Chapter 2.2 - Visualizing Correlation

> "Correlation doesn't imply causation, but it does waggle its eyebrows suggestively and gesture furtively while mouthing 'look over there'." - xkcd

Every year Chicago experiences an increase in homicides at the same time as ice cream sales increase. This is a morbid (and tired) example of a beautiful statistical concept: correlation.

Correlation

Violent crime increases as it gets hotter outside. There's a lot of rationale behind this but the simplest one I can provide is that it's hard to commit murder when you're shivering indoors wrapped in blankets. Ice cream sales also increase as it gets hotter outside. (Hopefully) That one makes sense on it's own.

Despite the fact that people aren't more inclined to shoot their neighbor when they've got a Choco Taco in their hands, the data paints that picture. It's all a product of each variable, homicides and ice cream sales, having a **shared cause** for their increased prevalence.

This is the general basis behind correlation: **two or more variables that have some shared trend due to unknown causation or a shared cause.**

Correlation is one of our most powerful tools in statistics because we can use a relatively small amount of evidence to build a foundation upon which we can develop a full scientific story. Variables being correlated doesn't imply they're causative, but with a little bit of scientific exploration we can explain the correlations we see across different variables.

Scatterplots

The fundamental goal of applied statistics is to describe the relationship between variables measured from a sample of individuals representative of a greater population. So far we've looked at the relationship within single variables, but rarely are we ever concerned about making inference from just one variable.

Consider a sample of 10 patients, of varying demographics, all who received testing of their cholesterol levels during their last visit with their doctor. There are two variables for each individual in the sample:

$$X = \text{Patient Age} \qquad Y = \text{Serum Cholesterol Level}$$

For the i^{th} patient, we'll denote it's observed values as:

- x_i = the age of the i^{th} patient in *years*
- y_i = the serum cholesterol level of the i^{th} patient in mmol/L

Age	41	62	54	52	40	64	52	61	65	44
Cholesterol	235	231	286	196	167	263	204	307	360	226

Our data is a collection of **ordered pairs**; two values from the same individual. We sometimes refer to data consisting of ordered pairs as **bivariate data**, although this language is more common in mathematical statistics topics like distribution theory. We can notate these variables in the form of a table or a **vector**; a collection of points.

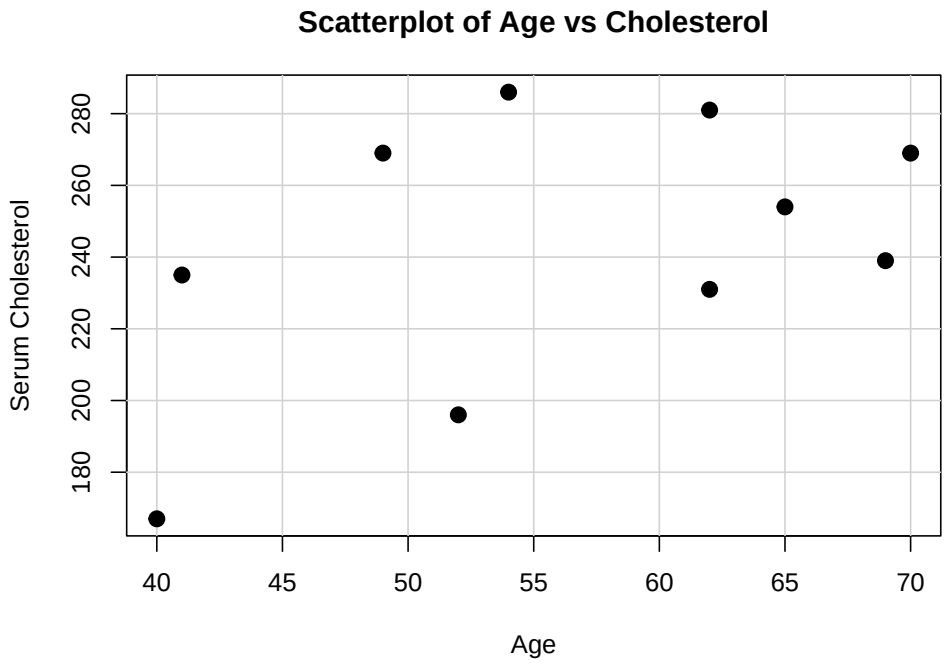
$$\boldsymbol{x} = (x_1, x_2, ...x_9, x_{10})'$$

$$\boldsymbol{y} = (y_1, y_2, ...x_9, y_{10})'$$

Then we can consider the data for any given individual to be a coordinate:

$$(x_1, y_1) = (41, 235), ..., (x_{10}, y_{10}) = (44, 226)$$

This concept of treating ordered pairs as coordinates isn't just a coincidence. We can plot these ordered pairs on an x and y coordinate plane by fiddling with the scale for each axis:



With this visualization alone we can ask ourselves real statistical questions: How are x and y related in this data? What happens to our cholesterol as we get older? How can we best describe the relationship between x and y?

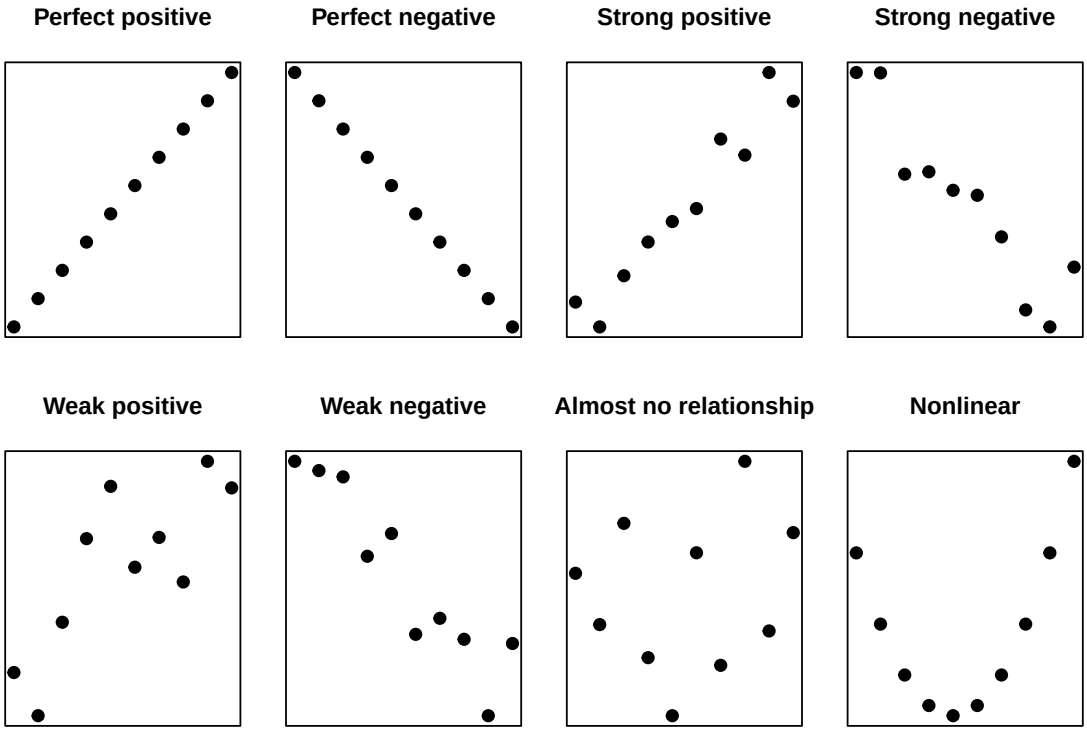
Instinct (or prior knowledge) might lead us to draw a line throw the plot to represent the trend in the data. When we can reasonably represent the relationship between two variables with a line we refer to it as a **linear association**.

Scatterplot Definitions

As always, it's important to develop vocabulary as we're improving our familiarity with a language. Scatterplots have their own cluster of vocabularies; we'll be sticking to a simple set for now.

For any two variables we can define their relationship as a:

- **Positive association** if large values of one variable are associated with large values of another
- **Negative association** if large values of one variable are associated with small values of another



Correlation Coefficient

Describing the relationship between two variables via visualization is simple, but it's not very convenient or reliable. Avoiding ambiguity is a good goal overall but it's especially important here. The **correlation coefficient** is statistics' solution to this problem.

Correlation coefficient: Numerical measurement of the strength (and direction) of the linear relationship between two quantitative variables

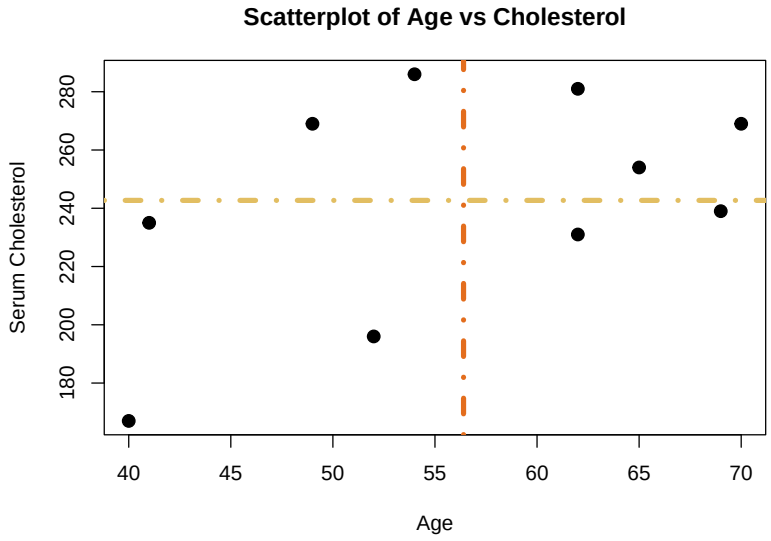
The correlation coefficient can be a cumbersome concept mathematically but as usual we'll be developing it piece by piece. We'll start with our cholesterol data:

Age	41	62	54	52	40	64	52	61	65	44
Cholesterol	235	231	286	196	167	263	204	307	360	226

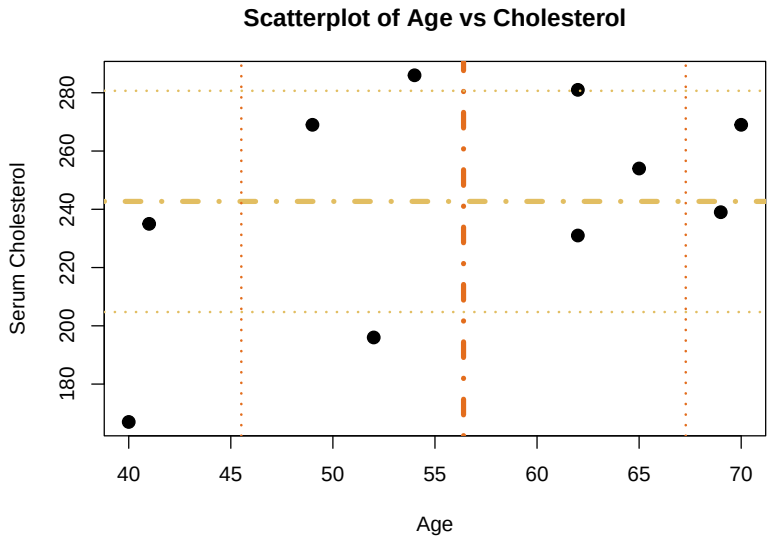
Let's continue to consider X = Age and Y = Cholesterol. In any data analysis it's best to start with the basics and calculate mean and standard deviation.

$\bar{x}$	$\bar{y}$	s_x	s_y
56.40	242.70	10.89	37.97

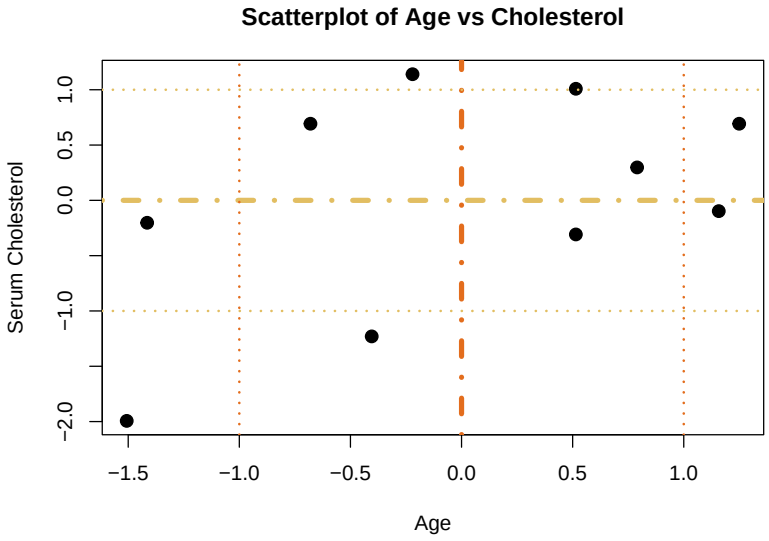
We can visualize the data on a scatterplot once again, and put segmented lines for the mean of each variable:



We'll add in ± 1 standard deviation for each variable with dotted lines to get an idea of what's going on with the data:



Currently, both variables (age and cholesterol) have different units and different scales. It's entirely possible to make comparisons between them using these units and scales as they are, but it's much more convenient to get them into the same format. This is easily done by standardizing them (a.k.a. converting both variables to z-scores).



We should recognize a few things right away:

$$\bar{x} = \bar{y} = 0$$

$$s_x = s_y = 1$$

That is, if we consider the new standardized values to be the same variables (which we will for now). All that's changed in our scatterplot is the values in the X and Y axis', everything else is in the same exact position.

If you were to take each variables "original" values and **multiply** them together ($\boldsymbol{x} \times \boldsymbol{y}$), and **average** them you would have calculated the **covariance** for the un-standardized variables. Covariance is (loosely) a measurement of the association between two variables, however statisticians tend to only concern themselves with the *sign* of covariance (the direction of the association). When we **standardize** the data as we've done here, covariance becomes something far more interpretable.

Covariance has a more involved (and accurate) proof behind it than the one being shown here. It should be pointed out that a proper proof of it's derivation requires the use of trigonometry, the annoying person showing up at every party despite nobody knowing who continues to invite them. I would sooner burn this text and destroy all evidence of its existence before I allowed trigonometry to formally involve itself in an example. Hence, we'll make some wild over-simplifications and proceed as if they're accurate.

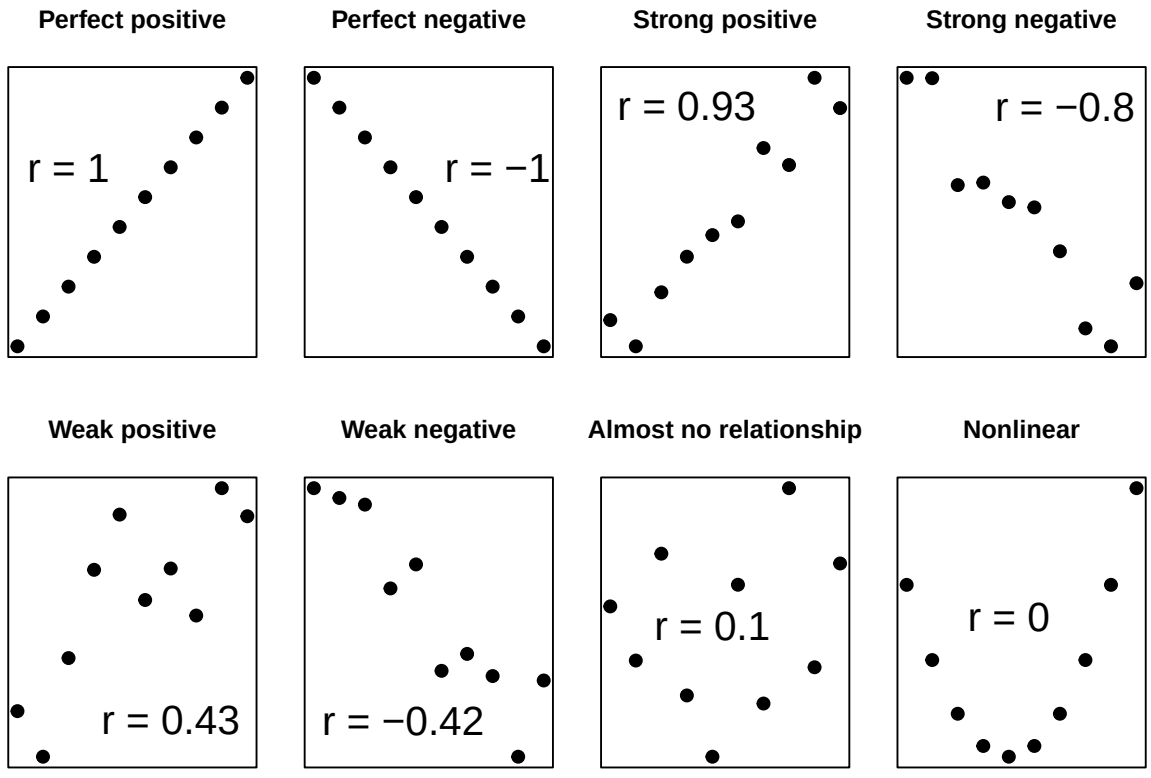
We have two vectors of z-scores, one for x and one for y. We'll multiply these vectors by one another and add the values up:

$$\sum(\boldsymbol{x} \times \boldsymbol{y}) = 4.414$$

We can now average this value, but once again we'll wave our hands and use the $n - 1$ trick to adjust for bias.

$$\frac{\sum(\boldsymbol{x} \times \boldsymbol{y})}{n - 1} = 0.49$$

This is the correlation coefficient, r, and is a rough measurement for linearly associated data on the **strength** and **sign** of their relationship. In this case the association is a **weak(er) positive** linear relationship, which tracks since Y is generally increasing as X is increasing (and vice versa) but they're not perfectly set on a line.



We'll recap with the proper formula below:

Given n ordered pairs (x_i, y_i), with sample means $\bar{x}$ and $\bar{y}$, sample standard deviations s_x and s_y; the correlation coefficient r is given by:

$$r = \frac{1}{n-1} \sum_i \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

Properties of the Correlation Coefficient

1. The value is always between $-1 \leq r \leq 1$
   - If $r = 1$, all of the data falls on a line with a positive slope.
   - If $r = -1$ all of the data falls on a line with a negative slope.
   - The closer r is to 0, the weaker the linear relationship between x and y.
   - If $r = 0$ *no linear relationship exists.*
   - Statisticians generally consider values between 0 and ± 0.6 to be **weak** relationships, however this isn't a perfect rule.
2. The correlation does not depend on the unit of measurement for the two variables
   - x is years and y is mmol/L, but they can still have r calculated.
3. Correlation is very sensitive to outliers.
   - One point that does not belong in the dataset can result in a misleading correlation.
   - Always plot your data!
4. Correlation measures only the linear relationship and may not (by itself) detect a nonlinear relationship
   - A nonlinear relationship won't always show up as $r = 0$ so you should be aware of what your data looks like whenever analyzing correlation.

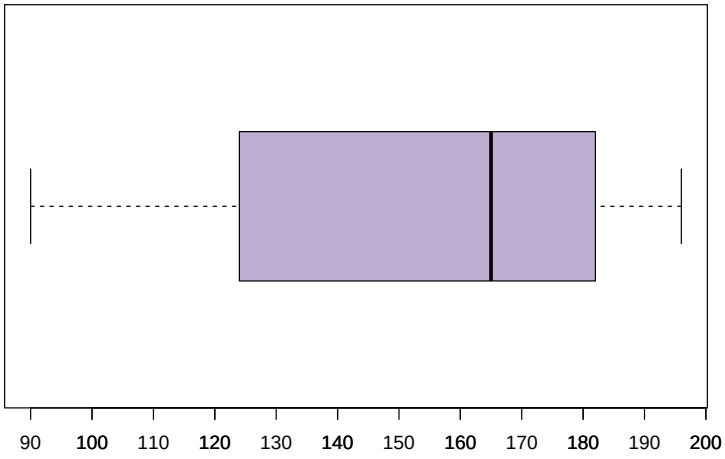
Boxplots

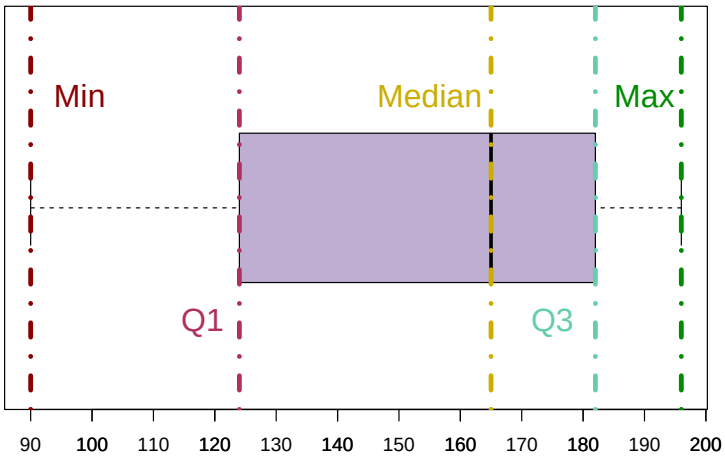
Another useful visualization for assessing data is the **boxplot**. While we're going to start by introducing **univariate** (one variable) boxplots, we're building towards using this graphic as a way of assessing multiple variables simultaneously.

A **boxplot** (or "box-and-whisker" plot if you're old like me) is a graphical display of a five number summary.

Given the five-number summary of the FMD data, the associated boxplot is shown below:

Min	Q_1	Median	Q_3	Max
90	124	165	182	196





Each component of the five number summary is on this boxplot (showcased below) but not all boxplots are made equally. It's not uncommon to remove the "whiskers" (min and max) for the purpose of comparing many boxplots in one figure.

How to Construct a Boxplot

1. Find the 5 values in the five number summary
    a. Compute the IQR
    b. Find the upper & lower bounds for outliers
2. Draw a number line to represent the scale
3. Above the number line, draw a box with one end at Q_1 and the other at Q_3
    a. Draw a verticle line across the box at the **median**
4. Draw horizontal lines ("whiskers") from the box to the smallest and largest values within the upper & lower outlier bounds
5. Plot observations outside the bounds with a "star" (*) to identify them as outliers

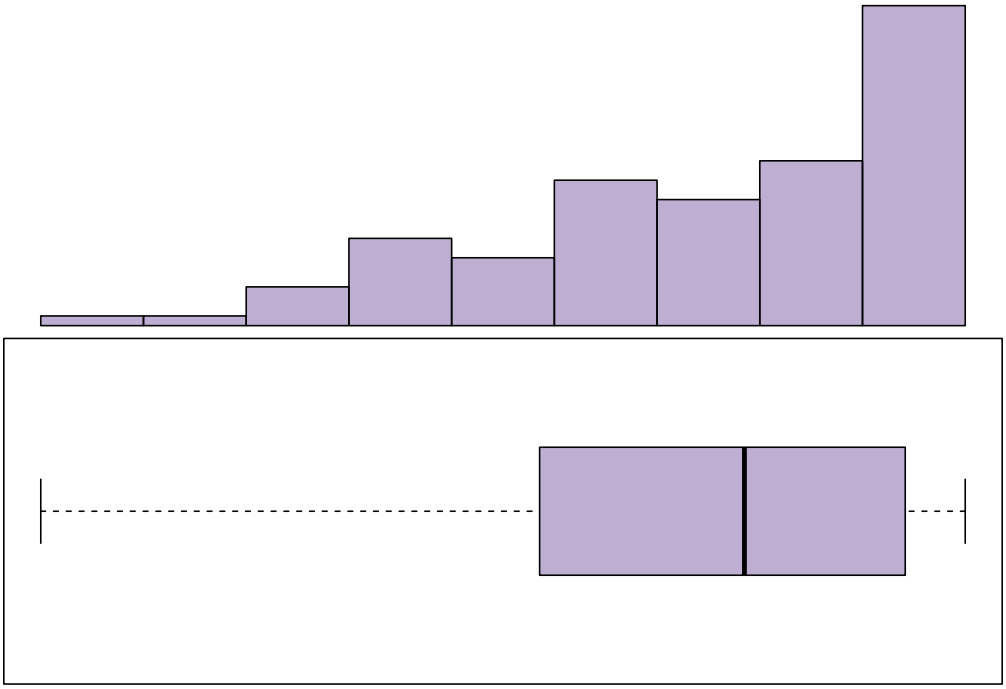
---

Skewness and Boxplots

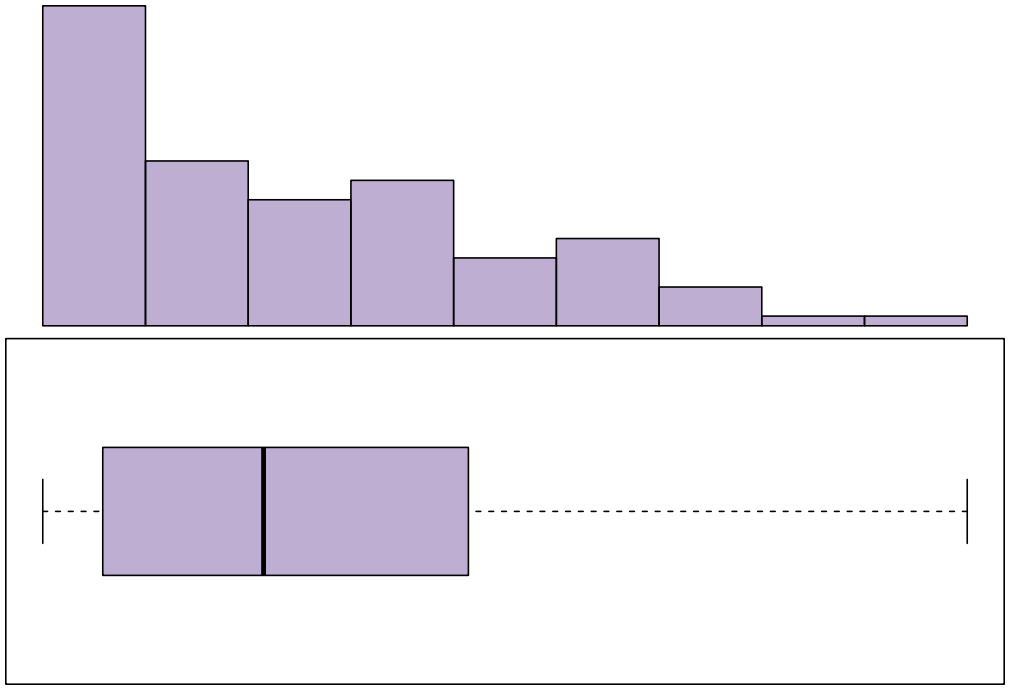
Skewness is a troubling topic, as it really doesn't matter that much. Higher level courses in applied statistics care about shape and visible skew just enough to make assumptions about the data and some more niche courses in mathematical statistics will enlighten students to the reality that skewness is completely different from what we're taught in introductory classes.

Still it's important to learn a skill even when we suspect it will become obsolete— even if the rational is simply to understand the history of our science. Skewness of boxplots mimics that of histograms directly.

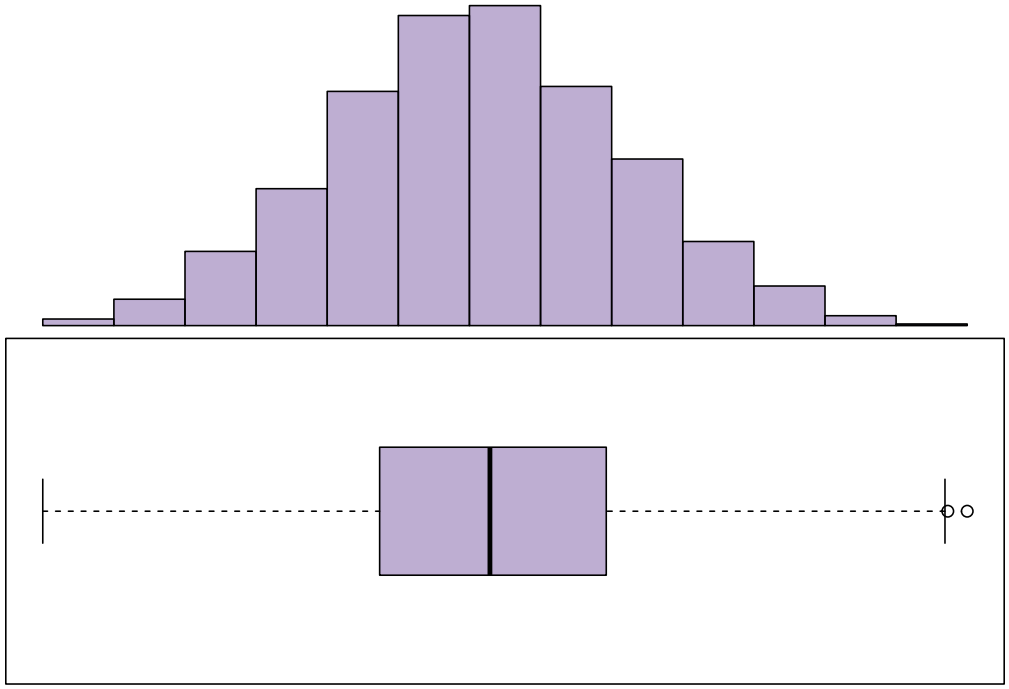
This would be **negatively-skewed**:



Positively-skewed is the opposite:



Approximately symmetric:



It should be noted that not **all** boxplots will use the same symbol to denote outliers. In the boxplot below I've used circles, but stars and crosses are just as common.

Comparative Boxplots

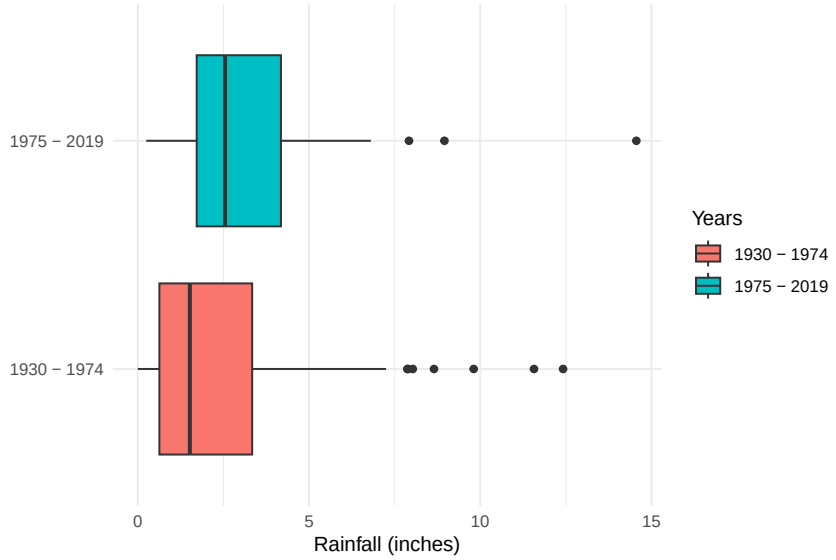
When we're working with multiple variables that are on the **same scale** we can make the task of comparison trivial by using boxplots. While histograms and scatterplots rule the graphical market for observational studies, boxplots have a monopoly in designed experiments.

We can come up with the reasoning for this easily, since an observational study only looks at phenomena as they naturally occur and a designed experiment is a complete control of nature. When the scientist is in charge of what happens they tend to go *out of their way* to make sure everything's similar units and scales.

Consider the data below of annual rainfall data (in inches) in LA during February: $1930 - 1974$

Year	Rainfall	Year	Rainfall	Year	Rainfall	Year	Rainfall	Year	Rainfall
1930	0.45	1939	1.13	1948	1.29	1957	1.47	1966	1.51
1931	3.25	1940	5.43	1949	1.41	1958	6.46	1967	0.11
1932	5.33	1941	12.42	1950	1.67	1959	3.32	1968	0.49
1933	0.00	1942	1.05	1951	1.48	1960	2.26	1969	8.03
1934	2.04	1943	3.07	1952	0.63	1961	0.15	1970	2.58
1935	2.23	1944	8.65	1953	0.33	1962	11.57	1971	0.67
1936	7.25	1945	3.34	1954	2.98	1963	2.88	1972	0.13
1937	7.87	1946	1.52	1955	0.68	1964	0.00	1973	7.89
1938	9.81	1947	0.86	1956	0.59	1965	0.23	1974	0.14

We can compare the data from $1930 - 1974$ with data from $1975 - 2019$ using boxplots:



Consider the following questions as an exercise:

1. What can you say about the shape of each dataset?
2. In which time period was the amount of rainfall generally greater?
3. On the whole, the rainfall was more variability in which time period?