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.

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*.

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

Income level	*Rel. frequency*
\$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%

Income level	*Frequency*
\$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.

And what message you are trying to convey.

Income level	*%*
\$0 - \$10000	40%
\$10000 - \$15000	26%
\$15000 - \$25000	26%
\$25000 - \$50000	8%
\$50000 and over	1%

Income level	*%*
\$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 and over	35%

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:

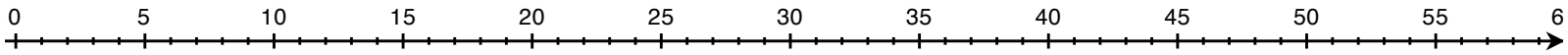
(*) The *frequency* – the number of all observations in a given bin.

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

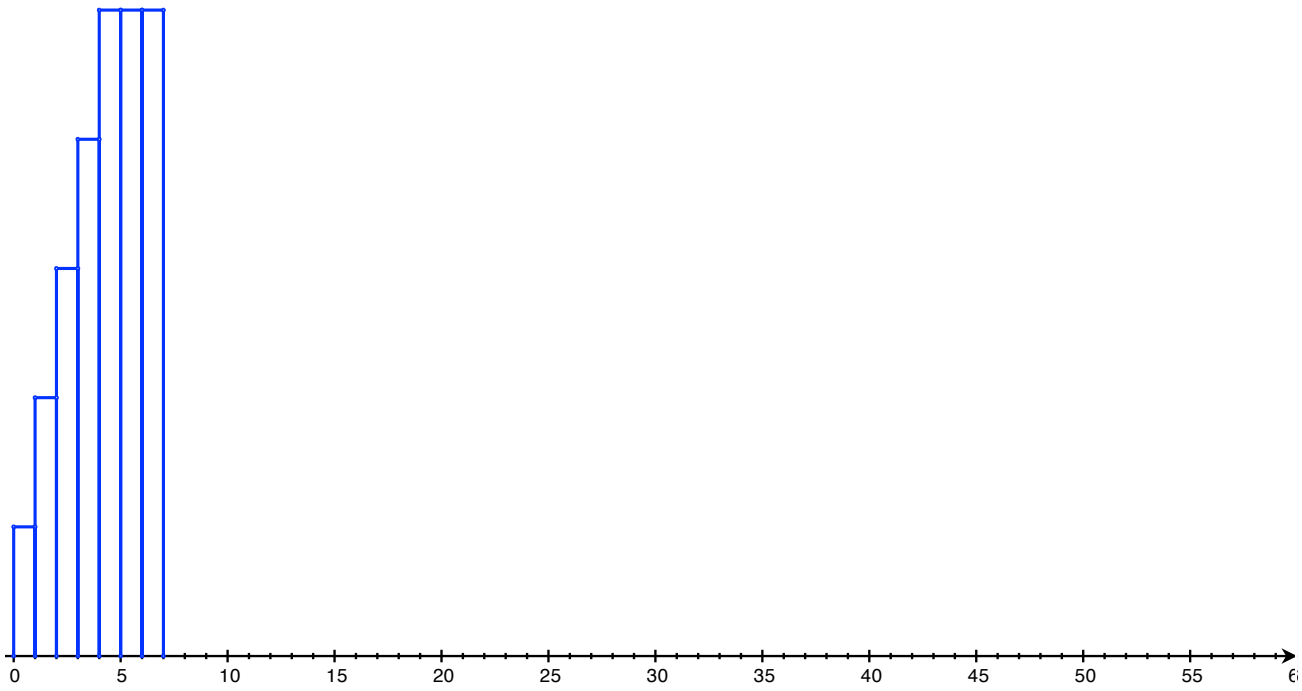
(*) The *density* – the relative frequency of the bin divided by its width.

- If a histogram uses the ***density scale*** (the only one we use in this class), then the ***area*** of the region drawn above a class interval represents the *relative frequency* of that class interval.

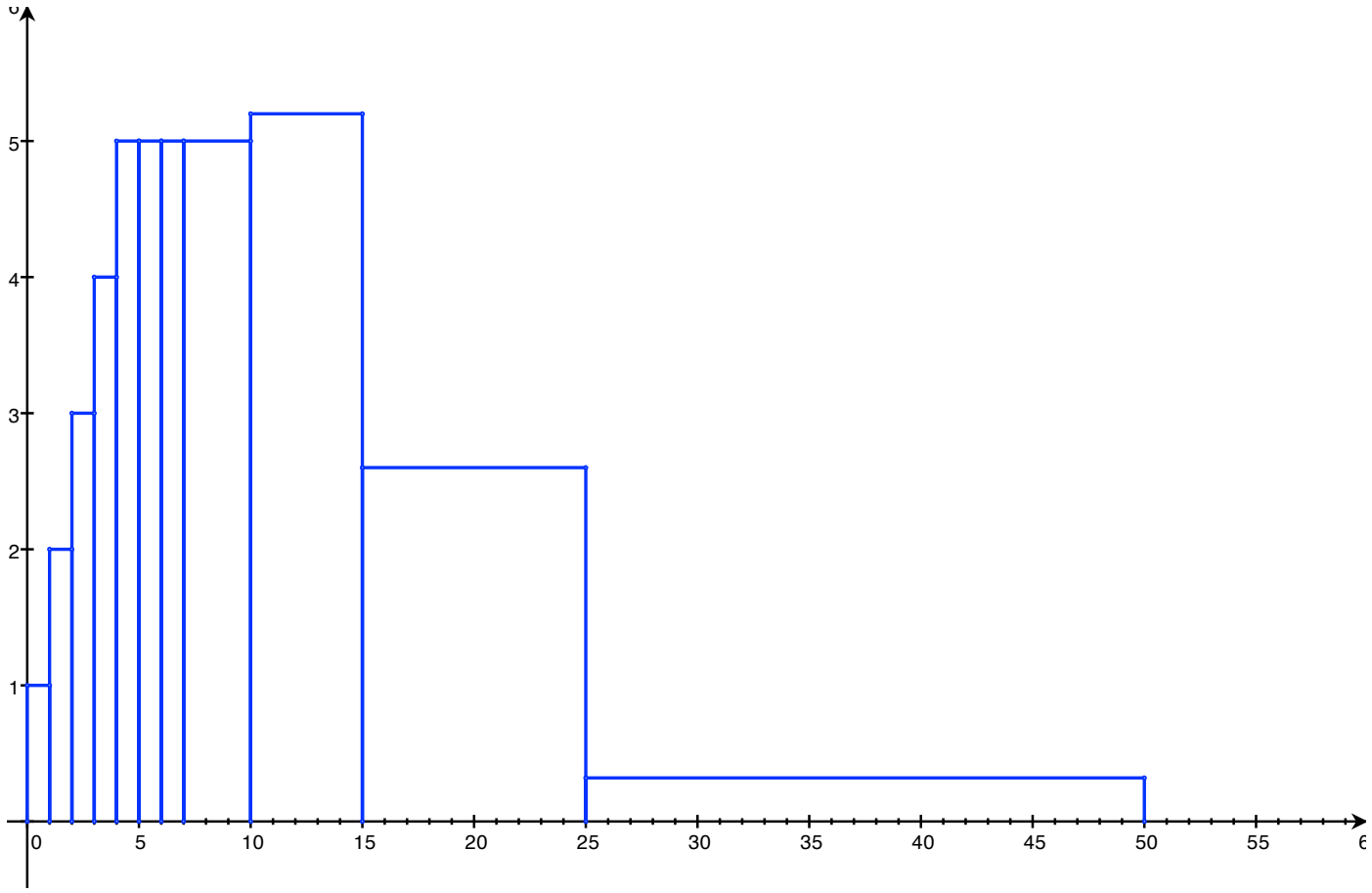
Example: Starting with the table of income distribution we saw earlier, we first draw the horizontal axis...



... 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 looks 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).

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

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.

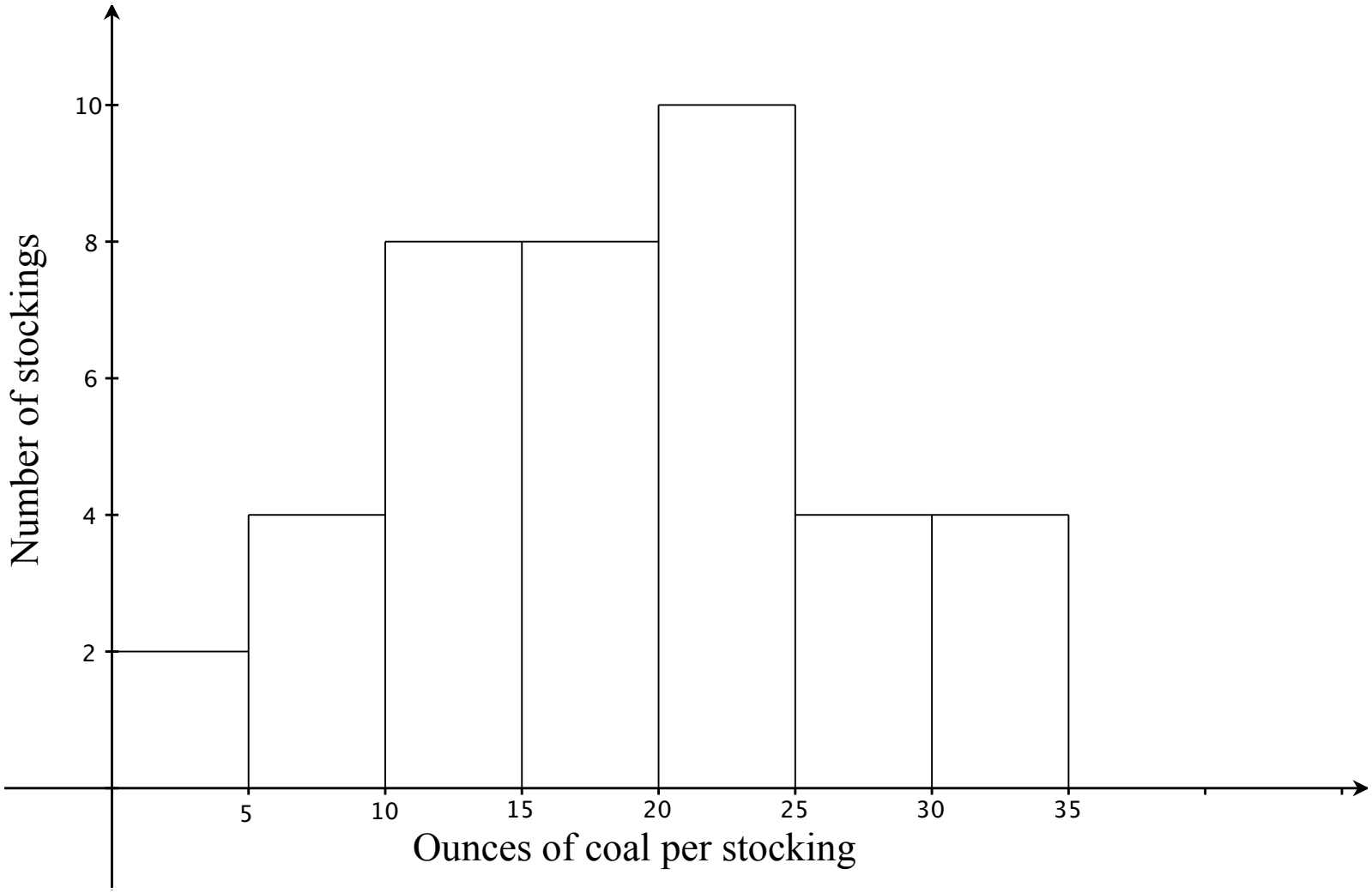
(*) ***If bins have different widths — use the density scale.***

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.

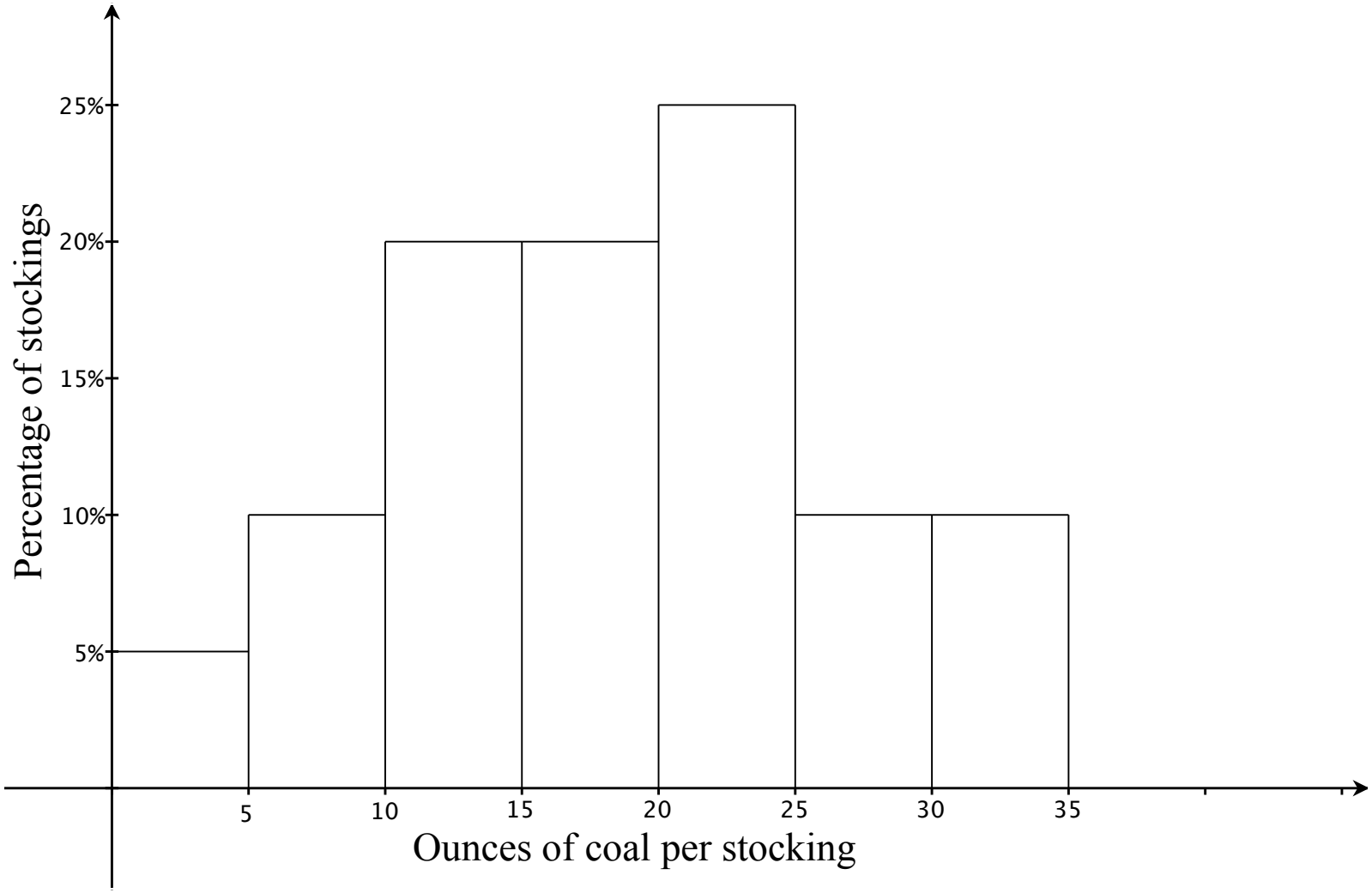
Example: Distribution of coal (by weight) in Christmas stockings of 40 children at Wool's orphanage.

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: