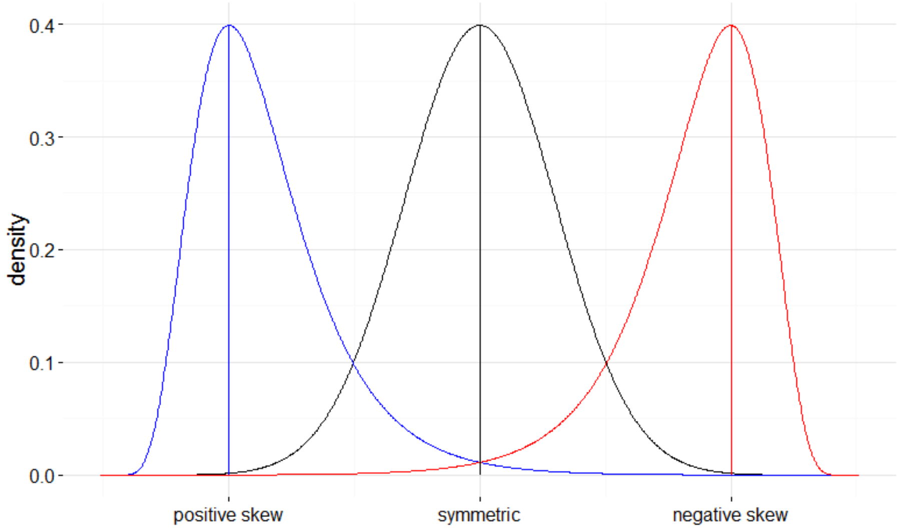
Describing Distributions Qualitatively

Or: why not every distribution can be described as ‘normal’

When we plot our data as a frequency distribution, oftentimes, we wish to describe the distribution by its visual qualities alongside reports of descriptive statistics. Here, we present a few ways to describe the shape of distributions, both with general terms and in relation to commonly-observed distributions.

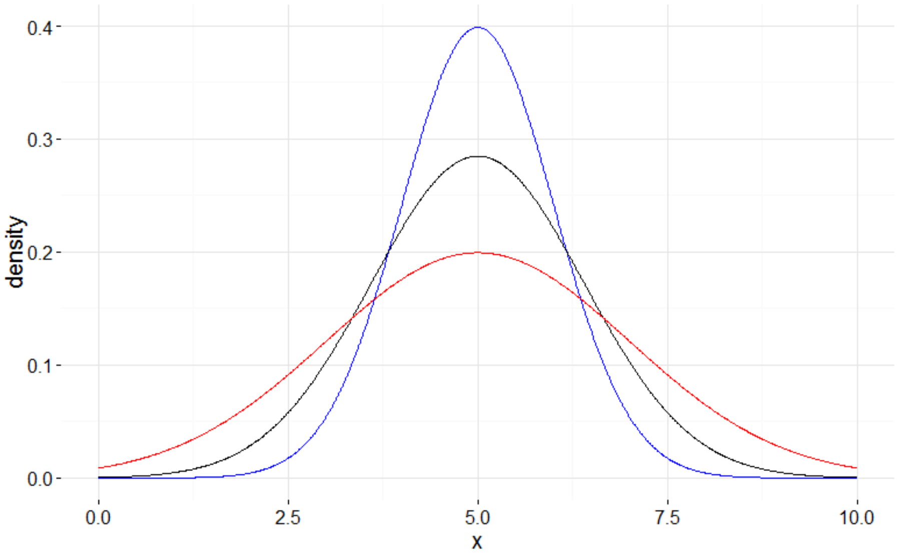
Generic ways to describe distributions

Two ways to describe distributions are in their modality and symmetry. Modality refers to the number of ‘peaks’ in the data’s distribution. Most commonly, we expect to see *unimodal* distributions, or those with single peaks. Sometimes we will see *bimodal* or *multimodal* distributions with two or more peaks; these can be indicative of multiple groups in the data.



For unimodal distributions, we also commonly refer to aspects of the distribution’s symmetry. If data is distributed equally about its peak, it is a *symmetric* distribution. On the other hand, data that is not symmetrically distributed is said to be *skewed*. If a distribution has more mass in its left tail or if its left tail is stretched long, it is said to have *negative skew*. Distributions with heavy or long right-tails are said to have *positive skew*. Skewness has importance for its effects on the mean and median. In asymmetric distributions, both the median and mean will be dragged in the direction of the long or heavy tail, but the mean will be affected by outliers more than the median will. This can have ramifications in the statistics we report and how they should be interpreted. If data is highly skewed, then we may favor reporting the median over the mean to represent a measure for centrality of the data distribution.

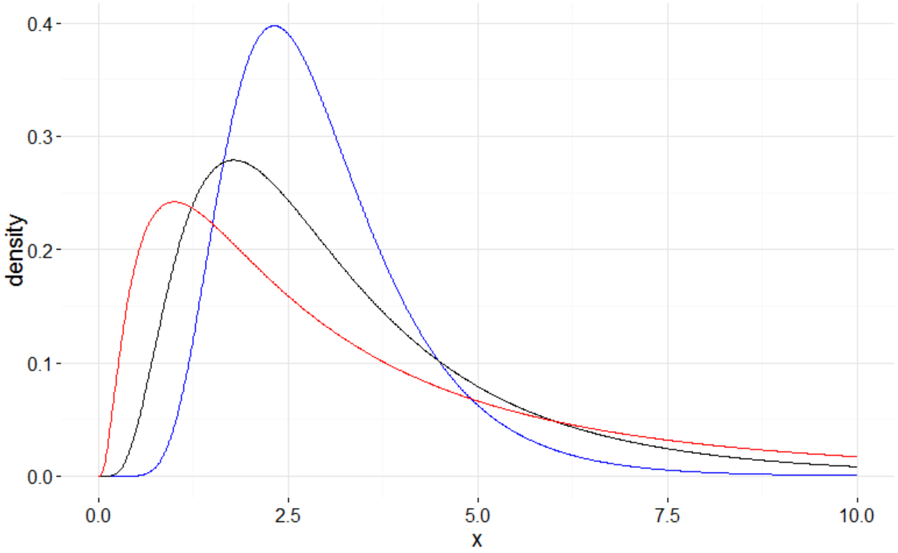
Describing distributions in relation to common functions



高斯！

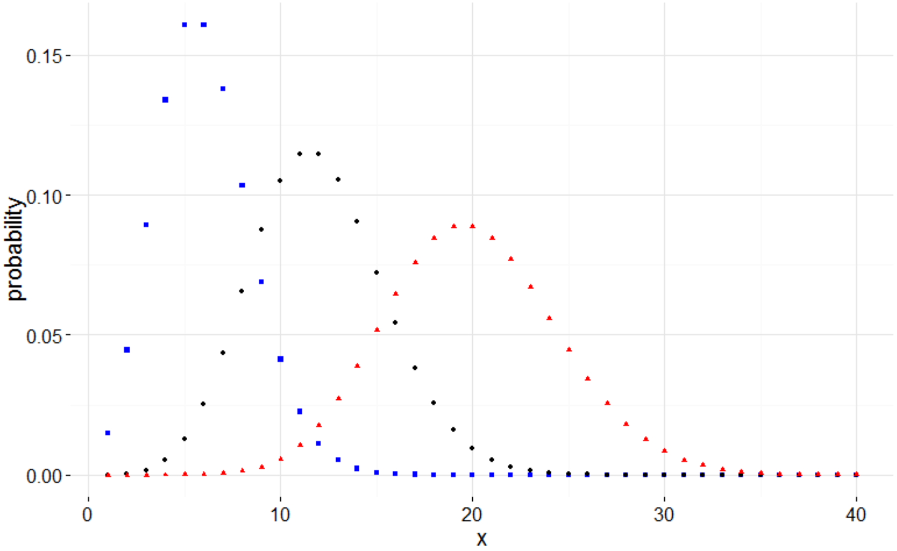
In many cases, we tend to expect data to follow approximately *normal* distributions (also known as a *Gaussian* distribution). Normal distributions are distinguished by their unimodal, symmetric, bell-shaped functions and have a number of well-behaved properties that make them useful for statistical analysis. Taking the sum of multiple independent normal distributions or performing a linear transformation of a normal distribution returns a new normal distribution. If we make multiple observations of a normally-distributed random variable, then if our estimate of the distribution's mean was normally distributed before, it remains normally distributed afterwards.

Perhaps the most important result is the central limit theorem, which states that if we have a number of observations drawn independently from the same distribution (and the distribution has a well-behaved mean and variance), the distribution of the observations' sum is approximately normal, becoming more normal with increased observations. From this, the mean of a set of observations takes on a normally distributed uncertainty. The theorem also helps explain why the normal distribution is so commonly observed, if observed data is actually generated from a process that acts as a sum of many smaller processes.

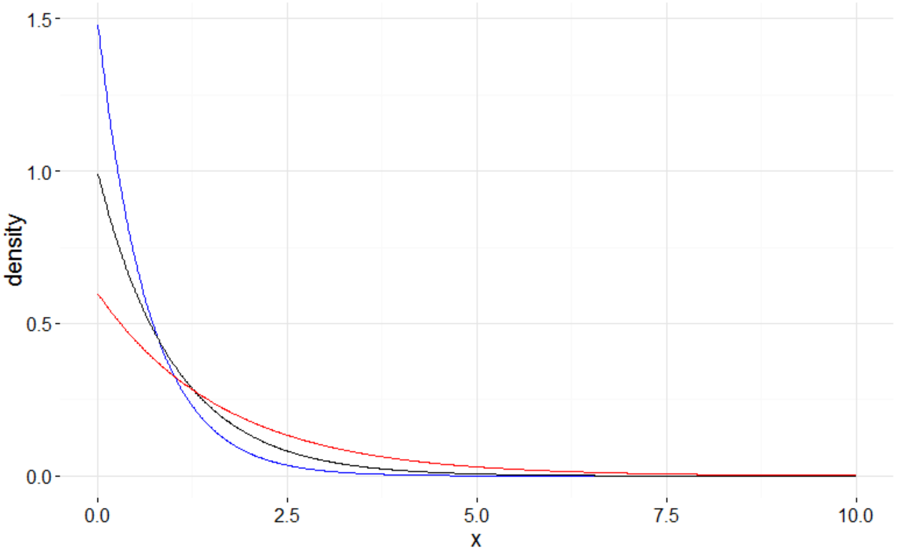


A common relative of the normal distribution is the *log-normal* distribution. In a normal linear scale, log-normal distributions are unimodal with a positive skew. When plotted on a logarithmic scale, the distribution function will look like a normal distribution: symmetric and bell-shaped. The log-normal distribution pops up in a number of places where scaling is exponential or growth is proportional to size. For instance, income data is often modeled as a log-normal distribution; various biological measures such as reference ranges (expected values) in blood tests are better modeled by the log-normal distribution than a standard normal distribution. Tests that can be performed on normally-distributed data may also be applied to log-normal data, so long as the data has been transformed first so its behaviour is normal.

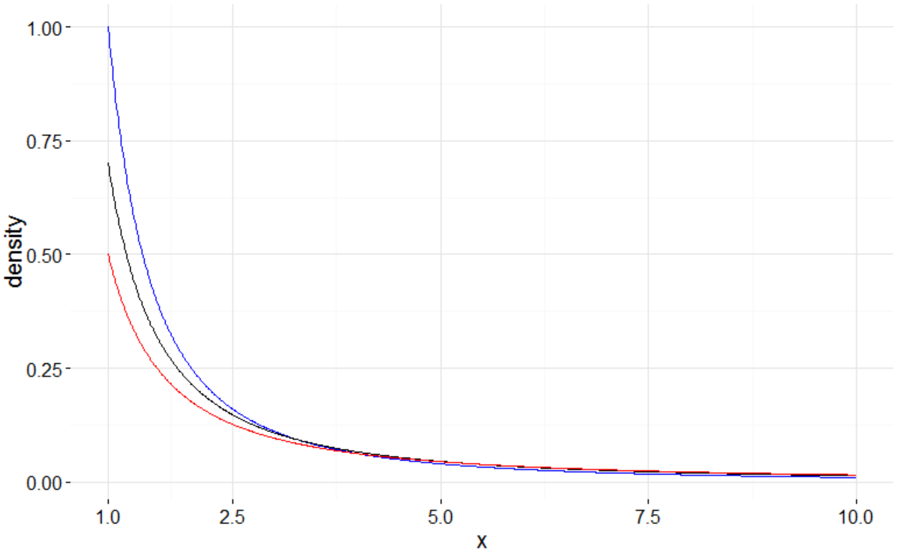
For data that deals with the occurrence of events, you may also encounter the Poisson and exponential distributions. In these distributions, we assume that events occur independently of each other and at a constant rate. For example, we can use the distribution to model earthquake frequencies or to model the volume of calls that go through a system. If we’re concerned about the number of events that occur within a constant time frame, then we will encounter a Poisson distribution. If we want to model the time between events, then we will observe an exponential distribution.



The Poisson distribution is unimodal and has a positive skew and is bounded on the left by 0 (i.e. the minimum observable count of events is 0). For larger event rates, the skew decreases, such that for a large enough event rate the Poisson distribution is well-approximated with the normal distribution. Note, however, that the Poisson takes discrete integer values, while the normal distribution is continuous - continuity corrections must be performed to keep the approximation good at smaller event rates.



The exponential distribution differs from the previous distributions in a number of ways: it is bounded on the left by 0 (i.e. the minimum time between events is 0) and its *mode* (peak) is at 0, so the distribution decreases at all points. This may be somewhat surprising, but it comes about as a result of modeling events as independent and constant rate. In terms of conditional probabilities, knowing the amount of time that has passed since the last event does not provide us additional information about how much longer we need to wait for the next event, compared to the estimated time to wait immediately after an event has occurred.



We close this document with the Pareto distribution. Similar to how the log-normal demonstrates a normal distribution on a logarithmic scale, the Pareto distribution demonstrates an exponential distribution on a logarithmic scale. The shape of the Pareto distribution is not too different from the exponential, with its peak and left-side bound at some value greater than zero, and steadily decreasing in value across the positive axis. However, the Pareto distribution exhibits a very long tail compared to the exponential distribution. The Pareto distribution is an example of a power-law distribution and is used to model populations with a majority amount of value in a minority of the population and a minority of value in the population’s majority. The canonical example of the Pareto distribution lies in modeling wealth inequality, but it can be used to model other things such as how software errors will affect users (there may be a few errors that affect the majority of users with problems and a large number of errors that are incredibly rare). The Pareto distribution is best interpreted through its *cumulative probability*:

a large majority of the points have small values, while an increasingly small minority takes larger and larger values.