THE CHOICE OF AGE GROUPINGS MAY AFFECT THE QUALITY OF TABULAR PRESENTATIONS

Carl E. Pierchala[1], **National Highway Traffic Safety Administration**
400 Seventh Street, SW, Room 6125, NPO-123, Washington, DC 20590

Abstract: Data on the ages of persons are often grouped in tables. However, unequal bin widths are frequently used in setting up the groupings. This can lead to a variety of distortions, depending on the degree to which adjustments are made for nonuniformity in the bin widths of the groups. Using motor vehicle crash data, the issues are illustrated via histograms and bar charts produced with SAS/INSIGHT visualization and analysis software and with SAS/GRAPH. The principles extend to any variable grouped with unequal bin widths. In particular, it is important to be aware of the difference between density and relative frequency. Finally, authors and software developers are urged to give greater attention to the concept of density and the proper labeling of the *y*-axis of a histogram.

[1]E-mail: cpierchala@nhtsa.dot.gov. The views expressed in this paper are those of the author and not necessarily those of the National Highway Traffic Safety Administration or of the U.S. Department of Transportation.

Key Words: Frequency Tables; Histograms; Unequal Bin Widths; Density; SAS/INSIGHT; Motor Vehicle Crash Data

1. INTRODUCTION

In the study of motor vehicle traffic crashes, person age is often an important consideration. In carrying out studies and presenting data concerning crashes, it is often useful to group age data. However, it is not always recognized that the choice of the bin widths in the groupings can be a source of distortion, especially when unequal bin widths are used and counts are reported without some sort of normalization.

Indeed, as can be seen in Table 1, it is common to use bins of unequal width in tables of basic counts for age in work at the National Highway Traffic Safety Administration (NHTSA). In addition to a listing produced by a coworker, Table 1 includes examples from *1996 Traffic Crashes, Injuries, and Fatalities – Preliminary Report* (U.S. Department of Transpor-

Table 1. Age Groupings Used in Selected NHTSA Reports and Work

1996 Traffic Crashes, Injuries, and Fatalities – Preliminary Report (Pages 17-19)		*Traffic Safety Facts 1999* (Table 100); 1994 State Data System Summary (Exhibits 118A-D)		December 2001 Listing Produced by Coworker	
Age Bin	Years in Bin	Age Bin	Years in Bin	Age Bin	Years in Bin
00-04	5	00-04	5	00-04	5
05-09	5	05-09	5	05-09	5
10-15	6	10-15	6	10-14	5
16-20	5	16-20	5	15-20	6
21-24	4	21-24	4	21-24	4
25-34	10	25-34	10	25-34	10
35-44	10	35-44	10	35-44	10
45-54	10	45-54	10	45-54	10
55-64	10	55-64	10	55-64	10
65-69	5	65-74	10	65-74	10
70-74	5				
75-79	5	75-97	25*	75-97	25*
80-84	5				
85-97	15*				

* Approximate, since age coded as 97 means "97 years old or older"

tation, 1997), *Traffic Safety Facts 1999* (U.S. Department of Transportation, 2000), and the 1994 State Data System summary (U.S. Department of Transportation, 1998). Bin widths of 4, 5, 6, 10, 15 and 25 years are all used, depending on the tabulation being considered.

Note that in some but not all cases the age bin 16-20 is used, which has a width of 5 years. This is a very natural bin because it includes youths who, generally speaking, are legally old enough to drive, but not legally old enough to drink alcoholic beverages.

Below I present various histograms and bar charts to give a pictorial representation of the corresponding frequency and relative frequency tables that might be used to summarize the age data. This quickly gives an intuitive representation of the effects of the various choices. Using SAS/INSIGHT software for interactive data exploration, analysis and visualization, it is easy to produce several kinds of histograms and bar charts of age data for persons involved in motor vehicle crashes.

2. RESULTS

As an example, below I use the ages of fatally injured persons as reported in the annual report version of the Fatality Analysis Reporting System (FARS) data on fatal crashes occurring in the United States in the calendar year 2000 (U.S. Department of Transportation, 1999). Unknown values (code 99) are excluded in producing the results given below. Also, note that in FARS the value 97 for age means "97 years old or older". This 'censoring' leads to ambiguity as to the true width of the last age bin.

Figure 1 is a histogram generated with SAS/ INSIGHT's Histogram/Bar Chart option using one-year bins. Note that this is the case of ungrouped data, since age is stored in the FARS data sets as age in years. This histogram shows the greatest level of detail.

One feature is that of *process variability*. Even though the data come from a census of fatal crashes, they are the result of a "process" that produces those crashes. Like all processes, it is subject at least in part to (for practical purposes) random variation. Thus, the frequencies in a given age bin are best viewed as being subject to process variability. For example, in the first dozen years, the frequencies are fairly low and fairly constant, but show fluctuations from one year of age to the next that are essentially random.

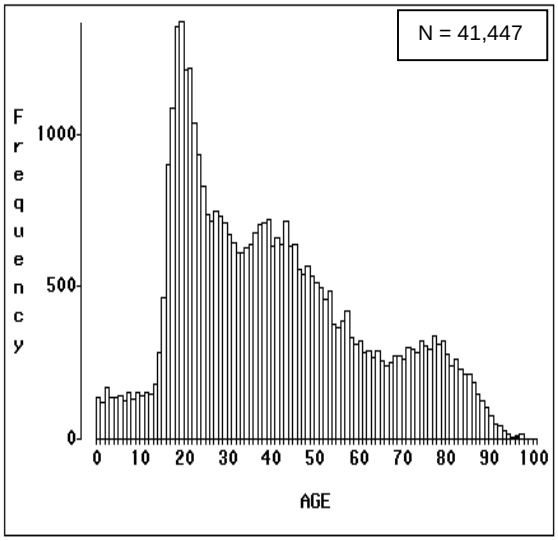

Figure 1. Histogram of age, using one-year bins.

Around thirteen years of age, the frequencies start to rise, reaching a peak at age 19. The frequencies then fall until around age 33. Interestingly, age 21 has a slightly higher frequency than age 20. It is not entirely clear if this is due to process variability or is due in part to persons who have attained legal drinking age. Beginning around age 33, there appears to be a slight increase in frequencies until about age 40, when they start to fall off again. There is another trough in the frequencies around age 67, when they begin to rise again until around age 78. After that the frequencies drop off at a fairly steady rate until they are nearly zero around age 96. There is a slight increase at 97 because this value really represents "97 and older".

The effect of using unequal bin widths in constructing frequency tables is graphically represented in the bar chart in Figure 2. This was produced by first preprocessing the age data to produce the categorical variable AGE_GRPT, which indicates the age category to which each observation belongs. That is, observations with age between 0 and 4 inclusive were assigned the literal '00-04', etc. SAS/INSIGHT's Histogram/Bar Chart option was then used to produce the bar chart.

Note that each category has the same physical width in the graph, even though the number of years is not the same for every category. The bar, with a fixed width, is the perceptual equivalent of one row in a table. From a psychological perspective, people tend

to view each row in a table as equivalent to every other row, even when they are not as is the case with our age data where some categories encompass more years than other categories.

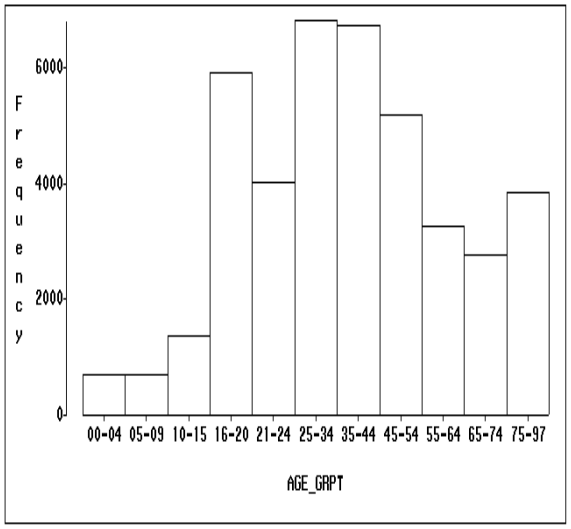

Figure 2. Bar chart of age, using TRAFFIC SAFETY FACTS groupings as bins.

Compared to Figure 1, the shape of the age distribution appears quite different in Figure 2. As examples, the category 21-24 in Figure 2 shows a drop relative to the adjoining categories that is not seen in Figure 1; the category 25-34 has the greatest height, unlike Figure 1 where age 19 has greatest height; and the category 75-97 shows a rise relative to the two categories to its left, quite different from the tailing off seen in Figure 1.

It is clear in Figure 2 that the frequencies in each bin depend in part on the number of years in the bin. It would make more sense to divide the frequency of a bin by the width of a bin to produce a (frequency) *density* for the bin, i.e., frequency per year of age. Indeed, when we first studied statistics, we learned that in a (relative frequency) density function for a distribution, *area under the curve*, not its height, *represents* relative *frequency*. The *height is* the *density*.

If it were possible for SAS/INSIGHT to apply as weights the reciprocal of the bin widths, then that would be equivalent to dividing the bin frequency by the bin width. Now SAS/INSIGHT's Histogram/Bar Chart option does not allow this, but its Distribution option does. However, the Distribution option gives relative frequency histograms/bar charts, rather than those with absolute frequency as in Figures 1 and 2. Happily, this is irrelevant since only the labeling for the *y*-axis changes in a relative frequency histogram/bar chart produced with the Distribution option.

To illustrate how relative frequency charts are different, the Distribution option using AGE_GRPT without weighting was selected. This gave the chart in Figure 3. Indeed, as can be observed there, the height and width of each bar is unchanged vis-à-vis Figure 2. The same is true for the *x*-axis labeling. Only the *y*-axis labeling is changed.

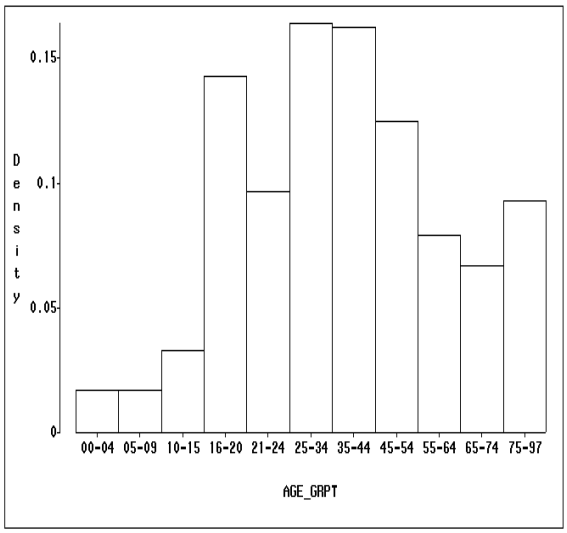

Figure 3. Relative frequency chart of age, using TRAFFIC SAFETY FACTS groupings.

Note in Figure 3 that the *y*-axis is labeled incorrectly by SAS/INSIGHT as 'density'; it is actually relative frequency. The distinction is important. (Frequency) *density* is the *bin frequency / bin width*; whereas *relative frequency* is the *bin frequency / total frequency*. As a final nuance, *relative frequency density* is the *relative frequency / bin width*. The latter concept is in analogy to what statistics texts normally refer to as a density function.

Having verified that the Distribution option gives results equivalent to those of the Histogram/Bar Chart Option, the next step was to apply weighting with the Distribution option. The reciprocal of the number of years in the age bin was used as a weight, and the Distribution option then produced the graph seen in Figure 4. Note that what SAS/INSIGHT refers to in Figure 4 as 'density' is (for each bin) really *weighted frequency* / *total weighted frequency*.

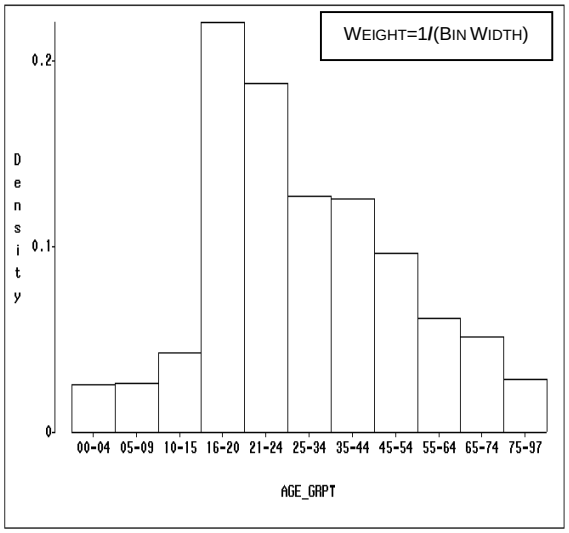

Figure 4. Relative weighted frequency chart of age, using TRAFFIC SAFETY FACTS groupings.

The graph in Figure 4 has a shape that better approximates that in Figure 1 than does the graph in Figure 3. In Figure 4, the category 16-20 has the highest bar. The category 21-24 has the next highest bar, and that is no longer shorter than the bar for the category 25-34, unlike Figure 3. And in Figure 4 the category 75-97 has a shorter bar than the category 65-74, again in contrast to Figure 3.

However, there are still distortions in Figure 4. For example, the horizontal position of the bar for the 16-20 age group is closer to the center of the *x*-axis than is the corresponding age range in Figure 1. The bar for 75-97 is further to the right than the corresponding values are in Figure 1. Furthermore, the relative area under the bar for the 75-97 age group is only 0.0289, much less than the actual proportion of 0.0932 in that age range.

The problem is that although the heights of the bars are better in Figure 4, the widths of the bars are all constrained to be equal, when in reality the widths should vary depending on the width of the age bin. This difficulty in the graph implies a corresponding difficulty when the data are summarized in a table.

The best graphical solution to the difficulty in Figure 4 is to produce a histogram with both proper heights and proper widths for the bars, so that area in the bars does indeed represent frequency. Unfortunately, such a histogram apparently cannot be produced by SAS/INSIGHT. So, the question becomes how to get SAS to produce the desired histogram. PROC GCHART in SAS/GRAPH produces bar charts with gaps between the bars, so they do not resemble histograms. PROC UNIVARIATE produces histograms, but only with fixed bin width. Fortunately, a bit of thought shows that SAS/GRAPH's PROC GPLOT can produce the desired histogram.

The key is to recognize that a histogram consists of a series of rectangles adjoining one another, and that a rectangle is a polygon. PROC GPLOT can easily draw polygons, by connecting consecutive points. Thus, we only need to provide the points determining the four corners of each rectangle in the histogram.

First, the bins along the base of the histogram can be characterized by the following half open intervals: $[X_1,X_2)$, $[X_2,X_3)$, …, $[X_b,X_{b+1})$, where b is the number of bins in the histogram. Let W_i, i=1,…,b, denote the width of the i-th bin. It follows that $X_{i+1}=X_i+W_i$. Also, let F_i denote the frequency in the i-th bin. Then the density D_i in the i-th bin is just $D_i=F_i/W_i$. It now follows that the four points determining the i-th rectangle in the histogram are $(X_i,0)$, (X_i,D_i), (X_{i+1},D_i) and $(X_{i+1},0)$.

Note that the above gives a frequency density histogram; the *y*-axis is labeled as frequency per year of age in our application of this theory. Thus, area in a bar gives the frequency in the corresponding bin. For a proper dimensional result, the frequency represented by the i-th rectangle is given by years in its base on the *x*-axis multiplied by frequency density (frequency per year) on its side on the *y*-axis. Indeed, many textbooks confuse this point by labeling the *y*-axis as 'frequency' rather than 'frequency/unit of *x*-axis'.

The resulting histogram for our FARS data is given in Figure 5. The unequal bar widths are quite obvious. The area in each histogram bar yields the correct frequency for the bin, eliminating the distortions noted previously. Since a value of 97 really means age '97 or greater' in FARS, I arbitrarily extended the last bin width to 25, to provide a rough adjustment for the censoring. Comparing the shapes, it is seen that the coarsely partitioned Figure 5 gives a reasonable approximation to the finely partitioned Figure 1, without the distortions noted in Figures 2-4.

However, there are still some difficulties. First, the local maxima noted around age 40 and age 78 in Figure 1 are lost in Figure 5. Second, the obvious picturing of the varying bin widths in the latter figure has no intuitive counterpart in a tabular display,

where one's mind is likely to inappropriately perceive the categories as equally spaced.

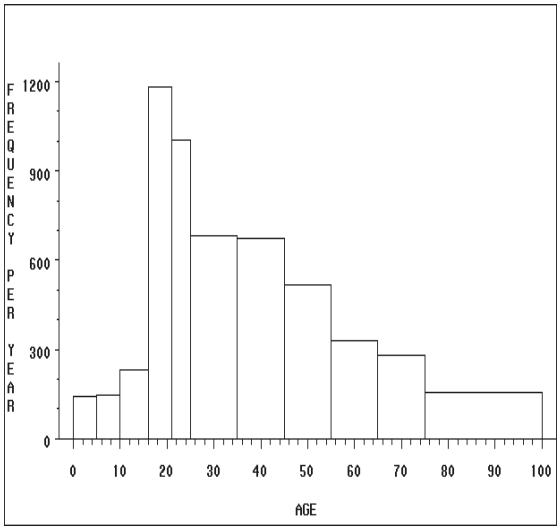

Figure 5. Frequency density histogram, using TRAFFIC SAFETY FACTS groupings.

Thus, it would appear that uniformly and relatively narrowly spaced bins would produce the best results. Now for the case of age, the age interval from 16-20, with width 5 years, is of particular importance in highway safety. Individuals in this age range are generally old enough to legally drive but not to legally drink alcoholic beverages. This suggests using bins of 5 years in width that include the 16-20 age interval. There is a difficulty of age 0 not fitting this scheme well. But age 0 seems to be similar to age 1 to 10 in Figure 1. Thus I would argue to make one exception, by using a six-year category 00-05, and then continue with 5-year categories 06-10, 11-15, 16-20, etc.

In Figure 6, this idea is implemented using SAS/INSIGHT. To get around the problem of unequal bins due to the 00-05 category, age 0 was arbitrarily changed to age 1. This compromise in the first category is clearly seen in a slightly higher than expected height for the first bin. Other than that, the shape in Figure 6 mimics that in Figure 1 quite well. Local maxima at around 40 and around 78 are apparent.

3. DISCUSSION AND RECOMMENDATIONS

The illustrations used above lead to one general recommendation. When making tables, charts and histograms of continuous variables, it is usually best to avoid nonuniform groupings.

The graphical approach used in this paper helps give insight into the consequences for interpretation caused by unequal age bin widths, in both tables and charts. The use of unequal bin widths mandates various adjustments and normalizations so as to avoid distortions in the results. Clearly, the findings apply to other variables and data besides age in traffic crashes.

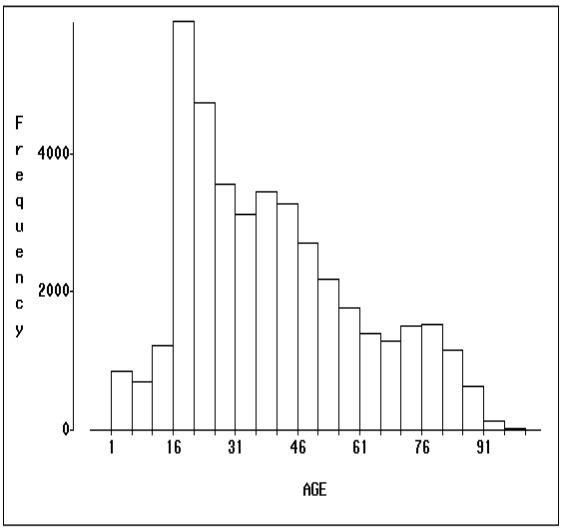

Figure 6. Histogram of age, using five-year bins (but six years in the first bin).

In particular, to report frequencies without any normalization can be very misleading when unequal bin widths are used. This appears to be analogous, loosely speaking, to failing to apply proper weights and analytical methods when using complex probability sampling and complicated experimental designs.

Software developers and authors, especially textbook writers, need to give greater attention to the proper labeling of the *y*-axis of histograms. Indeed, as noted above, SAS/INSIGHT's Distribution option confuses relative frequency with density. Software for producing a histogram needs to either automatically produce a dimensionally correct label for the *y*-axis or allow the user to insert the correct label. The latter should reflect density as frequency per unit of the *x*-axis, as was done in Figure 5.

It has been my observation that there has been a longstanding failure by most authors of textbooks to attend to this detail. When I first studied statistics, I was confused about the concept of probability density. This was finally cleared up when, as an

instructor trying to teach the concepts of histograms, I realized that the *y*-axis of the histogram is usually not properly labeled in textbooks. Note that histograms as described here are appropriate only for continuous variables. Finally, it is my view that the probability function for a discrete distribution should not be called a density function. Calling it a probability mass function would be reasonable.

Labeling the *y*-axis of a histogram incorrectly as 'frequency' fails to reinforce the concept that area under the density curve yields relative frequency, and it fails to make clear the notion of density and the latter's distinction from relative frequency. When teaching the concepts, exercises involving area under the histogram could be used to show students how to estimate frequencies in ranges other than those corresponding to the bases of the histogram's bars. Such examples would likely be a good prelude to finding relative frequencies as areas under a continuous density function such as the Gaussian (Normal).

Finally, the graphical analogies used above lead to a recommendation for work at NHTSA. Except for the first category, use 5-year age groupings when constructing tables involving age, at least in the initial tabulations. That is, when constructing tables involving highway crash age data, use as categories 00-05, 06-10, 11-15, 16-20, …, 96-100, at least initially. Depending on the findings and whether some kind of normalization of the data has been used, this approach might be relaxed as appropriate. Keep in mind that when there are several studies of the same issue that involves age, use of a standard set of age bins will ease comparison of results between studies.

REFERENCES

U.S. Department of Transportation (1997), *1996 Traffic Crashes, Injuries, and Fatalities – Preliminary Report*, DOT HS 808 543, National Highway Traffic Safety Administration, National Center for Statistics and Analysis, Washington, DC: author.

U.S. Department of Transportation (1998), *State Data System: A Summary of Motor Vehicle Traffic Crashes from State Crash Data Files*, DOT HS 808 626, National Highway Traffic Safety Administration, National Center for Statistics and Analysis, Washington, DC: author.

U.S. Department of Transportation (1999), *Fatality Analysis Reporting System, 2000 Coding and Validation Manual*, National Highway Traffic Safety Administration, National Center for Statistics and Analysis, Washington, DC: author.

U.S. Department of Transportation (2000), *Traffic Safety Facts 1999*, DOT HS 809 100 National Highway Traffic Safety Administration, National Center for Statistics and Analysis, Washington, DC: author.