Introduction to Statistics
Activity 2.3c **NAME:**

Diabetes Diagnoses

The National Health and Nutrition Examination Survey (NHANES) is a large-scale study conducted annually by the National Center for Health Statistics. It involves over 10,000 Americans, randomly selected according to a multistage sampling plan. All sampled subjects are asked to complete a survey and take a physical examination.

One of the questions asked in the 2003–2004 NHANES survey pertained only to subjects who had been diagnosed with diabetes. Subjects were asked to indicate the age at which they were first diagnosed with diabetes by a health professional. Responses for the 548 subjects with diabetes are summarized in the following frequency table:

Age	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Tally	1	3	7	5	5	2	4	5	1	5	1	2	1	0	2	3	1
Age	18	19	20	21	24	25	26	27	28	29	30	31	32	33	34	35	36
Tally	4	3	1	3	2	1	4	4	2	1	6	5	3	8	5	15	4
Age	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53
Tally	7	10	9	15	9	8	9	5	12	8	7	11	12	25	9	9	7
Age	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70
Tally	8	22	13	8	11	16	19	5	16	9	10	10	9	6	12	7	13
Age	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	87	88
Tally	4	4	9	4	8	7	5	6	3	2	3	2	1	2	1	1	1

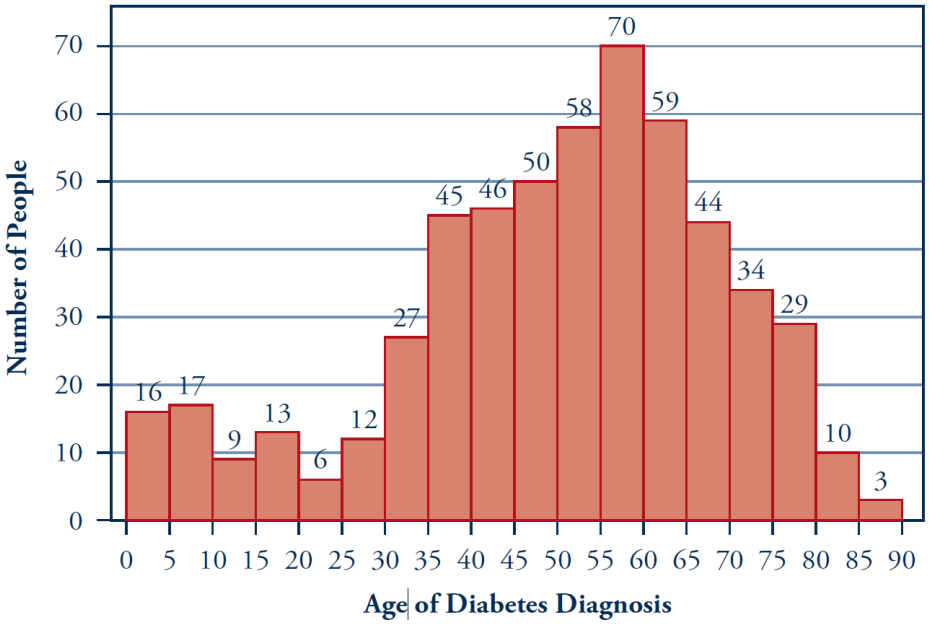
a) Identify the observational units and variable.
Observational units: Variable:

b) Is it practical to construct a dotplot or a stemplot to display this distribution of ages? Explain.

A **histogram** is a graphical display like a dotplot or stemplot, but histograms are more feasible with very large datasets; histograms also permit more flexibility than stemplots. You construct a histogram by dividing the range of data into subintervals (**bins**) of equal length, counting the number (**frequency**) of observational units in each subinterval, and constructing bars whose heights correspond to the frequency in each subinterval. The bar heights can also correspond to the proportions (**relative frequencies**) of observational units in the subintervals.

The following histogram displays the distribution of ages at which the 548 survey subjects were diagnosed with diabetes. Note that the endpoints of the subintervals are reported. For example, the first subinterval indicates that 16 people were diagnosed with diabetes before age 5, and the second subinterval represents

the 17 people who were at least 5 years of age but younger than 10 years of age when they were diagnosed with diabetes. (A person aged 5 years would appear in the 5–10 interval, not the 0–5 interval.)



c) How many and what proportion of the 548 people were diagnosed with diabetes before the age of 20?

d) How many and what proportion of the 548 people were diagnosed with diabetes at age 65 or older?

e) Comment on the shape and center of the distribution of these ages of diagnosis.

f) Use the Histogram Bin Width applet to create a histogram for these data. [The data are stored in the class GoogleDoc under Data: Diabetes. So, you can simply copy and paste the data (and variable name) using the **Edit/Paste Data** button.] What bin width does the applet automatically use? How does this compare to the preceding histogram? How does this affect the appearance of the histogram?

g) Now type 90 in the "Number of Bins" box (bin width of 1 yr) and press **Enter**. Then use 2 bins, (bin width of 45 yrs) and press **Enter**. Which of the four histograms do you think provides the most informative display? Explain. [*Hint*: You want to see the overall pattern but without missing important details.]

Watch Out

- A histogram is not the same as a bar graph. A histogram displays the distribution of a quantitative variable, whereas a bar graph displays the distribution of a categorical variable. Note that the horizontal axis of a histogram is a numerical scale. In fact, all the graphs that you have studied in this topic (dotplot, stemplot, and histogram) apply only to quantitative variables.
- The choice of subintervals (bins) can have a substantial effect on the visual impression conveyed by a histogram. You might want to try several choices to see which subintervals provide the most informative display.
- Histograms are much easier to construct with technology than by hand.
- Some students mistakenly think that the frequencies in the previous data table (1, 3, 7, 5, 5, and so on) are the actual data. In this case, the data are the *ages* at which the subjects were diagnosed with diabetes. Th e frequencies provide a convenient way to report the ages without having to type, for example, the age of 50 a total of 25 times. When you look at a histogram, make sure you are clear on the total number of observational units involved.