FREQUENCY TABLES, BAR GRAPHS, AND HISTOGRAMS

After taking the Student Political Attitudes Survey early in the semester, many students express an interest in seeing the results of the survey. In one sense, "results" are returned to you early on in the form of the Student Survey Data spreadsheet. You should recall that this is an array of [approximately] 35 columns (corresponding to 35 questions on the questionnaire) and [approximately] 50 rows (corresponding to 50 student respondents). [The actual number of questions varies a bit from semester to semester, and the actual number of student respondents varies considerably more.] Each cell (intersection of a row and column) in the array contains the (numerically coded) response of a particular student to a particular question. Generically, each column corresponds to a *variable*, each row to a *case*, and each cell to the *observed value* (of the variable in that case).

This data set with $35 \times 50 = 1750$ observed values is of course much smaller than most quantitative data sets. The SETUPS/NES data set has 70 variables and 17,650 cases (a total of 1,235,500 observed values). The full NES Cumulative Data File has upwards of 1000 variables and 50,000 cases (about 50 million observed values). Even so, one cannot just stare at the Student Survey data array and draw any substantive conclusions from it. It is necessary somehow to "boil down" this array of quantitative data in order to find meaningful information, patterns, and relationships. This "boiling down" process, sometimes referred to as "number crunching," can now be quickly accomplished for even the largest data sets by using computer software such as SPSS. For a small data set like the Student Survey, it is feasible (but tedious) to do this "boiling down" by hand.

The "boiling down" of quantitative data can proceed either one variable at a time, two-variables at a time, or multiple variables at a time. These processes are referred to as *univariate analysis*, *bivariate analysis*, and *multivariate analysis*, respectively. We will focus on univariate analysis for several weeks, then turn to bivariate analysis for about a month, and then (I hope) have time to take a quick look at multivariate analysis during the last couple of weeks of the semester.

Typically, "boiling down" quantitative data proceeds through two stages. First, we reduce the data to a single relatively compact table (frequency table, crosstabulation, control table, etc.) or corresponding chart (frequency bar graph, histogram, line chart, scattergram, etc.) and, second, we reduce it further to one or several summary statistical measures (measures of central tendency, dispersion, association, correlation and regression coefficients, etc.). Here we look at the process of boiling *univariate data* down to *frequency tables*, *frequency bar graphs*, and *histograms* (or *density graphs*).

Constructing Frequency Tables for Discrete Variables in the Student Survey Data

Recall that the first question in the Student Survey was the following:

> Generally speaking, do you think of yourself as a Republican, a Democrat, an Independent, or what?
>
> (1) Democrat
> (2) Independent
> (3) Republican
> (4) Other; minor party
> (5) Don't know

The data in the first column of the Student Survey spreadsheet gives the responses of all students to this question. The values are coded as shown in the question above. However, in general we need to add one more code (typically "9") for missing data.

To construct a frequency table of this data, we first make a table worksheet (or "template") in the following manner. The first step is to write down a descriptive name for the variable and then to list in the leftmost column all the possible values (note that we can do this only because the variable is either qualitative or, if quantitative, discrete) of the variable (code values and/or labels), including missing data. This creates three (if the variable is dichotomous and provision is made for missing data) or more rows in the frequency table. We should add one more "Total" row and then set up additional columns as shown immediately below.

FREQUENCY TABLE OF PARTY ID (V1)

Value	*Code*	*Tallies*	*IDs*	*Absolute Freq*	*Relative Freq*	*Adjusted Rel Freq*
Dem.	1					
Ind.	2					
Rep.	3					
Other	4					
DK	5					
NA	9					
Total						

We next turn to read the data. We find the column in the data set that corresponds to the variable of interest, read down the column of the spreadsheet, and put a tally mark adjacent to each value as it is observed in the data. When we are done, we should check that the number tally marks matches the number of cases. Unfortunately, it easy to skip or double count cases in this tallying process. For this reason, it is a good idea to record not just a tally mark for each case but the actual ID number of each case. If the number of observations tallied does not match the number of cases, we can review the ID numbers to determine which cases were missed or double counted. (This is one reason why it is highly desirable to assign each case in a data set a unique ID number.)

The next step is to count up the number of tally marks (or IDs) for each value and to enter this number in the *Absolute Frequencies* (or *Case Count*) column. At this point, we have created the basic frequency distribution table. However, several further embellishments are in order.

Often one wants to compare the frequency distributions of the same variable in different data sets, e.g., to compare the distribution of PARTY ID in the Fall 2006 Student Survey with those in earlier student surveys. Since each survey has a different number of respondents, we should focus not on absolute frequencies but on *proportions* or *percentages*, in order to make *valid* comparisons among them. (This point is even more evident if we want to compare Student Survey data, with a few dozen cases, with SETUPS/NES data, with thousands of cases.) Therefore frequency tables

commonly display *relative frequencies* (as percentages) instead of (or in addition to) absolute frequencies. The formula is simply:

$$\textit{Relative Frequency}\ (\%) = \frac{\textit{Absolute Frequency}}{\textit{Total Number of Cases}} \times 100\%$$

Beyond this, often we want to set some cases aside before calculating relative frequencies. Almost certainly we should set all missing data aside, and we may want also to set "don't know," "no opinion," "other," etc., cases aside also. The Fall 2005 Student Survey, every student answered the first PARTY ID question, so there is no missing data coded "9." However, there are several "DK" and "Other" responses that are effectively missing. We might exclude these cases and calculate *Adjusted Relative Frequencies* based on "valid" cases only:

$$\textit{Adjusted Rel. Frequency}\ (\%) = \frac{\textit{Absolute Frequency}}{\textit{N of Cases} - \textit{N of Missing/Invalid Cases}} \times 100\%$$

Putting everything together [and using data from an earlier semester], we get the following complete frequency table:

FREQUENCY TABLE OF PARTY ID (V1)

Values	*Code*	*Tallies*	*IDs*	*Absolute Freqs.*	*Relative Freqs.*	*Adjusted Rel. Freqs.*
Dem.	1			20	46%	49%
Ind.	2	[not		12	28%	29%
Rep.	3			9	21%	22%
Other	4	shown]		0	0%	
DK	5			2	5%	
NA	9			0	0%	
Total				43	100%	100%

The table format above provides a worksheet. In a finished "presentation grade" table (in a term paper, journal article, book, etc.), the variable would be given a fully descriptive name, the codes, tallies, and ID numbers would not be displayed, probably only one type of percentage would be shown, and the case counts might be shown only parenthetically. Such a table would look something like this:

PARTY IDENTIFICATION AMONG POLI 300 STUDENTS, FALL 2006

Democratic	49%
Independent	29%
Republican	22%
Total	100%
	(n = 41)

Source: POLI 300 Student Political Attitudes Survey, Fall 2006

Here is another example from the Student Survey (for which there is no missing or invalid data, so the relative and adjusted relative frequencies are the same):

FREQUENCY TABLE DEMOCRATIC THERMOMETER SCALE (V16)

Value	*Code*	*Abs. Freqs.*	*Rel. Freqs.*	*Adj. Rel. Freqs.*	↓ *Cumulative Rel. Freqs.* ↑	
0-20	1	8	19%	19%	19%	100%
21-40	2	4	9%	9%	28%	81%
41-60	3	16	37%	37%	65%	72%
61-80	4	18	19%	19%	84%	35%
81-100	5	7	16%	16%	100%	16%
Missing	9	0	0%			
Total		43	100%	99%		

Note that V16 is (at least) *ordinal* in nature. The list of values should (like the numerical codes) follow the natural ordering. If the ordering runs from "Low" to "High," standard practice is to put the lowest value at the top of the list and the highest value at the bottom (illogical though this seems), as has been done above. In this table we have added one other type of percentage — namely, *cumulative* (adjusted relative) *frequencies*, where the cumulation can proceed either downwards (↓) or upwards (↑). Thus the "61-80" row of the table shows that 25% of the respondents have 61 to 80 degrees of "warmth" toward the Democratic Party, 87% have this level of warmth or cooler (i.e., 80 degrees or less), and 37% have this level of warmth or warmer (i.e., 61 degrees or more). Note that cumulative frequencies make no sense if the variable in question is merely *nominal* in nature. Since cumulation proceeds from one "pole" of the variable to the other (e.g., "low" to "high" or vice versa), the values must fall into some kind of natural ordering, i.e., the variable must be measured at the *ordinal* level (or higher).

SPSS Frequency Tables for NES Discrete Variables

SPSS produces frequency distribution tables whose (default) format lies between worksheet and presentation-grade. Here are two examples:

V25 DEMOCRATIC CANDIDATE THERMOMETER SCORE

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0-20	2359	12.9	13.4	13.4
	21-40	2787	15.3	15.8	29.2
	41-60	5003	27.4	28.4	57.6
	61-80	3376	18.5	19.2	76.8
	81-100	4097	22.4	23.2	100.0
	Total	17623	96.5	100.0	
Missing	NA	638	3.5		
Total		18260	100.0		

V30 MOST IMPORTANT NATIONAL PROBLEM

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	economy	4581	25.1	36.9	36.9
	foreign affairs	2116	11.6	17.0	53.9
	social welfare	3029	16.6	24.4	78.2
	crime,public order	1889	10.3	15.2	93.4
	other	816	4.5	6.6	100.0
	Total	12430	68.1	100.0	
Missing	NA	5830	31.9		
Total		18260	100.0		

Note that SPSS uses somewhat different labels for different types of frequencies:

Frequency = *Absolute Frequency*
Percent = *Relative Frequency*
Valid Percent = *Adjusted Relative Frequency*

Notice also that SPSS calculates and displays (downward) cumulative frequencies even if they make no substantive sense (as with the nominal variable MOST IMPORTANT NATIONAL PROBLEM). This is because SPSS doesn't "know better": it operates on the code values and cannot tell the difference between different types of variables.

Frequency Charts

Frequency distribution information is often presented by bar (or other types of) charts. To construct a *frequency bar chart*, we first draw a *horizontal line* and place *tick marks* at *equal intervals* along the line. Each tick mark represents a possible value of the (qualitative or discrete) variable; if the variable is ordinal, the marks should follow the natural ordering of the values. Conventionally, the lowest value is to the left and the highest to the right. We then erect a vertical axis that represents (absolute or relative) frequency. Above each tick mark, we erect a bar with some standard width and the height of which is proportional to the frequency of that value. Conventionally the sides of the bars do not touch each other, representing the fact that the values of the variable are discrete.

Figures 1 and 2 are SPSS frequency bar charts for V25 and V30. Missing data was excluded from both charts. Note that the vertical axis can be calibrated in terms of either absolute (Figure 2) or relative (Figure 1) frequencies. Relative frequencies are more typically displayed, especially with sample data (where the actual number of cases is of no special interest). It is possible, of course, to have two axes (one at the left and the other at the right edge of the chart) displaying both absolute and relative frequencies.

Another mode of presenting frequency data is to use *pie charts*, such as are displayed in Figures 3 and 4 (also SPSS output) for the same variables. Since pie charts do not show values in a linear order, they are especially appropriate for displaying frequencies of nominal variables (Thus Figure 4 is a lot more appropriate than Figure 3.) Since such charts show how a "pie" is

"divided up," they are also especially appropriate for displaying "shares," such as how parties divide up popular votes, electoral votes, or seats in a legislature, or how a budget is divided up among different spending categories.

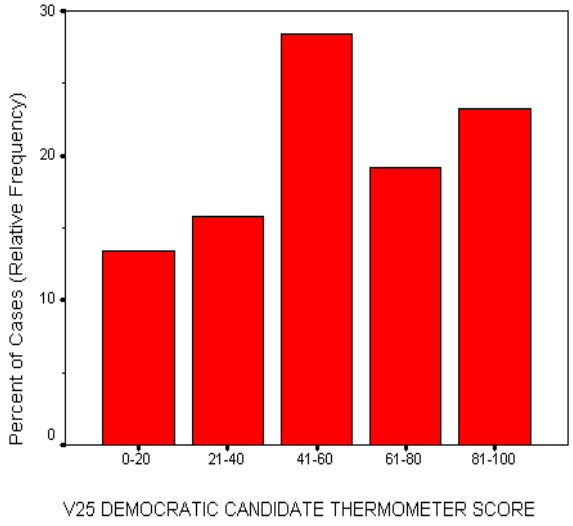

Figure 1

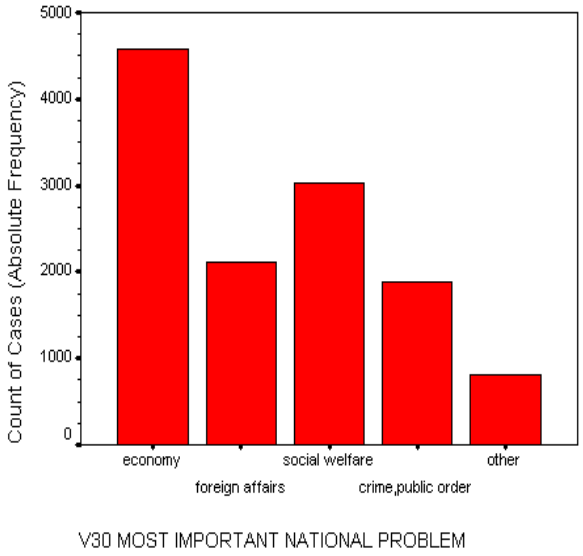

Figure 2

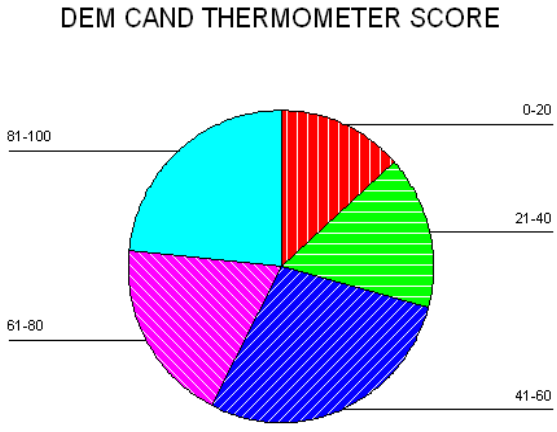

Figure 3

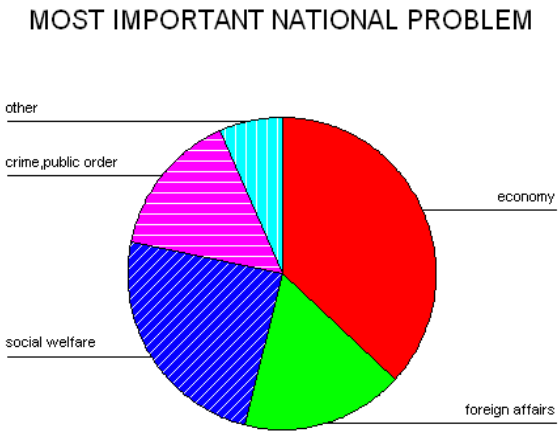

Figure 4

Often we want to compare frequency distributions from different data sets (e.g., compare the distributions of PARTY IDENTIFICATION in different student surveys) or for different subsets of cases (e.g., men vs. women) in the same data set. Figure 5 shows how gender-specific frequency distributions of the MOST IMPORTANT NATIONAL PROBLEM variable can be *merged* into a single bar chart in a way that facilitates comparison between men and women. (It shows that men tend to be more concerned with economic issues and women with social welfare issues, but in other respects there is little "gender gap.") Clearly such merged bar graphs should display *relative* frequencies if the data sets (or subsets) being compared are of different size. You might merge (hand drawn) bar graphs in this manner when you compare Student Survey and SETUPS/NES data in Problem Set #5A.

Figure 6 shows a similar but somewhat more elaborate graph pertaining to Presidential candidate "Thermometer Scores." To make it more compact, Figure 6 does not display the entire frequency distribution (over five levels of warmth). Only a single frequency is displayed for the two highest levels (61-80 degrees and 81-100 degrees) combined, with the Democratic and Republican candidate frequencies merged. The chart then further merges together graphs for each election year. The single overall graph "tells the story" of the balance of "warm feelings" toward the Democratic and Republican candidates over the last 30 years (and suggests one reason why the 1972, 1984, 1988, 1996, and perhaps 2004 elections came out as they did).

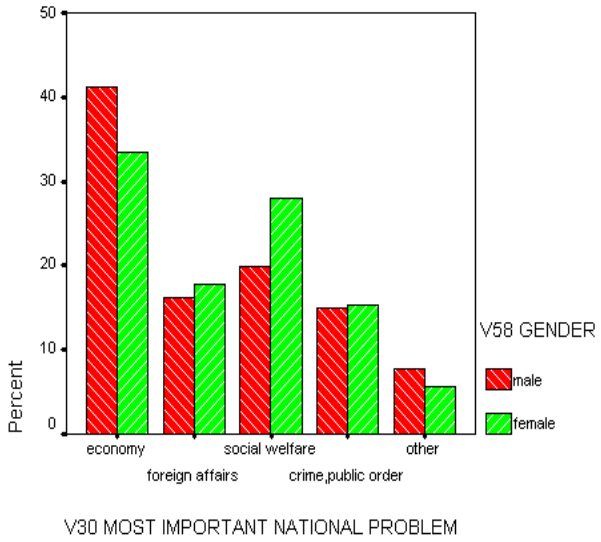

Figure 5

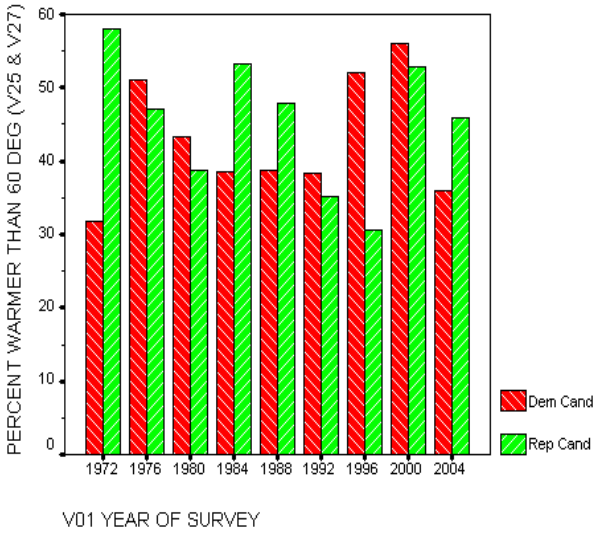

Figure 6

Another way to compress and merge bar graphs is displayed in Figure 7. Here we "stack" all the bars of the frequency distribution of MOST IMPORTANT PROBLEM on top of one another to form a single bar representing 100% of the cases. We then combine nine such stacked bars to "tell the story" of the changing perceived importance (or "salience") of different

types of issues in Presidential elections over the last 33 years. (Note for example that foreign policy concerns were much more prominent in 1972, 1980, 1984, and 2004 than in other years.)

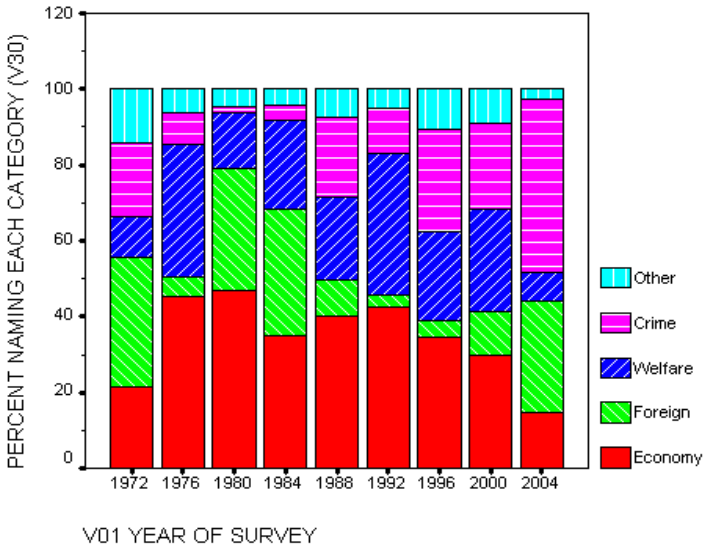

Figure 7

Continuous Variables, Class Intervals, and Histograms

Recall that the first step in constructing a frequency table is to list all possible values of the variable. But clearly we cannot do this if the variable of interest is quantitative and continuous in nature, because such a variable has an infinite number of possible values. How then should we proceed?

Remember that all points along (some interval of) the real number line represent possible values of a continuous (and interval) variable. One way to proceed is to divide the line representing values of the variable up into a (relatively small) number of segments called *class intervals*. We noted in Problem Set #3B that some of the variables in the SETUPS/NES data are "truly continuous" in nature but have been effectively turned into discrete variables. This was accomplished by creating class intervals for such variable as V60 (AGE), V65B and V65C (DOLLAR INCOME), and all the "Thermometer Scales." (See their entries in the SETUPS/NES Codebook.) Once such class intervals have been created, we can proceed to create frequency tables and charts in the same manner as with discrete variables. (Indeed, we have already done this with respect to V25 DEMOCRATIC CANDIDATE THERMOMETER SCORE.)

Table 1 shows the raw data set for the single continuous variable PERCENT OF POPULATION AGED 65 OR HIGHER pertaining to the U.S. states. As is inevitable, the data is not recorded entirely *precisely* but rather is rounded off to the nearest one-tenth of one percent. For

example, Illinois and Indiana almost certainly have different values on the variable, but this does not show up because of rounding. To boil the data down to a frequency table or graph, we might create class intervals one percentage point wide, i.e., 0-1%, 1-2%, etc.[1]

TABLE 1 – PERCENT OF POPULATION AGED 65 OR HIGHER IN THE 50 STATES (UNIVARIATE DATA ARRAY)

Alabama	12.4	Montana	12.5
Alaska	3.6	Nebraska	13.8
Arizona	12.7	Nevada	10.6
Arkansas	14.6	New Hampshire	11.5
California	10.6	New Jersey	13.0
Colorado	9.2	New Mexico	10.0
Connecticut	13.4	New York	13.0
Delaware	11.6	North Carolina	11.8
Florida	17.8	North Dakota	13.3
Georgia	10.0	Ohio	12.5
Hawaii	10.1	Oklahoma	12.8
Idaho	11.5	Oregon	13.7
Illinois	12.1	Pennsylvania	14.8
Indiana	12.1	Rhode Island	14.7
Iowa	14.8	South Carolina	10.7
Kansas	13.6	South Dakota	14.0
Kentucky	12.3	Tennessee	12.4
Louisiana	10.8	Texas	9.7
Maine	13.4	Utah	8.2
Maryland	10.7	Vermont	11.9
Massachusetts	13.7	Virginia	10.6
Michigan	11.5	Washington	11.8
Minnesota	12.6	West Virginia	13.9
Mississippi	12.1	Wisconsin	13.2
Missouri	13.8	Wyoming	8.9

SOURCE: *Statistical Abstract of the United States*, 1989

Figure 8 shows an SPSS *histogram* for the data with class intervals one percentage point wide. (However, the intervals are 3.5-4.5% and so forth and the value labels are the whole numbers at the *mid-point* of the interval. You can verify that the 11.5 and 12.5 observations are included in the 11.5-12.5 and 12.5-13.5 intervals respectively.) This histogram is *logically equivalent* to a frequency bar chart, with the merely *cosmetic difference* that the bars touch each other (reflecting the continuous nature of the variable).

[1] Note that the numerical bounds on adjacent intervals must "touch" each other so that every possible value is included in some interval. Note incidentally that the AGE intervals in the SETUPS/NES Codebook appear *not* to "touch" in this way. Presumably the 17-24 interval actually includes everyone who has not yet turned 25 (and so would be better be written as 17-25), and likewise for other AGE intervals. The same consideration applies to THERMOMETER SCORE intervals. The DOLLAR INCOME intervals, on the other hand, are consistent with the rule that intervals should touch. Finally, we need some rule (disclosed to readers) about whether (for example) a case with a rounded value of 1.0% goes into the 0-1% or 1-2% interval.)

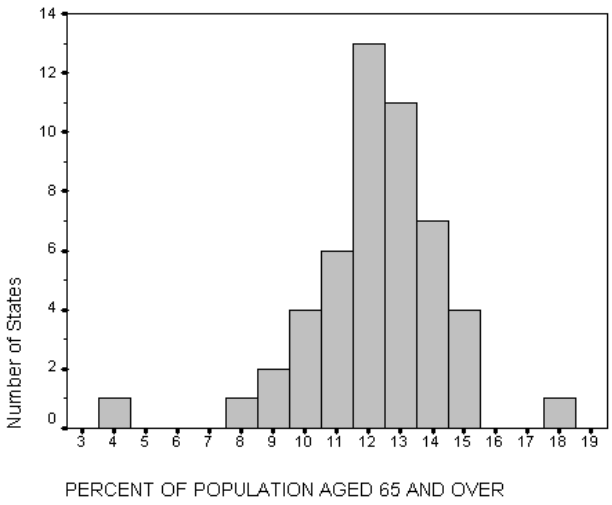

Figure 8

However, the histogram in Figure 8 is essentially no different from a frequency bar chart only *because the class intervals all have equal width*. Otherwise, a bar chart and a histogram of the same data may look quite different, and the bar chart may present a misleading picture of the data. This can be illustrated by focusing on the SETUPS/NES variable V65D (DOLLAR INCOME IN 2004), for which *unequal* class intervals were created. Here is the SPSS frequency table for V65D.

V65D DOLLAR INCOME (2004)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Less than$15,000	145	12.0	13.7	13.7
	$15,000 to $25,000	121	10.0	11.4	25.2
	$25,000 to $35,000	102	8.4	9.7	34.9
	$35,000 to $50,000	154	12.7	14.6	49.5
	$50,000 to $80,000	246	20.3	23.3	72.8
	$80,000 to $120,000	167	13.8	15.8	88.6
	More than $120,000	120	9.9	11.4	100.0
	Total	1055	87.0	100.0	
Missing	NA	157	13.0		
Total		1212	100.0		

Figure 9 shows the standard SPSS bar chart for this frequency distribution, which appears to display distribution of income that is approximately uniform — that is, all bar are approximately the same height — except for a peak (or "mode") in the third highest income category. Indeed, since the bars for the two highest income categories are about the same height as the the bars for the two lowest income categories, the impression the bar graph conveys to the eye is that there are as many well-off than not-so-well-off people. However, this impression is quite misleading, as you can begin to understand when you look more as you study the income class intervals and notice that they are not of equal width.

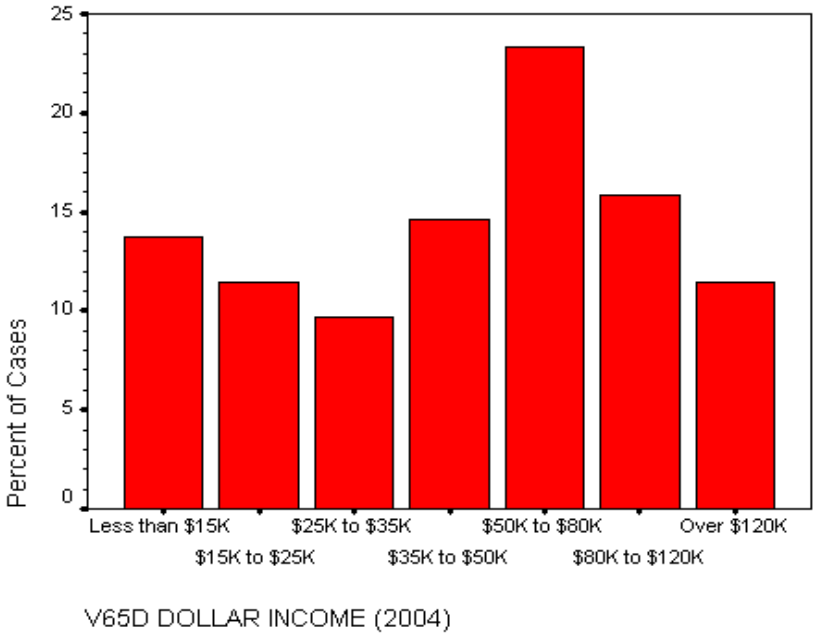

Figure 9

Figure 10 (hand-drawn since SPSS cannot readily create such a graph) presents a histogram of this data. The fundamental difference between a frequency bar graph in Figure 9 and a histogram in Figure 10 is that in the former *frequency is represented by height* while in the latter *frequency is represented by area*.

To draw a histogram, we first draw a horizontal line representing the possible values of a variable. Since the variable is interval, we can place tick marks at equal intervals to mark equal increments in the value of the variable, e.g., \$0K, \$20K, \$40K, etc. for INCOME. Next we put other marks along the scale at the points that separate the class intervals — in this case at \$15K, \$25K, \$35K, \$50K, \$80K, and \$120K.[2] Next we erect a rectangle (a "bar," if you wish) on each

[2] Note that the last class interval presents a problem, in that it is an *open interval* that has no upper bound and thus no definite width. (The first class interval may appear to be likewise open, with no lower bound. But INCOME is a ratio variable that cannot have a value below zero, so the first interval is simply \$0-15K.) Here I have addressed the problem presented by the last class interval by setting an upper bound more or less arbitrarily at \$250K. A typical NES sample of some 1500-2000 Americans would turn up a few respondents with incomes higher than this and probably none *much* higher than this, but of course such people do exist in the population from which NES samples are drawn. (Some CEOs — many recently disgraced — have received annual compensation on the order of \$250,000K [\$250 million]; such an income level would lie on the income scale in Figure 10 about one fifth of a mile beyond the right margin of the page.)

class interval, so that the *area* of each rectangle is proportional to the frequency associated with that class interval. Remember that the width of each rectangle is the width of the class interval. How tall should each rectangle be? Well, remembering what we learned in the third grade, we know that

$$Area = Height \times Width \quad \text{so} \quad Height = Area \,/\, Width.$$

Since *Area* here represents *Frequency*, we have

$$Height = Frequency \,/\, Width,$$

where *Width* is the width of the class interval. Thus we can calculate the following (relative) heights. (Since only relative magnitudes matter, we can ignore the \$000 = \$K in INCOME values.)

Class Interval	*Width*	*Frequency*	*Height*
0-15	15	13.7	13.7 / 15 = 0.913
15-25	10	11.4	11.4 / 10 = 1.140
25-35	10	9.7	9.7 / 10 = 0.970
35-50	15	14.6	14.6 / 15 = 0.973
50-80	30	23.3	23.3 / 30 = 0.777
80-120	40	15.8	15.8 / 40 = 0.395
120-250	130	11.4	11.4/130 = 0.088

To construct the histogram, we need first to construct a vertical axis so that we can set the height of the rectangles. The tallest rectangle has a height of about 1.14, so the axis should extend a bit higher than this, say to 1.2.[3] Such an axis is shown in Figure 10. However, this axis should be regarded as "scaffolding" — put up to help construct the histogram but taken down once the construction is finished. The reason for removing the axis from a "presentation-grade" histogram is that readers are likely interpret it as representing frequency, like the vertical axis in a bar graph.

Given that height in a histogram does *not* represent frequency, we might ask what height in a histogram *does* represent? The answer is that it represents *density* — that is, *how closely cases are "packed into" each class interval*. Note that the class interval \$50-80K includes about twice as many cases (23.3%) as the interval \$15-25K (11.4%). This fact is reflected in the *bar graph* in Figure 9 by the fact that the bar on the \$50-80K interval is about *twice as high* as the bar over the \$15-25K interval. It is reflected in the histogram in Figure 10 by the fact that the "bar" (rectangle) on the \$50-80K interval has about *twice the area* of the bar on the \$15-25K interval. But the 23.3% of the cases in the \$50-80K interval are spread over an income interval

---

[3] Usually the physical lengths of the horizontal and vertical axes of printed chart are set so that they form either (i) an approximate square with Height = Width or (ii) an approximate "golden rectangle" in which

$$\frac{\text{Height}}{\text{Width}} = \frac{\text{Width}}{\text{Height} + \text{Width}} \approx \frac{1}{1.6}$$

You will notice that SPSS charts follow the latter convention.

that is three times as wide than the interval into which the 11.4% of the cases in the \$15-25K interval are packed, so the former is only about two-thirds as tall as the latter. To take another example, notice that the \$15-25K and \$120-250K interval have bars of exactly equal height, reflecting their equal frequencies (11.4%) but the latter is 13 times as wide and its rectangle therefore has only 1/13 of the height as the former.[4]

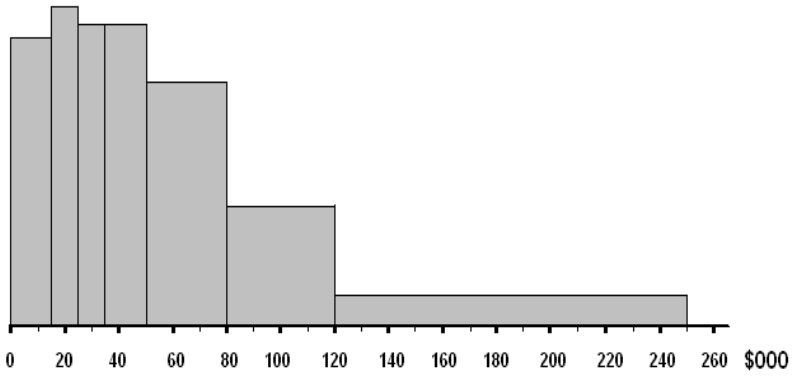

Figure 10

These considerations also explain why a histogram is logically equivalent to a frequency bar graph (and therefore can have a vertical axis representing frequency) only in the special case in which all class intervals have equal width. If all class interval have the same width, frequency depends only on density. Geometrically, if all rectangles have the same width, the area of the rectangles depends only their heights, so height represents frequency as well as area does.

The histogram in Figure 10 clearly presents a rather different (and more accurate) picture of the distribution of income from the bar graph in Figure 9.

As a concluding point, it might be mentioned that popular newspapers (especially *USA Today*), magazines, advertisements, etc., like to present bar graphs but usually can't resist the temptation of making them "cute" by letting figures of one sort or other take the place of simple bars. And often, as the heights of the figures vary, their widths also vary in a proportionate manner. The eye then tends to compare areas rather than heights, producing distinctly misleading impressions. Cuteness trumps clarity. See Figure 11 (taken from Moore, *Statistics: Concepts and Controversies*) for examples. (Other times, charts are simply incorrectly drawn; see Figure 12.

---

[4] It might be helpful to point out that a histogram type of diagram could be used to display the areas, populations, and population densities of the U.S. states. Each state would be represented by a segment of the horizontal axis proportional to its *area* [square miles]. The *population* of each state would be represented by the area of the rectangle erected on its interval. The height of the rectangle would represent the each state's *population density* [people per square mile]. Only if all states had the same area would their populations would depend solely on their population densities.

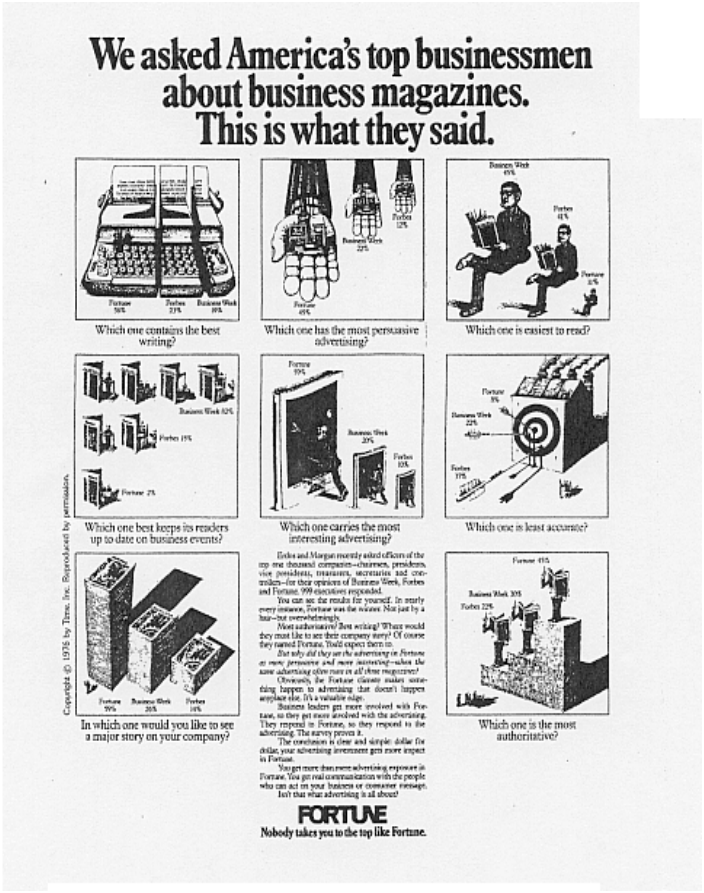

Figure 11

Continuous Densities

The histogram in Figure 10 is based on a small number of (rather wide) class intervals and a modest number of cases (n = 1212). Suppose we actually have INCOME data that is recorded very *precisely*, e.g., to the near dollar or even cent. (This data would not be any more *reliable*, of course; no respondent could tell a survey interviewer what his or her annual household income is to the nearest dollar.) Suppose also we have a huge number of cases, perhaps even census data. We could then refine income into narrower and narrow class intervals, redrawing the histogram accordingly. If we pushed this process to the limit, we would probably end up with what would be an essentially *continuous* (and probably fairly smooth) *density curve* such as is enclosed with Problem Set #5C. Such a curve can be sketched in fairly plausibly even in the very crude histogram with only seven income class intervals, in the manner of Figure 12.[5]

[5] Contrary to the (hypothetical) continuous density curve for INCOME with Problem Set #5C, the SETUPS/NES data suggests that the distribution of household income has two "peaks" (or *modes*), one at about $18K and another at about $43K, with a slight "valley" between them. This probably results from the fact that there are two types of households: family or multi-person households (typically two or more adults and often children as well) and single-person households (typically widows/widowers or young adults who have recently "flown the nest" but are not yet married with children). On average, the former type of household has (and needs) higher income than the latter. This tends to produce two peaks in the overall distribution of household income.

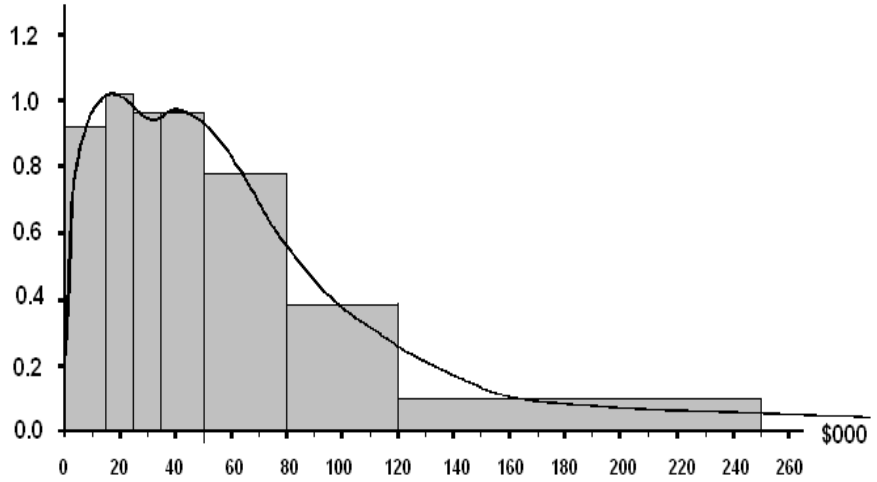

Figure 12

A continuous density of this sort is a histogram with minutely small class intervals, so it remains true that frequency is represented by area. Thus, given the hypothetical income density curve enclosed with Problem Set #5C, you can estimate the proportion or percent of cases that lie within any income interval by making an "eyeball" estimate of the area under the curve that lies within the interval as a proportion or percent of the total area under the curve (the latter of course representing 100% of the cases). The height of the density curve at any point along the income scale indicates relatively how many cases are packed into the vicinity of that level of income. For example the PS#5C curve appears to be about twice as high at $30K as at $40K. This does *not* mean that twice as many people have household incomes of exactly $30,000.00 as have incomes of exactly $40,000.00. (Quite likely, *nobody* has *exactly* either level of income.) It does mean that about twice as many people have household incomes of "just about" $30K (say $29-31K) as have incomes of "just about" $40K (say $39-41K). This follows directly from the "area represents frequency" principle. We are comparing the areas of two very narrow and almost rectangular strips under the density curve. (They are not rectangles, because their tops are not level straight lines, but they are so narrow that this makes almost no difference.) Since they have the same width, their area depends only on their heights.

A density curve like the income curve is said to be *skewed*, because it has a long thin "tail" in one direction (high income) and a short thick "tail" in the other direction (low income). The most famous and common continuous density curve is the *normal curve*, which is not skewed but rather *symmetric* about its peak (so that it is identical to its mirror image). The normal distribution will be discussed in Handout #8.