2.02% |
| 1.00 - 1.49 | 4.12% |
| 1.50 - 1.99 | 11.84% |
| 2.00 - 2.49 | 25.26% |
| 2.50 - 2.99 | 24.21% |
| 3.00 - 3.49 | 17.72% |
| 3.50 - 3.99 | <u>8.51%</u> |
| | 100.00% |

9. The cumulative frequencies are determined by repeated addition of successive frequencies to obtain the combined number in each class and all previous classes. NOTE: Consistent with the emphasis that has been placed on class boundaries, we choose to use upper class boundaries in the "less than" column. Conceptually, heights occur on a continuum and the integer values reported are assumed to be the nearest whole number representation of the precise measure of height. An exact height of 59.7, for example, would be reported as 60 and fall in the second class. The values in the first class, therefore, are better described as being "less than 59.5" (using the upper class boundary) than as being "less than 60." This distinction becomes crucial in the construction of pictorial representations in the next section. In addition, the fact that the final cumulative frequency must equal the total number (i.e., the sum of the frequency column) serves as a check for calculations. The sum of cumulative frequencies, however, has absolutely no meaning and is not included.

| <u>(#9) height</u> | <u>cumulative frequency</u> |
|------------------------|-----------------------------|
| less than 59.5 | 1 |
| less than 64.5 | 4 |
| less than 69.5 | 53 |
| less than 74.5 | 99 |
| less than 79.5 | 100 |

| <u>(#11) GPA</u> | <u>cumulative frequency</u> |
|----------------------|-----------------------------|
| less than 0.495 | 72 |
| less than 0.995 | 95 |
| less than 1.495 | 142 |
| less than 1.995 | 277 |
| less than 2.495 | 565 |
| less than 2.995 | 841 |
| less than 3.495 | 1043 |
| less than 3.995 | 1140 |

11. The cumulative frequencies are determined by repeated addition of successive frequencies to obtain the combined number in each class and all previous classes. NOTE: Consistent with the emphasis that has been placed on class boundaries, we choose to use upper class boundaries in the "less than" column.

10 Chapter 2

13. Assuming that "start the first class at 0.7900 lb" refers to the first lower class limit produces the frequency table below and violates none of the guidelines for constructing frequency tables.

| <u>weight (lbs)</u> | <u>frequency</u> |
|---------------------|------------------|
| .7900 - .7949 | 1 |
| .7950 - .7999 | 0 |
| .8000 - .8049 | 1 |
| .8050 - .8099 | 3 |
| .8100 - .8149 | 4 |
| .8150 - .8199 | 17 |
| .8200 - .8249 | 6 |
| .8250 - .8299 | 4 |
| | <hr/> |
| | 36 |

NOTE: The class boundaries above are .78995, .79495, .79995, etc. Using 0.7900 as the first lower class boundary produces boundaries of .7900, .7950, .8000, etc. This is not acceptable, as these are possible data values. This introduces subjectivity about where to place a value that falls on the boundary and violates the guideline that each of the values must belong to only one class.

14. Assuming that "start the first class at 0.7750 lb" refers to the first lower class limit produces the frequency table below and violates the guideline that frequency tables should have between 5 and 20 classes.

| <u>weight (lbs)</u> | <u>frequency</u> |
|---------------------|------------------|
| .7750 - .7799 | 4 |
| .7800 - .7849 | 13 |
| .7850 - .7899 | 15 |
| .7900 - .7949 | 4 |
| | <hr/> |
| | 36 |

NOTE: That this frequency table has only 4 categories, which is usually not sufficient to give a picture of the nature of the distribution, is allowable in this context -- since the class limits employed work well with the other cola data and permit meaningful comparisons across the data sets.

15. Assuming that "start the first class at 0.8100 lb" refers to the first lower class limit produces the frequency table below.

| <u>weight (lbs)</u> | <u>frequency</u> |
|---------------------|------------------|
| .8100 - .8149 | 1 |
| .8150 - .8199 | 6 |
| .8200 - .8249 | 16 |
| .8250 - .8299 | 8 |
| .8300 - .8349 | 3 |
| .8350 - .8399 | 1 |
| .8400 - .8449 | 1 |
| | <hr/> |
| | 36 |

While similar to the frequency table in exercise #13, this table differs in two ways. (1) In exercise #13 [Coke], there were 5 classes below the modal class and 2 above; in exercise #15 [Pepsi], there are 2 classes below the modal class and 4 above. (2) The weights in exercise #13 appear to be less than those in exercise #15.

16. Assuming that "start the first class at 0.7700 lb" refers to the first lower class limit produces the frequency table below.

| <u>weight (lbs)</u> | <u>frequency</u> |
|---------------------|------------------|
| .7700 - .7749 | 1 |
| .7750 - .7799 | 6 |
| .7800 - .7849 | 14 |
| .7850 - .7899 | 13 |
| .7900 - .7949 | <u>2</u> |
| | 36 |

While similar to the frequency table in exercise #15, this table differs in two ways. (1) In exercise #15 [regular Pepsi], there were 2 classes below the modal class and 4 above; in exercise #16 [diet Pepsi], there are 2 classes below the modal class and 4 above. In both cases, however, there are more values above the modal class than below it. (2) The weights in exercise #15 appear to be greater than those in exercise #16.

17. For 11 classes to cover data ranging from a beginning lower class limit of 0 to a maximum value of 514, the class width must be at least $(514 - 0)/11 = 46.7$. A convenient class width would be 50, which produces the frequency table given below.

| <u>weight (lbs)</u> | <u>frequency</u> |
|---------------------|------------------|
| 00 - 49 | 6 |
| 50 - 99 | 10 |
| 100 - 149 | 10 |
| 150 - 199 | 7 |
| 200 - 249 | 8 |
| 250 - 299 | 2 |
| 300 - 349 | 4 |
| 350 - 399 | 3 |
| 400 - 449 | 3 |
| 450 - 499 | 0 |
| 500 - 549 | <u>1</u> |
| | 36 |

19. Assuming that "start the first class at 200 lb" refers to the first lower class limit produces the frequency table below.

| <u>weight (lbs)</u> | <u>frequency</u> |
|---------------------|------------------|
| 200 - 219 | 12 |
| 220 - 239 | 9 |
| 240 - 259 | 18 |
| 260 - 279 | 84 |
| 280 - 299 | <u>52</u> |
| | 175 |

Yes. Since the lowest recorded weight before collapse is over 200 and most of the weights are over 260, it appears the cans will withstand pressure that varies between 158 and 165.

20. Assuming that "start the first class at 200 lb" refers to the first lower class limit produces the frequency table given at the right.

Yes. Most of the thicker cans support a weight of 280 before collapse, and they appear to be stronger. Since the thinner cans already meet the criterion given in exercise #19, however, the added strength of the thicker cans may not be worth the added cost.

| <u>weight (lbs)</u> | <u>frequency</u> |
|---------------------|------------------|
| 200 - 219 | 6 |
| 220 - 239 | 5 |
| 240 - 259 | 12 |
| 260 - 279 | 36 |
| 280 - 299 | 87 |
| 300 - 319 | <u>28</u> |
| | 174 |

12 Chapter 2

21. Assuming that "start the first class at 200 lb" refers to the first lower class limit produces the frequency table below.

| <u>weight (lbs)</u> | <u>frequency</u> |
|---------------------|------------------|
| 200 - 219 | 6 |
| 220 - 239 | 5 |
| 240 - 259 | 12 |
| 260 - 279 | 36 |
| 280 - 299 | 87 |
| 300 - 319 | 28 |
| 320 - 339 | 0 |
| 340 - 359 | 0 |
| 360 - 379 | 0 |
| 380 - 399 | 0 |
| 400 - 419 | 0 |
| 420 - 439 | 0 |
| 440 - 459 | 0 |
| 460 - 479 | 0 |
| 480 - 499 | 0 |
| 500 - 519 | <u>1</u> |
| | 175 |

In general, an outlier can add several rows to a frequency table. Even though most of the added rows have frequency zero, the table tends to suggest that these are possible valid values -- thus distorting the reader's mental image of the distribution.

23. The two frequency tables are given below.

a.

| <u>height</u> | <u>frequency</u> |
|---------------|------------------|
| 66 - 67 | 4 |
| 68 - 69 | 3 |
| 70 - 71 | 10 |
| 72 - 73 | 10 |
| 74 - 75 | 0 |
| 76 - 77 | <u>1</u> |
| | 28 |

b.

| <u>height</u> | <u>frequency</u> |
|---------------|------------------|
| 66 - 67 | 6 |
| 68 - 69 | 4 |
| 70 - 71 | 4 |
| 72 - 73 | 4 |
| 74 - 75 | 4 |
| 76 - 77 | <u>6</u> |
| | 28 |

Data set (b) appears to be the phony data for two reasons. (1) The frequencies in set (b) follow a regular pattern unlikely to be achieved by chance, while the frequencies in set (a) follow the type of irregular pattern expected by chance. (2) The pattern in (b) [heights fairly uniformly distributed with more at the extremes than near the middle] disagrees with the generally accepted pattern in (a) [many heights near the middle values and fewer at the extremes].

2-3 Pictures of Data

- 42, the height of the bar centered at 0.0
- 2, the one Monday represented by the bar centered at 1.0 and the one Monday represented by the bar centered at 1.4