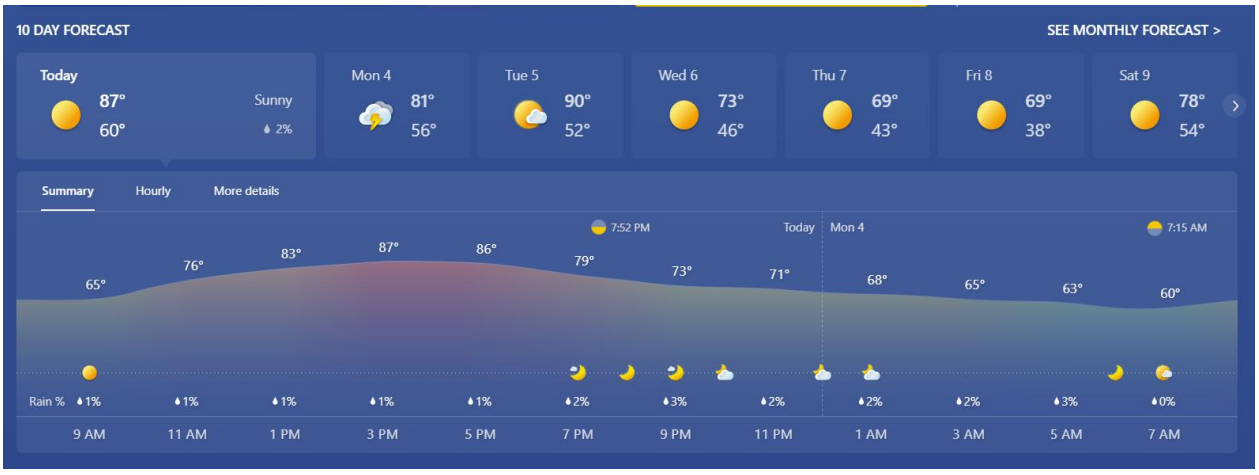
Examples of Using Statistics

Weather Prediction

Statisticians use models to predict the weather.

- The Automated Surface Observing Systems (ASOS) run by the National Weather Service (NWS), the Federal Aviation Administration (FAA), and the Department of Defense (DOD), records surface temperatures in the U.S. and uses hundreds of monitoring stations.
- Over 150 Doppler Radar stations give up-to-date information about current and emerging weather patterns.
- A radiosonde is launched twice daily from nearly 100 locations in the U.S. (closer to 800 worldwide) and gathers data about the conditions of the upper atmosphere. This small instrument (250 grams or so) is carried by a small balloon.
- NOAA, the National Oceanic and Atmospheric Administration, uses three different types of satellites.
- Geostationary satellites, which orbit the earth at approximately 22,800 miles. As their name suggests, they keep one area in view.
- Polar-orbiting satellites, which circle the earth 14 times a day. These satellites orbit approximately 500 miles and see every location on earth twice per day.
- NOAA's Deep Space Climate Observatory (DSCOVR) is approximately one million miles away and monitors the solar energy impacting the earth as well as aerosols and ozone levels.

All of this information can be used to create ten-day forecasts, such as the one below.



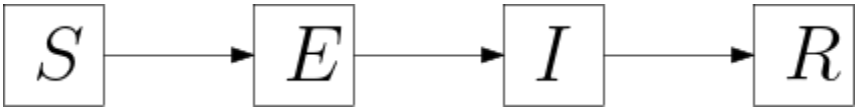
Some of this information is fed to supercomputers (according to NOAA, 2.8 quadrillion mathematical calculations per second around the clock) and forecast models are created. The accuracy of these models varies. According to SciJinks,

- Five-day forecasts have about 90% accuracy
- Seven-day forecasts have about 80% accuracy

- Ten-day or longer have poor accuracy, maybe 50%.

Mathematical Biology

Consider the spread of an illness such as the flu. It might follow the following model as shown in the below compartmental diagram.



This diagram indicates the movements of individuals throughout the disease cycle.

- SS represents individuals who are *susceptible* to the disease but have not yet contracted it.
- EE represents individuals who have been *exposed* and have contracted the disease but are not yet exhibiting symptoms.
- II represents *infected* individuals who have the disease and are exhibiting symptoms.
- RR represents the individuals who have *recovered* from the disease and are no longer infectious.

Statistics can be used to help answer the following questions.

- How long, on average, will an individual be infected?
- What is the chance of contracting the disease?
- How many other individuals will a single, infectious individual infect?
- What is the chance of dying?

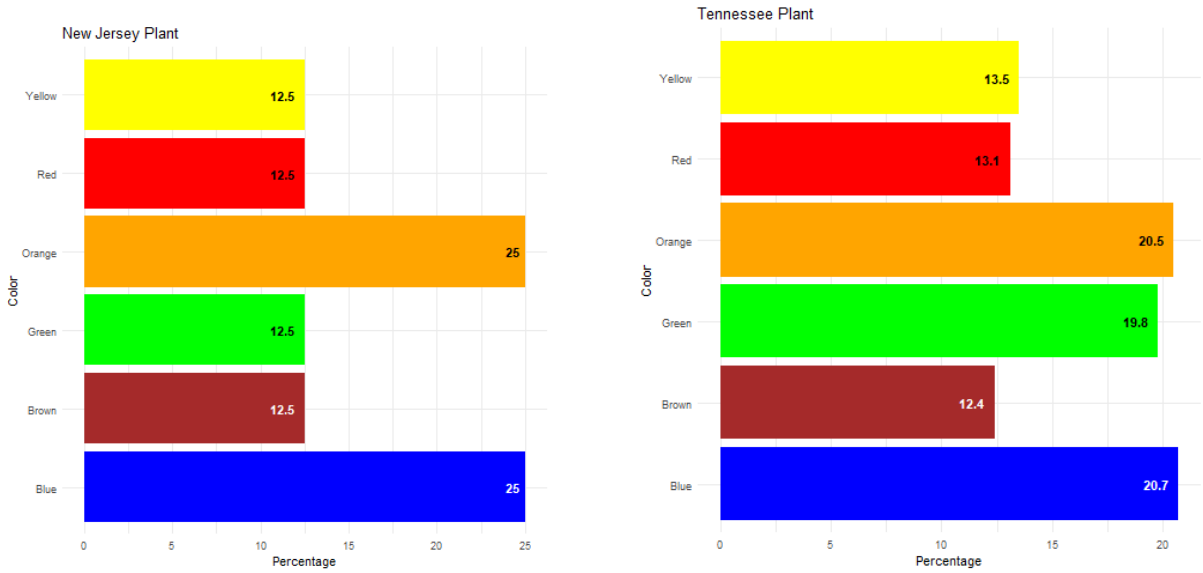
Netflix Recommendations

A large percentage of shows are discovered through Netflix's recommendation system. This system generates recommendations through

- User profiles report information about
- what you watched before a show
- what you watched after a show
- what time of day did you watch
- what you watched a year ago
- Taggers who watch the shows and report on content such as
- how cerebral the show might be
- type of actors
- genre
- many, many more
- Machine learning algorithms are used with the above data to generate recommendations.
- There are several thousand taste communities or groups
- Some algorithms differentiate between countries, languages, and cultures.
- Explicit data - your thumbs up on a show, has a direct impact
- Implicit data - you binge-watched an entire series recently also impacts your recommendations.

Quality Control

Ever wondered how many of each color M&M should appear in a given bag? In 2017, Mars reported that two plants (in the U.S.) are manufacturing M&Ms (according to Rick Wicklin who investigated these distributions). The color distributions are different, as shown below.



There are a number of reasons someone might want to investigate the distribution of colors of M&Ms.

- You could use a random sample to determine which plant manufactured the M&Ms. This might be useful in tracking a shipment of M&Ms when it has some sort of flaw such as contamination.
- As a quality control expert working for a plant, a random sample could be used to ensure that your plant is manufactured according to the proportions stated.
- Any excuse to eat M&Ms.

Linear Regression

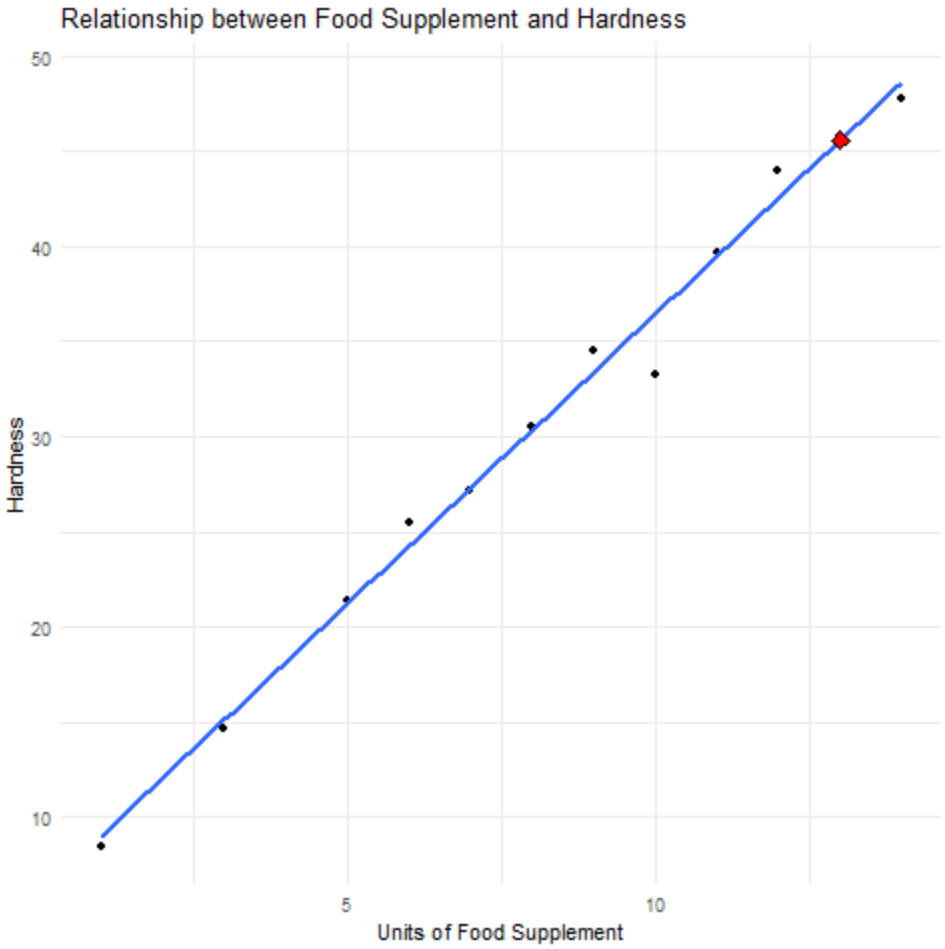
The following example is taken from Graybill's ***Theory and Application of the Linear Model***. A new food supplement (xx units) was fed to a breed of chickens for six weeks to determine the effect on the hardness of the eggshells. A linear model was assumed for each breed. Can we predict the hardness of the shell given the number of units given? The linear regression line is the line which is closest **at the same time** to every data point. The line is plotted below. Notice how close the line is to all of the data. The linear regression equation is given by

$$y\hat{}=5.9662x+3.0494.y\hat{}=5.9662x+3.0494.$$

We can use the line to estimate the hardness given reasonable values for the food supplement, xx. Let's try $x=13x=13$, which is not one of our data points. Then

$$=5.9662(13)+3.0494=45.61.=5.9662(13)+3.0494=45.61.$$

So, given 13 units of the food supplement, we should expect an egg with a hardness of around 45. This point is plotted on the line and is represent by the red diamond.


Relationship between Food Supplement and Hardness
Hardness
50
40
30
20
10
5
10
Units of Food Supplement