DSA 8010 - Exploratory Data Analysis, Part 1

Summarizing and visualizing one variable

Observations and variables

Rectangular data:

	A	B	C
1	Department	log(property)	log(violent)
2	Lower Salford Twp Po	6.126432408	3.277145
3	Village Of Port Washin	6.406879986	2.351375
4	Duxbury Police Dept	6.432779308	2.97553
5	Wyckoff Police Dept	6.683986532	3.569533

- The rows contain *observations* measured for each of the n *observational units*.
- The columns contain *variables*, or characteristics that are measured on each observational unit.

Notation

y_i denotes the ith observation of the variable y. A sample of n observations will be denoted by $y_1, \ldots, y_n$.

The subscript i is called an index.

Types of variables

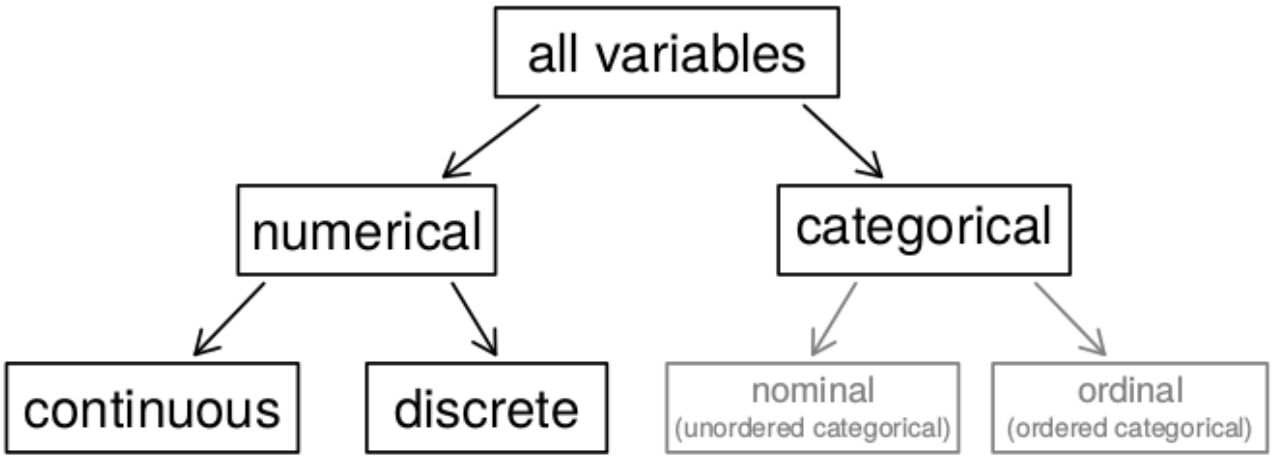


Figure 1.7: Breakdown of variables into their respective types.

Image from *Open Intro Statistics, 4th edition, Diez et al*

Categorical variables

Categorical variables are divided into *nominal* and *ordinal* variables.

Examples of ordinal variables: satisfaction level, education level.

Examples of nominal variables: race, major. (Binary variables take only two possible values, such as responses to yes/no questions.)

Watch out for categorical variables that are coded with numbers!

Categorical variables

Categorical variables are divided into *nominal* and *ordinal* variables.

Examples of ordinal variables: satisfaction level, education level.

Examples of nominal variables: race, major. (Binary variables take only two possible values, such as responses to yes/no questions.)

- Frequency tables and proportions are the most common ways to summarize categorical variables.
- Bar charts and pie charts are the most common ways to visualize categorical variables.

Summarizing one categorical variable

A survey was administered to a random sample of 750 student members of a statistical association. Each student reported their undergraduate major.

Summarizing one categorical variable

A survey was administered to a random sample of 750 student members of a statistical association. Each student reported their undergraduate major.

Frequency table:

mathematics/statistics	business	economics	other	total
325	75	250	100	750

Summarizing one categorical variable

A survey was administered to a random sample of 750 student members of a statistical association. Each student reported their undergraduate major.

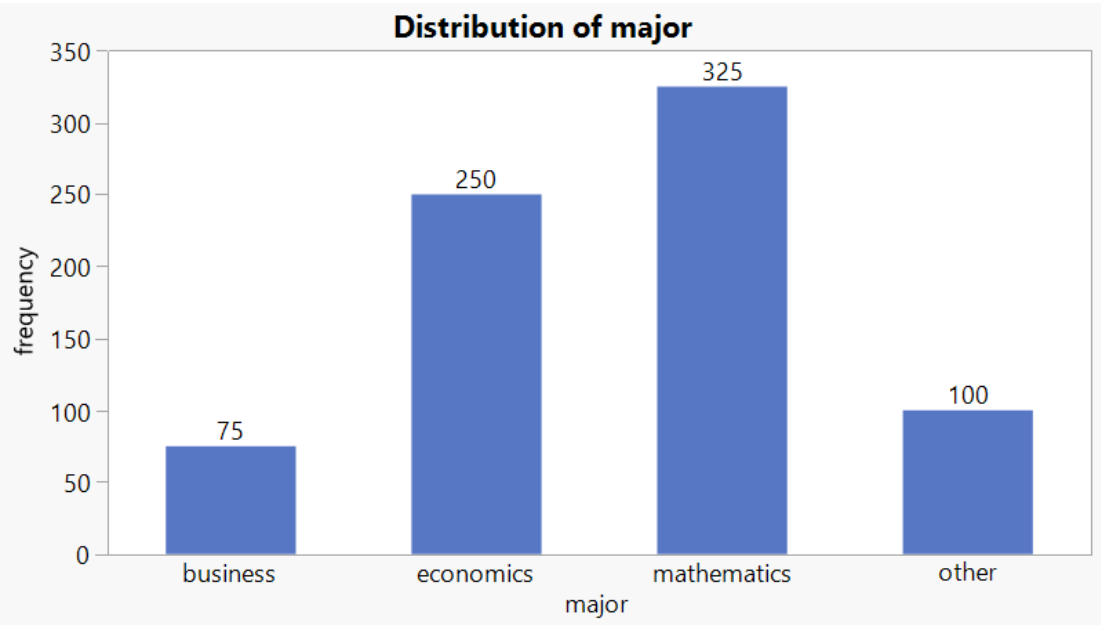
Relative frequency table:

mathematics/statistics	business	economics	other
0.433	0.100	0.333	0.133

Summarizing one categorical variable

Frequency table:

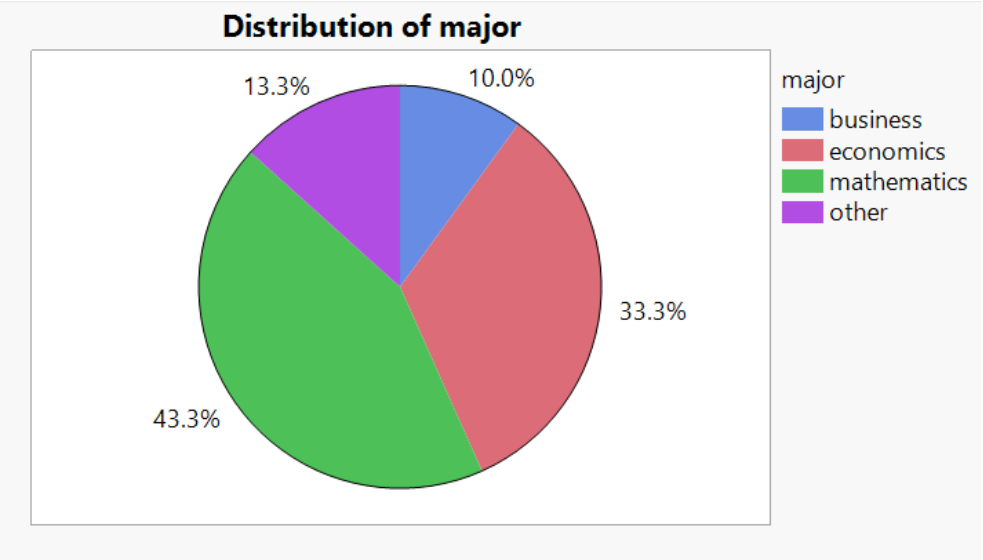
mathematics/statistics	business	economics	other
325	75	250	100



Summarizing one categorical variable

Frequency table:

mathematics/statistics	business	economics	other
325	75	250	100



Quantitative/numeric variables

Types of numeric variables:

discrete. countable (often integers).
Examples. number of children in a household, SAT scores.

continuous. can take any real number, possibly within some interval.
Example. temperature, time, weight.

Note: sometimes for mathematical convenience, we treat variables that are technically discrete as continuous.

Summarizing quantitative variables: measuring center

Three popular measures of center:

mean. average of the values.

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

median. The midpoint of the ordered values.

mode. The value that appears most frequently.

Summarizing quantitative variables: measuring center

Example: find the mean, median, and mode of the following data set ($n = 12$).

14 15 15 15 16 16 17 19 20 25 27 29

Summarizing quantitative variables: measuring center

Example: find the mean, median, and mode of the following data set ($n = 12$).

14 15 15 15 16 16 17 19 20 25 27 29

Summarizing quantitative variables: measuring variability

Four popular measures of variability:

sample variance. $s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$

sample standard deviation. $s = \sqrt{s^2}$

range. maximum value - minimum value.

IQR. Q3 - Q1.

Measures of variability are always positive, with higher values indicating higher variability.

Summarizing quantitative variables: measuring variability

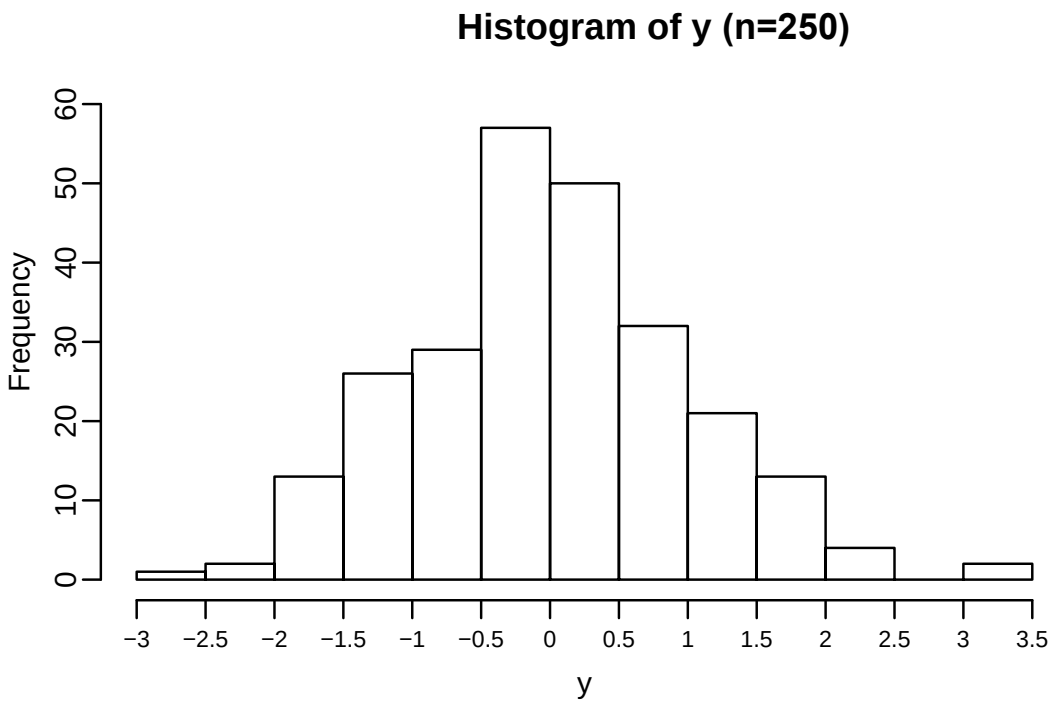
Find the standard deviation of the following data set.

-1 -1 2 4

Summarizing quantitative variables: histograms

A **histogram** displays one quantitative variable.

- The x-axis shows values of the variable divided into bins.
- The y-axis shows the frequency (or relative frequency) of observations in each bin.



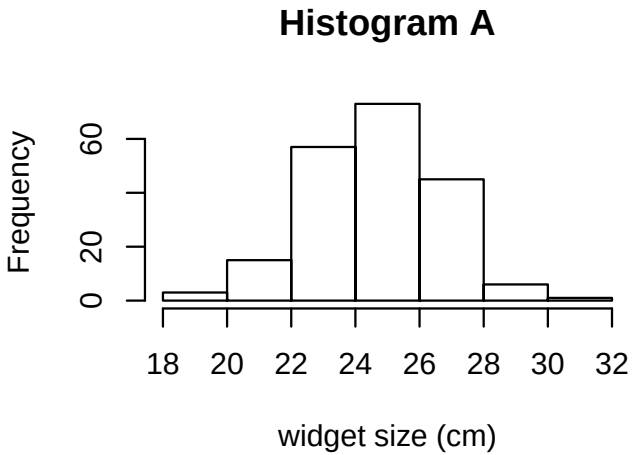
Summarizing quantitative variables: histograms

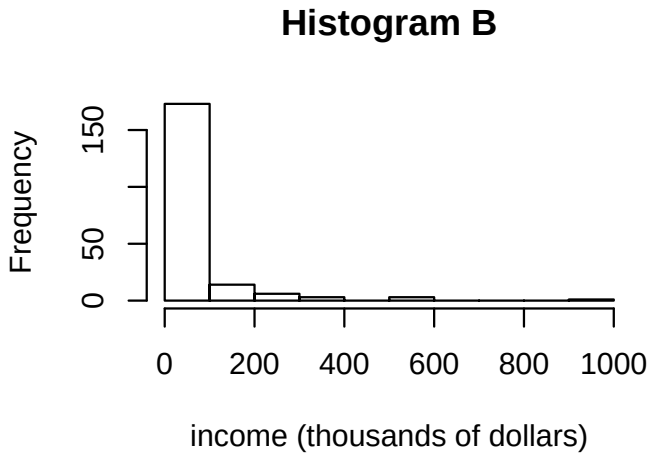
The histogram's **shape** indicates characteristics of the variable it displays.

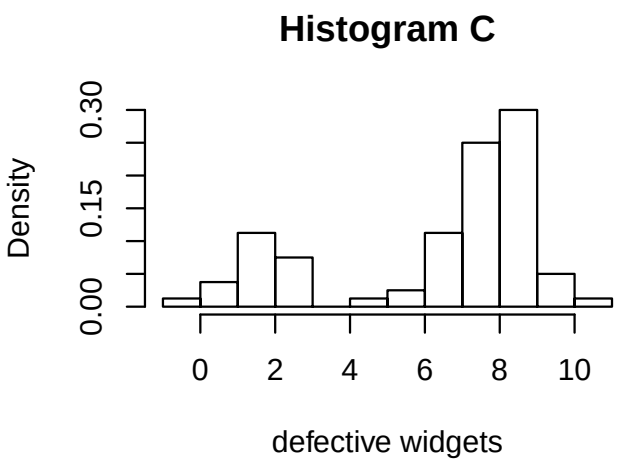
- symmetric - observations are equally likely to be above or below the center
- bell - symmetric with a peak in the center.
- uniform - rectangular; all possible values are equally likely.
- right-skewed - long tail to the right. Most values are low; a few are high. Extremely high values occur more often than extremely low values (e.g. incomes).
- left-skewed - long tail to the left. Most values are high; a few are low. Extremely low values occur more often than extremely high values.
- bimodal (multimodal) - two (or more) separate peaks. There are two (or more) intervals occur with high frequency

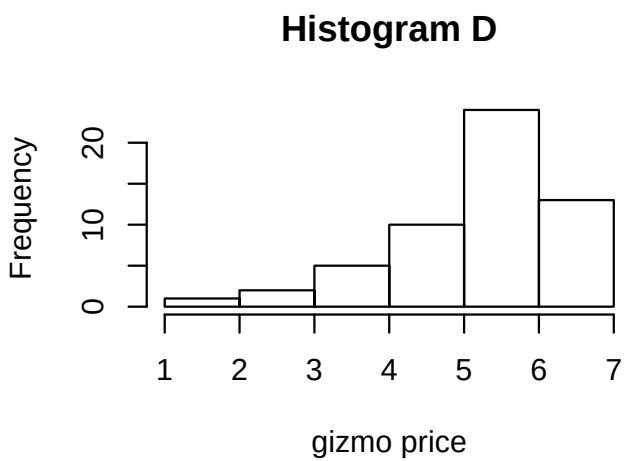
Also look for extreme values and unusual features.

Summarizing quantitative variables: histograms





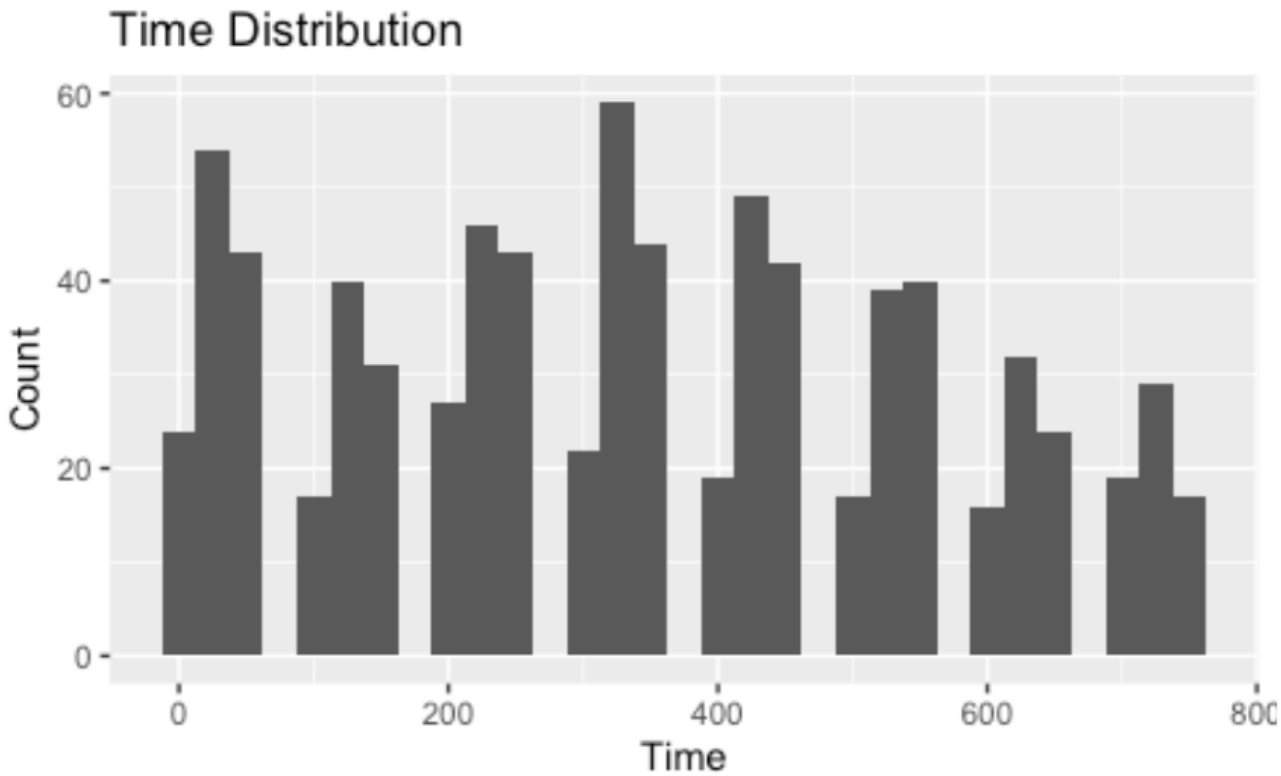




Summarizing quantitative variables: histograms

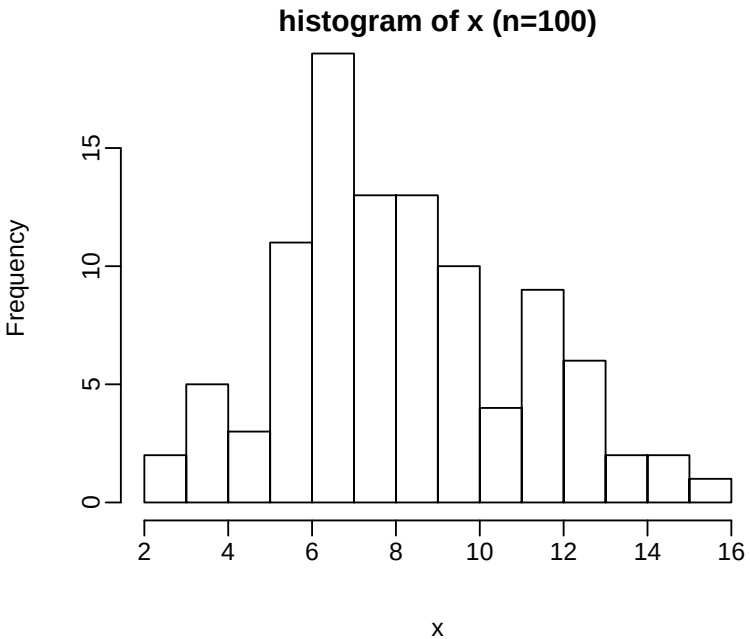
Always plot your data. Sometimes histograms can reveal unusual features that summary statistics miss!

Histogram of UTC time of amateur radio spots:



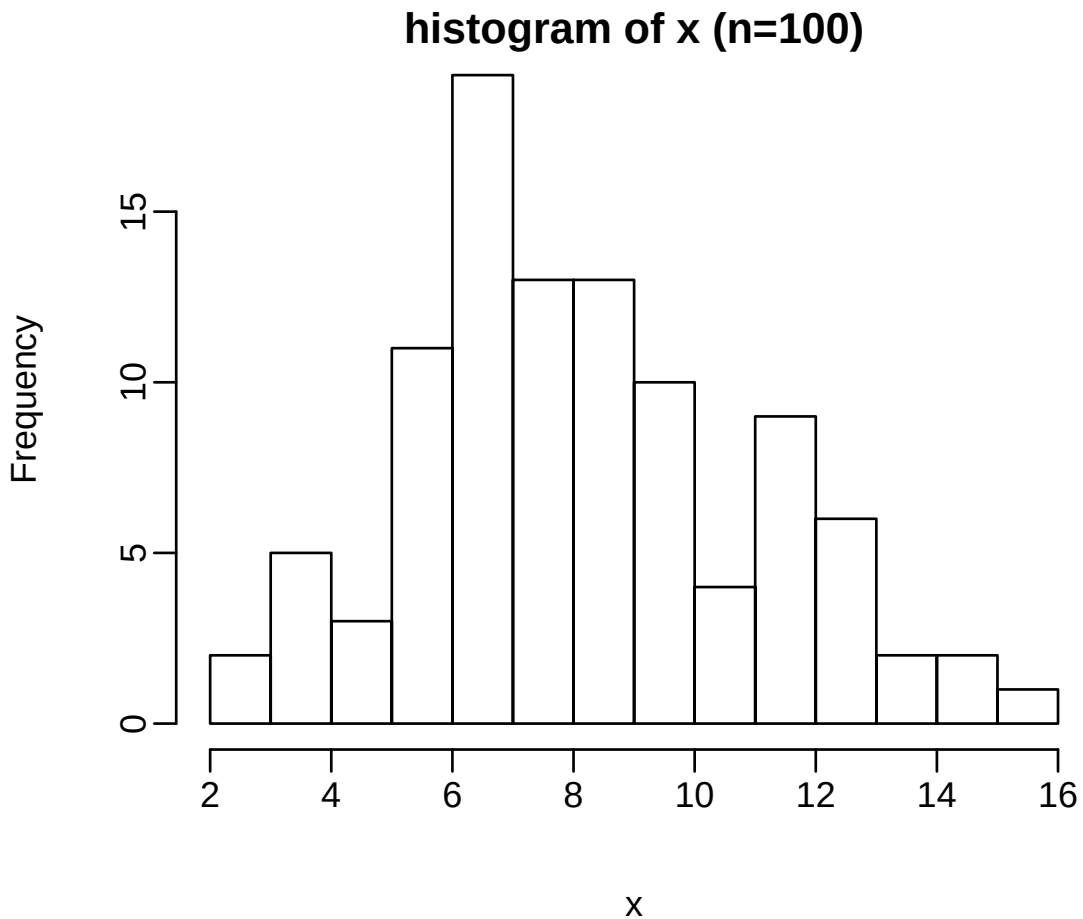
Summarizing quantitative variables: percentiles

The *p*th *percentile* (quantile) of sample is the value such that *p*% of the observations are less than [or equal to*] that value.



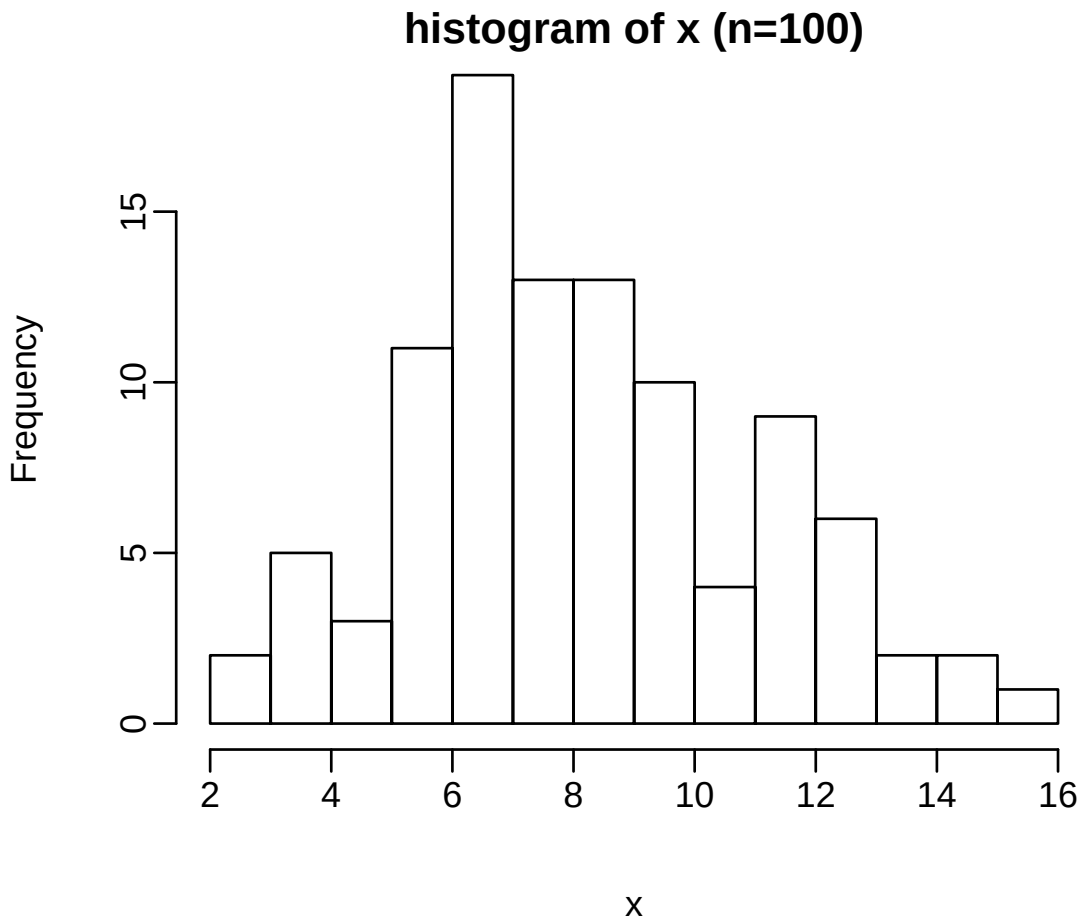
Summarizing quantitative variables: percentiles

Based on the histogram below, what is the 20th is percentile of the distribution of x?



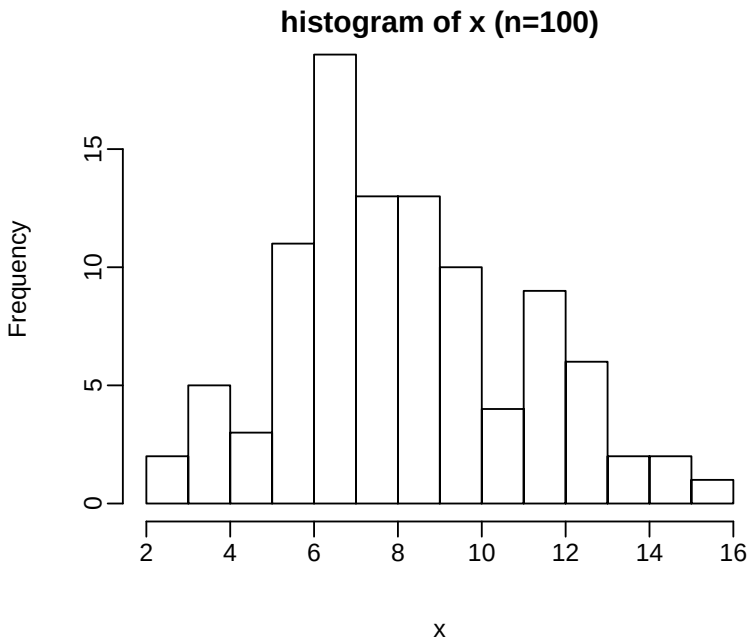
Summarizing quantitative variables: percentiles

Based on the histogram below, 12 is what percentile of the distribution of x?



Summarizing quantitative variables: percentiles

The *p*th *percentile* (quantile) of sample is the value such that *p*% of the observations are less than [or equal to*] that value.



- Software programs differ in their methods for calculating the percentiles. With large samples, these differences will be negligible.

Summarizing quantitative variables: percentiles

The **quartiles** are three special percentiles of a distribution.

Q_1	25th percentile
Q_2	50th percentile
Q_3	75th percentile

The middle 50% of values fall between Q_1 and Q_3.

The *interquartile range*, or IQR, is defined as Q3 - Q1 and gives the range of the middle 50% of values.

Summarizing quantitative variables: percentiles

To find the quartiles*:

Q_1	take the median of the lower half of the data set (excluding the median).
Q_2	median
Q_3	take the median of the upper half of the data set (excluding the median).

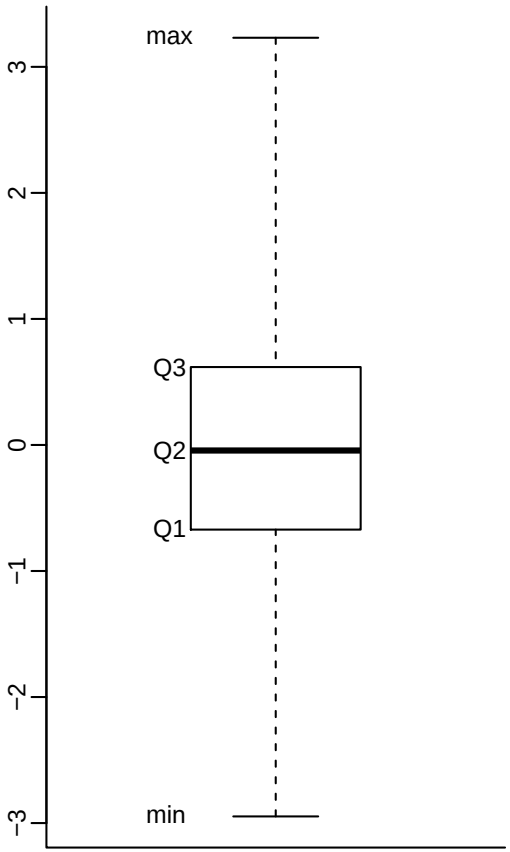
* `type 2` method in R.

Summarizing quantitative variables: percentiles

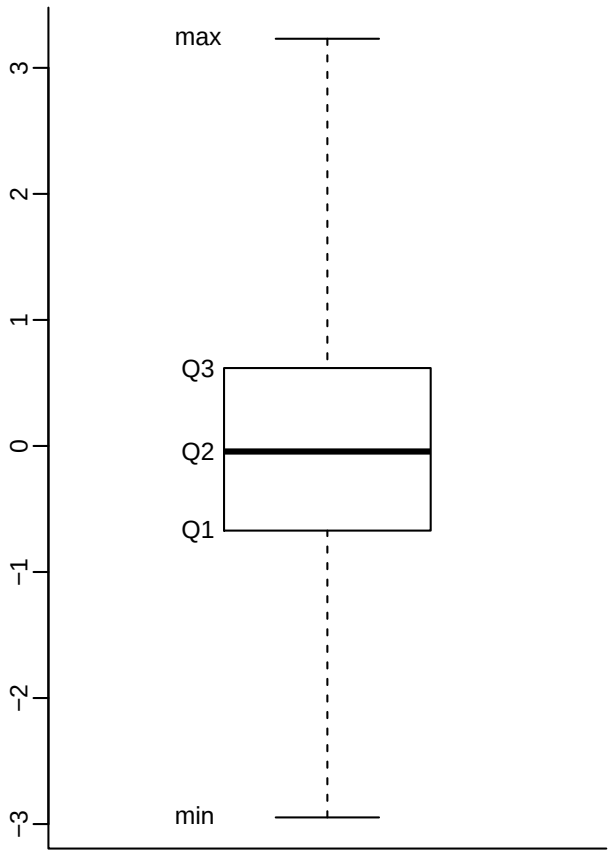
Example: find the quartiles of the following data set ($n = 12$).

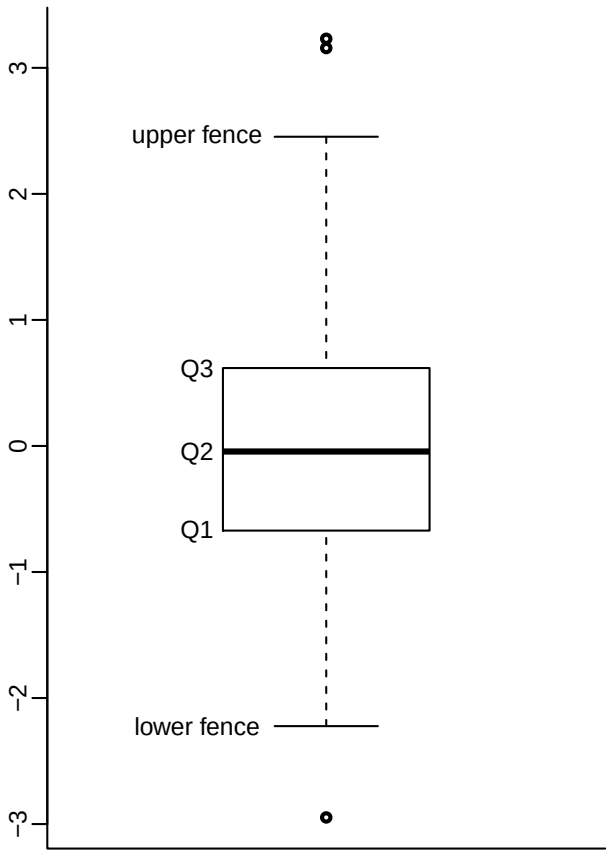
14 15 15 15 16 16 17 19 20 25 27 29

Summarizing quantitative variables: boxplots



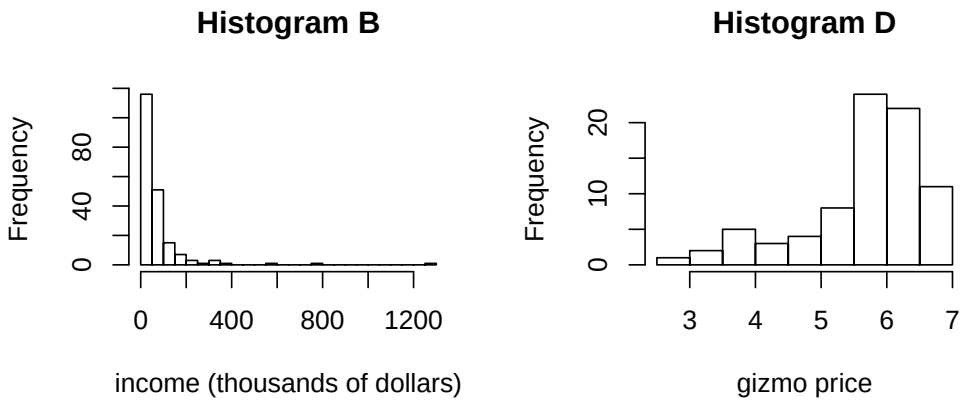
Summarizing quantitative variables: boxplots





Robust measures of center and variability

- The sample mean is sensitive to outliers and skewness.
- In right-skewed data, the mean $>>$ median. In left-skewed data, the mean $<<$ median.



- The sample standard deviation and variance are also sensitive to outliers.
- The median, mode, and IQR are more robust statistics. When data are skewed, report these in addition to / instead of the mean and standard deviation.