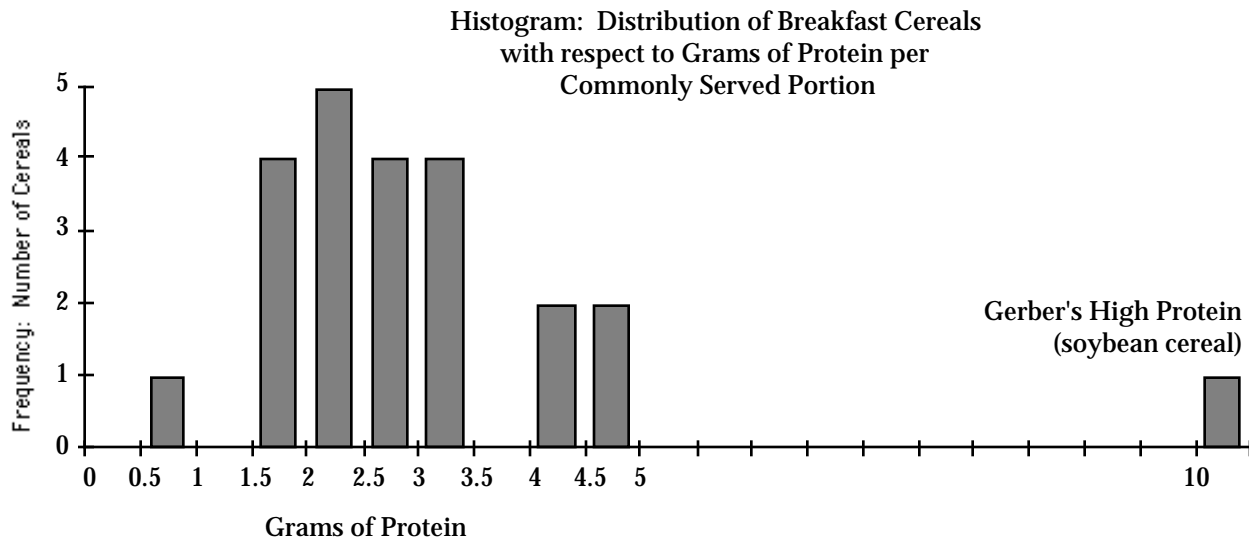


# Histograms

The most widely used graphical device for for the *formal* display of a distribution the histogram. For example, Figure 1 is a histogram of distribution of protein count from breakfast cereals.



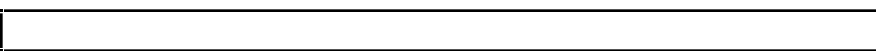
The histogram is a stylized version of the stem and leaf — cleaned-up for public inspection, with much of the information removed. It’s the kind of thing you might use after the analysis is completed — when you are constructing a “pretty” final report — or use when the number of objects involved is massive as, for example, the distribution of family incomes for 100-million families, or when you use a computer program that takes some care with its graphics.

My only objection to these things is that, because they are “clean,” and often computer drawn, the impression is created that they are somehow better or, at least, more scientific than the stem and leaf — when exactly the opposite is the case: That “mess” in the stem and leaf is information, information that is not available in the histogram. And for that reason, the histogram is not the kind of thing that a real human being does early, as a first resort — when you are in hot pursuit of information. For that, by hand, most people would use tallies, *###*, or digits or labels, as I’ve done in the stem and leaf drawings.

That being said, and the warning having been sounded, histograms are often used for income distributions; they play an important role in numerical work on ecology,<sup>1</sup> and they are useful for visually contrasting two different distributions — where they serve much the same function as a stem and leaf diagram.

I want to examine the construction of a histogram in painful detail because there are a few tricks to it, a few places where its construction is different from the construction of the stem and leaf. In particular, you should focus on two questions: First, what is the vertical axis of the histogram, the “height?” The height is not always proportional to the simple count (as it was in the stem and leaf). And, second, what is the area of the histogram, the “area” under the curve? (Usually, these things are simply unlabelled — the computer doesn’t know what they are. But you have to know what they are in order to build a histogram for yourself, and you have to be able to build one for yourself in order to be sure you understand it.)

For inspection, here is a set of histograms: A histogram of gross national products (by nation), a histogram of numbers of animals per species within a \_\_\_\_, and a pair of histograms showing the family income by “race” in the United States.



---

<sup>1</sup> See *Ecological Diversity and Its Measurement* by Anne E. Magurran, Princeton University Press, 1988.

---

---

Use the SocPol data for Gross National Products or new data from the World Bank.

Use the biodiversity data from Magurran

Use the U.S. census, family incomes

Histograms look simple, and *are* simple for data like the distribution of protein content of breakfast cereals. But when you get to data like the income distributions, you have to be clear about the construction rules. The problem is that the widths of the categories, \$1,000 at one end (e.g., from \$1,000 to \$1,999) and \$4,000 at the other (from 10,000 to 14,999). To get it right, let me look at the construction of this histogram in painful detail.

The key point to remember in constructing a histogram is that the shaded area is proportional to the number of “things” whose distribution is being described. Knowing the meaning of the area, answers both questions at once: If the “things” are thirty breakfast cereals, then the shaded area is proportional to thirty breakfast cereals. If the “things” are 100 percent of the income earning families, then the shaded area is proportional to 100 percent.

For the first detailed example, I’m going to build the histogram for U.S. family income. Figure   , the first two columns on the left,

shows the data exactly as they come from the book (The U. S. Book of Facts Statistics, and Information, 1972, page 316.)

**Note: All I have handy is old data, for some reason. Revise to something more recent.**

Money Income Percent Distribution of Families by Income Level, Table 500, page 316, U.S. Book of Facts Statistics and Information, 1972.

Income Level	Population in Percent %	Income Interval in dollars \$	Height in percent per dollar %/\$	Height in simple numbers (without units) used for drawing the histogram
Under \$1,000	1.6%	~\$1,000	~.0016 %/\$	16
\$1,000 to \$1,999	3.1%	\$1,000	.0031 %/\$	31
\$2,000 to \$2,999	4.6%	\$1,000	.0046 %/\$	46
\$3,000 to \$3,999	5.3%	\$1,000	.0053 %/\$	53
\$4,000 to \$4,999	5.4%	\$1,000	.0054 %/\$	54
\$5,000 to \$5,999	5.9%	\$1,000	.0059 %/\$	59
\$6,000 to \$6,999	6.4%	\$1,000	.0064 %/\$	64
\$7,000 to \$9,999	21.7%	\$3,000	.0072 %/\$	72

\$10,000 to \$14,999	26.7%		\$5,000	.0053 %/\$	53
\$15,000 or over	19.2%		Ill-defined guess \$100,000	Ill-defined guess .00019 %/\$	02

Now, to construct the histogram. First I get out my graph paper and mark off the income levels, left to right, marking \$1,000, \$2,000, and so forth. These marks give me the left and right boundaries of each piece of the histogram. But note, the intervals are not equal in size.

supply graph in progress.

Now, all I have to do is supply a height for the shaded area over each of the intervals, a height for the area between 0 and \$1,000, a height for the area between \$1,000 and \$2,000, and so forth. Here's where I fill in the last three columns of the table:

For the first area, here's what I know: I know that the area is 1.6% and I know that the width of the interval is approximately \$1,000. And I also know, from simple geometry that

$$\text{Area} = \text{Width times Height}$$

So, what's the height? Simple: If area equals width times height then height equals area divided by width

$$\text{Height} = \text{Area} / \text{Width}$$

Using the equation for this particular problem, the first equation, area equals width times height means

$$1.6\% = \$1,000 \text{ times Height}$$

And so, the height is

$$\text{Height} = \frac{1.6\%}{\$1,000}$$

which is

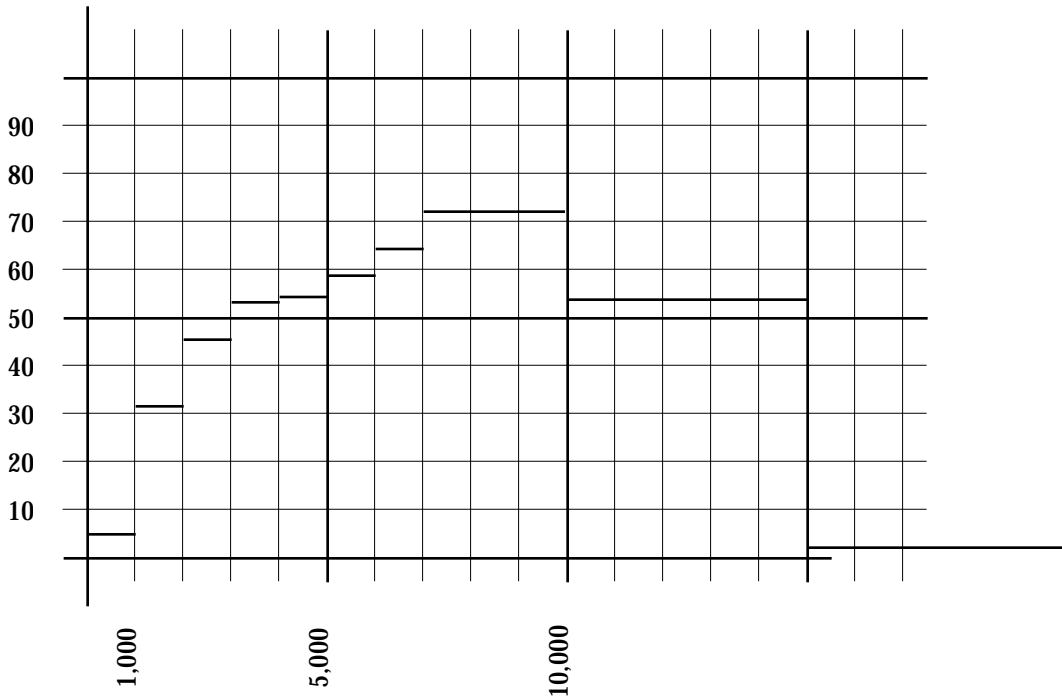
$$.0016 \text{ percent per dollar}$$

That is the height of this first part of the histogram, and those are the units on the vertical axis: The height measures percent of the population per dollar of family income.

Incomplete graph, with one piece of the histogram, and the vertical axis labelled in units from 0 to 100 percent of the population per dollar of family income.

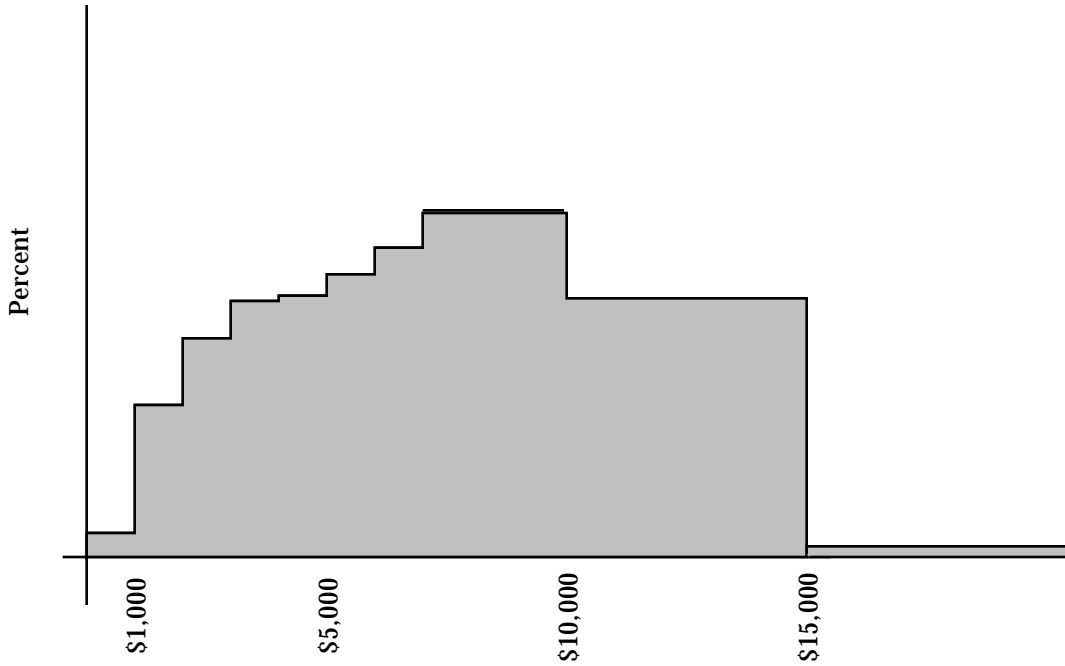
The thing you have to compute in order to draw the histogram is the height even though the data that you get describe the base and the area. So you have to compute the height which is, in this case, percent of the population per dollar of family income. The thing that you have to get right is the height, while the data describe the base and the area. So we compute the height that makes the area right.

Having computed the height, I get out my graph paper and translate the numbers onto a grid.



And, then I begin to clean things up. At the low end, I'm not really sure that \$0 is really the bottom on income — you can do worse than break even. But I'll assume that "0" is the left end of the first interval. At the high end, for the interval above \$15,000, I have a more difficult problem. In fact, it is an insoluble problem and there's not much I can do about it: Fact is, I don't know what to use for the highest income, which makes the width of the last piece of the graph entirely arbitrary. And since the width is arbitrary, the height of this piece of the histogram is also arbitrary. You could reasonably ask for better data: Lumping everyone above \$15,000 dollars is wiping out an awful lot of detail. But even with better data, there's always going to be a last category and often, with income data, there's no answer. I've arbitrarily chosen \$100,000 as my top, and forged ahead with my arbitrary choice. And here's the result

(fix the label on the vertical axis -- in fact, do the whole thing over again on more recent data)



To take a look at a more “classical” shape, consider this histogram of total population of nations, using 1975 data from the *World Handbook of Social and Political Indicators*

(check World Bank for more recent data. Show both the chart and the histogram.)



**Homework:**

Family Income 1991, General Social Survey, National Opinion  
Research Center, University of Chicago

Income Level	Number of Families (Sample size = 1517) families	Income Interval in dollars \$	Height in number per dollar families/\$	Height (for use on graph)
Under \$1,000	11			
\$1,000 to \$2,999	26			
\$3,000 to \$3,999	31			
\$4,000 to \$4,999	35			
\$5,000 to \$5,999	39			
\$6,000 to \$6,999	29			
\$7,000 to \$7,999	23			
\$8,000 to \$9,999	38			
\$10,000 to \$12,499	76			
\$12,500 to \$14,999	82			
\$15,000 to \$17,499	97			
\$17,500 to \$19,999	60			

**Histograms****Joel H. Levine**

\$20,000 to \$22,499	60				
\$22,500 to \$24,999	68				
\$25,000 to \$29,999	112				
\$30,000 to \$34,999	94				
\$35,000 to \$39,999	86				
\$40,000 to \$49,999	149				
\$50,000 to \$59,999	86				
\$60,000 to \$74,999	86				
\$75,000 and higher	80				
Refused	85				
Don't Know	47				
No Answer	17				

Household size: Number of Household Members (General Social Survey, 1991, Var "HOMPOP, #33")

Household Size	Number of Respondents (n=1517)
1	377
2	476
3	275
4	241
5	98
6	29
7	14
8	2
9	2
10	2
No Answer	1

Respondent's Education General Social Survey, 190, Variable "EDUC", #15

Grade or Years	Number of Respondents (n=1517)
Schooling	2
1st grade	0
2nd grade	0
3rd grade	5
4th grade	5
5th grade	6
6th grade	12
7th grade	25
8th grade	68
9th grade	56
10th grade	73
11th grade	85
12th grade	461
1 year of Coll	130
2 years	175
3 years	73
4 years	194
5 years	43
6 years	45
7 years	22
8 years	30
Don't Know	0
No Answer	0

Respondents Age VARIABLE: AGE #12

Age	Number of Respondents (n=1517)
10 - 19	12
20 - 29	293
30 - 39	382
40 - 49	280
50 - 59	165
60 - 69	171

---



---

70 - 79	148
80 or over	63
No answer,	
Don't know	3

Variable SIBS #10: How many brother and sisters did you have? Please count those born alive, but no longer living, as well as those alive now. Also include stepbrothers and stepsisters, and children adopted by your parents.

Number of Siblings	Number of Respondents (n=1517)
0	74
1	236
2	276
3	236
4	209
5	118
6	80
7	81
8	58
9	47
10	34
11	22
12	11
13	9
14	5
15	3
16	1
17	2
18	1
19	0
20	0
21	1
22	0
23	0
24	0
26	1
Don't know	4
No Answer	8

VARIABLE: HRS1, If working, full or part time: How many hours did you work last week, at all jobs

Number of Hours	Number of Respondents (n=1517)
0 - 09 hours	25
10 - 19 hours	55
20 - 29 hours	75
30 - 39 hours	117
40 - 49 hours	397
50 - 59 hours	114
60 - 69 hours	65
70 - 79 hours	17
80 or more	18
No ans, don't know	1
Not Applicable	633

Ed: Get each of the above separated on "race", gender, and perhaps, as a table, age, educ, income.

The SRC data for education is better, showing military, trade school, etc.