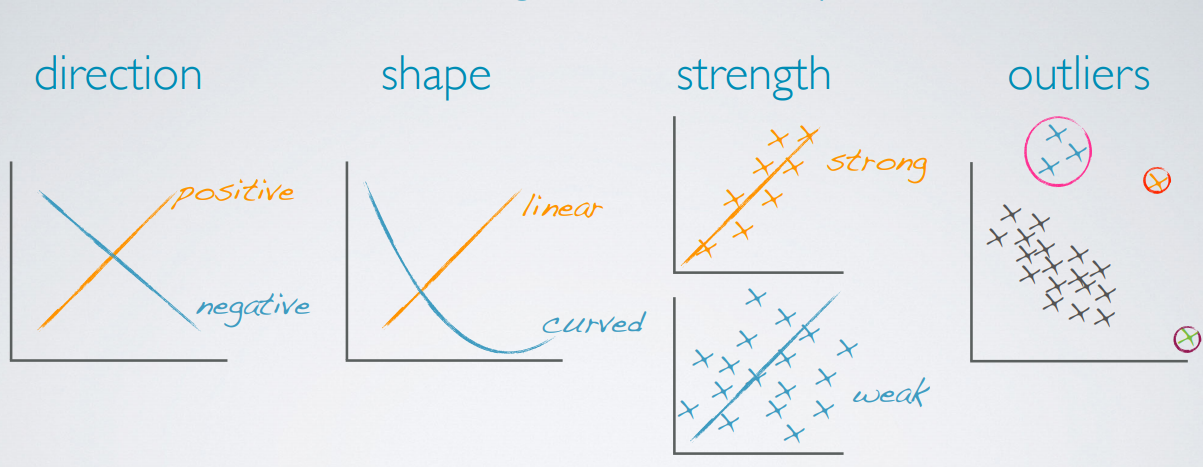
# Exploring Numerical Data

## Overview

* When describing the distribution of a numerical variable, mention its shape, center, and spread, as well as any unusual observations.
* there are three commonly used measures of center and spread:
* center: mean (the arithmetic average), median (the midpoint), mode (the most frequent observation).
* spread: standard deviation (variability around the mean), range (max-min), interquartile range (middle 50% of the distribution).
* Identify the shape of a distribution as symmetric, right skewed, or left skewed, and unimodal, bimodal, multimodal, or uniform.
* Define a robust statistic (e.g. median, IQR) as a statistic that is not heavily affected by skewness and extreme outliers, and determine when such statistics are more appropriate measures of center and spread compared to other similar statistics.

## Evaluating the relationship



## Modality



Bimodal might indicate there are two distinct groups in our data.

## Variance

* Sample variance
* Population variance

Roughly the average squared deviation from the mean

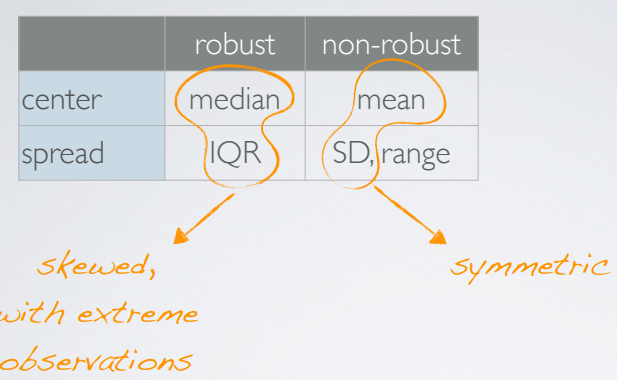
## Standard deviation

* Sample standard deviation
* Population standard deviation

Roughly the average deviation around the mean, and has the same unit as the mean.

## Robust statistics

We define robust statistics as measures on which extreme observations have little effect.



## (natural) Log transformation

* Often applied when much of the data cluster near zero (relative to the larger values in the data set) and all observations are positive.
* Sometimes to make the relationship between the variables more linear and hence easier to model with simple methods.

Other transformations:

* Square root
* Inverse

Goals of transformations

* To see the data structure differently
* To reduce skew assist in modeling
* To straighten a nonlinear relationship in a scatterplot

# Exploring Categorical Data and Introduction to Inference

## Difference between bar plot and histogram

* Barplots for categorical variables, histograms for numerical variables
* X-axis on a histogram is a number line, and the ordering of the bars are not interchangeable.

## Plot types

