# Data Science from Scratch

First Principles with Python

Stephen CUI

March 20, 2023

# Contents

# Chapter 1

# Statistics

## 1.1 Describing a Single Set of Data

One obvious description of any dataset is simply the data itself. For a small enough dataset, this might even be the best description. But for a larger dataset, this is unwieldy and probably opaque. (Imagine staring at a list of 1 million numbers.) For that reason, we use statistics to distill(提取) and communicate relevant features of our data.

As a first approach, you put the data counts into a histogram using Counter and plt.bar (Figure 1.1).

### 1.1.1 Central Tendencies

Usually, we'll want some notion of where our data is centered. Most commonly we'll use the mean (or average), which is just the sum of the data divided by its count.

We'll also sometimes be interested in the median, which is the middle-most value (if the number of data points is odd) or the average of the two middle- most values (if the number of data points is even).

Notice that—unlike the mean—the median doesn't fully depend on every value in your data. For example, if you make the largest point larger (or the smallest point smaller), the middle points remain unchanged, which means so does the median.

At the same time, the mean is very sensitive to outliers in our data. I If outliers are likely to be bad data (or otherwise unrepresentative of whatever phenomenon we're trying to understand), then the mean can sometimes give us a misleading picture.

A generalization of the median is the quantile, which represents the value under which a certain percentile of the data lies (the median represents the value under which 50% of the data lies).

Less commonly you might want to look at the mode, or most common value(s).

### 1.1.2 Dispersion(离散度)

Dispersion refers to measures of how spread out our data is. Typically they're statistics for which values near zero signify not spread out at all and for which large values (whatever that means) signify very spread out. For instance, a very simple measure is the range, which is just the difference between the largest and smallest elements.

A more complex measure of dispersion is the variance.

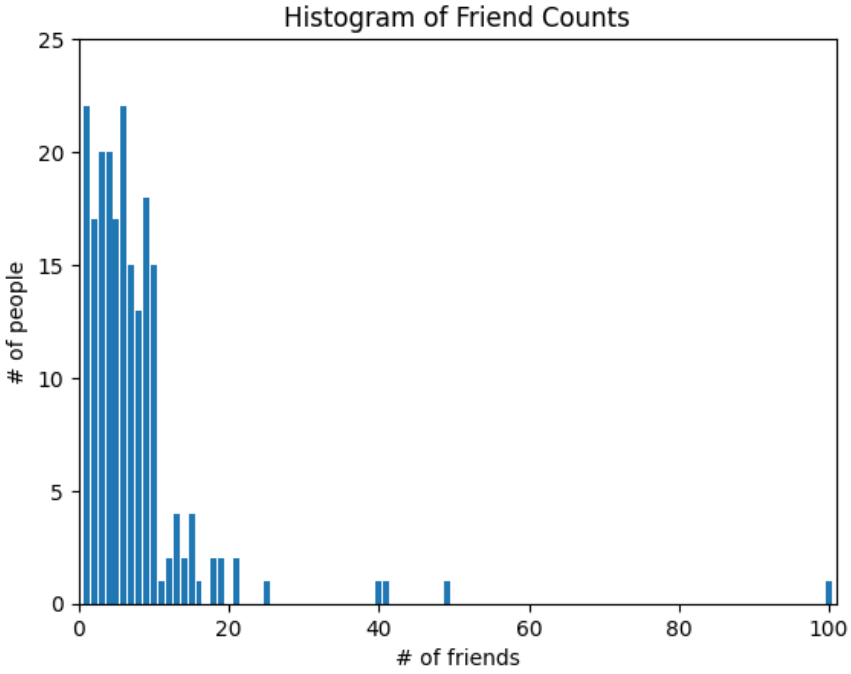


Figure 1.1: A histogram of friend counts

Now, whatever units our data is in, all of our measures of central tendency are in that same unit. The range will similarly be in that same unit. The variance, on the other hand, has units that are the square of the original units. As it can be hard to make sense of these, we often look instead at the standard deviation.

Both the range and the standard deviation have the same outlier problem that we saw earlier for the mean. A more robust alternative computes the difference between the 75th percentile value and the 25th percentile value, which is quite plainly unaffected by a small number of outliers.

## 1.2 Correlation

We'll look at covariance, the paired analogue of variance. Whereas variance measures how a single variable deviates from its mean, covariance measures how two variables vary in tandem from their means.

Nonetheless, this number can be hard to interpret, for a couple of reasons:

1. Its units are the product of the inputs' units, which can be hard to make sense of.
2. If one data had twice as many friends (but the same number of the other), the covariance would be twice as large. But in a sense, the variables would be just as interrelated. Said differently, it's hard to say what counts as a "large" covariance.

For this reason, it's more common to look at the correlation, which divides out the standard deviations of both variables.

The correlation is unitless and always lies between –1 (perfect anticorrelation) and 1 (perfect correlation). Correlation can be very sensitive to outliers.

## 1.3 Simpson's Paradox

One not uncommon surprise when analyzing data is Simpson's paradox, in which correlations can be misleading when confounding variables are ignored. The key issue is that correlation is measuring the relationship between your two variables all else being equal. If your dataclasses are assigned at random, as they might be in a well-designed experiment, "all else being equal" might not be a terrible assumption. But when there is a deeper pattern to class assignments, "all else being equal" can be an awful assumption.

The only real way to avoid this is by knowing your data and by doing what you can to make sure you've checked for possible confounding factors. Obviously, this is not always possible. 但是有些时候没有必要的数据，可能得出的结论就会有问题。

## 1.4 Some Other Correlational Caveats

A correlation of zero indicates that there is no linear relationship between the two variables. However, there may be other sorts of relationships.

In addition, correlation tells you nothing about how large the relationship is. 因为有些强关系，但是实际中没有任何的意义。

## 1.5 Correlation and Causation

You have probably heard at some point that "correlation is not causation.". Nonetheless, this is an important point—if x and y are strongly correlated, that might mean that x causes y, that y causes x, that each causes the other, that some third factor causes both, or nothing at all.

One way to feel more confident about causality is by conducting randomized trials. If you can randomly split your users into two groups with similar demographics and give one of the groups a slightly different experience, then you can often feel pretty good that the different experiences are causing the different outcomes.