**Generalised Linear Models**

*NB. See notebooks for details on how to code these*

Key features

* These are applications of linear algebra (ie. Linear regression) that are tailored to more accurately mirror data from the real world.
* Linear regression assumes a normally distributed dataset, however in reality this occurs relatively infrequently in the real world, and is also often inappropriate.
* Generalised linear models apply different distributions to linear regression to reflect the distributions of the data in the real world.
* Once these distributions have been applied, a link function is then required in order to convert these adjusted datasets back into a suitable linear form.

A generalized linear model consists of 3 parts:

1. An exponential family probability distribution
2. A linear predictor
3. A link function

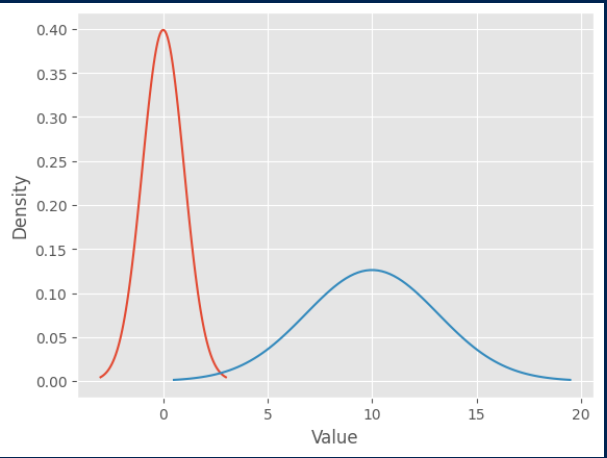
The \*\*exponential family probability distribution\*\*, is the probability distribution that our response variable follows, for instance, Normal, Binomial, Poisson, and Negative Binomial.

The \*\*linear predictor\*\* is a linear combination of a set of coefficients and explanatory variables used to predict our response variable. This is essentially the same form used to specify our linear model, which means we can use the same formula set up for our GLM.

The \*\*link function\*\* allows us to use the linear predictor by providing the link to connect the exponential distribution to the linear predictor. There will often be defaults for the link when you call the model family, but you can also specify this yourself.

For instance, the Poisson distribution is a discrete distribution where values can take on integers between 0 and infinity. However, the linear predictor is based on the normal distribution that is continuous and can go from -infinity to infinity. It is therefore common to use a log link to log the response variable, which effectively limits the normal distribution to 0 to infinity, thereby making it suitable for the Poisson distribution. This means that we often need to do a back transformation when we interpret the parameter coefficients in the actual scale. For example, when using the log-link we would exponentiate the coefficients.

**Normal (Gaussian) Distribution**



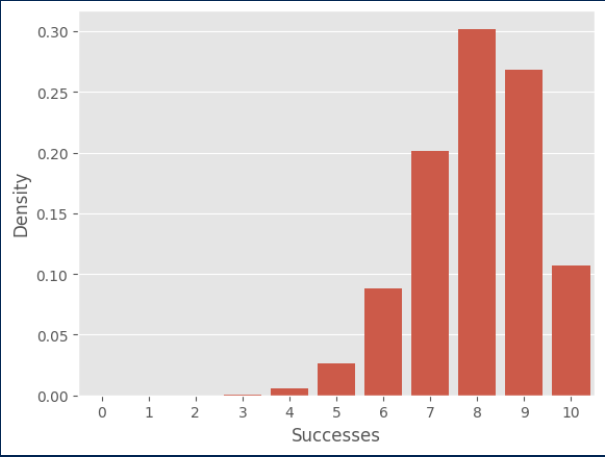
Key features

* Mean, median and mode are all equal
* Unimodal – only has one mode
* Only requires the mean and the variance to be specified
* Symmetrical distribution
* Applicable to continuous data

Common applications

* Finance and Economics: In finance, the normal distribution is used to model stock market returns, interest rates, and risk assessments. It helps in portfolio optimization and in the pricing of options and other financial derivatives.
* Quality Control and Manufacturing: Manufacturers use the normal distribution to understand variability in product dimensions, weights, and other measurable qualities. This helps in maintaining quality control, reducing defects, and optimizing production processes.
* Psychology and Education: Test scores, like IQ tests or standardized academic tests (SAT, ACT), often follow a normal distribution. This helps in understanding and interpreting the performance of individuals relative to the average.
* Healthcare and Medicine: Many physiological and biological measurements (like blood pressure, cholesterol levels, etc.) follow a normal distribution in a healthy population. This assists in diagnosing abnormalities and in drug development.
* Environmental Science: Various environmental data, such as rainfall amounts and temperature, often exhibit normal distribution, aiding in climate modeling and environmental risk assessments.
* Social Sciences: Data related to human behavior and societal trends, like income distribution in a population (though this can sometimes follow other distributions), often follow a normal pattern.
* Anthropometry: Body measurements such as height and weight in a specific population usually follow a normal distribution, which is useful in fields like ergonomics, clothing design, and nutrition.
* Astronomy: Measurement errors and observational errors in astronomy often follow a normal distribution, aiding in data analysis and interpretation.
* Sports Analysis: Performance metrics in sports, like scores or timings, can sometimes be modeled using normal distributions, helping in player evaluation and strategy development.

**Binomial Distribution**



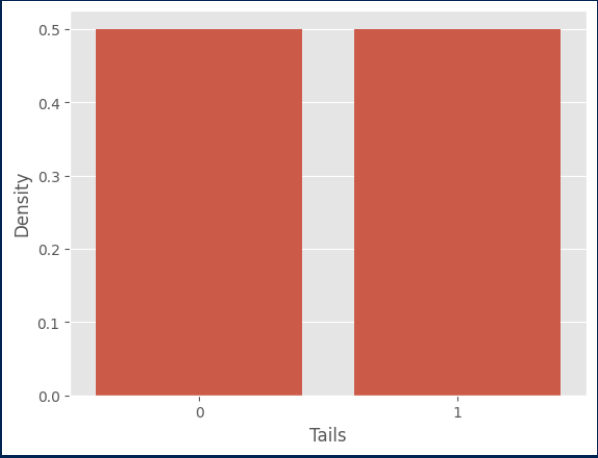
Key features

* Applies where the outcome of a trial can only have two outcomes (eg. head/tail, alive/dead, present/absent, buyer/non-buyer)
* Applicable to discrete data
* Where the number of outcomes is greater than 2, e.g. (single, married, divorced), then we can use a multinomial distribution which allows for more than 2 possibilities. We can get our clue for the number of outcomes from the name (bi = 2, multi = multiple).
* Applications leverage the clear-cut nature of binomial outcomes (success/failure, yes/no, win/lose) and the ability to model scenarios with a fixed number of trials and a constant probability of success in each trial.

Common applications

* The number of individuals with malaria out of the number of individuals in the population (e.g. an individual can have malaria (1) or not (0))
* The number of customers who bought item A out of total number of shop customers (a customer can have bought item A (1) or not (0))
* Quality Control and Manufacturing: In industrial settings, the binomial distribution is used to model the number of defective items in a batch of products. This helps in assessing and controlling the quality of products.
* Healthcare and Medicine: It's used to estimate the success rate of a medical treatment or surgery, or the likelihood of a particular side effect occurring in a group of patients.
* Survey Sampling and Market Research: When analyzing survey results, the binomial distribution can model the number of people who might respond positively to a question out of a sample size.
* Genetics and Biology: In genetics, it helps in determining the probability of inheriting certain genetic traits, based on known probabilities of genetic variation.
* Sports Analysis: In sports, the binomial distribution can predict the likelihood of a team winning a certain number of games out of a series, assuming the probability of winning each game is known.
* Finance and Risk Management: For investments with two possible outcomes (like default/no default, gain/loss), the binomial distribution can be used to model the probability of different outcomes.
* Educational Testing and Evaluation: In education, it can be used to model the number of questions a student might answer correctly in a multiple-choice test, assuming a certain probability of guessing correctly.
* Environmental Studies: It's applicable in ecological studies, such as estimating the probability of a certain number of animals in a population having a specific characteristic.
* Political Science: For analyzing voting behavior, the binomial distribution can model the likelihood of a certain number of votes for a candidate given a known voting preference ratio.
* Information Technology: In computer science, it's used in algorithms for data transmission, modeling the number of successful data packets transmitted over a network.

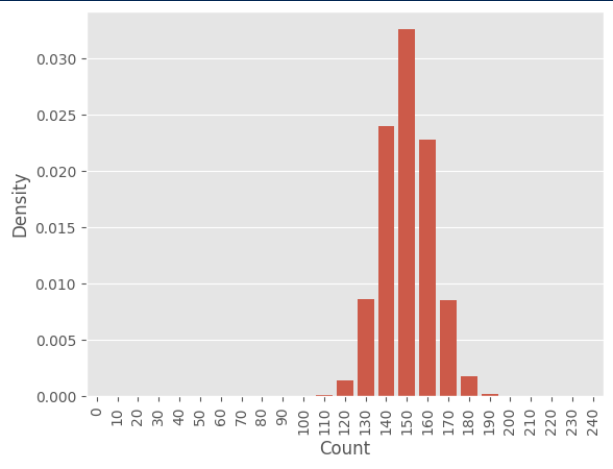
**Bernoulli Distribution**



Key Features

* This is a special case of the binomial distribution, where there is only one trial

**Poisson Distribution**



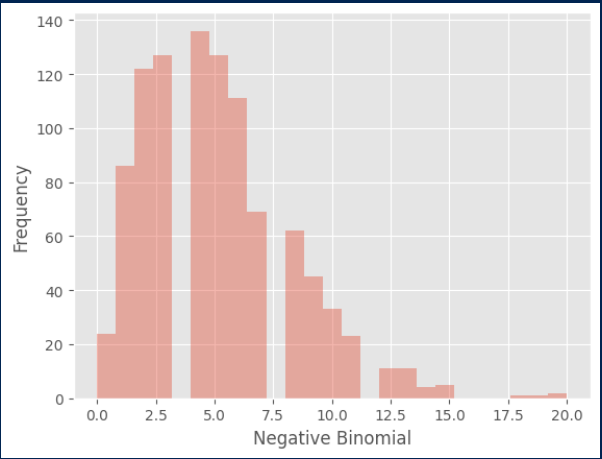
Key features

* Gives the distribution of the number of individuals, arrivals, events, counts, etc., in a given time/space/unit of counting effort.
* This is the common distribution to use whenever things are counted and when they are discrete (e.g. you cannot have 1.5 of a person).
* Only need to specify the mean, as the mean and the variance are equal to each other.
* We are constrained by the fact that the mean and variance are the same. This means that more counts, and as a result larger means allow for greater variation, whereas fewer counts with a smaller means allow for less variation.
* If we need to allow for more variation (i.e. wider distribution of counts) we can look to the negative binomial

Common applications

* Traffic Flow and Queuing Theory: The Poisson distribution is used to model the number of cars passing through a checkpoint in a given time period or the number of phone calls received by a call center. It helps in designing roads, managing traffic, and optimizing customer service resources.
* Retail and Sales Forecasting: Businesses use it to predict the number of customers who might visit a store in a given time period, or the number of sales transactions that might occur. This assists in staffing and inventory management.
* Biology and Medicine: It can model the number of mutations in a given stretch of DNA, the number of bacteria in a certain volume of liquid, or the incidence of a particular disease in a population over time.
* Insurance and Risk Management: The Poisson distribution is used in insurance to model the number of claims or losses that might occur within a given period. This is crucial for risk assessment and policy pricing.
* Environmental Science: It can model rare events, like the number of earthquakes in a region over a period, or the number of times a certain species is sighted in a wildlife survey.
* Sports: In some sports, the Poisson distribution can model the number of goals or points scored in matches, aiding in prediction and analysis.
* Astronomy: The distribution is used to model the number of stars in a given volume of space or the number of meteorites of a certain size that hit the Earth over time.
* Industrial Engineering and Failure Analysis: It can predict the number of system failures or breakdowns in a factory or industrial process over a given time period, helping in maintenance planning.
* Public Services: For example, it can predict the number of emergencies, like fires or accidents, that a city's public service department might have to respond to in a given time frame.

**Negative Binomial**



Key features

* It is discrete, like the Poisson distribution, however it’s overdispersed (variance is larger than its mean).
* It is often followed when the data largely follows a Poisson distribution, but there are a large number of zeros in the data, which create an increased variance
* The negative binomial is actually a generalisation of the Poisson distribution, where the mean is dealt with separately to the variance.
* You need to be able to specify the mean and the variance
* Useful for modelling the number of trials required to observe a certain outcome

Common applications

* Epidemiology and Public Health: In epidemiology, the negative binomial distribution is used to model the number of disease cases needed to occur before a certain number of them are resolved or result in another outcome (like hospitalization).
* Quality Control: It can model the number of units that need to be sampled or tested before finding a predetermined number of defective items, especially when the defects are rare and randomly distributed.
* Genetics and Biology: In ecological and biological studies, this distribution can model the number of trials needed to observe a certain number of rare mutations or to capture a specific number of a rare species.
* Insurance and Risk Analysis: The negative binomial distribution is used in actuarial science to model the number of claims or losses that might occur before reaching a certain threshold, especially in cases where claims are overdispersed.
* Sports Analysis: For certain types of sports data, like the number of at-bats before a certain number of hits in baseball, the negative binomial distribution can provide a good model.
* Manufacturing and Process Control: It's used to model the number of items produced before achieving a set number of failures, particularly in processes where the probability of failure is not constant.
* Traffic Flow and Transportation: This distribution can be applied to model the number of vehicles passing a point before observing a certain number of a specific type of vehicle (e.g., trucks).
* Retail and Business Analytics: In customer purchase behavior analysis, it might model the number of customer visits before a certain number of purchases are made.
* Environmental Studies: For instance, in rainfall modeling, it can model the number of days before a certain amount of rainfall is accumulated.
* Telecommunications: In network traffic, it can model the number of data packets transmitted before a certain number of packets are dropped or lost.