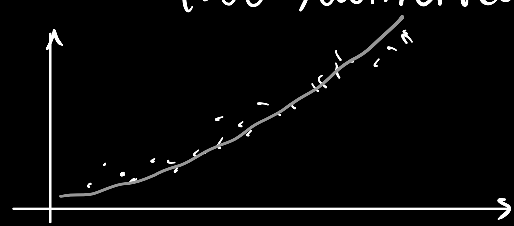
		a	b	c	d
county	1	x	x	x	x

- variable
  - numerical
    - contineous
    - discrete
  - categorical
    - regular
    - ordical

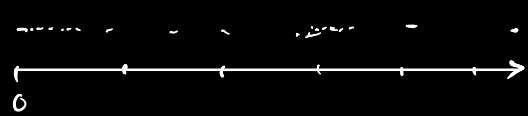
independent ⟷ associated

无关 有关

ScatterPlot Two numerical variables

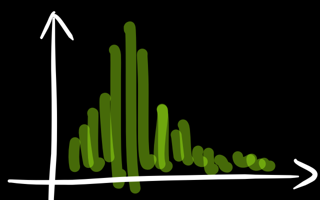


Dot Plot One variable

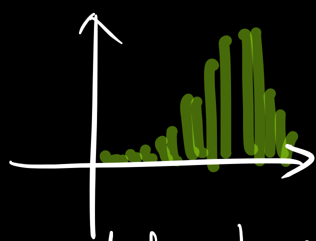


Mean (average) – common way to measure the center of the distribution of data

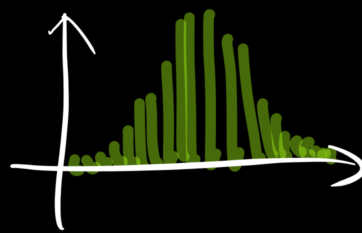
Histogram binned data – measure shape of the data distribution



right skewed



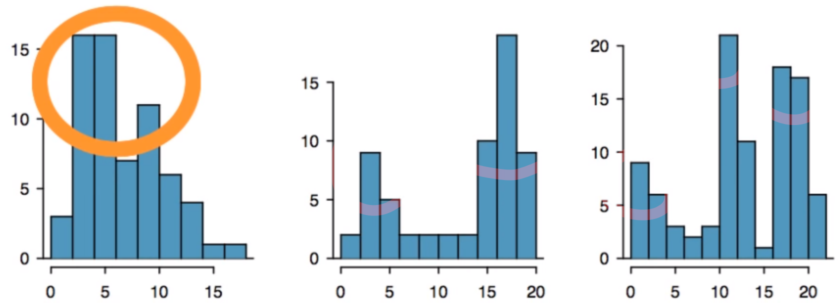
left skewed



symmetric

Histograms and shape

A **mode** is represented by a prominent peak in the distribution. The histograms below have one, two, or three prominent peaks. Such distributions are called **unimodal**, **bimodal**, and **multimodal**, respectively.



1.6.4 Variance and standard deviation

The mean was introduced as a method to describe the center of a data set, but the variability in the data is also important. Here, we introduce two measures of variability: the variance and the standard deviation. Both of these are very useful in data analysis, even though their formulas are a bit tedious to calculate by hand. The standard deviation is the easier of the two to understand, and it roughly describes how far away the typical observation is from the mean.

We call the distance of an observation from its mean its **deviation**. Below are the deviations for the 1^{st}, 2^{nd}, 3^{rd}, and 50^{th} observations in the `num_char` variable. For computational convenience, the number of characters is listed in the thousands and rounded to the first decimal.

$$x_1 - \bar{x} = 21.7 - 11.6 = 10.1$$
$$x_2 - \bar{x} = 7.0 - 11.6 = -4.6$$
$$x_3 - \bar{x} = 0.6 - 11.6 = -11.0$$
$$\vdots$$
$$x_{50} - \bar{x} = 15.8 - 11.6 = 4.2$$

[30] There might be two height groups visible in the data set: one of the students and one of the adults. That is, the data are probably bimodal.

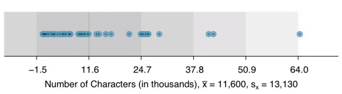


Figure 1.24: In the `num_char` data, 41 of the 50 emails (82%) are within 1 standard deviation of the mean, and 47 of the 50 emails (94%) are within 2 standard deviations. Usually about 70% of the data are within 1 standard deviation of the mean and 95% are within 2 standard deviations, though this rule of thumb is less accurate for skewed data, as shown in this example.

If we square these deviations and then take an average, the result is about equal to the sample **variance**, denoted by s^2:

s^2 sample variance

$$s^2 = \frac{10.1^2 + (-4.6)^2 + (-11.0)^2 + \cdots + 4.2^2}{50 - 1} = \frac{102.01 + 21.16 + 121.00 + \cdots + 17.64}{49} = 172.44$$

We divide by $n - 1$, rather than dividing by n, when computing the variance; you need not worry about this mathematical nuance for the material in this textbook. Notice that squaring the deviations does two things. First, it makes large values much larger, seen by comparing 10.1^2, $(-4.6)^2$, $(-11.0)^2$, and 4.2^2. Second, it gets rid of any negative signs.

The **standard deviation** is defined as the square root of the variance:

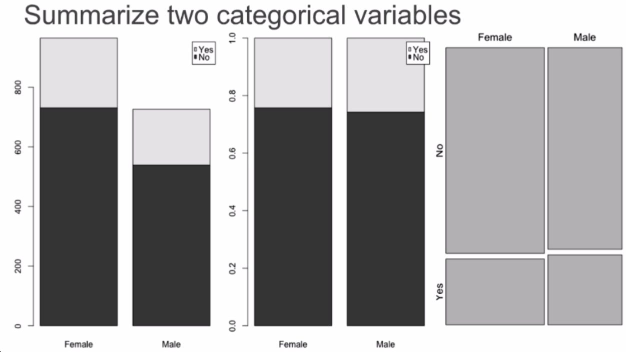
$$s = \sqrt{172.44} = 13.13$$

s sample standard deviation

The standard deviation of the number of characters in an email is about 13.13 thousand. A subscript of $_x$ may be added to the variance and standard deviation, i.e. s_x^2 and s_x, as a reminder that these are the variance and standard deviation of the observations represented

Stacked Bar Plot / Mosaic Plot

Summarize two categorical variables



Quartiles and IQR

The median is the value m such that 50% of observations are < m

Extending this to other percentages:

- First quartile: the value Q1 such that 25% of observations are < Q1
- Third quartile: the value Q3 such that 75% of observations are < Q3

(What is the second quartile?)

Interquartile range (IQR) = Q3 - Q1

The “middle 50%” of observations fall between Q3 and Q1

Small IQR: most observations fall close to the median

Large IQR: many observations fall far from the median

Five number summary

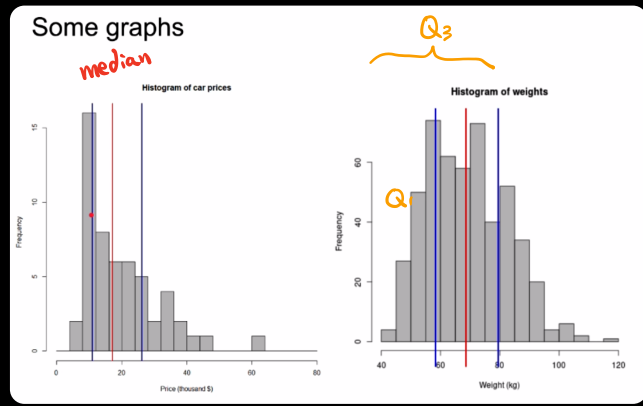
minimum

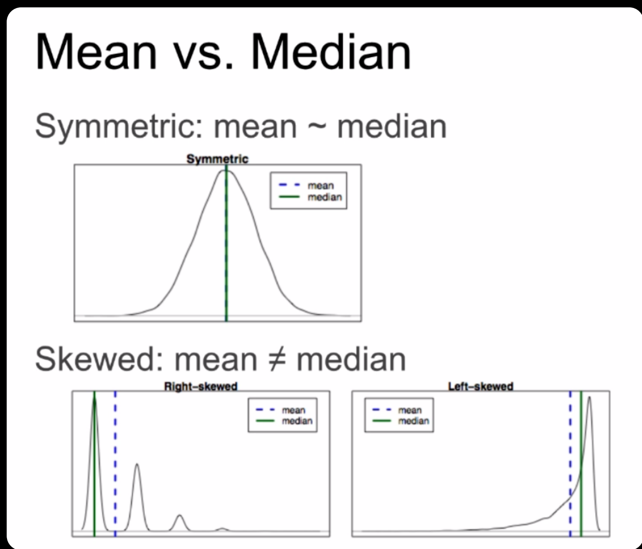
Q_1

median

Q_3

maximum



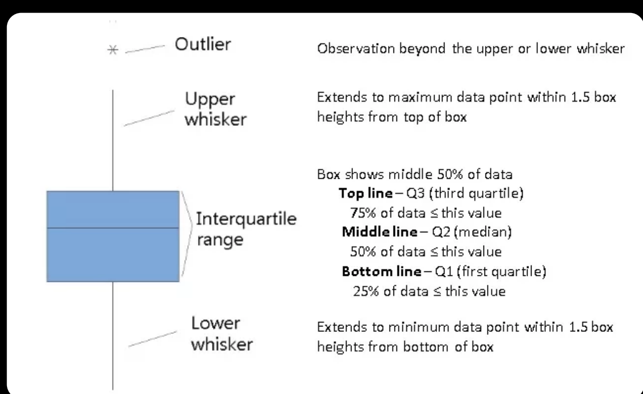


Robustness

A statistic is *robust* if it does not change much when an extreme observation is added to or removed from the data

Robust statistics are often more appropriate in the presence of strong skew or extreme outliers

- Median, quartiles *are* robust
- Mean, variance, standard deviation *are not* robust



勇哦