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 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. Continuous variables come in two flavors. ***Interval*** variables — the difference between two values has a reasonable interpretation, but the ratio does not. E.g., temperature. ***Ratio*** variables — the ratio between two values has a reasonable interpretation. E.g., height.

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

(*) Categorical data can be ***summarized*** in tables by recording the ***frequency*** or ***relative frequency*** of the data in each category.

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).

Distribution tables – quantitative data

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

(*) The ***range*** of the observed values is divided into ***class intervals***, also called ***bins***. The bins play the role of the categories.

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

- The frequency of a bin is the ***number*** of the observations that fall into that bin.
- The relative frequency of a bin is the ***proportion*** of the observations that fall into that bin. Proportions are often recorded as ***percentages***.

Comment: Data can be divided into class intervals in different ways. How this is done can affect the way that the data is perceived.

Example: US household incomes from 2015 — frequency distribution.

Income level	*Frequency*
$0 - $14,999	14,595,004
$15,000 - $24,999	13,210,995
$25,000 - $34,999	12,581,900
$35,000 - $49,999	15,979,013
$50,000 - $74,999	21,011,773
$75,000 - $99,999	15,224,099
$100,000 - $149,999	17,740,479
$150,000 - $199,999	7,800,778
$200,000 and over	7,674,959

Example: US household incomes from 2015 — *relative* frequency distribution.

Income level	*Relative frequency*
$0 - $14,999	11.6%
$15,000 - $24,999	10.5%
$25,000 - $34,999	10%
$35,000 - $49,999	12.7%
$50,000 - $74,999	16.7%
$75,000 - $99,999	12.1%
$100,000 - $149,999	14.1%
$150,000 - $199,999	6.2%
$200,000 and over	6.1%

Comment: A distribution table makes it much easier to absorb large amounts of data. The price we pay is the loss of information inherent in *summarizing*.

When determining the class intervals for the table, you have to decide how much of the fine detail you are willing to lose.

And what message you are trying to convey.

Income level	*Relative frequency*
\$0 - \$24,999	22.1%
\$25,000 - \$49,999	22.7%
\$50,000 - \$99,999	28.8%
\$100,000 and over	26.4%

Cross-tabulation

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

(*) 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 (for quantitative data).

- Histograms *for data* are usually drawn as bar-charts.
- The horizontal axis of the chart is divided into class intervals.
- The scale on the vertical axis of the chart is typically one of following three:

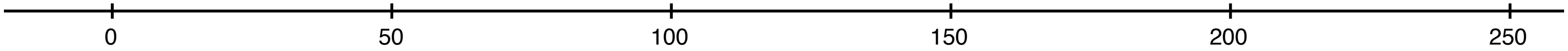
(*) The *frequency* – the number of all observations in a given bin.

(*) The *relative frequency* – the percentage of all observations in a given bin.

(*) The *density* – the relative frequency of the bin divided by its width.

- If one uses the ***density scale*** (as we will be doing in this class), then the ***area*** of the region drawn above a class interval represents the *relative frequency* of (pecentage of data in) that class interval.

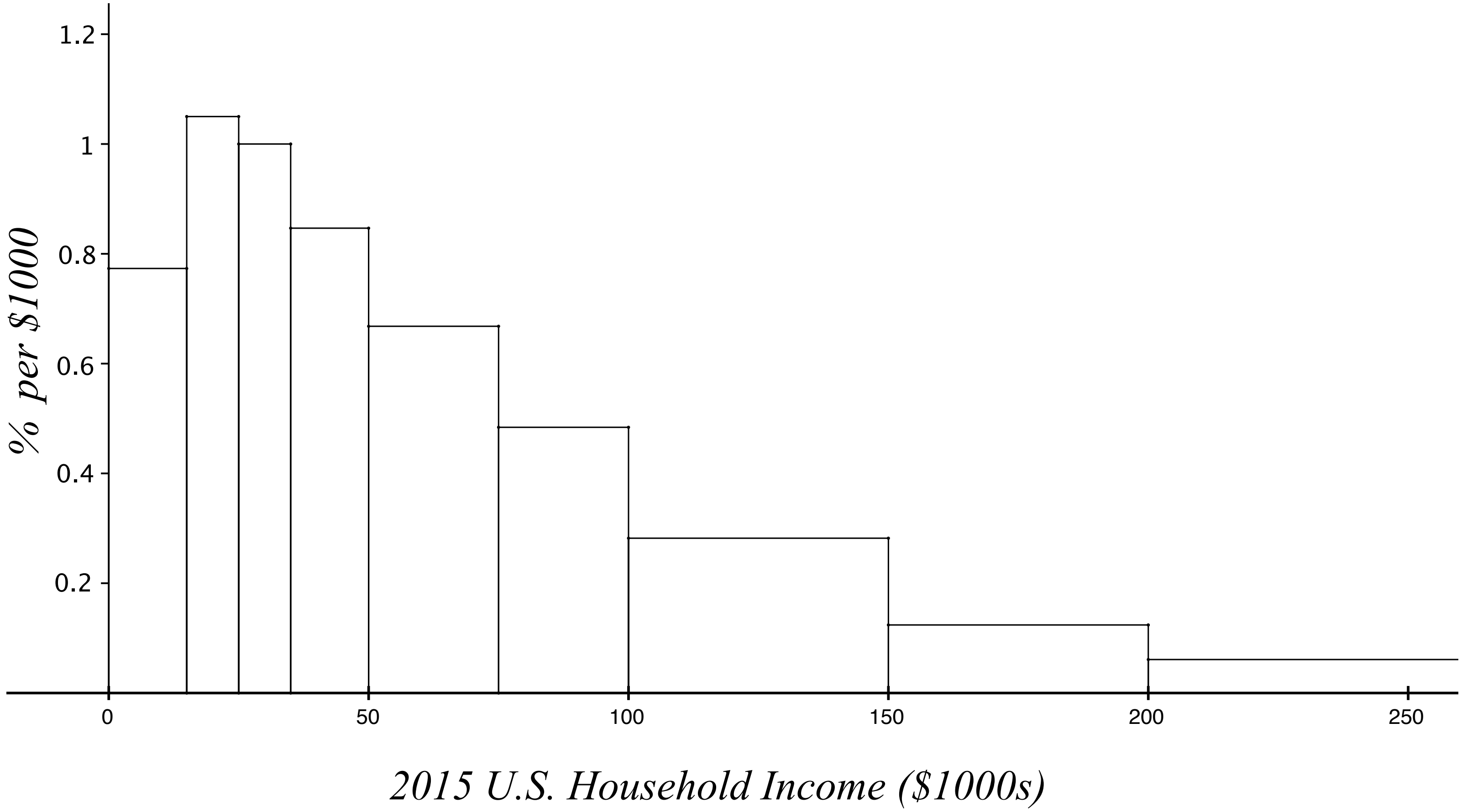
Example: Starting with the table of income distribution we saw earlier, we first draw the horizontal axis...



... Using a density scale, we draw rectangles over each class interval whose areas equal the percentages of the families in those intervals.

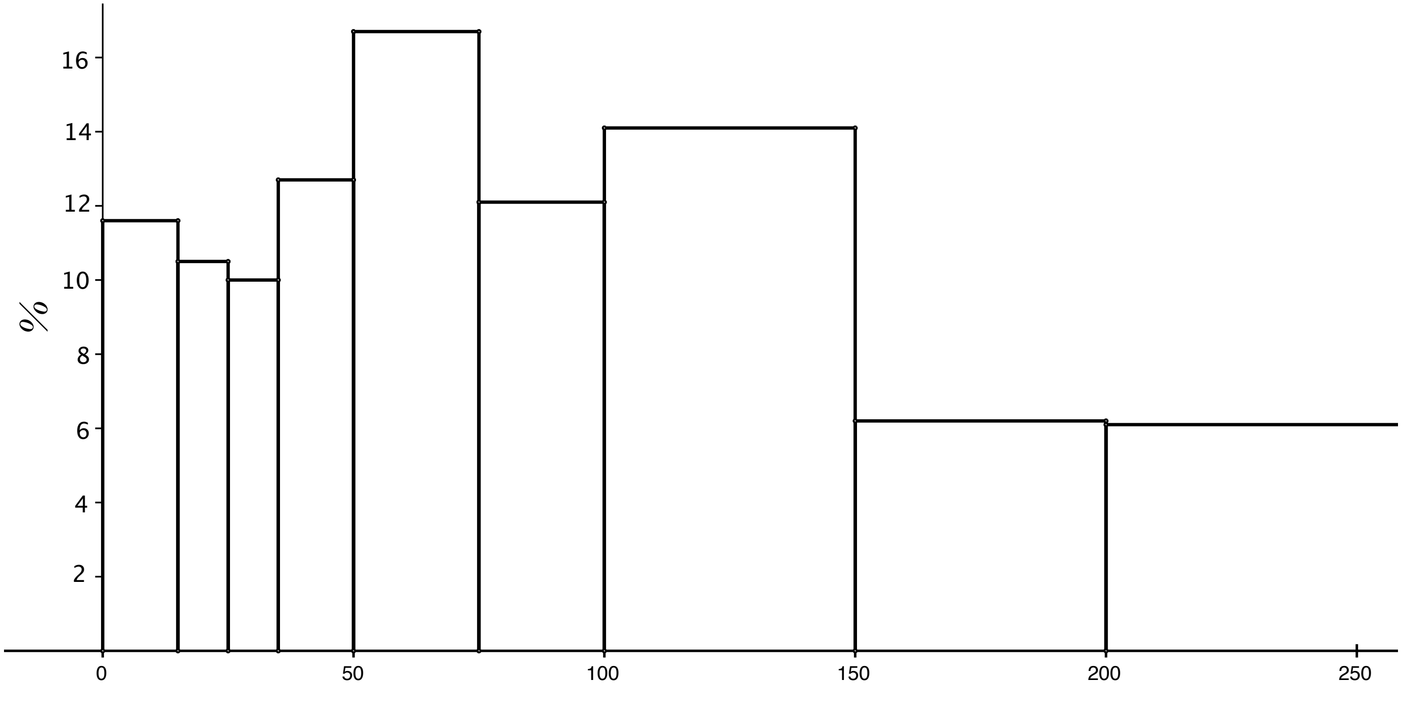
The height of each rectangle is equal to the percentage of the observations in the corresponding class interval divided by the length of the class interval (the width of the rectangle).

The end-result should look like this



The vertical scale here is *percent per $1000* – i.e., it is the relative frequency (percentage) divided by the width of the intervals (which in this case are measured in $1000s). It's always a good idea to label the axes.

If, for example, we use the *relative frequency* scale instead of the *density* scale, the histogram looks like this:



2015 U.S. Household Income ($1000s)

This histogram reports the information accurately, but it is misleading. The bins for the higher incomes seem to be much bigger than the bins for the lower incomes *because they are wider*.

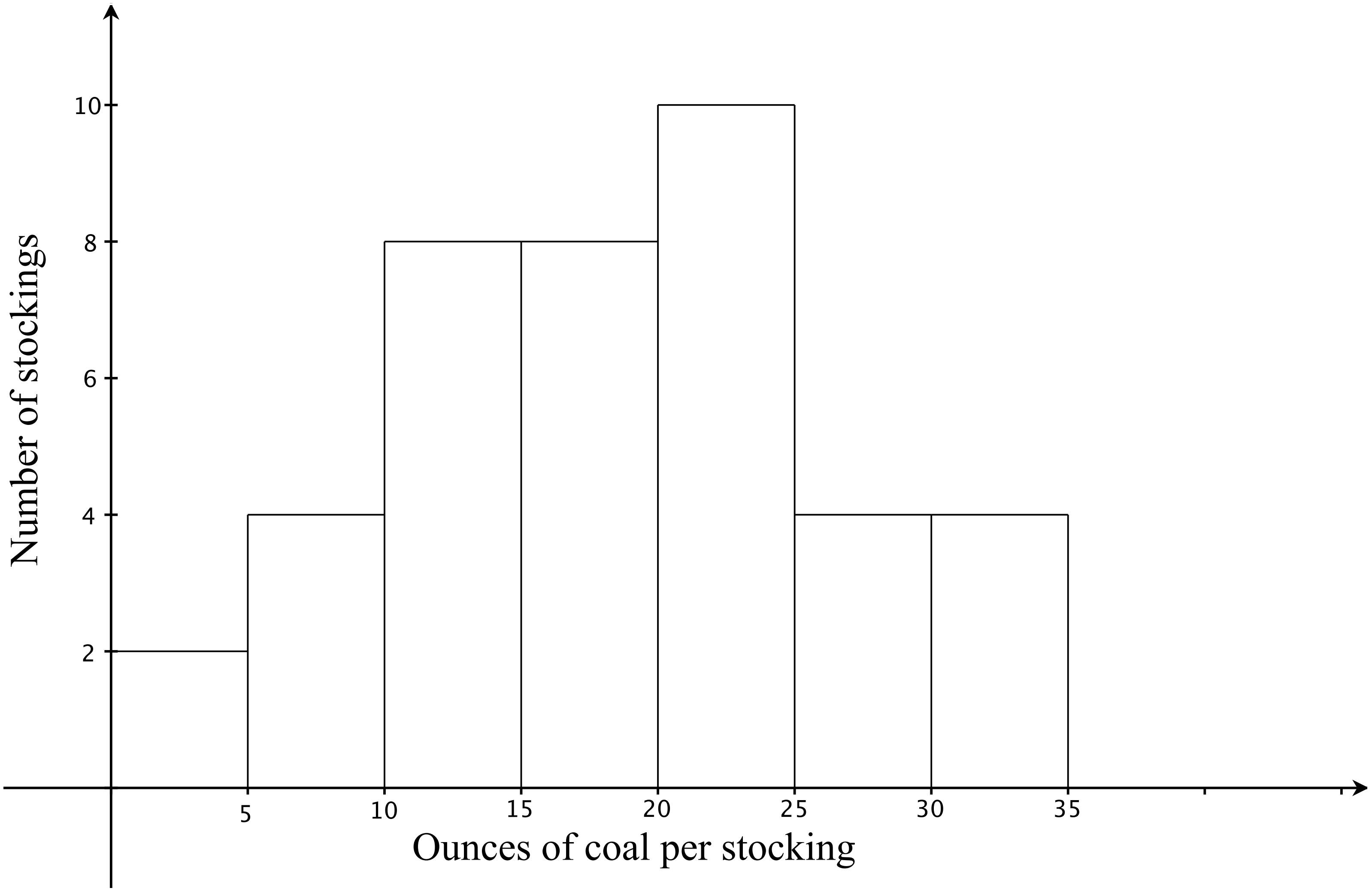
(*) ***If bins have different widths — use the density scale.***

Comment: If all the bins in the distribution table have the same width, then the appearance of the histogram will be the same for all three scales. Only the units (and numbers) on the vertical scale will change.

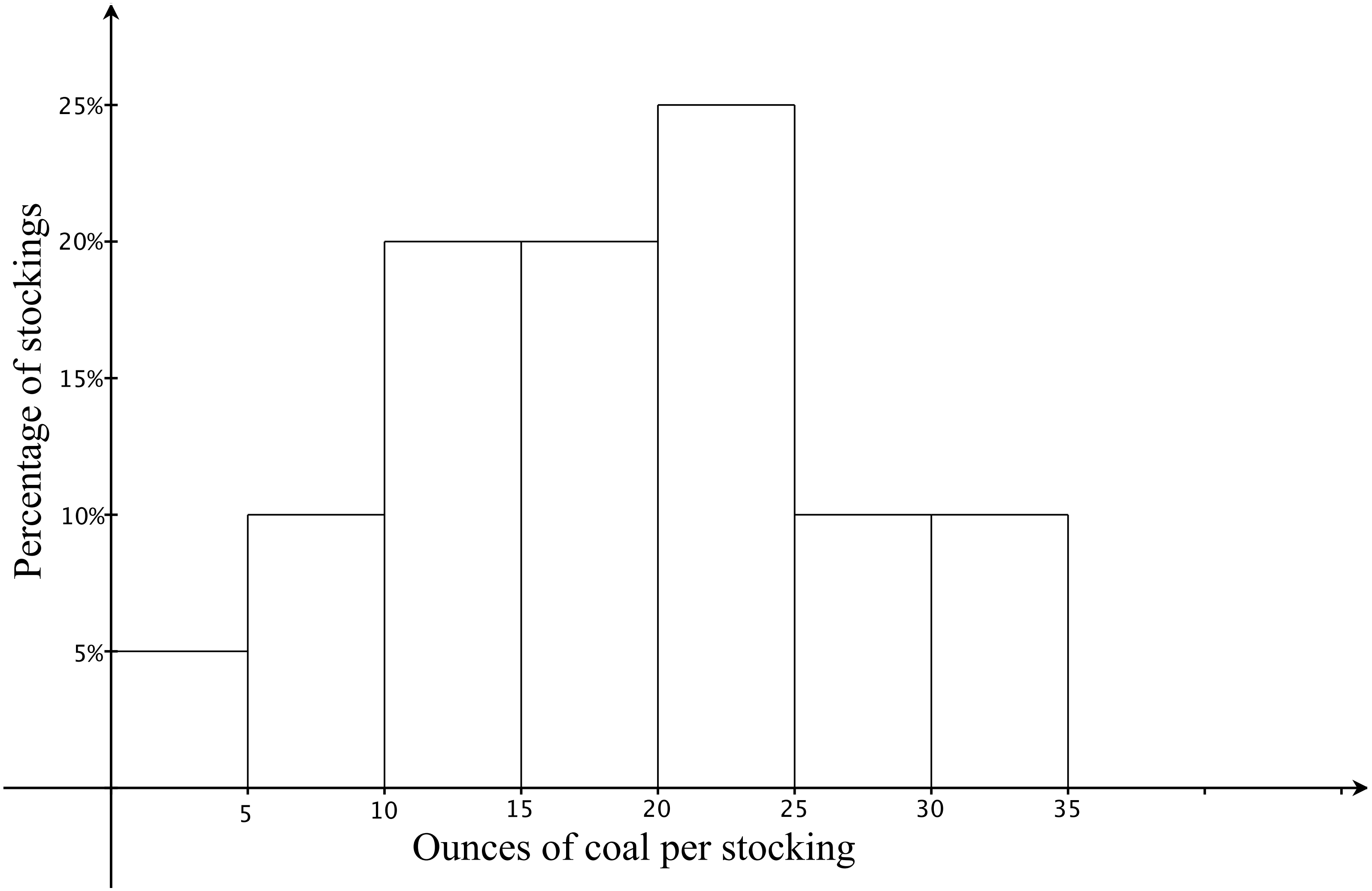
Example: Distribution of coal (by weight) in Christmas stockings of 40 children at Wool's orphanage.

ounces of coal	number of stockings
0 – 5	2
5 – 10	4
10 – 15	8
15 – 20	8
20 – 25	10
25 – 30	4
30 – 35	4

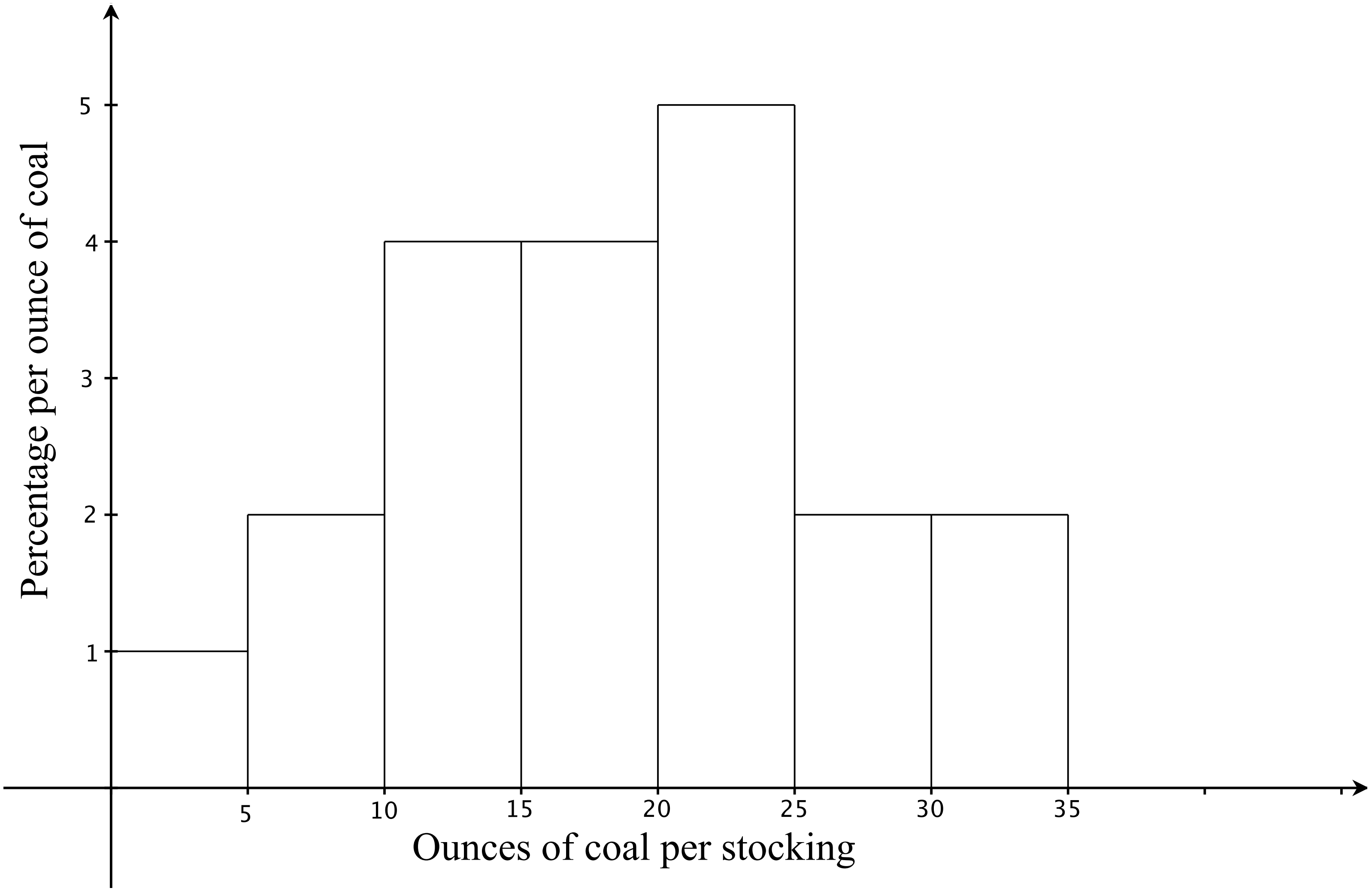
Histogram with frequency scale:



Histogram with relative frequency scale:



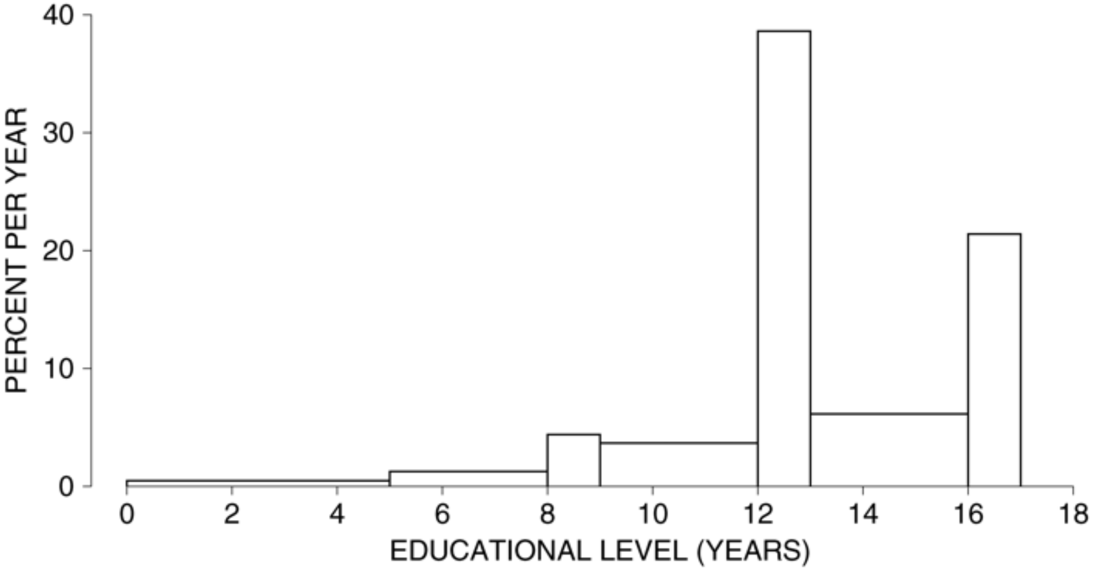
Histogram with density scale:



To ‘read’ a histogram, you need to remember that it comes from a distribution table. You also need to know the ‘endpoint convention’ and what scale is being used on the vertical axis.

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 10%.
- 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?