The Epic Guide to Probability Distributions

A random variable allows us **to model some random event**, for which we'd use a probability distribution to define how it behaves. In this guide, we explore the well-known **classes of distributions** that statisticians often use for a variety of practical purposes.

There are 3 different things you should always consider when using a distribution or making your own:

1 **The Outcomes:** (known as the 'support vector' or the 'sample space') When we draw a sample from a random variable with this distribution, what outcomes can we get? e.g. The result of a single throw of a dice, the number of heads we see in 10 throws, the number of blue cars that pass in the next hour.

2 **The Parameters**: These are variables we can tweak that affect the probability of each outcome for a particular type of distribution. e.g. The total number of times the coin was thrown, the average rate at which blue cars pass.

3 **The Probability Mass/Density Function (the PMF/PDF):** A fancy way of saying 'what's the probability of a particular outcome?' These use the parameters in some expression to yield a value between 0 and 1. We have a PMF if our outcomes are discrete (e.g. the throw of a die), and PDF if it's continuous (e.g. height).

Bernoulli Distribution

Sounds complicated, but it just models **a single trial of an event with two outcomes**, e.g. the flip of coin! Each trial is known as a '**Bernoulli Trial**'.

1) **Outcomes**: 0 to represent failure (e.g. tails), 1 to represent success (e.g. heads).
2) **Parameters**: A value p for the probability of success (e.g. we can control how unfair the coin is)
3) **PMF**: $p(0) = 1-p$, $p(1) = p$

Generalise to k outcomes

Generalise to n trials

Multivariate Distribution

Represents **a single trial of an event with multiple outcomes**, e.g. the throw of a die.

1) **Outcomes:** Whatever we like, e.g. 1 to 6 for a die. Let's call these x_1 to x_k .
2) **Parameters:** A value p_i for the probability of each outcome.
3) **PMF:** $p(x_i) = p_i$

Binomial Distribution

Models **multiple trials of an event with two outcomes.** e.g. If I toss a coin *n* times (i.e. n trials), what's the probability of getting 1 head, 2 heads, etc.

1) **Outcomes**: An integer in the range 0 to n, representing the number of 'successes' (e.g. heads).
2) **Parameters**: Again, p for the probability of success on each trial, but we also need the number of trials *n*.
3) **PMF**:

$$p(x) = \binom{n}{x}(1-p)^x p^{n-x}$$

Don't panic! Suppose n=4 and we're interested in the probability of x=3 heads with an unfair coin. Then for a given sequence of 4 throws in which 3 heads appear, there's a p x p x p x (1-p) chance of it occurring. But there's multiple ways of getting 3 heads, for example HHHT, or THHH. The term at the front (known as the '**binomial coefficient**' therefore tells us the number of ways of getting x heads in n throws.

Generalise to n trials

Geometric Distribution

This is a fun one! This distribution models **the number of failures before we have a success.**

Imagine trying to vanquish a foe in battle, and that we have a p chance of being successful in each shot. Then we might interested in the chance of one shot being required before being successful, or two shots, and so on.

1) **Outcomes:** Between 0 (i.e. being immediately successful) and infinity (we may be very unfortunate!).
2) **Parameters:** Like the Bernoulli Distribution, a value p for the probability of success.
3) **PMF:** $p(x) = (1-p)^{x-1}\, p$

Example application: The 'Coupon Collector's Problem' says: Given say 200 possible Pokemon cards (and assuming each is equally 'rare'), how many cards would we expect to buy before collecting all of them? We can solve this by stringing together a bunch of geometric distributions to represent successfully getting each card. (The final answer is 1176 by the way!)

Multinomial Distribution

Similarly to above, this models **multiple trials of an event with multiple outcomes**.
For example, what's the chance of getting two 1s, three 2s, one 1 and no 4s, 5s, or 6s if we throw an unfair die five times?

Related: Dirichlet Distribution

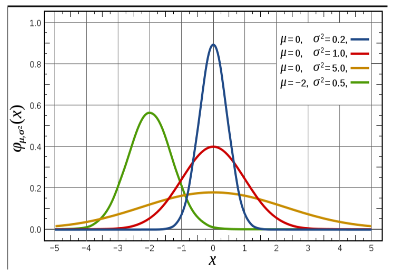
The conceptual opposite. Given observed counts, what is the probability of each event in a single trial? For example, after 12 throws of a die, we observe two 1, two 2s and so on. What is the probability it is a fair die? It's most likely to be fair, but equally we might have had a slightly biased dice. Useful in a field known as 'Bayesian statistics'.

$\lambda = np,\ n \to \infty$

Gaussian Distribution

The big daddy of them all. Perfect **when our data has a bell-shaped curve about some mean**, e.g. IQ scores, heights of people in the class, inaccurate sensor measurements.

Parameterised by a mean μ and a standard deviation σ.



Arises due to the **Central Limit Theorem**. To see it in action, try adding the scores from 5 throws of a die, repeat, and plot total scores with frequency. What shape do you get?

χ^2 Distribution

Useful for answering questions like:

- Given my population is 50% male and 50% female, and someone provides me with a sample of 50 people in which 23 are male and 27 are female, how confident am I that their sample is reliable? (i.e. **goodness of fit**)
- If I provide a sample of boys and girls with English and Maths tests, how confident am I based on this sample that gender affects performance at these tests? (known as using a **contingency table**)

Poisson Distribution

Fish jokes have no plaice here! Models **the number of events that occur in a given time, given a rate λ at which they occur.**

For example, what's the probability of 100 cars passing your house in the next hour given you typically see 50 on average?

1) **Outcomes**: An integer in the range 0 (no events occur) to infinity.
2) **Parameters**: A rate λ at which events occur.
3) **PMF**:

$$p(k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Related to the Binonomial Distribution via **The Law of Raw Events**—imagine each tiny slither of time as a Bernoulli trial in which our event may or may not occur.

Related: Exponential Distribution

Given a volcano erupts on average 5 times a millennium, what's the probability it'll erupt within the next 100 years?

Student's t-Distribution

Imagine we picked 50 people at random and took the mean of their heights. If we'd repeated this we would have likely ended up with a slightly different mean each time. We can use a Gaussian distribution to model the uncertainty of the mean of our sample (known as the **sampling distribution**). But had we used a small sample (under 30 people), we'd typically use the Student's t-distribution instead.

DrFrostMaths.com