Variables and data types

(*) Data comes from ***observations***.

(*) Each observation yields values for one or more ***variables***.

(*) ***Qualitative variables:*** The characteristic is ***categorical***. E.g., gender, ethnicity, treatment group vs. control group.

(*) ***Quantitative variables:*** The characteristic is ***numerical***. E.g., income level, age, blood pressure. Quantitative variables can be *discrete* or *continuous*.

- ***Discrete*** variables can take values that differ by fixed amounts, usually usually used to *count* things. E.g., number of children.
- ***Continuous*** variables can take values that differ by arbitrarily small amounts. E.g., height or temperature.

Example: 500 households are surveyed by a marketing research firm. The investigators collect data on: size of each household; monthly household income; occupation of head-of-household ; number of computers in house; type of internet connection.

(*) 500 observations, each producing data for five variables.

(*) Household size, monthly income and number of computers — these are quantitative variables.

- Income is a continuous variable.
- Household size and number of computers are discrete variables.

(*) Occupation of head of household and type of internet connection are qualitative variables.

Tables – categorical data

(*) Data, is often ***summarized*** in tables.

Example. The table below describes the results of the randomized, double-blind field test of the Salk polio vaccine.

Group	***size***	***infections/100,00***
Treatment	200,000	28
Control	200,000	71
No consent	350,000	46

(*) Observations: children.

(*) Variables: group to which child belongs (three categories) and infection status (two categories).

(*) The categorical data is ***summarized*** in two ways here: *number* of children in each group; *relative frequency* of infection in each group.

Distribution tables – quantitative data

(*) To summarize quantitative data in a table, the typical approach is to transform it into *categorical data*.

(*) The quantitative data is divided into ***class intervals***.

(*) The ***size***, or ***relative size*** of each class interval is recorded in the *distribution table*.

(*) Class intervals are also called ***bins***.

- The size of a bin is the ***number*** of data points that it contains.
- The relative size of a bin is the ***proportion*** of the data that it contains. Proportions are typically recorded as ***percentages***.

Summary: A distribution table treats numerical data as if it were categorical: the ranges of values we group together—the bins—are the categories.

Example: Family incomes for 50000 US families, from the *Current Population Survey* of 1973.

The table on the left is Table 1, in chapter 3 of FPP (p.35). The table on the right lists the same information, using class *sizes* instead of class *percentages*.

(I estimated the class sizes from the table in the book, using the fact that there were 50000 observations in the study.)

Income level	*Percent*
\$0 - \$1000	1
\$1000 - \$2000	2
\$2000 - \$3000	3
\$3000 - \$4000	4
\$4000 - \$5000	5
\$5000 - \$6000	5
\$6000 - \$7000	5
\$7000 - \$10000	15
\$10000 - \$15000	26
\$15000 - \$25000	26
\$25000 - \$50000	8
\$50000 and over	1
Total:	101%

Income level	*Number*
\$0 - \$1000	500
\$1000 - \$2000	1000
\$2000 - \$3000	1500
\$3000 - \$4000	2000
\$4000 - \$5000	2500
\$5000 - \$6000	2500
\$6000 - \$7000	2500
\$7000 - \$10000	7500
\$10000 - \$15000	13000
\$15000 - \$25000	13000
\$25000 - \$50000	4000
\$50000 and over	500

(*) ***The endpoint convention*** says which bin contains the data that lies on the border between two intervals.

(*) The endpoint convention for the preceding tables:

The *left-hand* endpoint of the class interval belongs to the class, but the right-hand endpoint belongs to the next one. E.g., a family earning exactly $5000 a year is included in the 6th class, not the 5th class.

Comment: A distribution table makes it much easier to read and understand large amounts of data. The price we pay is that there is a loss of information. When determining the class intervals for the table, you have to decide how much of the fine detail you are willing to lose.

Cross-tabulation

(*) In studies with more than one category, we can produce different distribution tables for different categories. The separate distribution tables can be combined into one table.

(*) The result of this process is called a ***cross-tabulation***, and it helps to ***control for*** (observe the effect of) confounding variables.

Example. Oral contraceptives and blood pressure. The following table summarizes the results of the study on the effects of oral contraceptives on the blood pressure of women who use them done by the Kaiser clinic in Walnut Creek, CA.

(*) Qualitative variable: ***user/nonuser***

(*) Quantitative variable: ***blood pressure***.

(*) Variable controlled for: ***age***.

Table 2. Systolic blood pressure by age and pill use, for women in the Contraceptive Drug Study, excluding those who were pregnant or taking hormonal medication other than the pill. Class intervals include the left endpoint, but not the right. – means negligible. Table entries are in percent; columns may not add to 100 due to rounding.

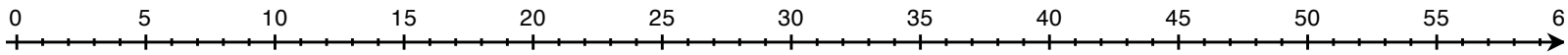
Blood pressure (millimeters)	*Age 17–24*		*Age 25–34*		*Age 35–44*		*Age 45–58*	
	Non-users	*Users*	*Non-users*	*Users*	*Non-users*	*Users*	*Non-users*	*Users*
	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
under 90	–	1	1	–	1	1	1	–
90–95	1	–	1	–	2	1	1	1
95–100	3	1	5	4	5	4	4	2
100–105	10	6	11	5	9	5	6	4
105–110	11	9	11	10	11	7	7	7
110–115	15	12	17	15	15	12	11	10
115–120	20	16	18	17	16	14	12	9
120–125	13	14	11	13	9	11	9	8
125–130	10	14	9	12	10	11	11	11
130–135	8	12	7	10	8	10	10	9
135–140	4	6	4	5	5	7	8	8
140–145	3	4	2	4	4	6	7	9
145–150	2	2	2	2	2	5	7	9
150–155	–	1	1	1	1	3	2	4
155–160	–	–	–	1	1	1	1	3
160 and over	–	–	–	–	1	2	2	5
Total percent	100	98	100	99	100	100	99	99
Total number	1,206	1,024	3,040	1,747	3,494	1,028	2,172	437

Histograms

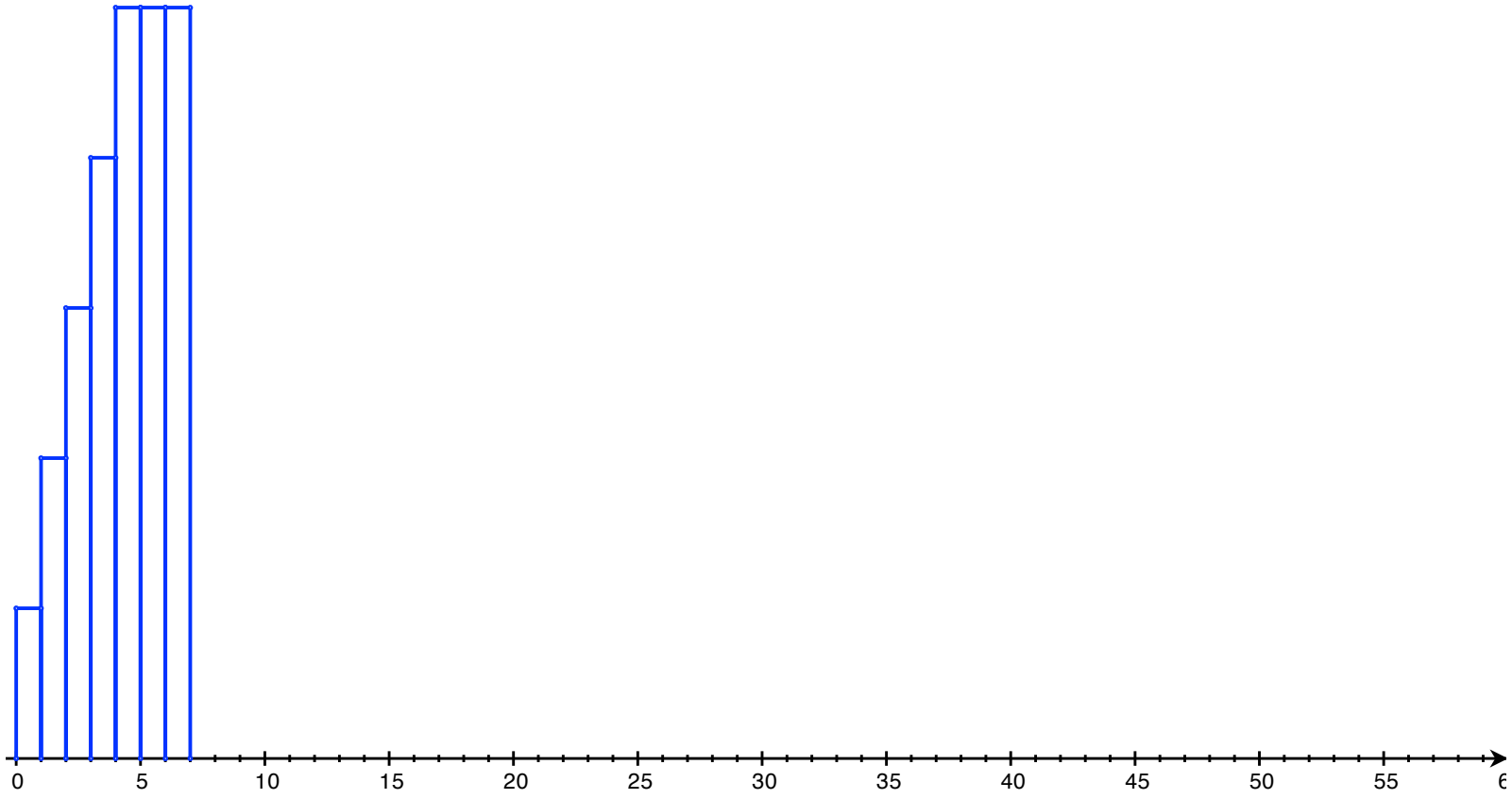
A ***histogram*** is a graphical representation of a distribution table, usually one that reports the relative size (percentages) of the class intervals.

- Histograms for data are usually drawn as bar-charts.
- The horizontal axis of the chart is divided into class intervals (bins).
- The ***area*** of the bar (rectangle) drawn above each interval represents the *relative size* of that class interval.

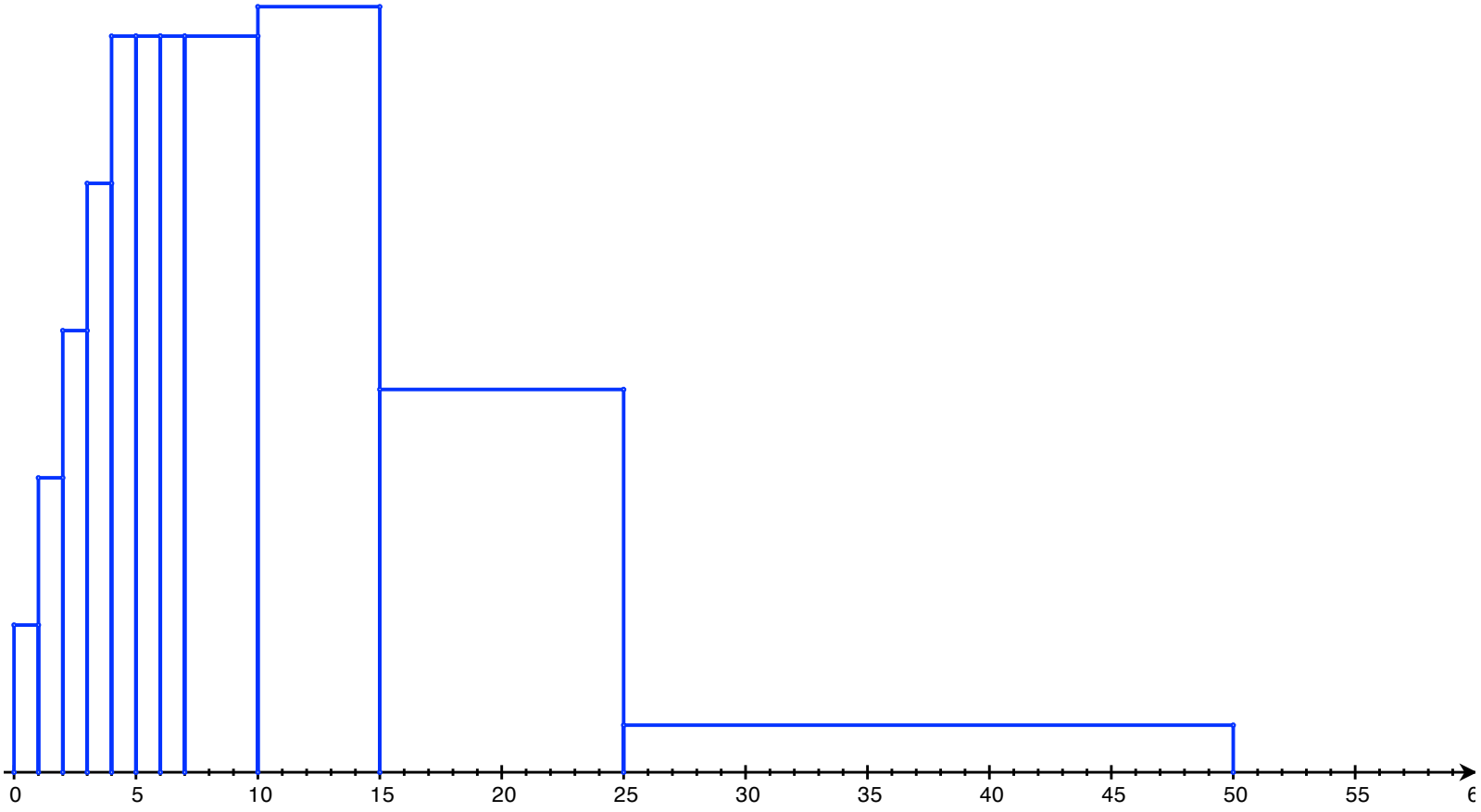
Example: Starting with the table of income distribution we saw earlier, we first draw the horizontal axis...



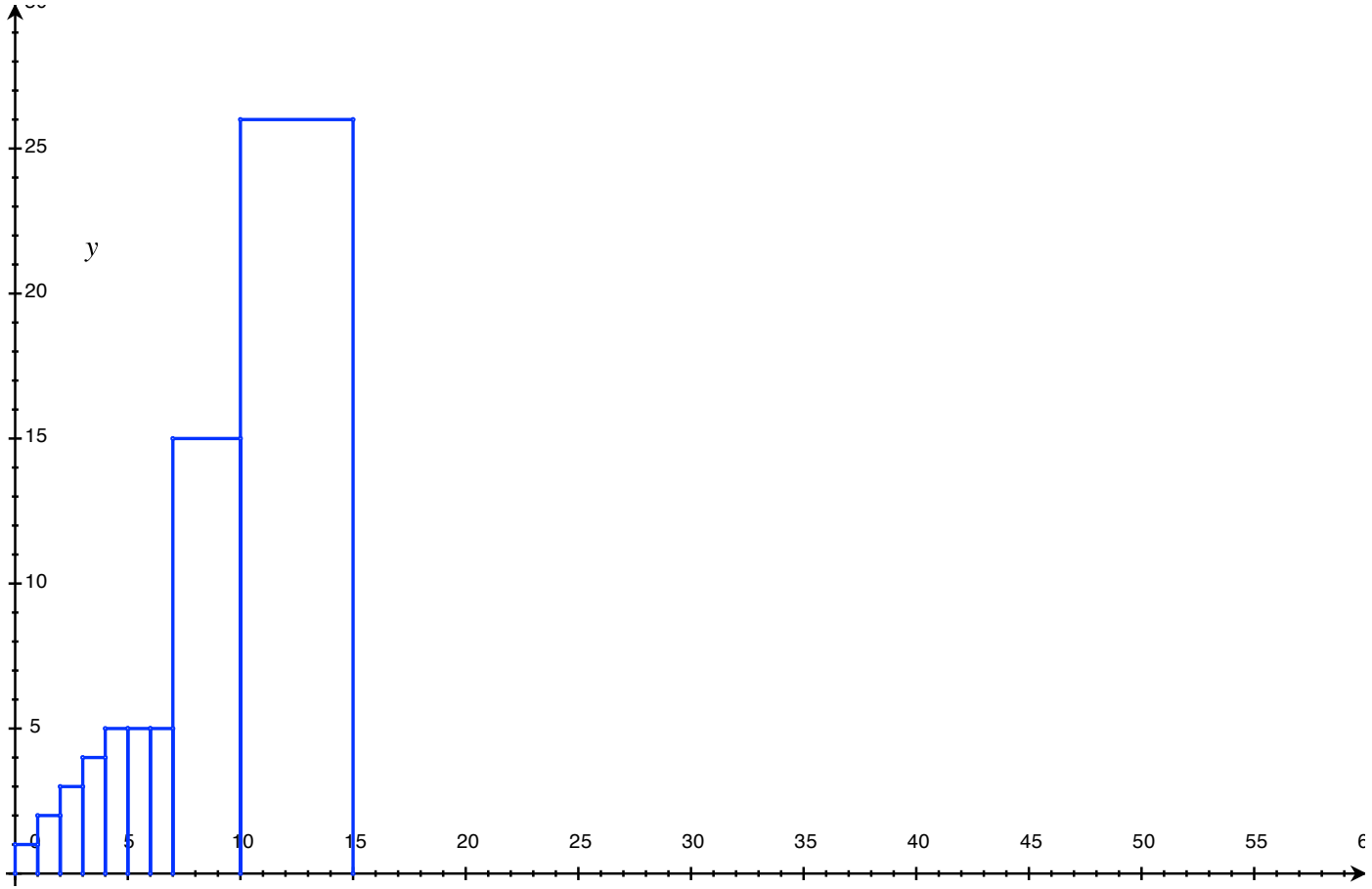
... Then we draw rectangles over each class interval whose areas equal the percentages of the families in those intervals...



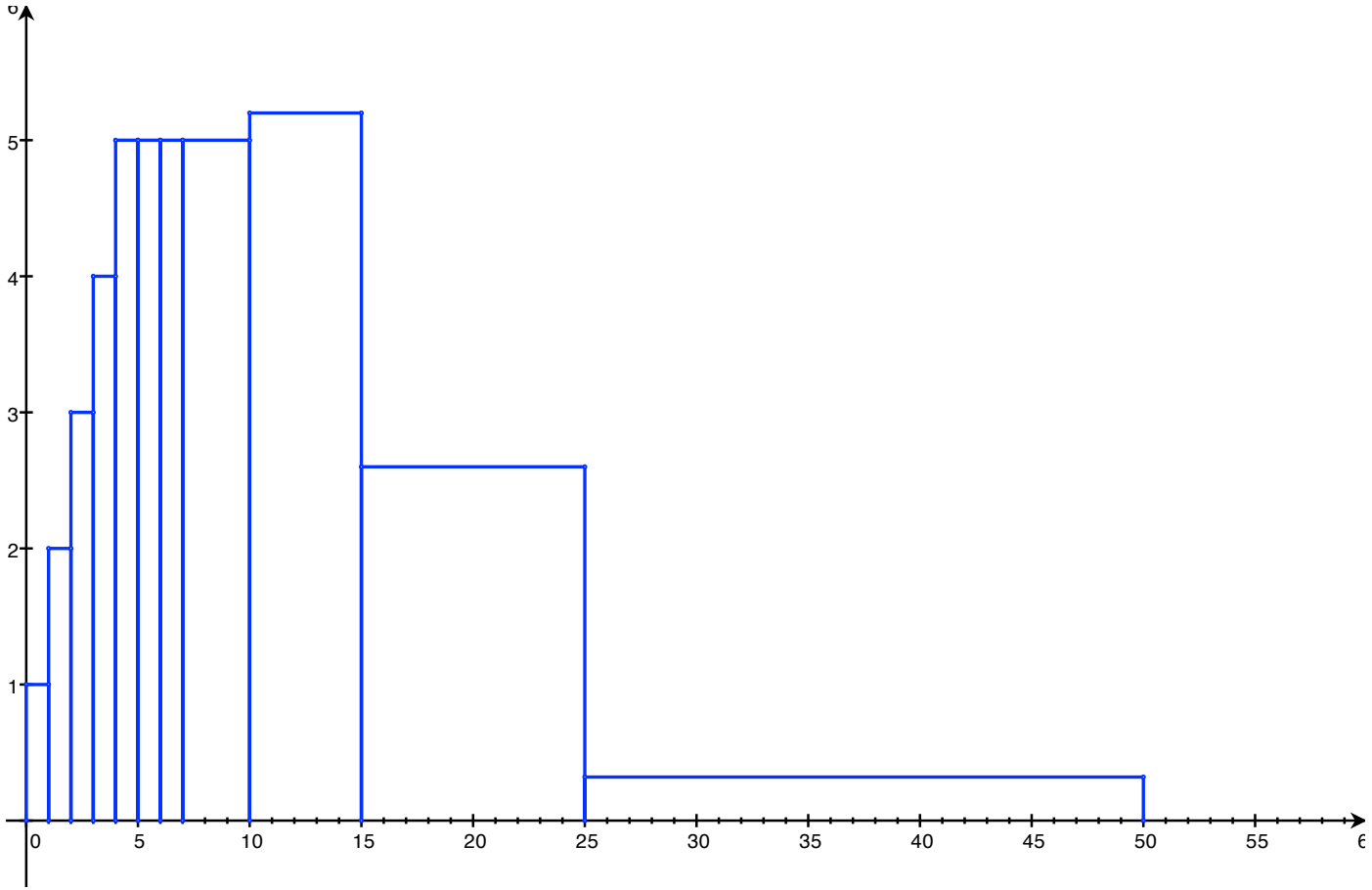
... If we do this correctly, the end result looks like this:



Remember: it is the area of the rectangle that should equal the percentage, **NOT** the height of the rectangle... I.e., you don't want your histogram to look like this:



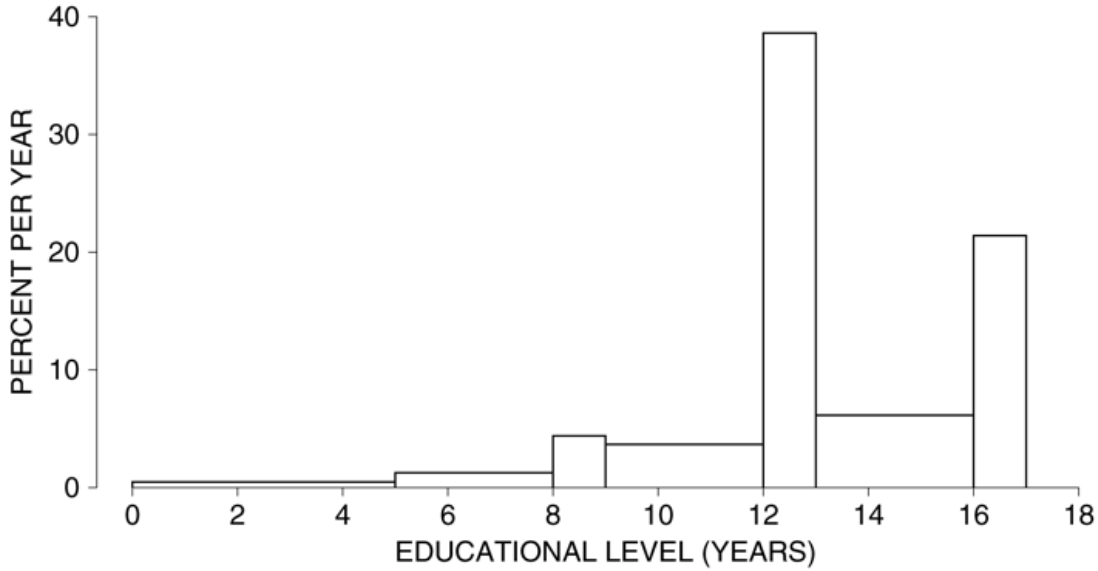
The vertical scale of a histogram is called the ***density scale***. In the case of income distribution, it is measured in units of *percent per \$1000*:



To 'read' a histogram, you need to remember where it came from, namely from a distribution table. You also need to know the 'endpoint convention'.

Example: The histogram below gives the distribution of persons age 25 and over in the U.S. in 1991 by education level.

Figure 5. Distribution of persons age 25 and over in the U.S. in 1991 by educational level.



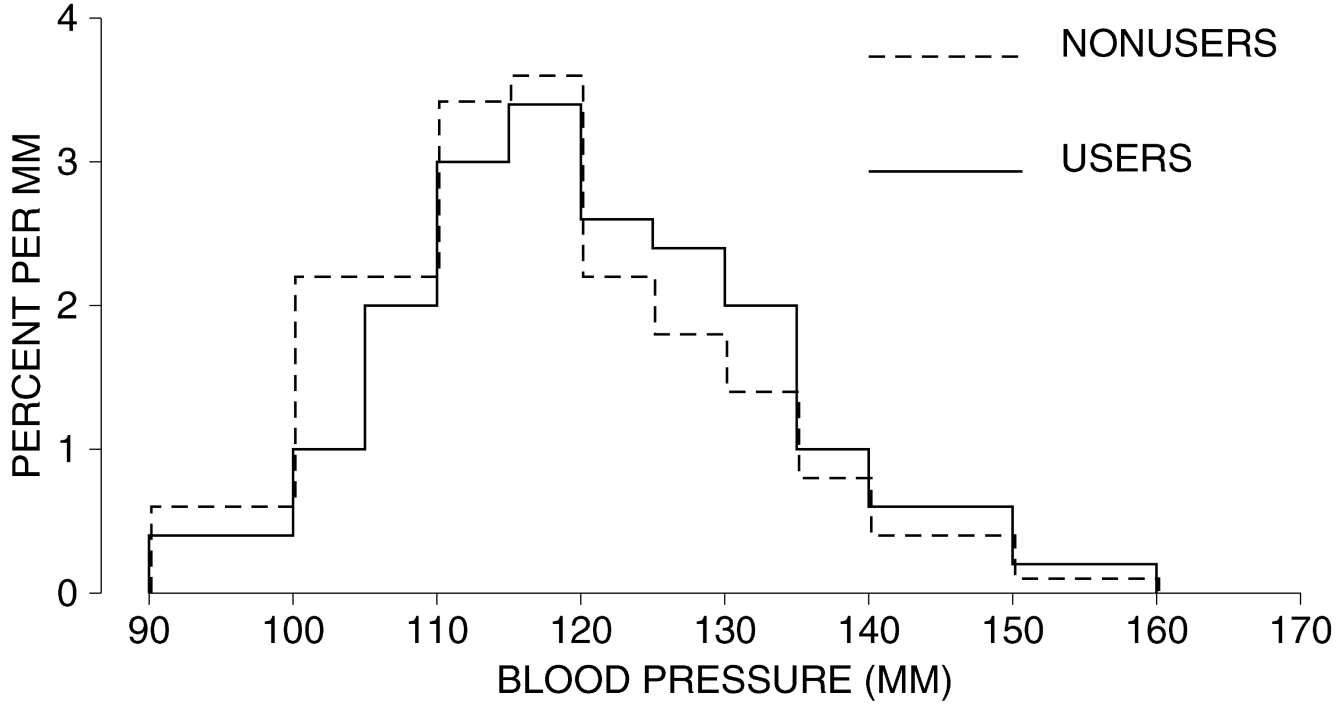
Source: *Statistical Abstract*, 1992, Table 220.

The endpoint convention in this case is that the right endpoint is not included. E.g, the block that starts at 12 and ends at 13 includes everyone who finished 12 years of school but did not finish 13.

- The percentage of persons 25 and older with fewer than 9 complete years of education is equal to the sum of areas of the first 3 blocks — about 9%.
- The percentage of people who finished high school is the sum of the areas of the last three blocks — about 78%.
- What percentage of this population attended college, but did not complete a degree?
- What percentage of this population completed between 8 and 10 years of schooling?

Comparing two histograms.

Figure 7. The effect of the pill. The top panel shows histograms for the systolic blood pressures of the 1,747 users and the 3,040 non-users age 25–34 in the Contraceptive Drug Study. The bottom panel shows the histogram for the non-users shifted to the right by 5 mm.



Nonusers histogram shifted 5mm to the right.