This notebook aims to summarize my key learnings from year 1 at NYU Center for Data Science's MS in Data Science program. I hope it also serves as a sampler of topics that are covered in the program and a general guide to the building blocks that make up the knowledge of a data scientist without being too technical.

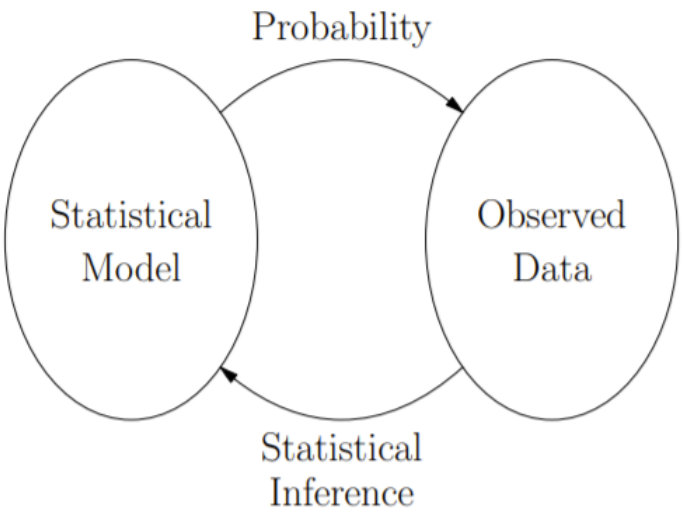
Probability and Statistics

Introduction

Throughout this notebook (and in data science generally), much of the theory is based on the fact that there is uncertainty in either the data, the model, or the solution. Afterall, if the problem was deterministic (A and B always lead to C), we would probably consider this more of a math problem than a data science problem.

Probability and statistics is important as a common language that we use to describe this uncertainty. They also help quantify a way of thinking that our brain is inherently bad at (for example, see this article (https://builtin.com/data-science/probability)).

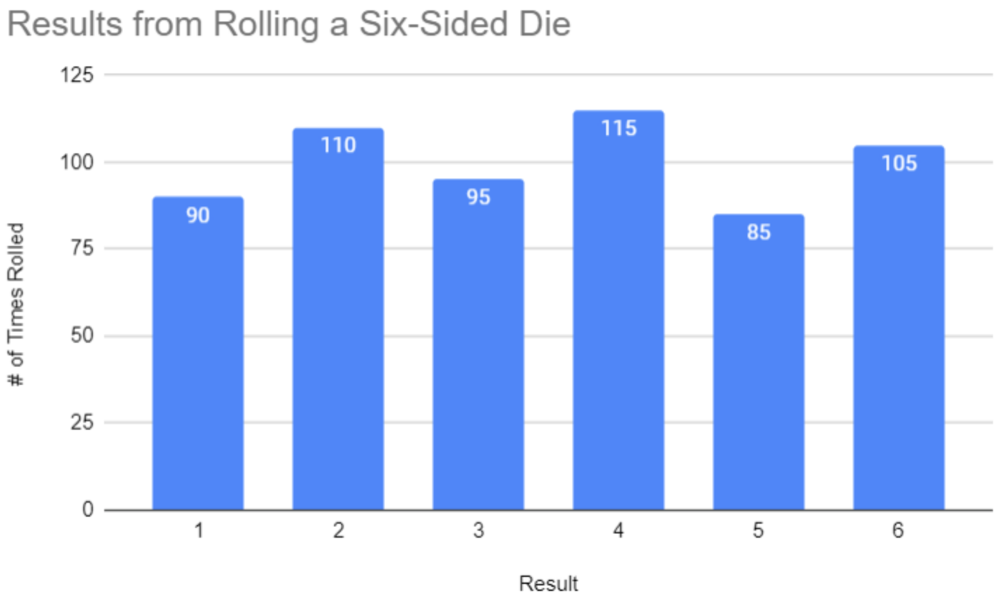
Let's talk about on a high level what these are and why we need both.



Credits to Brett Bernstein.

Staistics helps us estimate properties of a probability distribution given the data generated from it. What does this mean?

A simple example: we roll a standard six-sided die. If we roll it enough times, we get the below results.



Of course, we know that the die should have an equal probability of rolling any particular result. But in practice with any given data set, we often don't know what that distribution *should* be. If we didn't know that these results came from a standard six-sided die, what conclusions could we draw from these results? Would we still conclude that each number has an equal probability? How confident are we in the previous question? We aim to answer these questions via statistics, particularly via a statistical model that we hope estimates the true distribution accurately.

On the other hand, probability solves the reverse problem. Given a statistical model (i.e. a probability distribution), how do we calculate properties about the data? Particularly things such as probabilities of events of interests, or expectations (i.e. expected average over long run)? If we have a standard six-sided die, we can use probability theory to solve problems such as the probability of getting three 6s in a row or the standard deviation.

In data science, we can see why both are important. We use statistics to build a model from a data set, and we use probability to calculate useful information from that model. In the next couple sections, we go over key concepts from both in greater detail.

Key Concepts in Probability

To get started talking about probability, we need to define some basic terminology.

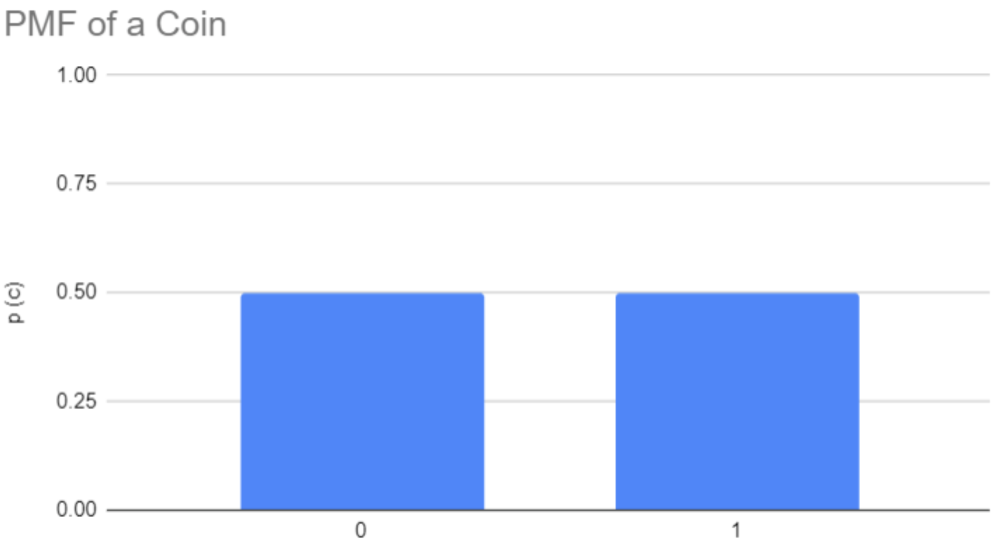
Random Variable : A variable that represents an uncertain quantity. Unlike a normal variable, it does not represent a fixed number, but a distribution.

It should map every possible outcome to a real number.

A simple example: let's name a random variable $\tilde{c}$. Given a coin, $\tilde{c} = 1$ if the coin lands on heads and $\tilde{c} = 0$ if the coin lands on tails.

Probability mass function / Probability density function : Random variables start to become very useful once we have tools to describe how they are distributed probabilistically. That's where probability mass functions and probability density functions come in.

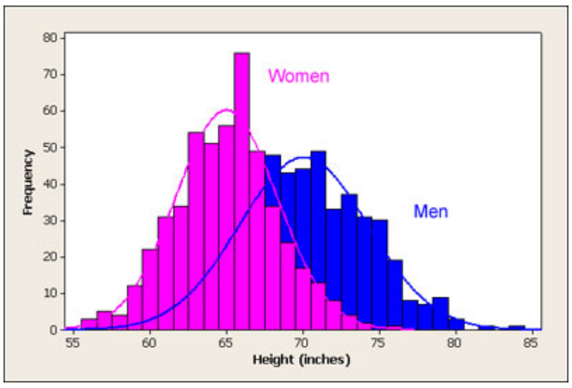
Let's take our random variable $\tilde{c}$ and describe how its outcomes are distributed.



A probability mass function (pmf) maps each and every outcome (tails --> 0 and heads --> 1) to a probability (50% for each side of the coin). We say that $\tilde{c}$ is distributed according to this pmf, which we can call $p_{\tilde{c}}$.

A probability density function (pdf) works in a similar way but for continuous random variables, which have an infinite number of outcomes. To find the probability of a set of events, we integrate to find the area under the pdf. Things like temperature, time, or distance are usually described by pdfs.

It's very useful to be familiar with the common distributions which help us describe the probabilistic behavior of a given phenomena. For example, the Gaussian distribution (also known as normal distribution or bell curve) can approximate many things well such as human height or the sum of outcomes from rolling two dice.



An example where we see the observations of male and female heights represented by the bars can be approximated well by the normal distribution.

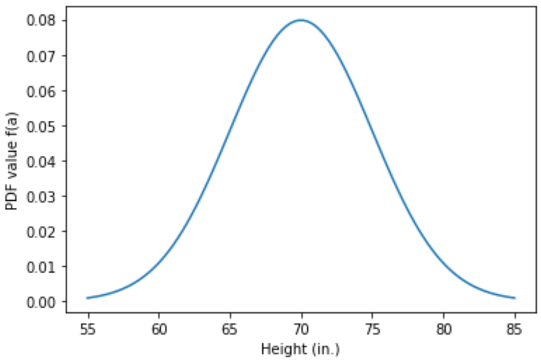
Mathematically speaking, the pmf or pdf will take the form of a function which takes an input (a given outcome) and returns the probability. For the Gaussian distribution, the pdf (for a Gaussian random variable we call $\tilde{a}$) looks like:

$$f_{\tilde{a}}(a) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(a-\mu)^2}{2\sigma^2}}$$

where μ is the average and σ is the standard deviation of the distribution.

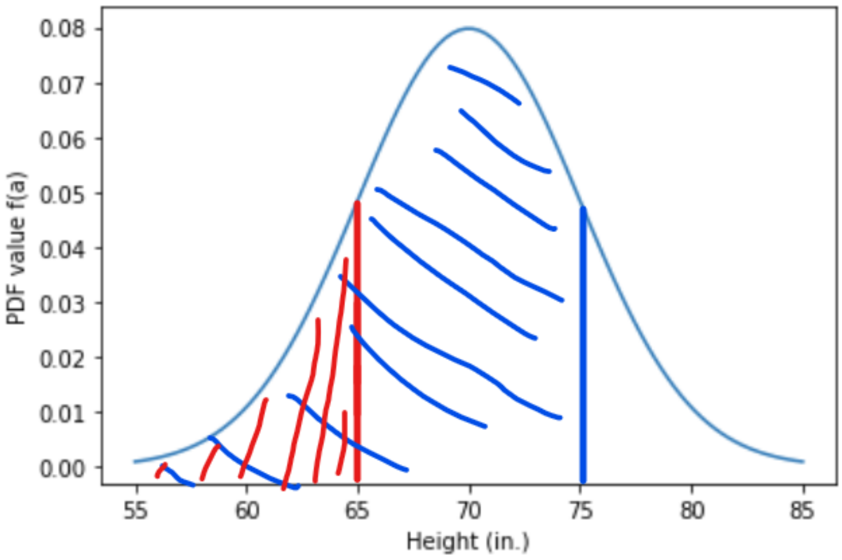
Let's take the blue line above and use a normal distribution to approximate it. Below, I recreate the graph (estimating average and standard deviation) with probability density ($f_{\tilde{a}}(a)$) on the y-axis.

In [11]:
```
# Python is the language of choice in the program, and for many data scientists.

import matplotlib.pyplot as plt
import numpy as np
import scipy.stats as stats
import math

mu = 70
sigma = 5
x = np.linspace(mu - 3*sigma, mu + 3*sigma, 100)
plt.plot(x, stats.norm.pdf(x, mu, sigma))
plt.xlabel("Height (in.)")
plt.ylabel("PDF value f(a)")
plt.show()
```



Using the graph above, we see that the probability density is the highest at 70 inches. However, if we want to find the probability of an event using this graph, we must select a range and integrate to find the area under it. For example, we could calculate that the probability that the height of a male is between 65 and 75 is 68.3% (see graph and calculations below).

Alternatively, we can use the **cumulative distribution function (cdf)**, which tells us the probability that $\tilde{a}$ is lower than a given height a.



In [15]:
```
probability_under_65 = stats.norm.cdf(65,mu,sigma)
probability_under_75 = stats.norm.cdf(75,mu,sigma)

probability_between_65_and_75 = probability_under_75 - probability_under_65
print(probability_between_65_and_75)
```

```
0.6826894921370859
```

With that, we have the basic probability tools we need to dive deeper. For readers unfamiliar with basic probability rules related to conditional probability and independence, I would recommend reviewing these as they are fairly straightforward rules related to computation and will not be restated here.

Expectation and Variance

Averaging is an important tool in mathematics, and it is no different with random variables. We can take averages of random variables by taking their expected values. This works slightly differently with discrete and continuous random variables.

For the discrete case, let's go back to our pmf of flipping a coin. For a discrete r.v. $\tilde{c}$ where it takes an outcome c, we multiply each possible outcome by the probability of the random variable taking that value. Generally speaking, the expectation or mean, $E(\tilde{c})$, is calculated:

$$E(\tilde{c}) = \sum_{c \in C} c p_{\tilde{c}}(c)$$

In the case of the coin flip (which is represented by a distribution called the Bernoulli distribution), our calculation is:

$$0 * p_{\tilde{c}}(0) + 1 * p_{\tilde{c}}(1) = .5$$

For the continuous case, since we have infinite outcomes, the calculation is slightly different. Given a continous r.v. $\tilde{b}$, our expectation is:

$$E(\tilde{b}) = \int_{b=-\infty}^{b=\infty} b f_{\tilde{b}}(b)$$

What other useful things can we do with expectation? How about understanding the typical variation around the mean we just calculated, aka variance?

In many practical cases, you have probably estimated variance by taking the average squared deviation from the sample mean. Expectation allows us to calculate the actual variance if we know how the r.v. is distributed. We can find the variance by calculating the following:

$$Var(\tilde{b}) = E([\tilde{b} - E(\tilde{b})]^2) = E(\tilde{b}^2) - E^2(\tilde{b})$$

We won't go too much into the details of how the variance of different distributions are derived here, but see below for a visual representation of the expectation and variance given some common distributions with the stated parameters.

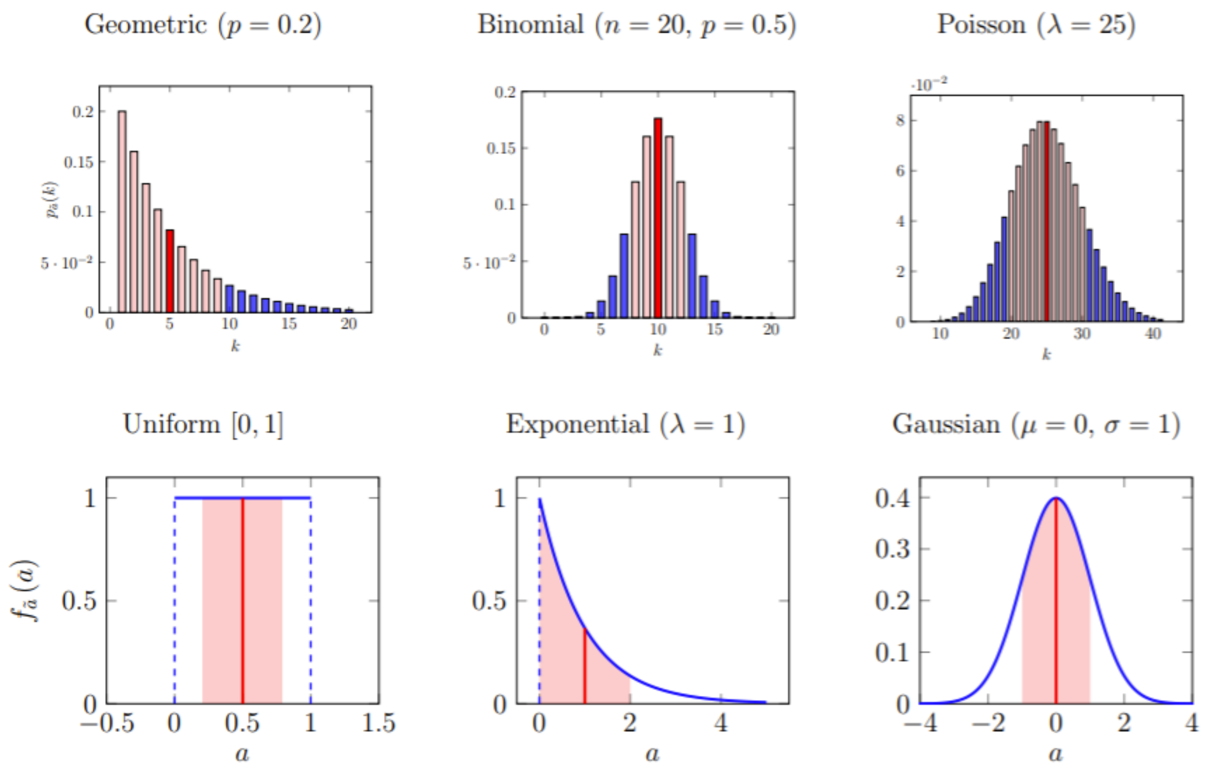


Figure 3: **Visualization of mean and standard deviation of parametric distributions.** The graphs show the pmfs of the discrete parametric models (top row) and the pdfs of the continuous models (bottom row). The mean of the random variable is marked in red. Values that are within one standard deviation of the mean are marked in pink.

Credits to Carlos Fernandez-Granda

Bonus Topics: Law of Large Numbers & Central Limit Theorem

The final (bonus) topics we consider within probability theory (though many of these ideas have overlap and use cases in other fields) is the law of large numbers and central limit theorem. You won't need these to follow along for the rest of the document, but are very useful results from probability.

Law of large numbers

Much of the concepts discussed above require either knowledge of the way data is distributed or at least an assumption about it. The (weak) law of large numbers is one of the first results discussed that helps us uncover information about the distribution from simply observing data. It states that given a sequence of independent and identically distributed (iid) random variables (a common simplifying assumption about observed data) $\tilde{a}_i$ for $i = 1, \ldots, n$, the sample mean of the sequence will converge to the true mean μ as n increases.

$$\lim_{n\to} \mathrm{E}\left(\left(\mu - \frac{1}{n}\sum_{i=1}^{n} \tilde{a}_i\right)^2\right) = 0.$$

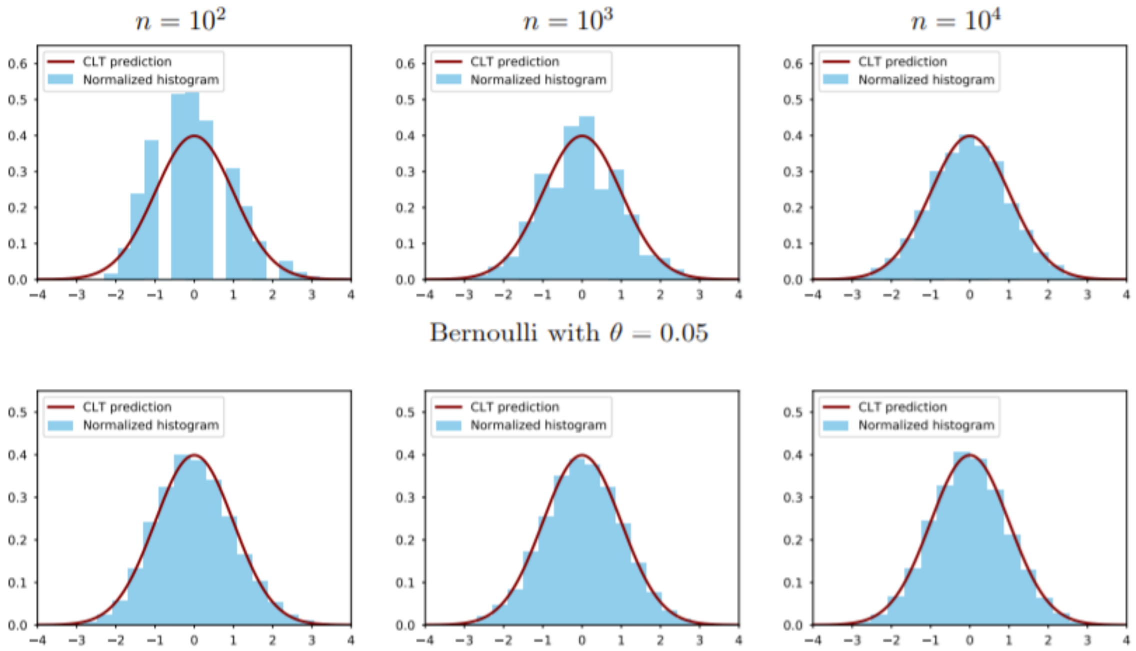
The most useful application of the law of large numbers is its justification of using empirical probabilities to estimate the value of a random variable taking a certain value. Going back to our example of the coin and the random variable $\tilde{c}$, let's say we wanted to check if it was truly a fair coin (50/50 probability). We flip the coin n times and receive the results $\tilde{c}_1, \tilde{c}_2, \ldots, \tilde{c}_n$. If we take enough samples, we expect the sample mean to be equal to the true probability that the coin lands on heads (recall that $\tilde{c} = 1$ for heads and $\tilde{c} = 0$ for tails).

Central Limit Theorem

Finally, we reach one of the most interesting results in probability - the central limit theorem. The law of large numbers establishes that the sample mean will converge to the real mean, but it does not provide information about its distribution. The central limit theorem tells us that the distribution of the sample means will converge to a Gaussian with specific properties as stated below:

Given a sequence of iid random variables $(\tilde{a}_1, \tilde{a}_2, \ldots, \tilde{a}_n)$ with mean μ and variance σ^2 The running average of the sequence $\frac{1}{n}\sum_{i=1}^{n} \tilde{a}_i$ converges to a Gaussian with mean μ and variance $\frac{\sigma^2}{n}$ in distribution, which means that its pdf is arbitrarily well approximated by the pdf of the Gaussian. In the case of discrete random variables, it means that the normalized histogram is arbitarily well approximated by the pdf of the Gaussian.

This gives us an idea of much the sample mean we calculate might deviate from the actual mean. Interestingly, $\tilde{a}$ can be distributed according to any distribution, but the distribution of the sample mean will converge to a Gaussian regardless. See below for examples where the samples are Bernoulli or Exponential.



Bernoulli with $\theta = 0.05$

Exponential with $\lambda = 2$

If you're interested in testing this our for yourself, try the simulation here: https://onlinestatbook.com/stat_sim/sampling_dist/index.html (https://onlinestatbook.com/stat_sim/sampling_dist/index.html). Alter the first graph representing the parent population however you want, then take samples of that distribution using the second graph. You will see that the third graph will resemble a Gaussian no matter the original distribution with enough samples.

Key Concepts in Statistics

We move onto statistics next, where many of the topics are very practically relevant. As data scientists, we are generally limited to finding insights from our collected data, which represents some subset of the actual population. We need to understand the limitations of working from a sample, what conclusions we can draw from it, and how confident we can be in those conclusions. Statistics helps us do this. For those job searching, these are also prime topics for interview questions!

We start with some terminology once again:

Statistical Model : A collection of random variables of interest, and a family of possible joint distributions for those random variables. A model is called *parametric* if there is a finite list of numbers, called **parameters**, that determine which joint distribution from the family is used.

For example, consider an additive normal error model $\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n$ where each r.v. is independent and normally distributed with mean μ and variance σ^2. This is used to model the variability in n independent samples of a quantity with true value μ. This is a very common base model with various possible adjustments added on such as bias (if μ cannot be measured accurately in practice).

Note that the discussion on properties of and techniques related to models in this section can apply both to simple models and more complex models that are commonly thought of when data science comes to mind (with varying levels of usefulness depending on the model setup).

Statistic : A quantity (which is random because the data is also random, such as in the additive normal error model) computed from the observable data of a statistical model. If r is a function, then $T := r(\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n)$ is a statistic. The distribution of T is called the **sampling distribution**. The sample mean (often written as $\bar{X}$) is an example of a statistic, and the central limit theorem (discussed above) gives us information about its sampling distribution.

That's enough for now to get us started on some useful statistics tools!

Confidence Intervals

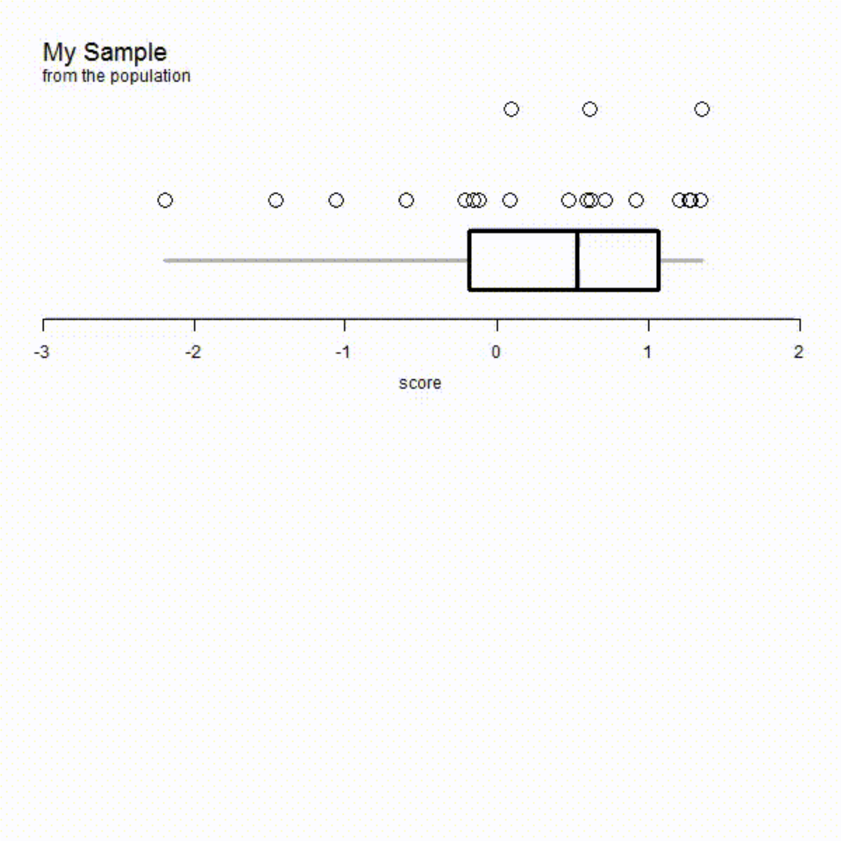
When we calculate a statistic, we often do so to estimate a quantity associated with the statistical model. Confidence intervals are a method by which we can express the probability of the quantity (which we can call θ) falling within a certain range.

If L, U are statistics, and $P(L \leq \theta \leq U) = \alpha$, then $[L, U]$ is a level α confidence interval for θ.

A common example used in problem sets is estimating the 95% confidence interval for the sample mean. Let's look at this from the perspective of the additive normal error model we descrived earlier. Note that the sample mean (being a sum of normal r.v. divided by n) is distributed with mean μ and variance $\frac{\sigma^2}{n}$ (Note that $Var(\bar{X}) = Var(\frac{\tilde{x}_1+\tilde{x}_2+\ldots+\tilde{x}_n}{n}) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$). With some knowledge about the normal distribution CDF, we know that we can find roughly 95% of the probability density within 2 standard deviations of the mean. So our final answer is $[\bar{X} - 2\frac{\sigma}{\sqrt{n}}, \bar{X} + 2\frac{\sigma}{\sqrt{n}}]$.

One very important note on the interpretation of confidence intervals once calculated - it does NOT represent the probability that μ falls in that interval. This might be confusing, but keep in mind that we view our sample as random, and our confidence interval can change from one experiement to the next depending on our data. The correct framing is to say: if we followed this procedure for creating confidence intervals for many experiments, 95% of the intervals calculated would contain the true μ.

As a final final note, bootstrapping is a very common and convenient method for generating confidence intervals. There are several ways to approach this which we will not into detail on, but essentially, we can resample our data to simulate data from new experiments. The distribution of the statistic of interest from these simulations can allow us to easily calculate confidence intervals (e.g. See below GIF. In 100 bootstrap simulations, 95% of the sample means calculated roughly fall within -.1 and .8. Therefore, this is our 95% confidence interval).



Source: https://www.stat.auckland.ac.nz/~wild/BootAnim/ (https://www.stat.auckland.ac.nz/~wild/BootAnim/)

Maximum Likelihood Estimation

We have some idea of what models are, how to calculate statistics associated with them, and how confident we can be in those statistics. But how do we define the actual model? Generally speaking, the practitioner will choose the family of functions that best models the phenomena, but the specific parameters will need to be estimated somehow.

Let's take a very simple example using our additive normal error model. We want to use it to simulate male height, similar to the examples mentioned earlier. We believe a normal distribution can approximate the true distribution effectively as most peoples' heights $(\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n)$ will converge around a certain number μ with some variance σ^2. But how can we actually guess what μ and σ^2 are?

Cue **maximum likelihood estimation** - an extremely common and intuitive method to estimate those parameters. At a high level, it states that we set the parameters according to what maximizes the probability (or density for continuous variables) of the observations. In our example, if we see 10 people with a height of exactly 70 inches, then our best guess at μ and σ^2 are 70 and 0, respectively.

We can be more mathetically precise in our definition:

$\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n$ are distributed to some joint distribution with pdf or pmf f, where f is determined by parameter values $\theta_1, \theta_2, \ldots, \theta_p$. The **likelihood function** is given by:

$$L(x_1, x_2, \ldots, x_n; \theta_1, \theta_2, \ldots, \theta_p) = f_{\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n}(x_1, x_2, \ldots, x_n; \theta_1, \theta_2, \ldots, \theta_p)$$

Note that $x_1, x_2, \ldots, x_n$ represent actual observed values. $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_p$ are **maximum likelihood estimators** (or MLEs) for $\theta_1, \theta_2, \ldots, \theta_p$ at the observed values $x_1, x_2, \ldots, x_n$ if they maximize L.

Hypothesis Testing

We will broach one more substantial topic in statistics - hypothesis testing. Hypothesis testing is a statistical procedure for choosing between alternative scenarios (i.e. parameters) for a model. There is a ton of terminology associated specifically with testing which we will carefully introduce with examples.

Let's start by elaborating a bit more on the purpose. We go back to our example of the coin flip.

definitions/terminology

hypothesis testing

bayesian inference (touched on more in machine learning) misc.

simple linear models (touched on more in linear algebra)

```
linear algebra

makes everything makes sense via math

if you interpret operations with geometry, things have a very elegant explanation

also needed to understand the language of data science literature
```

In [ ]:

```
3 blue 1 brown

vectors/matrices
    kernel/image/rank

inner product

eigenvalues/eigenvectors and markov chains

spectral theorem/pca/sva

convexity/least squares (linear model)

gradient descent

misc
clustering
```

In [ ]:

```
data science/ML concepts

bias variance
```