Introduction to Statistics for Astronomers and Physicists

Section 2c: Probability Distributions

Dr Angus H Wright

Updated 2022-02-09

Section 2: Introduction

Section 2: Probability & Decision Making (Weeks 3-5)

For all aspects of modern science, an understanding of probability is required. We cover a range of topics in probability, from decision theory and the fundamentals of probability theory, to standard probabilistic distributions and their origin. From this module, students will gain an insight into different statistical distributions that govern modern observational sciences, the interpretation of these distributions, and how one accurately models distributions of data in an unbiased manner.

Topics include:

- Decision theory
- Fundamentals of probability
- Statistical distributions and their origins

Probabilistic Distributions

Todays lecture is going to be all about understanding some useful probabilistic distributions. This list is not exhaustive, there are other standard distributions not listed here that you may want to investigate yourself, or come across in your own career. Nonetheless, the distributions that we will discuss today are all useful for understanding various problems in physics, astronomy, and statistics.

Discrete Distributions

- Bernoulli Random Variables
- Geometric Distribution
- Binomial Distribution
- Poisson Distribution

Continuous Distributions

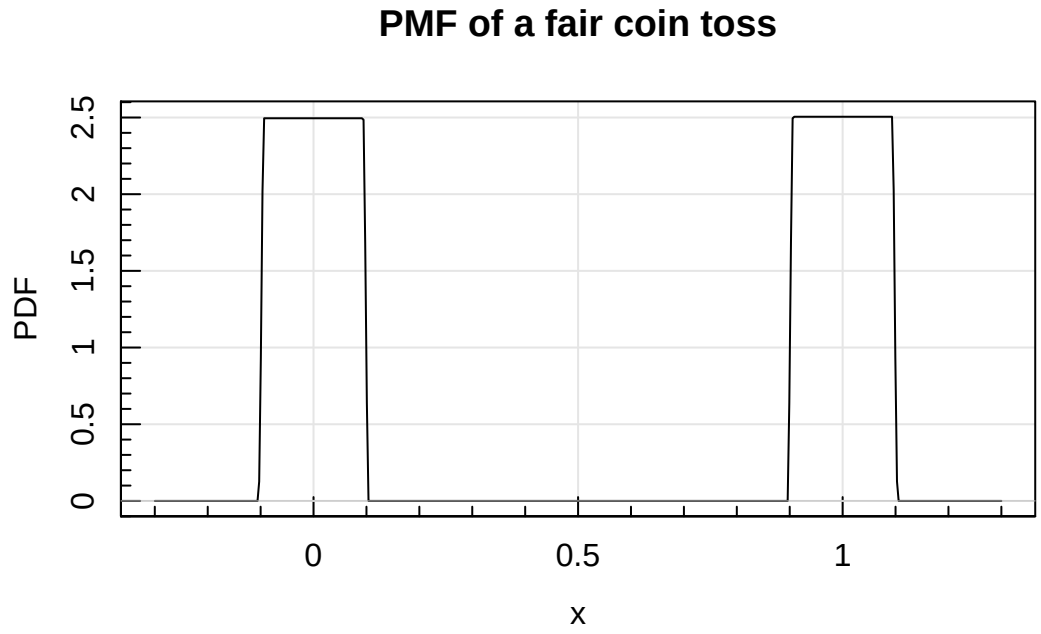
- Uniform Distribution
- Gaussian Distribution
- Gamma Distribution
- Beta Distribution
- Student t-Distribution
- χ^2-Distribution
- Pareto Distribution

The PMF, PDF, CDF, Quantile Function

Before we begin our exploration of statistical distributions, though, we first need to recall some information from previous lectures, and learn two new concepts.

The Probability Mass and Density Functions (PMF/PDF)

We've discussed probability mass and density functions previously in the course already, even if we haven't described them as such. The probability mass function (or PMF) is defined as the probability of observing a discrete value x given an arbitrary sample space Ω. Going back to our initial definition of probability, this can be considered to be (in the limit of large n), the relative frequency of observing the value x in the set of all outcomes.



For non-discrete distributions, the probability of observing any particular single value is always 0 (because there are an infinite number of possible values in any finite interval). As such a probability mass function is non-sensical for continuous data. Instead, the probability distribution of continous data is defined as the relative probability of finding a value x *over some finite range*. This is the **probability density function** or PDF.

The formal definition of the PDF is therefore:

$$P(a \leq X \leq b) = \int_a^b p(x)\mathrm{d}x$$

In practice we use infinitesimally small intervals to define the PDF.

Some important properties of the PDF:

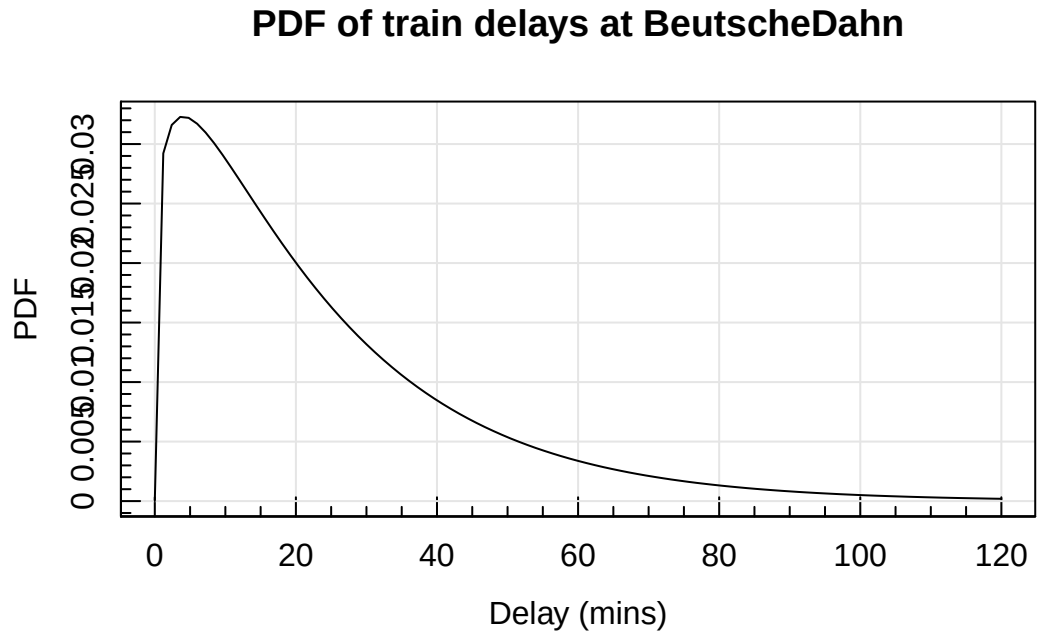
- The PDF is non-negative, because the $P(u) < 0$ would require

$$\int_{u-\epsilon}^{u+\epsilon} p(x)\mathrm{d}x < 0$$

  , which is impossible.
- The PDF must integrate to 1, which we can see if we make the interval $[a, b] = [-\infty, \infty]$:

$$\begin{aligned} P(-\infty \leq X \leq \infty) &= \int_{-\infty}^{\infty} p(x)\mathrm{d}x \\ &= 1 \end{aligned}$$

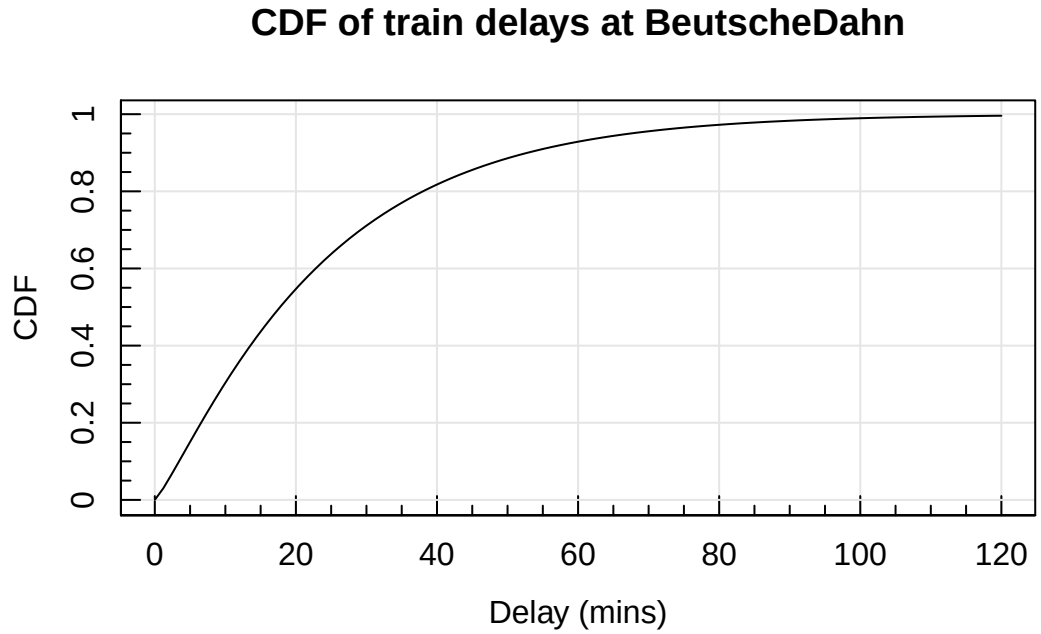
For example the PDF of train delays at a fictitious railway company that is totally fabricated and does not exist in Germany (or any other country) is:



So while in reality the probability of having two delays that are *exactly* the same is ~ 0, the probability of seeing a a delay within some interval is a well defined function. For example, the probability of seeing a train arrive with no delay is $P(X = 0) = 0$.

Cumulative Distribution Function (CDF)

We've discussed the CDF somewhat at length previously. For our train delays, the CDF is:

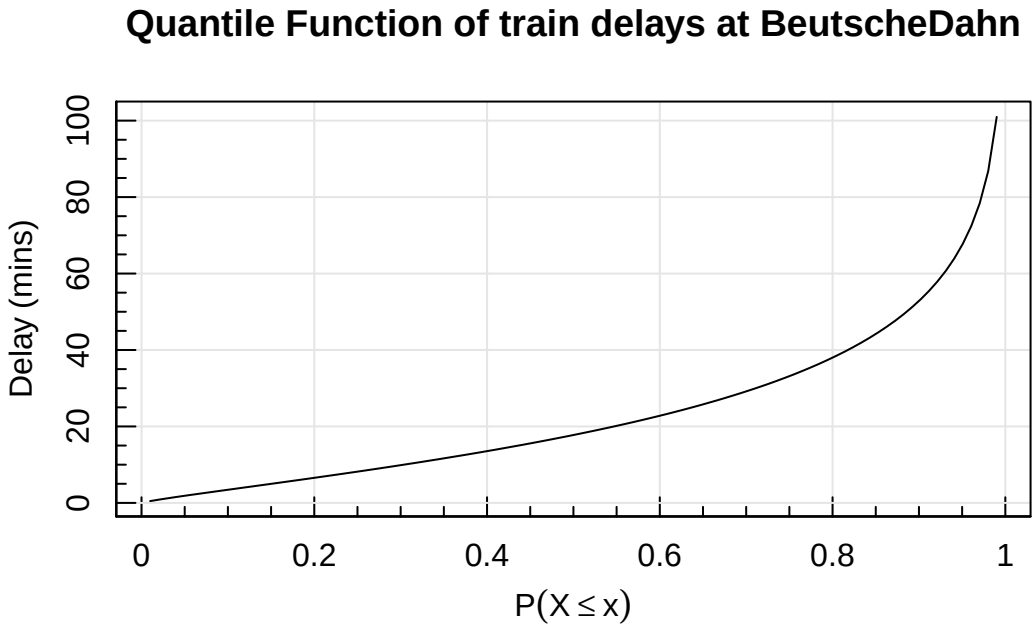


The CDF is extremely useful for measuring the fraction of probability mass that is being below or equal to some value X. In the above case, the probability that our train arrives with less than a 20 minute delay is $\sim 50\%$.

Thanks BeutscheDahn!

The Quantile Function

Again, we've looked at this somewhat at length. For analytic (invertable) functions, the quantile function is the inverse of the cumulative distribution function.



Expectation and Expected Values

In the first section of this course, we presented formulae for the mean, standard deviation, variance, and covariance of a sample of observations. These are *descriptive statistics*, in that they describe a dataset.

We can use these same statistics to describe the properties of a *probability distribution* from which we draw random variables. These are **expectations**.

The **expectation value** of a discrete variable X, with sample space Ω, and probability distribution P is defined as:

$$\mathbb{E}[X] = \sum_{x \in \otimes} xP(X = x)$$

For continuous data, we simply reformulate this sum as an integral:

$$\mathbb{E}[X] = \int_{x \in \otimes} xp(x)\mathrm{d}x$$

Let's take an example of a fair dice roll, and compute the expectation value of each roll. Assume a fair six-sided die ($p_i = 1/6 \ \forall \ i \in [1, 6]$).

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{x \in \otimes} xP(X = x) \\
&= 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + \cdots + 6 \times \frac{1}{6} \\
&= \frac{1}{6}(1 + 2 + \cdots + 6) \\
&= \frac{1}{n} \sum_{i=1}^{n} x_i \\
&= \bar{x}
\end{aligned}$$

So the expectation value is just the arithmetic mean of the probability distribution!

We can then also calculate other statistics, like the variance:

$$\begin{aligned}\text{var}[X] &= \mathbb{E}[(X - \mathbb{E}[X])^2] \\ &= \mathbb{E}\left[X^2 - 2X\mathbb{E}[X] + (\mathbb{E}[X])^2\right] \\ &= \mathbb{E}\left[X^2\right] - 2\mathbb{E}[X]\mathbb{E}[X] + (\mathbb{E}[X])^2 \\ &= \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2\end{aligned}$$

Now:

$$\begin{aligned}\text{var}[X] &= \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2\end{aligned}$$

Or the covariance:

$$\begin{aligned}\text{cov}[X, Y] &= \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right] \\ &= \mathbb{E}\left[XY + X\mathbb{E}[Y] - \mathbb{E}[X]Y + \mathbb{E}[X]\mathbb{E}[Y]\right] \\ &= \mathbb{E}\left[XY\right] - 2\mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y] \\ &= \mathbb{E}\left[XY\right] - \mathbb{E}[X]\mathbb{E}[Y]\end{aligned}$$

Recall once again that, if X and Y are independent, $\mathbb{E}\left[XY\right] = \mathbb{E}[X]\mathbb{E}[Y]$, and so $\text{cov}[X, Y] = 0$.

These formulae can be used to compute the mean and variance of arbitrary probability distributions. We will start looking at examples of these now.

The Bernoulli Random Variable

We started our discussion of probability and statistics with a simple experiment: a single toss of a coin. When we have an experiment whose outcome is restricted to one of two outcomes ("success/failure", "yes/no") the variable X is said to be a Bernoulli random variable.

The Bernoulli random variable is defined as taking the value 1 with a probability p, and the value 0 with a probability $1 - p$. Therefore the PMF of a Bernoulli Random Variable is:

$$X \sim \text{Bern}(p)$$

$$P(X = x) = \begin{cases} 1 - p & \text{if } x = 0 \\ p & \text{if } x = 1 \end{cases}$$

The related cumulative distribution function is therefore:

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - p & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x \geq 1 \end{cases}$$

The Bernoulli Random Variable has the following useful properties:

- Expectation: $\mathbb{E}[X] = p$

$$\begin{aligned}\mathbb{E}[X] &= \sum_{x \in \otimes} xP(X = x) \\ &= 0 \times (1 - p) + 1 \times p \\ &= p\end{aligned}$$

- Variance: $\text{var}[X] = p(1 - p)$.

$$
\begin{aligned}
\text{var}[X] &= \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2 \\
&= \sum_{x \in \otimes} \left[x^2 P(X = x^2)\right] - \left[\sum_{x \in \otimes} x P(X = x)\right]^2 \\
&= \left[0^2 \times (1-p) + 1^2 \times p\right] - p^2 \\
&= p - p^2 \\
&= p(1-p)
\end{aligned}
$$

If $p = 0$ or $p = 1$, then the Bernoulli random variable follows the **degenerate distribution**; a distribution with only one outcome (and so no randomness is possible). The degenerate distribution can be trivially seen from the above to have mean p and variance 0, which makes sense!

The Geometric Distribution

We have a biased coin, that throws heads with a probability $P(\{H\}) = p$. We flip this coin *until* the first head appears. The number of flips that we have to make is a discrete random variable that takes integer values greater than or equal to 1.

Each toss of the coin has the probability p of being a head, and so has a probability $(1-p)$ of being a tail. In order for us to observe n tosses, we must have thrown $(n-1)$ tails in a row, and then a head. Therefore the probability of seeing n tosses is:

$$P(X = n) = (1-p)^{(n-1)} p.$$

This is the PMF of the **geometric distribution**. It is defined only for positive integers $n \geq 1$ and $0 \leq p \leq 1$.

The geometric distribution describes probability of observing a number of independent trials, which have some probability of success p, prior to observing a success (at which point your experiment ends).

- Expectation: $\mathbb{E}[X] = p$
- Variance: $\text{var}[X] = p(1-p)$.

The Binomial Distribution

A different statistic that may be of interest is how often you expect your biased coin to throw heads given n tosses. In this case we're asking how many times k do I observe 'success' in n trials. The probability of seeing k successes from n trials: p^k. Consequently the probability of seeing the $n-k$ failues is: $(1-p)^{n-k}$. Finally, there are:

$$\frac{n \times (n-1) \times (n-2) \times \cdots \times (n-k+1)}{k \times (k-1) \times (k-2) \times \cdots \times 1}$$

ways of constructing k successes from n trials (when order of appearance doesn't matter). Therefore, the probability of observing k successes from n trials with our coin is:

$$
\begin{aligned}
k &\sim \text{Bin}(n, p) \\
P(k; n, p) &= \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k},
\end{aligned}
$$

This is the PMF of the **binomial function**.

The observant here will recognise the coefficient from last lecture. This is called the *binomial coefficient*, which we presented as being read as 'n choose k':

$$P(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k}.$$

The binomial function has the properties:

- Expectation: $\$\mathbb{E}[X] = Np$
- Variance: $\text{var}[X] = Np(1-p)$

The Multinomial Distribution

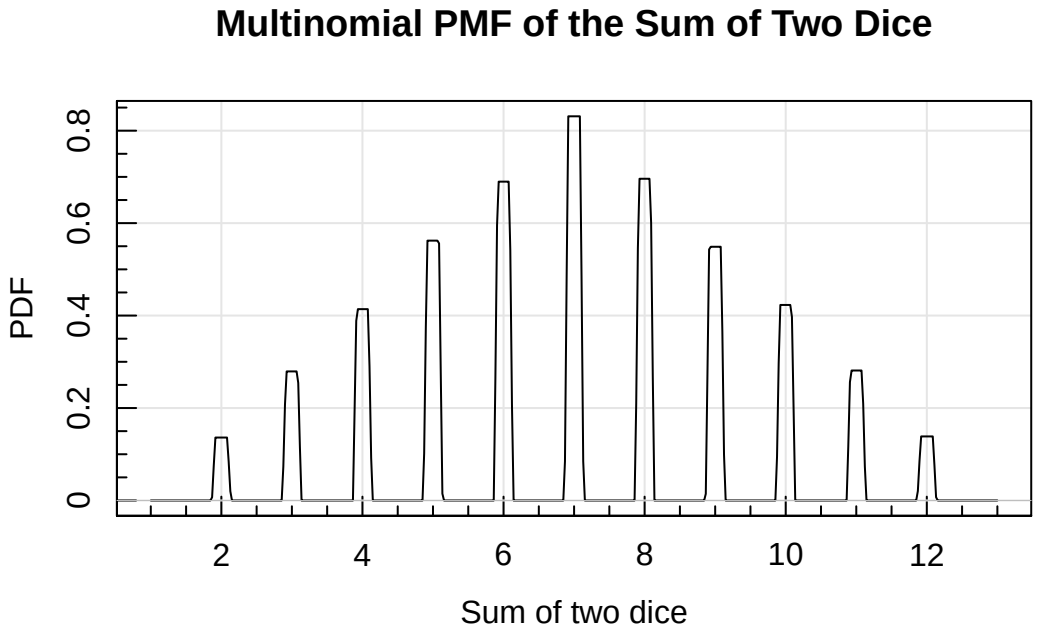
The natural extension of the binomial distribution is to allow for multiple binomial outcomes, like in our urn model.

If we perform N independent realisations of an experiment with $k \in \Omega$ outcomes, where the i^{th} outcome has probability p_i: the probability of observing outcome 1 n_1 times, outcome 2 n_2 times, etc, follows the multinomial distribution PMF:

$$n_1, n_2, \ldots, n_k \sim \text{Multi}(N, p_1, p_2, \ldots, p_k)$$
$$P(n_1, n_2, \ldots, n_k; N, p_1, p_2, \ldots, p_k) = \frac{N!}{n_1! n_2! \ldots n_k!} p_1^{n_1} p_2^{n_2} \ldots p_k^{n_k},$$

for non-negative integers $n_1, n_2, \ldots, n_k$ and where $\sum_{i=1}^{k} n_i = N$.

We can use the multinomial distribution, for example, so get the PMF of our 2-dice roll that we studied in previous lectures:



The multinomial distribution has the properties:

- Expectation: $\mathbb{E}[X_i] = Np_i$
- Variance: $\text{var}[X_i] = Np_i(1-p_i)$

The Poisson Distribution

Consider an experiment where we want to know the number of occurrences of an event within a fixed time window: say, the number of atomic decay processes in a box per hour. Because we are interested in counts per hour, the value of our outcomes are non-negative integers.

We can make two additional constraints:

- The decay processes happen with a fixed rate; and
- The time between decay processes is independent of the interval between previous decays.

Given this experiment, the resulting observations will follow the **Poisson distribution**, which has the

PMF:

$$k \sim \text{Po}(\lambda)$$
$$P(k;\lambda) = e^{-\lambda}\frac{\lambda^k}{k!},$$

where λ is the so-called 'intensity' parameter which governs the shape of the distribution. We can calculate the expectation of the Poisson distribution using our standard formula (but this is a bit more involved than previously...):

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{x\in\otimes} xP(X=x) \\
&= \sum_{x=1}^{\infty} x\frac{e^{-\lambda}\lambda^x}{x!} \\
&= \sum_{x=1}^{\infty} \frac{e^{-\lambda}\lambda^x}{(x-1)!} \\
&= \lambda e^{-\lambda}\sum_{x=1}^{\infty}\frac{\lambda^{(x-1)}}{(x-1)!} \\
&= \lambda e^{-\lambda}\left(\frac{\lambda^0}{0!}+\frac{\lambda^1}{1!}+\dots\right) \\
&= \lambda e^{-\lambda}\sum_{x=0}^{\infty}\frac{\lambda^x}{x!} \\
&= \lambda e^{-\lambda}e^{\lambda} \\
&= \lambda
\end{aligned}$$

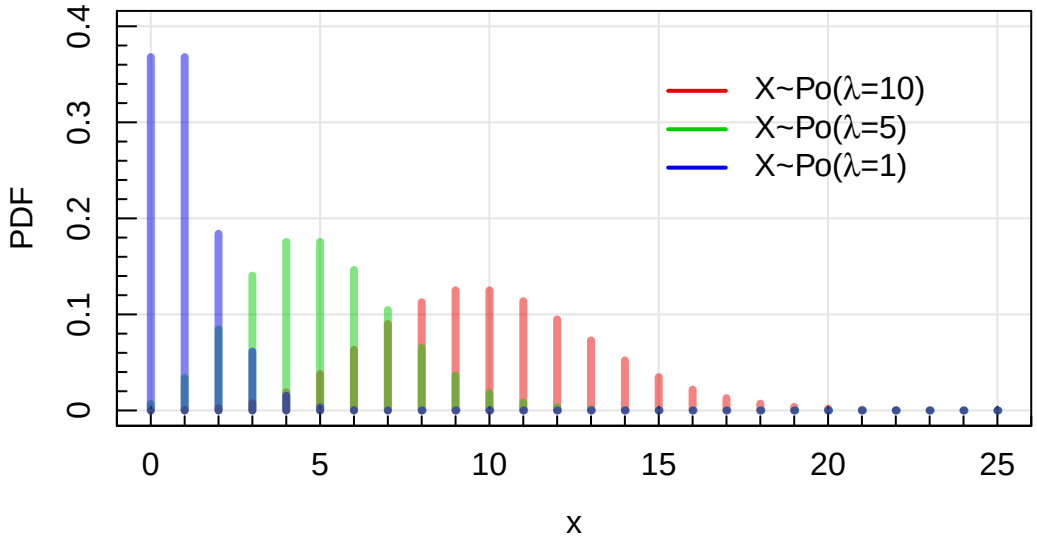
So the λ parameter describes the **expectation** of the Poisson distribution!

We can perform the same exercise to calculate the variance (you can try this yourself), and we find that the **variance** of the Poisson distribution is *also* λ.

The Poisson distribution has the property:

- $\mathbb{E}[X] = \text{var}[X] = \lambda$.

The PMF of a Poissonion random variable, for a range of values of λ is:



The Poisson distribution is very important in the natural sciences, as signals that we measure are frequently actually integrated counts over a period of time. In astronomy, for example, observed fluxes of galaxies are determined by the number of photons received at the detector during an observation. Therefore the minimal uncertainty is determined by the Poisson distribution. This is particularly relevant

in the high-energy regime, because γ-ray photons are relatively rare. As such, flux uncertainties at these short-wavelengths are generally Poisson-noise dominated.

One further important point about the Poisson distribution we can determine using some intuition.

Consider a Poisson variable. The expected number of occurances of an event in some period t is λ. How many events would you intuitively expect to observe in a period $2t$?

- 2λ is the sensible result, given both intuition and what we know about the expectation value.

But the poisson variable need not be constrained to *temporal* events; I can just as easily construct a poisson random variable that determines the number of events observed in a spatial interval x. Take, e.g., the number of insects per unit area in a field. If the insect movements are random and independent, then the number of insects per unit area ought to be a random poisson variable.

But this means that in twice the area we ought to find twice the number of insects. Putting this into more rigorous mathematical context: In some multi-dimensional domain $\mathcal{D}$, the number of spatially random observations within a subset s of $\mathcal{D}$ is a Poisson random variable with intensity $m(s)\lambda$, where $m(s)$ is the area/volume of s.

Models such as these are useful in Physics and Astronomy because they capture the properties that:

- the within the domain are randomly distributed; and
- the probability that you find a point doesn't depend on where you are in the domain.

Continuous Distributions

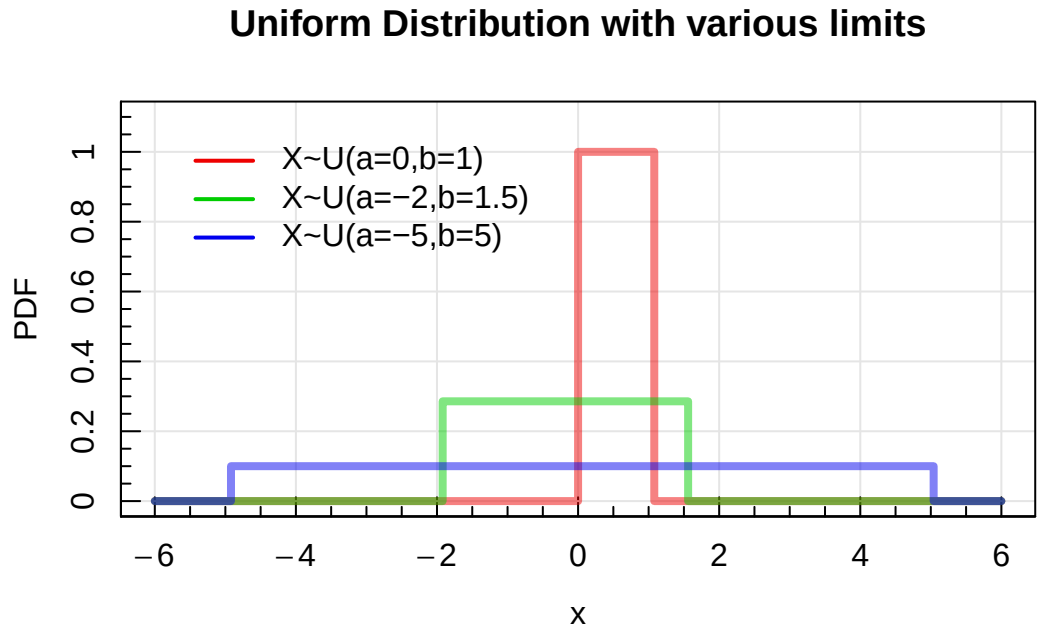
Until now we have looked exclusively at discrete distributions. We will now shift focus to continuous distributions.

The Uniform Distribution

The first continuous distribution that we will examine is the most simple. The **uniform distribution** is characterised by a lower bound a and upper bound b (where $a < b$), with constant probability density in between:

$$X \sim U(a, b)$$

$$p(X = x) = \begin{cases} 0 & \text{if } x < a \\ \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{if } x > b \end{cases}$$



The Uniform distribution has properties:

- Expectation $\mathbb{E}[X] = \frac{a+b}{2}$
- Variance $\text{var}[X] = \frac{(b-a)^2}{12}$

The uniform distribution is useful when considering random events over a fixed interval. For example: Consider the scenario where the tram to RUB arrives at Bochum Hauptbahnhof every 5 minutes during peak travel times. If people arrive at the station randomly and without knowing the schedule, how long can they expect to wait for a tram?

The waiting times in this scenario is a random uniform variable:

$$X \sim U(0, 5)$$

$$p(X = x) = \begin{cases} \frac{1}{5} & \text{if } 0 \leq x \leq 5 \\ 0 & \text{otherwise} \end{cases}$$

The expectation of this distribution is $\mathbb{E}[X] = \frac{5}{2}$, so the expected waiting time is 2.5 minutes. Similarly, we can compute the probability of waiting less than 1 minute using the cumulative probability distribution (which for the uniform distribution is trivial):

$$P(X < x) = \begin{cases} 0 & \text{if } x < a \\ \frac{x-a}{b-a} & \text{if } a \leq x \leq b \\ 1 & \text{if } x > b \end{cases}$$

So for our train waiting times:

$$P(X < x) = \begin{cases} 0 & \text{if } x < a \\ \frac{x}{5} & \text{if } 0 \leq x \leq 5 \\ 1 & \text{if } x > 5 \end{cases}$$

which means that there is a 20% chance of waiting less than 1 minute.

The Beta Distribution

The second continuous distribution that we're going to look at is the **beta distribution**. The beta distribution is not one that you are likely familiar with, but it is useful for two reasons:

- It is an extremely flexible function that is defined over a finite range; and
- It is widely used in congugate analyses (which we will explore in the next section of the course).

The Beta distribution is defined as:

$$X \sim \mathrm{Be}(a, b)$$

$$p(X = x; a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} x^{a-1}(1-x)^{b-1},$$

where Γ is the Gamma function, defined as:

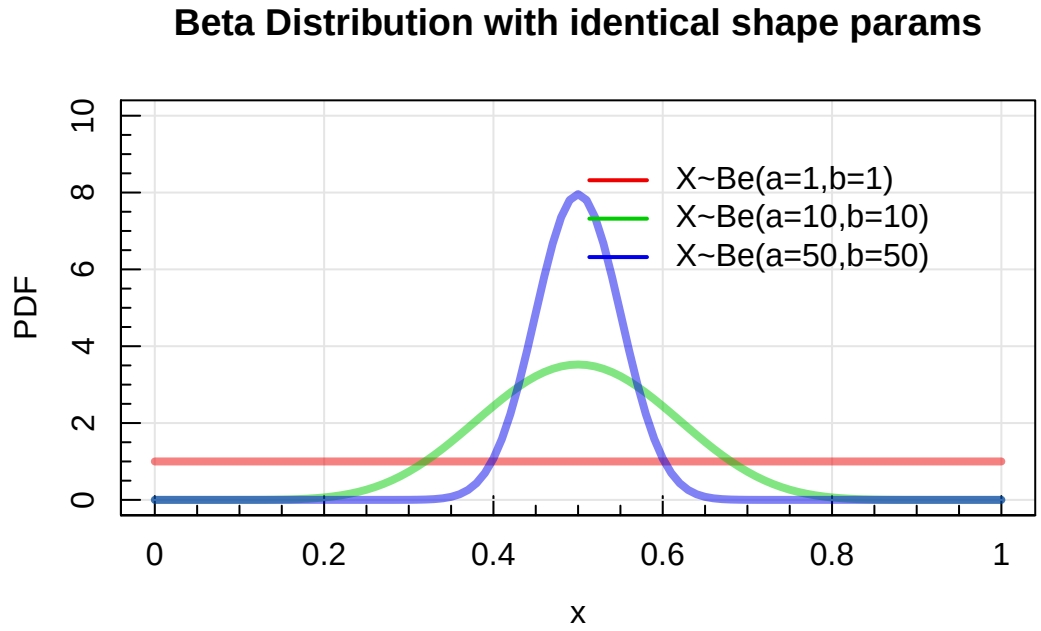
$$\Gamma(n) = (n-1)!$$

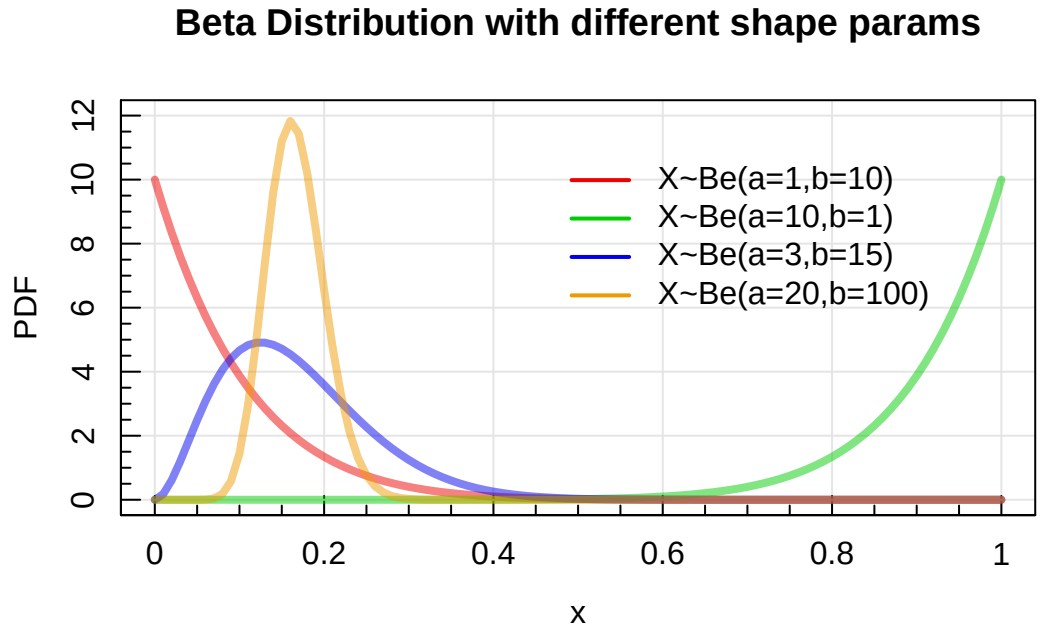
for integer values n, and:

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} \mathrm{d}x$$

for any complex number with a positive real component $\mathcal{R}(z) > 0$.

The variability of the Beta distribution as a function of shape parameters a and b are shown below:





The Beta distribution has the properties:

- Expectation $\mathbb{E}[X] = \frac{a}{a+b}$
- Variance $\mathrm{var}[X] = \frac{ab}{(a+b)^2(a+b+1)}$

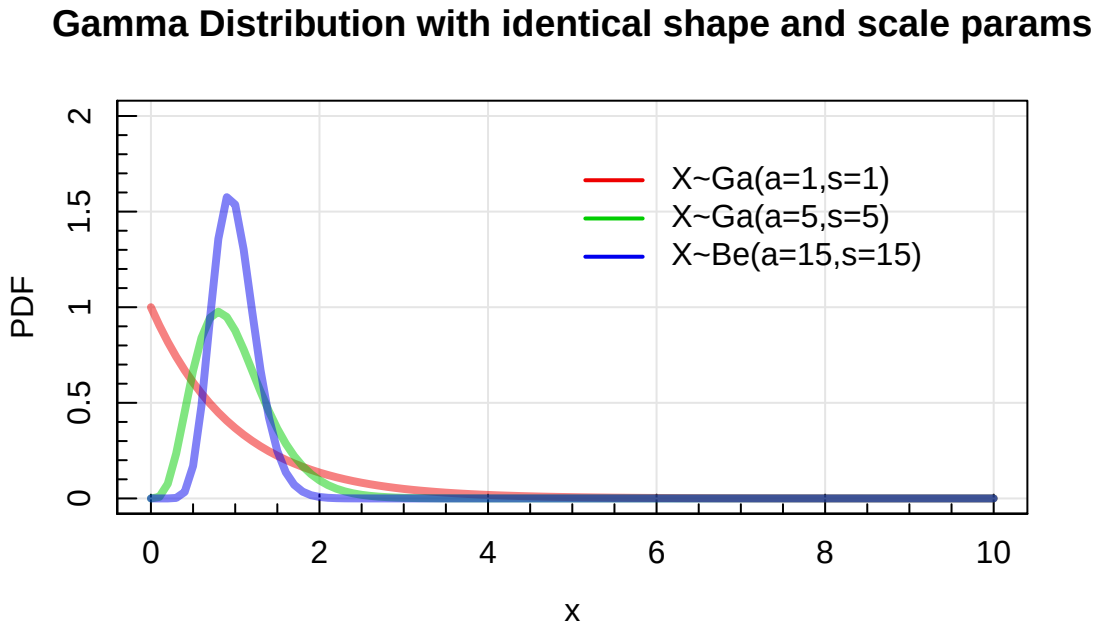
The Gamma Distribution

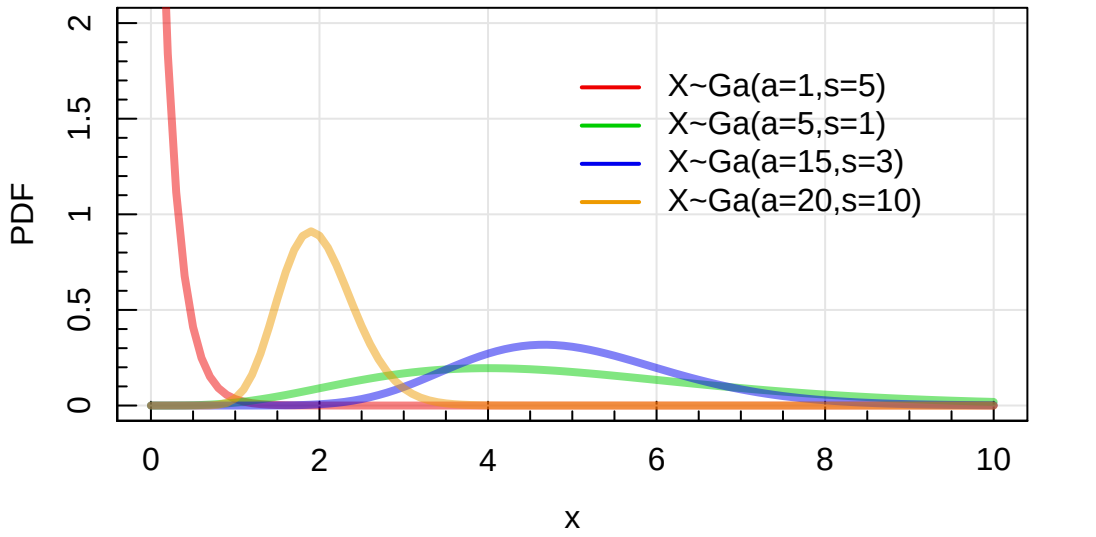
Similar to the Beta distribution, the **gamma distribution** will become relevant later in the course due to its use in conjugate analyses. The gamma distribution is defined by two parameters, the shape parameter a and the scale parameter s. The gamma distribution PDF is defined as:

$$X \sim \text{Ga}(a, s)$$
$$p(X = x; a, s) = \frac{s^a x^{a-1} e^{-\beta x}}{\Gamma(\alpha)}.$$

As with the Beta distribution, the Gamma distribution is extremely flexible. In fact, the Gamma distribution encompasses (as special cases with particular parameter combinations) many interesting distributions that you will likely already be familiar with, such as the exponential distribution and the χ^2-distribution (the latter of which we will discuss in a few slides time!).

The variability of the Gamma distribution as a function of shape parameter a and scale parameter s are shown below:





The Gamma Distribution has properties:

- Expectation $\mathbb{E}[X] = \frac{a}{s}$

- Variance $\text{var}[X] = \frac{a}{s^2}$

The Gaussian Distribution

Everyone in this course is undoubtably familiar with the **Gaussian Distribution**, and it is one of the most important distributions in statistics.

A random variable X which follows a Gaussian distribution with mean μ and variance σ^2 has the PDF:

$$X \sim N(\mu, \sigma^2)$$
$$p(X = x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

The cumulative distribution function of the Gaussian distribution is widely used because of the frequency with which variables are modelled as being Gaussian. The CDF is defined as:

$$\Phi(X = x) = \int_{-\infty}^{x} p(t)\mathrm{d}t$$

There is no analytic formulae for the CDF of the $N(\mu, \sigma^2)$ gaussian, and so the values are computed using numerical and computational methods.

A useful special case of the Gaussian distribution is the so-called "standard" Gaussian: $X \sim N(0, 1)$:

$$X \sim N(0, 1)$$
$$p(X = x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}},$$

It is possible to transform any gaussian variable into a standard gaussian variable using the so-called Z-transformation:

$$X \sim N(\mu, \sigma^2)$$
$$Z = \frac{X - \mu}{\sigma}$$
$$\therefore Z \sim N(0, 1)$$

This is useful for many reasons, particularly in modelling and understanding biases, but here we will use the transformation to derive some useful properties of the gaussian distribution:

$$\begin{aligned} P(X \leq b) &= P\left(\frac{X - \mu}{\sigma} \leq \frac{b - \mu}{\sigma}\right) \\ &= P\left(Z \leq \frac{b - \mu}{\sigma}\right) \\ &= \Phi(\frac{b - \mu}{\sigma}). \end{aligned}$$

- The probability of a Gaussian random variable being less than some value b, for an arbitrary mean and variance, is equal to the probability mass below the equivalent Z-score in the standard gaussian.

By symmetry and integral constraints we can equally demonstrate that:

$$P(X > a) = 1 - \Phi(\frac{a - \mu}{\sigma});$$
$$P(a \leq X \leq b) = \Phi(\frac{b - \mu}{\sigma}) - \Phi(\frac{a - \mu}{\sigma});$$
$$\Phi(-a) = 1 - \Phi(a);$$
$$P(-a < Z < a) = 2\Phi(a) - 1.$$

The Gaussian Distribution has properties:

- Expectation $\mathbb{E}[X] = \mu$
- Variance $\text{var}[X] = \sigma$

The Multivariate Gaussian Distribution

One can extend the Gaussian Distribution to multiple dimensions, forming the **multivariate gaussian distribution**. The multivariate Gaussian distribution is characterised by a vector of means μ and a **covariance matrix** Σ:

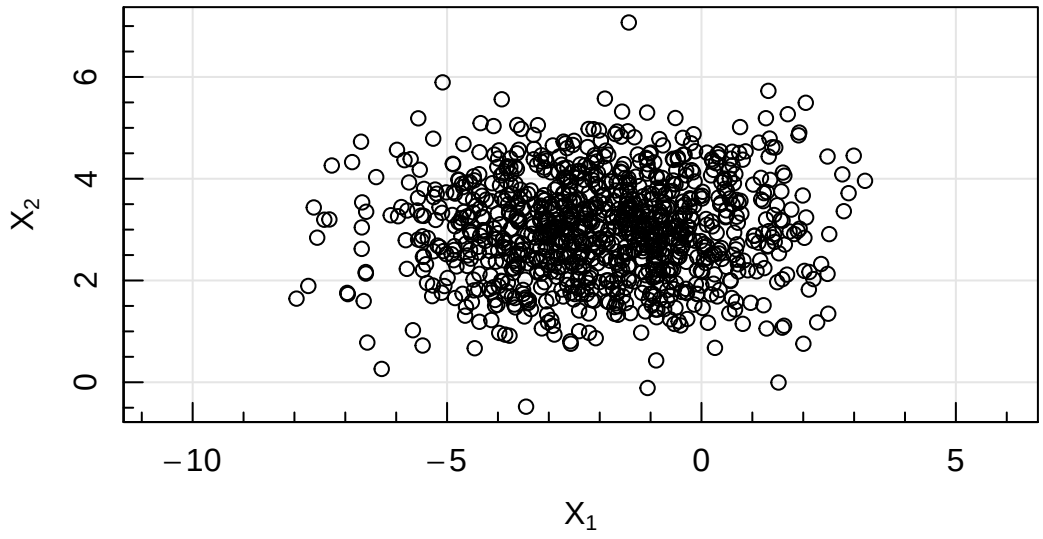
$$X \sim N_k(\mu, \Sigma),$$

The primary difference between the description of the univariate and multivariate Gaussian distributions is therefore that the multivariate gaussian requires the specification of the covariance matrix, which we discussed in previous lectures:

$$\Sigma = \begin{pmatrix} \mathrm{var}(X_1) & \mathrm{cov}(X_1, X_2) & \dots & \mathrm{cov}(X_1, X_n) \\ \mathrm{cov}(X_2, X_1) & \mathrm{var}(X_2) & \dots & \mathrm{cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(X_n, X_1) & \mathrm{cov}(X_n, X_2) & \dots & \mathrm{var}(X_n) \end{pmatrix}$$

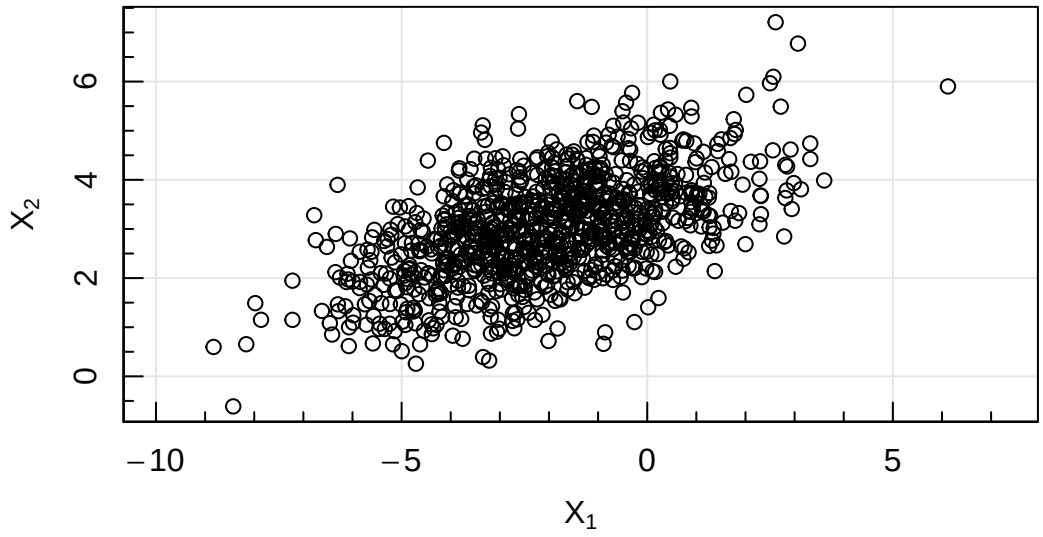
The covariance matrix encodes the degree of mutual variance between the different gaussian components. In cases with zero covariance:

$$X \sim N_2(\mu, \Sigma)$$
$$\mu = \{-2, 3\}$$
$$\Sigma = \begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix}$$



Or the case of significant covariance:

$$X \sim N_2(\mu, \Sigma)$$
$$\mu = \{-2, 3\}$$
$$\Sigma = \begin{pmatrix} 4 & 2 \\ 2 & 1 \end{pmatrix}$$



What's with all the Gaussians?!

In the natural sciences, we see the Gaussian distribution *a lot*. We typically formulate our uncertainties as Gaussian, we model physical processes as Gaussian, and we quote significances in terms of Gaussians. **Why do we do this??**

Central Limit Theorem (CLT)

In a simplified manner, the central limit theorem states that, for an arbitrary distribution with mean μ and variance σ^2, we can draw n observations at random and calculate their mean M_n. The distribution

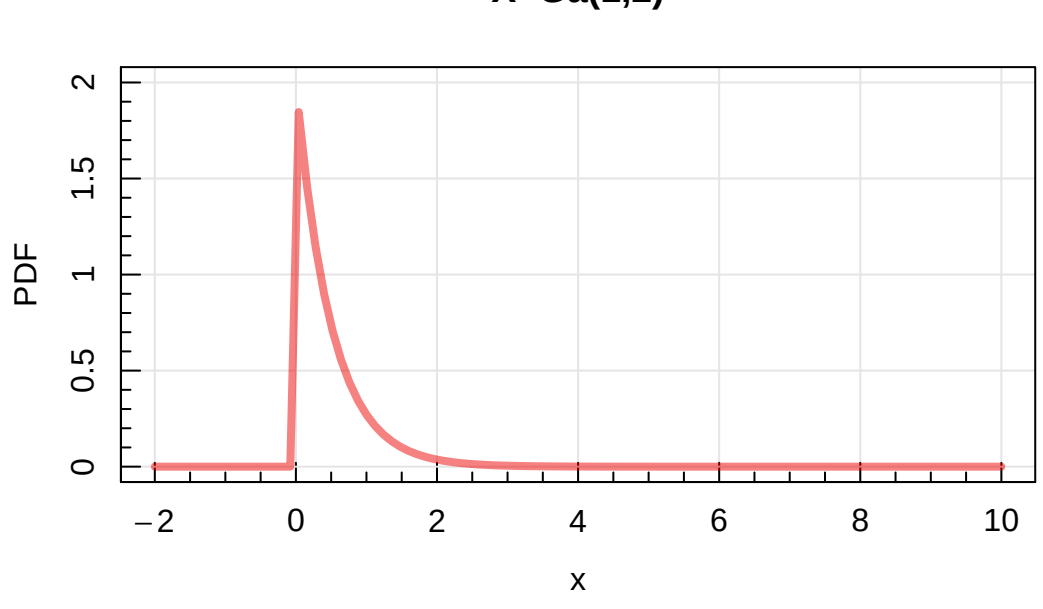
$$\left[\frac{M_n - \mu}{\sigma/\sqrt{n}}\right] \rightarrow N(0,1)$$

as $n \rightarrow \infty$.

This is a pretty remarkable statement. It means that (in certain conditions that I don't raise here, but just know that they are met in *many* circumstances), averaging over random variables will produce a Gaussian distribution of results *regardless of the shape of the distribution the samples are drawn from*!

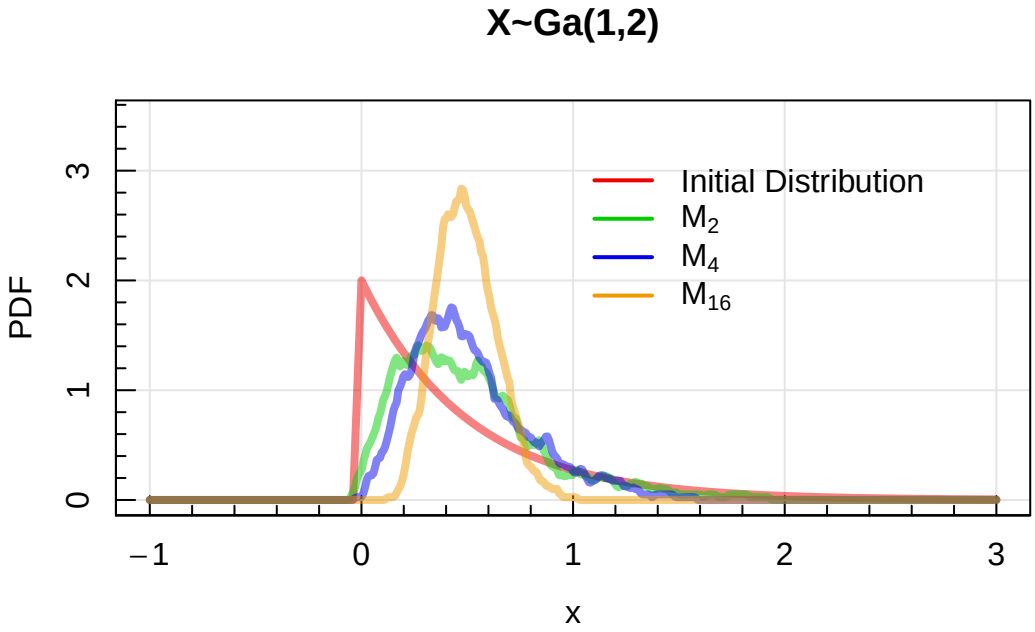
Let's do an experiment. In our experiment, we observe a random variable that follows a Gamma distribution:

$$X \sim Ga(1,2)$$

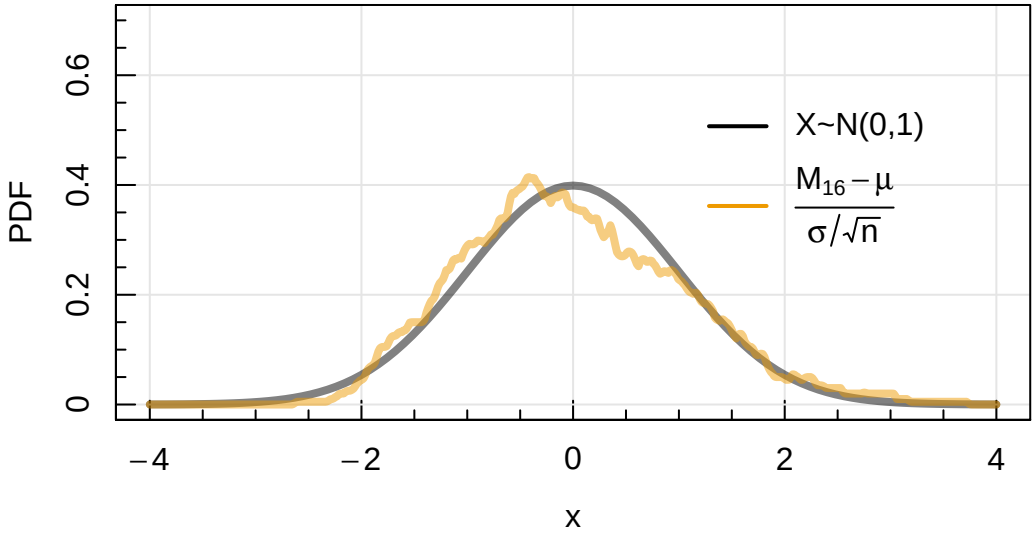


It should go without saying: the gamma distribution we have picked doesn't look very gaussian. But what happens if we are interested in the properties of sub-samples of our population, say, the mean of sets of 2 observations? Or 4? Or 16?

The expectation of this distribution is $\mathbb{E}[X] = \frac{a}{s} = 0.5$ (from earlier). If we take our observations and compute the expectation of discrete subsets of the observations we find the below:



To demonstrate just *how good* the approximation is, we can standardise the distributions and overplot a $N(0,1)$ gaussian:



This result has profound implications for the experimental sciences. It is the reason why we model complex systems as Gaussian distributions. It gives rise to Gaussian distributions in systems where repeat random events are the norm (e.g. repeated scattering of particles drives their energy to the expectation), causes us to model our uncertainties as gaussians (because final observed quantities are the cumulative result of many underlying processes), and leads us to compare models by quoting "sigma" differences.

Finally, one last comment about the CLT. The tails of the gaussian distribution are remarkable tight: we expect more than 95% of gaussian observations to reside within 2 standard deviations of the mean. However in the CLT, the convergence happens most rapidly in the *centre* of the distribution. This means that, in the limit of small numbers of observations, we expect the CLT converge on the Gaussian distribution fastest at the centre of the distribution, and slowest in the wings.

- As a result, gaussian modelling distributions of objects which only enact few averagings

will tend to ignore the presence of the outlier that are expected, leading to problems. **Always pay attention to the tails of your distributions**.

The Student t-Distribution

Given our observation about the wings of the convergence to a gaussian in the CLT, we are motivated to explore different distributions that account for this behaviour.

We can define a new statistic t which is our mean statistic from before:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{N}},$$

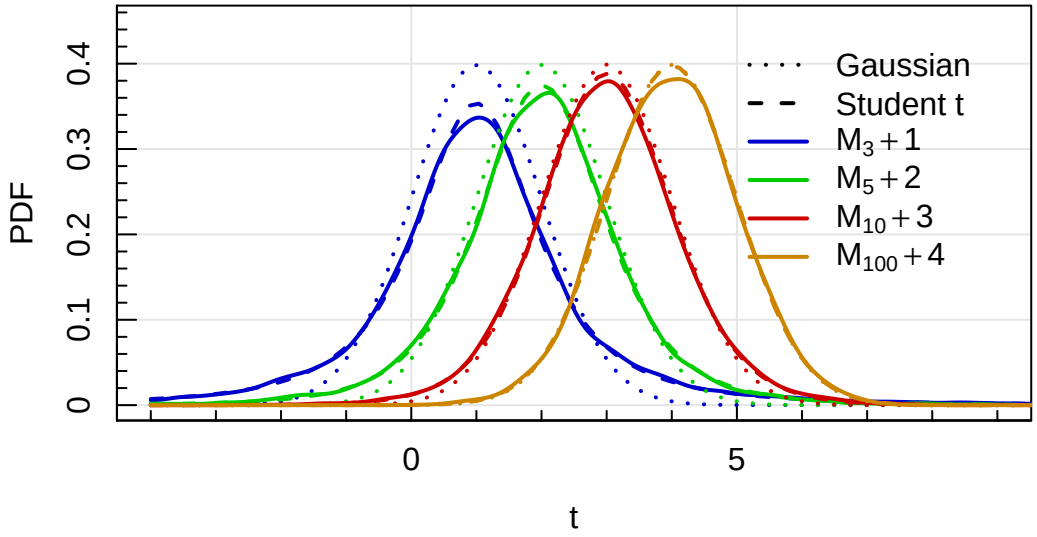
Given the t variable, we can compute the **student t-distribution**:

$$p(t; N) = \frac{\Gamma(\frac{1}{2}N)}{\sqrt{\pi(N-1)}\Gamma\left[\frac{1}{2}(N-1)\right]} \left(1 + \frac{t^2}{N-1}\right)^{-N/2}$$

Frequently, though, we will see this distribution parameterised in terms of the 'degrees of freedom' $\nu = (N-1)$.

Nonetheless, this still looks complicated, but the limits are the most relevant for understanding. When N is large, the student t-distribution converges to the gaussian distribution, as we wanted. However, for small N, the distribution converges to the Lorentzian (or Cauchy) distribution, which contains the strong wings that we expect.

We can demonstrate the above limiting behaviour easiest when computing the t-statistic for initially gaussian observations (rather than our one-sided Gamma distribution from last slide). We perform exactly the same game, though. We average n observations from an initially Gaussian distribution, and plot the distribution of our t statistic above (with their centroids shifted for clarity):



The Student t-Distribution has properties:

- Expectation $\mathbb{E}[X] = \frac{a}{s}$
- Variance $\text{var}[X] = \frac{a}{s^2}$

The χ^2 Distribution

Again related to the CLT is the χ^2 **distribution**. The χ^2 distribution is that which we find when summing random *squared deviations* from a Normal distribution. The χ^2 *distribution* is extremely common in the experimental sciences, as it is the classic metric that is used in the quantification of *goodness-of-fit* of a model to data. In this application it is frequently referred to as the Pearson χ^2 statistic.

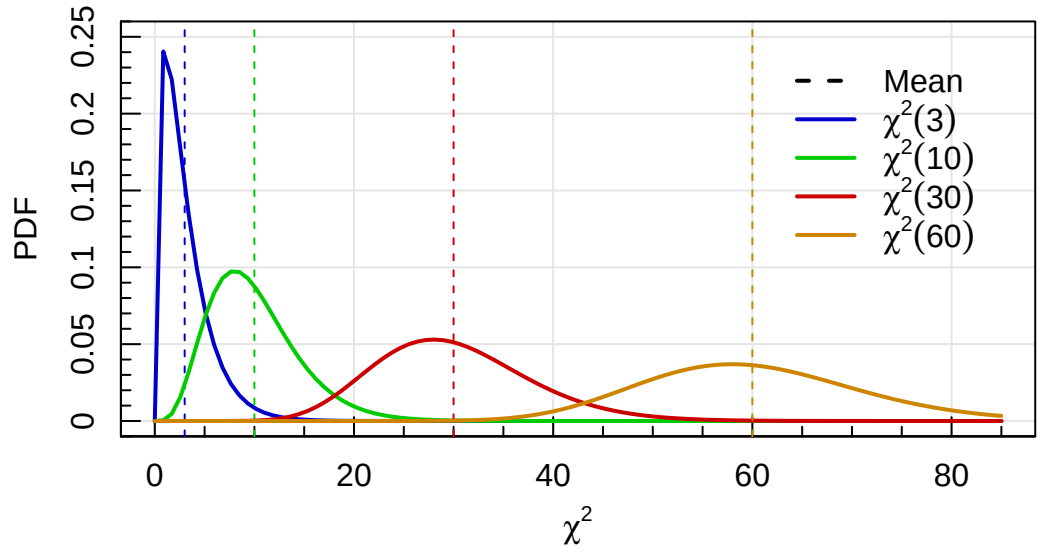
The χ^2 statistic is defined as:

$$\chi^2 \equiv C = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{\sigma^2}.$$

For the χ^2 distribution the degrees of freedom is given by $\nu = N$. The χ^2 distribution is then defined as:

$$p(C;\nu) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} C^{\nu/2-1} e^{-C/2}$$

Use of the χ^2 statistic in quantifying goodness-of-fit is frequently summarised in the literature as *reduced* χ^2, generally formulated as $\chi^2_{\mathrm{r}} = \chi^2/\nu$. The choice of this statistic is that the expectation of the χ^2 variable is ν: $\mathbb{E}[\chi^2/\nu] = 1$. However, model fits in the literature generally perform maximum likelihood optimisation, which aims to find the *mode* of the PDF, rather than to produce samples that report the mean/expectation. Looking at the χ^2 distribution, it's clear that the distribution is asymmetric, and so these results will converge on different answers:



The **mode** of the distribution is actualy at $\nu - 2$. Therefore, the appropriate formulation of the reduced χ^2 is for maximum likelihood fits is:

$$\chi^2_{\mathrm{r}} = \frac{\chi^2}{\nu - 2}$$

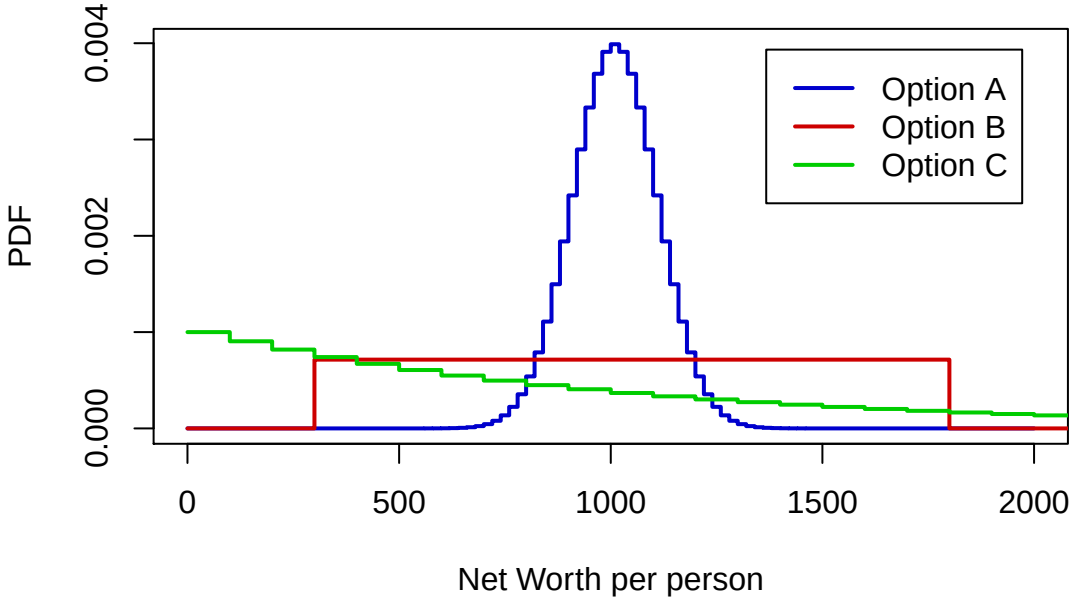
The χ^2 Distribution has properties:

- Expectation $\mathbb{E}[X] = \nu$
- Variance $\mathrm{var}[X] = 2\nu$

Wealth and equality

Consider $1,000$ people who each have $\$1,000$. We take two of these people at random, and give a fixed fraction of their total worth to the other (think of this a transaction of some kind). We then repeat this process $100,000$ times.

What will the final distribution of wealth look like?



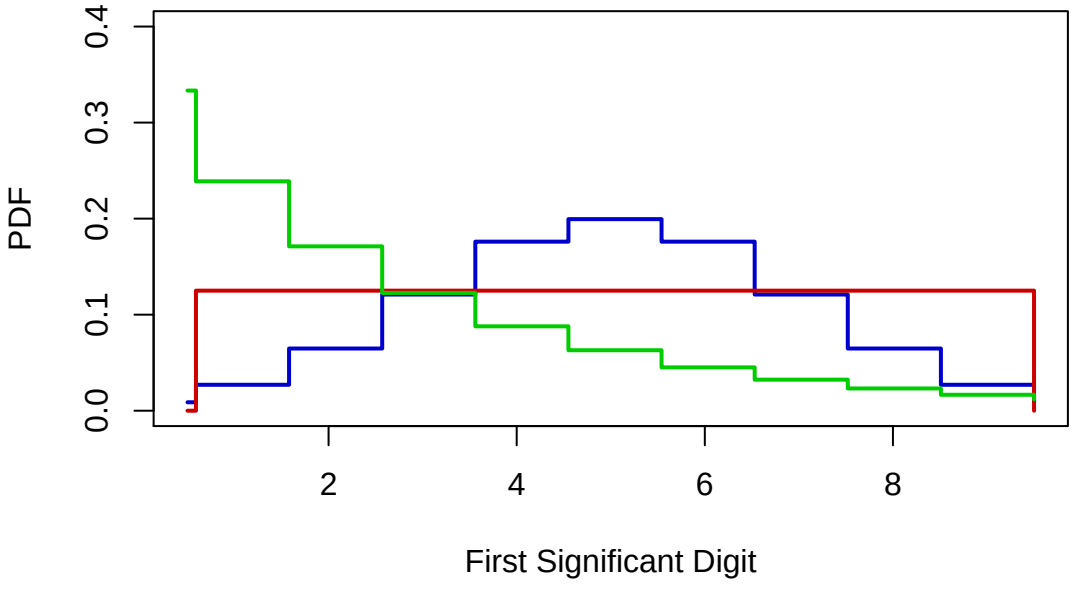
Can you get away with tax fraud?

For totally legitimate, not-illegal reasons: you need to fabricate a series of transactions to "pad out" a table of real transactions, such that your fabricated numbers are indistinguishable from "real" transactions. The transaction values range from large (many thousands of dollars) to small (a few cents).

What distribution would the values of the first significant digit of the numbers take? e.g. if the numbers are all written in scientific notation:

$$1.3 \times 10^{1}$$
$$7.6 \times 10^{-2}$$
$$5.1 \times 10^{2}$$
$$\vdots$$
$$9.0 \times 10^{-1}$$
$$2.3 \times 10^{3}$$

What will the histogram of the numbers in front of the "." look like?



The Pareto Distribution

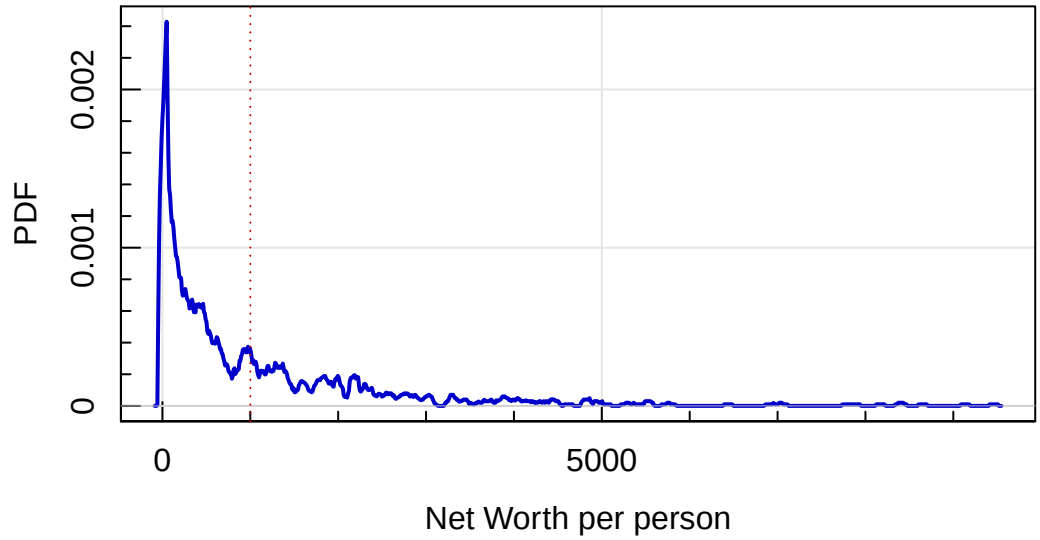
The last distribution that we explore in this section is the Pareto distribution, which is a probability function that follows a truncated power-law. The Pareto probability distribution is defined as:

$$p(x) = \begin{cases} \frac{\alpha x_m^\alpha}{x^{\alpha+1}} & x \geq x_m \\ 0 & x < x_m \end{cases}$$

for $x_m > 0$.

The Pareto distribution is named for Vilfredo Pareto, who first used the probability distribution to demonstrate wealth disparity. We can repeat this demonstration quite simply.

Consider $1,000$ people who each have $\$1,000$. We take two of these people at random, and give a fixed fraction of their total worth to the other (think of this a transaction of some kind). We then repeat this process $100,000$ times.



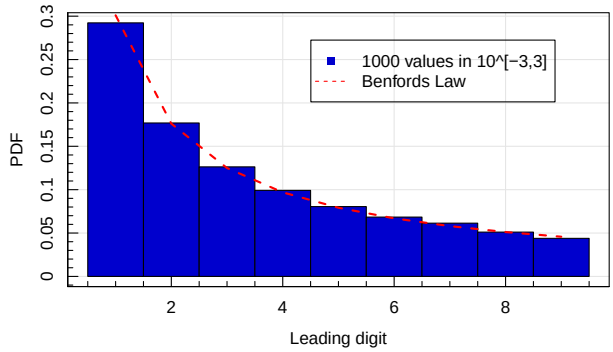
Power-law functions are reasonably common in natural sciences, as many natural processes tend toward the power law. The represent relationships between parameters where a relative change in one quantity produces a proportional relative change in another quantity, regardless of the magnitude of the various parameters. Notable astronomical examples include Keplers third law, and the stellar initial mass function. Notable physical examples include the Stefan-Boltzman law and all inverse-square laws (electrostatics, gravitation, etc).

One interesting demonstration of the power law is in Benfords Law. This is not directly related to physics or astronomy (unless you are doing something *very* questionable with your grant applications...).

Benfords Law, also called the first-digit law, states that the leading digit of listings of numbers that span multiple decades (i.e. many orders of magnitude) will tend to be preferentially valued 1; and the probability of finding larger values is inversely proportional to the digits value:

$$P(D) = \log_{10}\left(1 + \frac{1}{D}\right).$$

```
x<-10^runif(1e4,-3,3)
y<-gsub('0','',gsub(fixed=T,'.','',as.character(x)))
y<-helpRfuncs::vecsplit(y,"",1)
maghist(as.numeric(y),breaks=1:10-0.5,verbose=F,col='blue3',freq=F,
xlab='Leading digit',ylab='PDF')
lines(col='red',lty=2,1:9,log10(1+1/(1:9)),lwd=2)
legend('topright',inset=0.1,legend=c('1000 values in 10^[-3,3]','Benfords Law'),
       pch=c(15,NA),lty=c(NA,2),col=c('blue3','red'))
```



The Pareto Distribution has properties:

- Expectation

$$\mathbb{E}[X] = \begin{cases} \infty & \alpha \leq 1 \\ \frac{\alpha x_m}{\alpha - 1} & \alpha > 1 \end{cases}$$

- Variance

$$\operatorname{var}[X] = \begin{cases} \infty & \alpha \in (1, 2] \\ \left(\frac{x_m}{\alpha - 1}\right)^2 \frac{\alpha}{\alpha - 2} & \alpha > 2 \end{cases}$$