Data Science Essentials

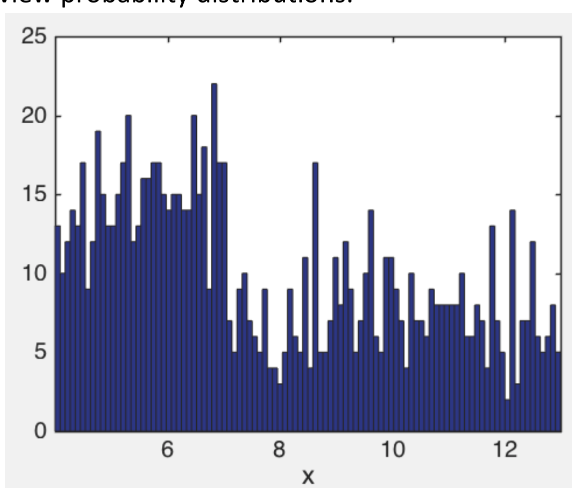
Statistics

Data science is largely concerned with statistical relationships and distributions of data.

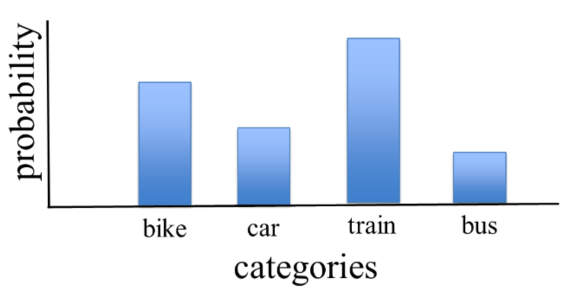
Visualizing Statistics

One of the first things you should do with data, is to look at it – often by creating visualizations that show the comparative frequency with which different data values occur or plot relationships between different variables.

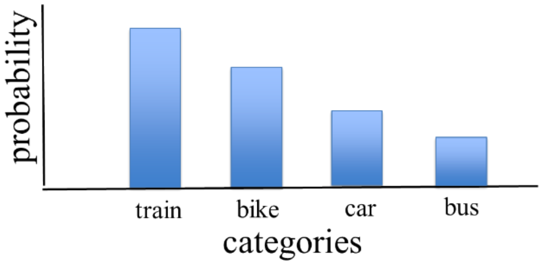
You can use a histogram to view probability distributions.



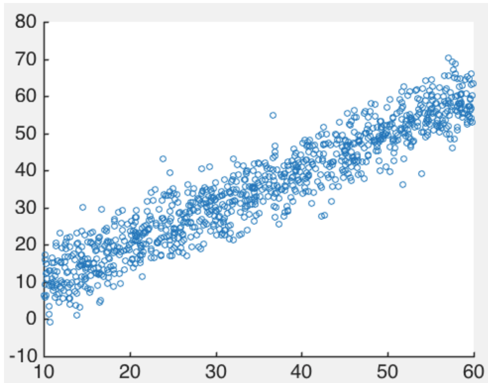
Bar charts are useful for plotting categorical data.



A Pareto chart is a bar chart with the data in descending order.



A scatter plot is used to show the intersections of two numeric variables.



Summary Statistics

Some common statistics can be calculated to identify the central tendency, positionality, and distribution of the data points.

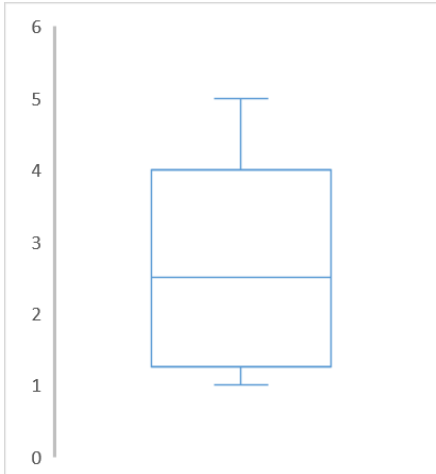
- The *mean* is the average of a distribution of data points.
- You can order the data and divide into *quantiles* that describe the ranked position of a value in relation to the other values in the dataset. For example, you could divide the dataset into 100 quantiles (technically *percentiles*) so that for example the data point at the 20th percentile (or P_{20}) has a value greater than 20% of the data. When you divide the data into four ranges, they are referred to as *quartiles* and the boundaries of these ranges are indicated as Q_i (for example Q_1 indicates the boundary between the first quarter and the second quarter, and is the same value as P_{25}).
- The *median* value is the middle value of the data points (that is, Q_2 or P_{50}).
- A common *five number summary* consists of the minimum (min), maximum (max), and the first, second, and third quartile boundaries (Q_1, Q_2, and Q_3).
- The *range* is a measurement of the overall distribution of the data values, and is measured as the min - max.
- The inter-quartile range (IQR) is measured as Q_3 - Q_1.
- The *variance* measures how widely the values in the dataset vary from one another. The calculation for variance produces a squared value.
- The *standard deviation* is the square root of the variance.

For example, consider the following data values:

```
1, 1, 2, 2, 3, 4, 4, 5
```

- The mean value of this sample is (1+1+2+2+3+4+4+5) ÷ 8, which yields the result 2.75.

- The value 1 is at the 0th percentile (that is, the minimum), and 5 is at the 100th percentile (the maximum). The value 3 is greater than 57.1% of the data points, so it is at the 57.1th percentile.
- The median value (or the 50th percentile) is 2.5 (the average of the middle values, 2 and 3) 50% of the values are less than this, and 50% are greater.
- The five number summary for this dataset is 1, 1.25. 2.5, 4, and 5. This can be summarized visually in the following box plot:



- The range is as 5 - 1, which is 4; and the IQR is 4 - 1.25, which is 2.75.
- The variance is 2.214, and the standard deviations is 1.488.

Note: You need to be careful to use the appropriate measurement and terminology depending on whether you're measuring the entire population or a sample of the data. For example, the *sample mean* ($\bar{x}$ or *xbar*) is an approximation of the *population mean* (μ or *mu*). Similarly, the formula to calculate the *population variance* (σ^2) is slightly different from that of the *sample variance* (s^2) to account for bias in the sample data.

Z-Scores

The Z-score of a data point measures how many standard deviations above or below the mean the value of the data point represents. The closer the Z-Score is to zero, the closer the value is to the mean. For example, the following table shows the data points discussed previously along with their Z-scores based on a mean of 2.75 and a standard deviation of 1.488:

Value	**Z-Score**
1	-1.17608
1	-1.17608
2	-0.50403
2	-0.50403
3	0.168011
4	0.840054
4	0.840054
5	1.512097

Covariance and Correlation

You can measure the relationship between two numeric variables (let's call them X and Y) to determine covariance or correlation.

Covariance is calculated by finding the mean of X - x̄ multiplied by Y - ỹ. This yields a positive result in cases where X increases while Y increases, a negative result in cases where X increases when Y decreases, and zero when a change in X appears to have no relationship to a change in Y.

The values calculated by covariance are absolute values that have no readily interpretable meaning. *Correlation* is scaled version of covariance that always returns a value between -1 and +1.

In both of these approaches, it is important to remember that correlation (or covariance) does not imply causation. You can determine that when X is high, Y tends to be low; but you cannot draw the conclusion that Y is low *because* X is high.