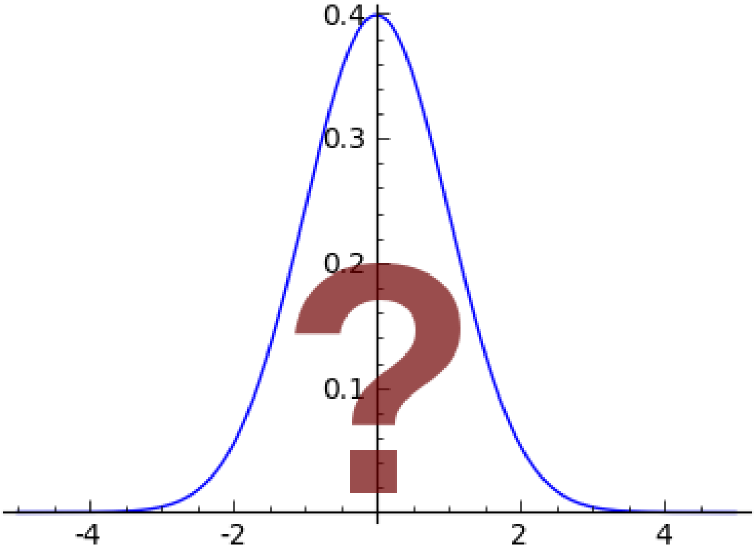
# Logs, Tails, Long Tails

moultano.wordpress.com/2013/08/09/logs-tails-long-tails/

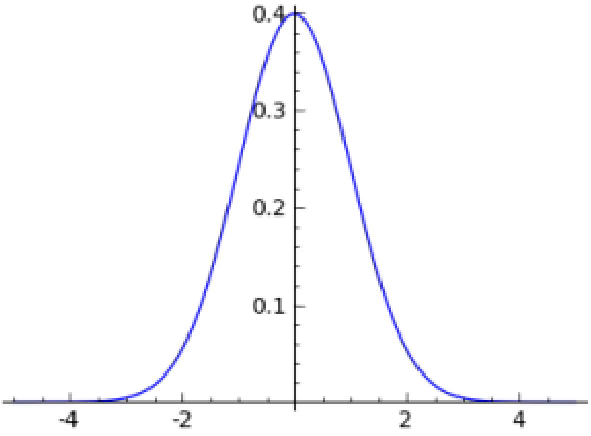
Ryan Moulton

August 9, 2013



## Gaussian Guessing Game

Guess which of these graphs shows the normal/gaussian distribution.

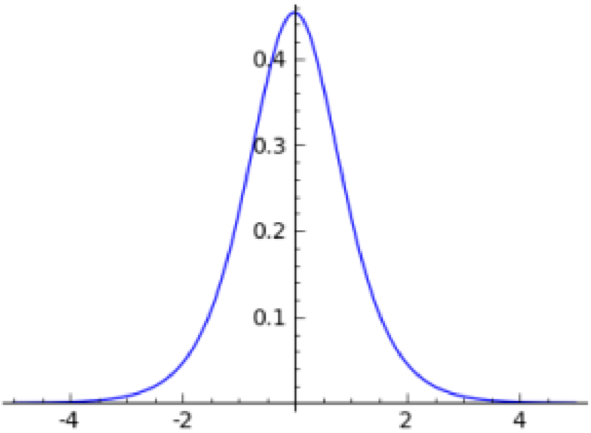


A

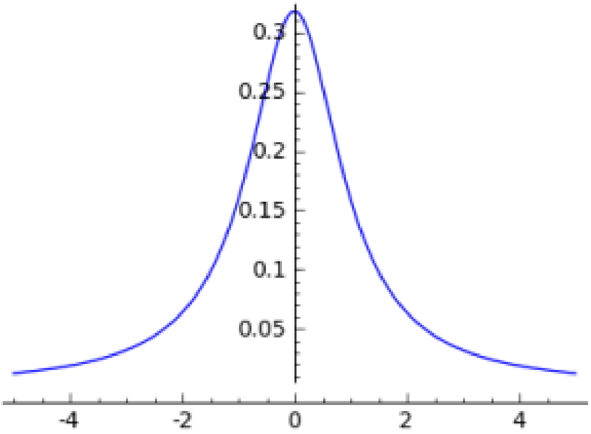
The correct answer is A. The rest are the logistic, cauchy, and beta distributions respectively.

Even if you picked it out correctly, it isn't easy to figure out. All of them look like "bell curves," they are all symmetric, they all taper off at what looks like a similar rate. In that case, why do we care about the difference between them? Why don't we just pick whichever one is most mathematically convenient and not worry about it?

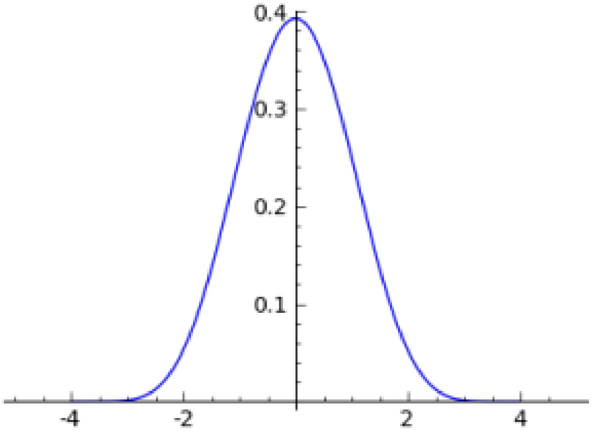
The difference lies in their tails, the parts of the graph where the plot disappears into the x axis. That difference, imperceptible in these graphs, makes these distributions behave incredibly differently, as I'll show below.



B



C



D

## The Basic Operation of Probability: Multiplication

Probabilities are very rarely added together, and probability distributions even more rarely. The basic operation of probability is multiplication. This arises from Bayes' Rule, which describes the relationship between the joint and conditional distributions of random variables:

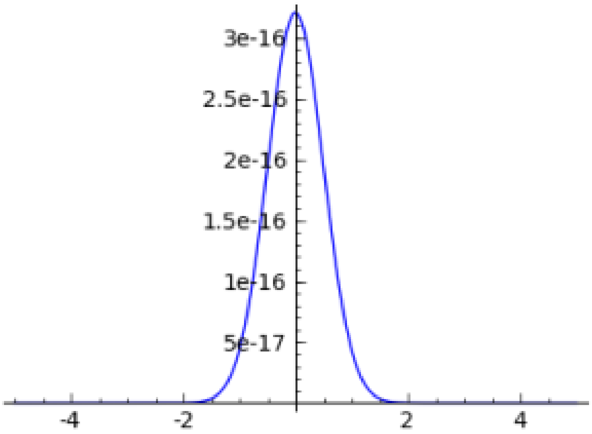
$$P(A, B) = P(A|B)P(B) = P(B|A)P(A)$$

This makes it important to understand how each distribution combines via bayes rule with other distributions; that is, how it multiplies.

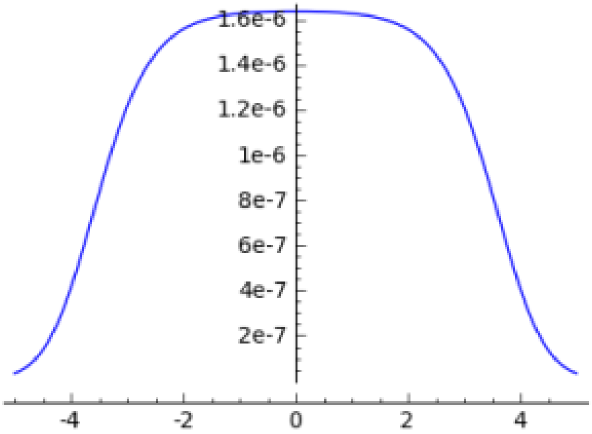
Before you look at the graphs below, try to guess what each one will look like. I've made two copies of each distribution, shifted their centers to -4 and +4, and multiplied the two together. This simulates having two sources of evidence about the same variable that substantially disagree.

Despite the fact that these distributions had very similar looking shapes, their products are entirely different. The distributions are shifted so that the center of one is 8 standard deviations from the center of the other, well out in the range where the plots are indistinguishable from the x axis. Clearly there's a lot going on in this invisible part of the graph!

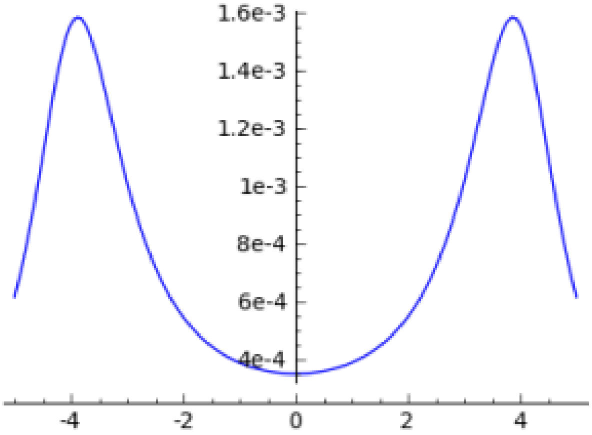
What should we plot to give us an intuition about this? The basic operation shouldn't be so surprising!



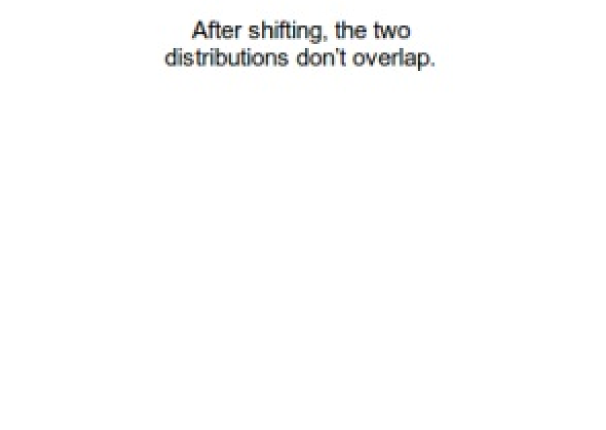
Product of Normal Distributions



Product of Logistic Distributions



Product of Cauchy Distributions



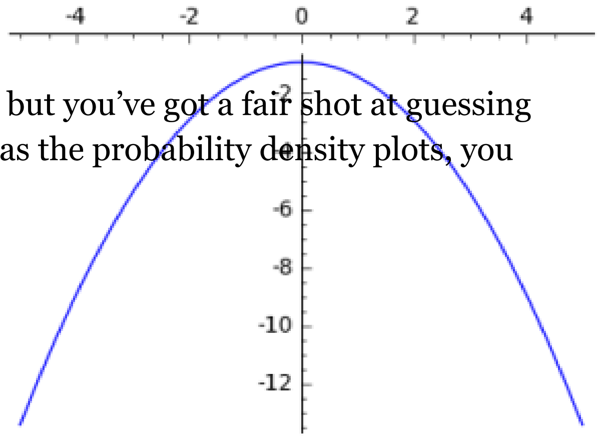
Product of Beta Distributions

## Log Probabilities

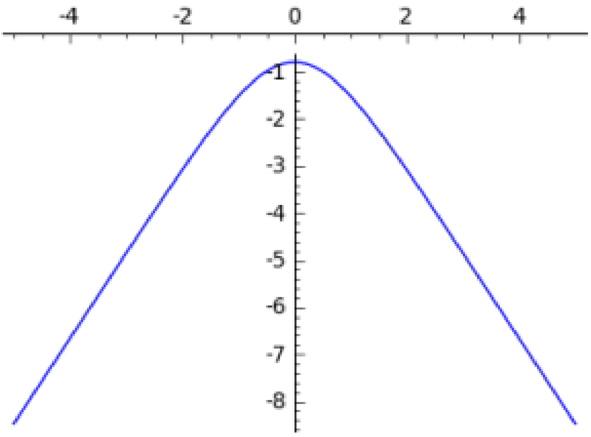
The log function asymptotes at 0. This makes it a perfect choice to visualize very small values of a function by expanding their range. These plots show the log of each probability distribution, and they are all easily distinguishable at a glance. They also make it plausible to predict how the distributions will multiply together. Since multiplying the variables is equivalent to adding the log of the variables, all we have to do is figure out what happens when you add two of these curves to each other.

And as you can see, these make the logs of the products of the distributions make a lot more sense

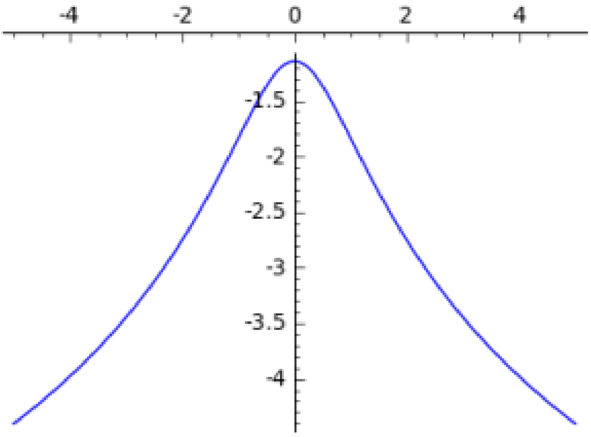
These shapes still aren't immediately obvious, but you've got a fair shot at guessing them from the log plots alone. If all you had was the probability density plots, you wouldn't have any recourse but to go immediately to the math.



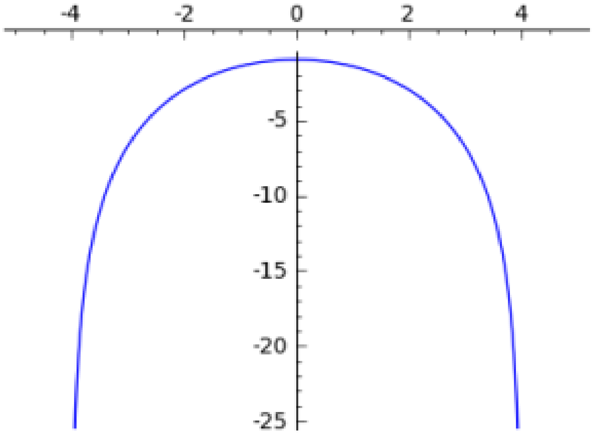
The log of a gaussian distribution is a parabola. Any two parabolas added together form another parabola, so the result of multiplying two gaussians must be a gaussian.



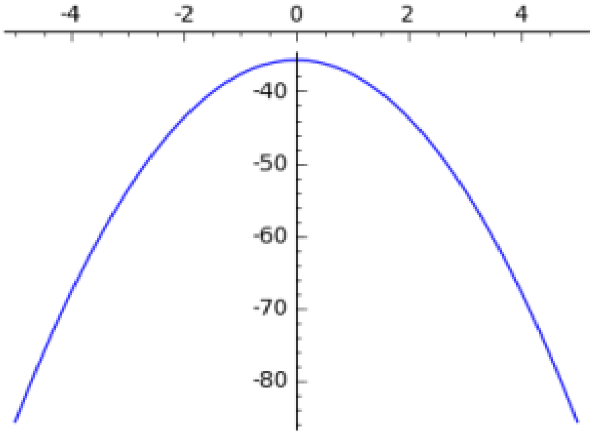
The log of a logistic distribution interpolates between two lines with opposite slopes. If we add two of these, the slopes between the means will cancel out and the slopes outside of the means will reinforce each other. We expect the log distribution to be constant between the two means and have twice the slope outside of them.



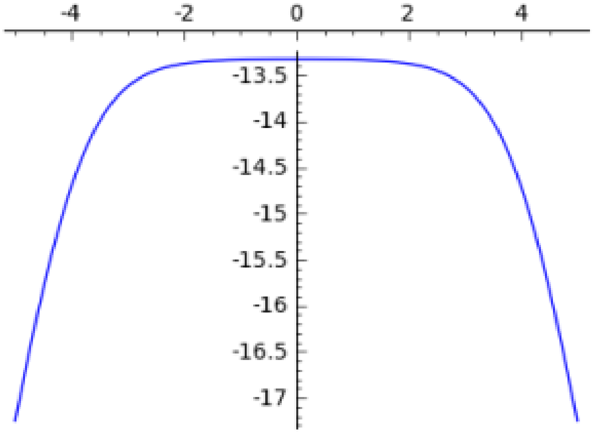
The log of the cauchy has tails that decrease very slowly so it's reasonable to presume that the peaks of the two shifted distributions will overpower the tails, and we'll have something bimodal.



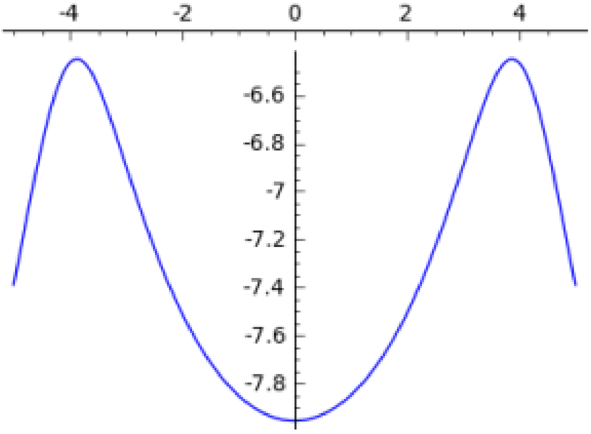
The log of the beta asymptotes at its bounds. It assigns 0 probability to anything outside of them.



Log of the Product of Normal Distributions



Log of the Product of Logistic Distributions



Log of the Product of Cauchy Distributions

## Fitting to Data

Log probability distributions also show up when you are fitting a model to data. Regardless of whether you are using Bayesian or Frequentist methods, fitting a distribution to a set of data is going to involve maximizing a likelihood function (potentially with some additional multiplied terms.) You'll be selecting the distribution by maximizing $P(X|\theta)f(\theta) = f(\theta)\prod_{x\in X} P(x|\theta)$ where $X$ is the vector of data, $\theta$ is the vector of parameters, and $f(\theta)$ is a regularization term, typically a prior distribution for $\theta$. Again we find that we're dealing with a product of probabilities.

The first step of maximizing a function is typically to take its derivative. To make the derivative simpler, it's easier to work with a sum of many terms than a product of many terms. Since log is monotonic, maximizing a nonnegative function is equivalent to maximizing its log, so we can instead maximize the log probability and turn the product into a sum. We'll be maximizing

$$\log P(X|\theta)f(\theta) = \log f(\theta) + \sum_{x\in X} \log P(x|\theta)$$

. The $\log P(x|\theta)$ term is exactly what we've been plotting. The log probability density arises for both mathematical convenience and intuitive convenience!

To determine what impact each sample will have on the likelihood, we can just read off the value from the log plot. Consider an outlier that is well away from the rest of the distribution. For a gaussian, the impact on the likelihood will be proportional to $-x^2$. For a logistic distribution, the impact will be roughly proportional to $-x$. This is what is meant when the gaussian is described as "sensitive to outliers." Samples far away from the mean will yank the whole fit around because they have such a large impact on the likelihood.

Log probabilities pervade in Information Theory as well. The negative log probability is called the “surprisal,” and every other quantity is defined in terms of an expected surprisal.

## Tails and Consequences

How much all this matters depends on what type of question you want to answer with your model. If you just want to know where the bulk of the data is, then your choice of distribution doesn’t matter that much. If you are trying to determine the probability of exceptional events, events far away from the mean, then the size of the tails is the only thing that matters. An event six standard deviations from the center of a logistic distribution is several thousand times more likely than an event six standard deviations from the center of a gaussian distribution.

Here are some real world examples where answering an important question requires estimating the probability of an exceptional event.

**A professional chess player has an off game and loses to an amateur. How much should their rating decrease?**

The United States Chess Federation has switched from the gaussian distribution to the logistic in order to make ratings more stable.

**A widely sold security is backed by a large number of mortgages with a low rate of default. What rating should this security receive?**

The financial crisis that brought the world economy to its knees was caused largely by bad statistics. Analysts assumed that the values of mortgages are Gaussian, and that the tails of mortgage values aren’t more correlated than than mortgage values are on average. In reality, the state of the housing market and the overall economy ties defaults together, and thus the value of mortgages together. A large number of them defaulting and losing value is much more likely than a normal distribution would predict. This left the world’s financial institutions completely unprepared when large numbers of them did default at once.

**What are the chances of a magnitude 9.5 earthquake hitting San Francisco in the next 10 years?**

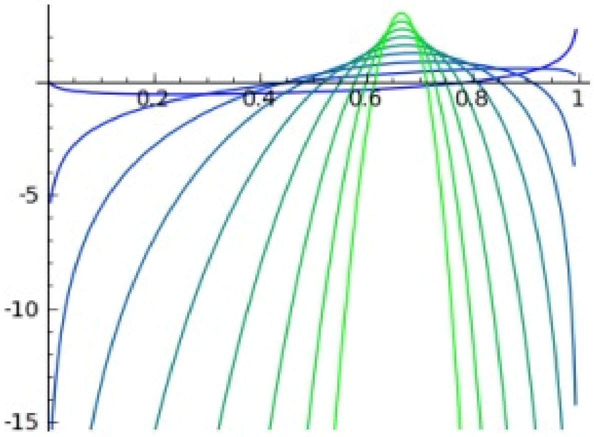
The Gutenberg-Richter law describes the probability distribution of earthquakes of different magnitudes. The log probability of an earthquake is linear in its magnitude. Disagreements about the slope of that line make an exponential difference in the likelihood of a large earthquake.

**A Climate model predicts a 1°C increase in global mean temperature. If the climate were to get several standard deviations hotter than that, the soil would lose the ability to hold moisture, and terrestrial plant life would end.**

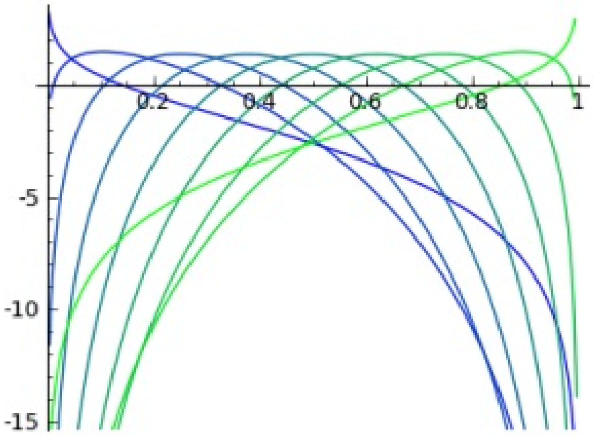
**How likely is the end of the world?**

Unfortunately for us, climate outcomes are fat-tailed.

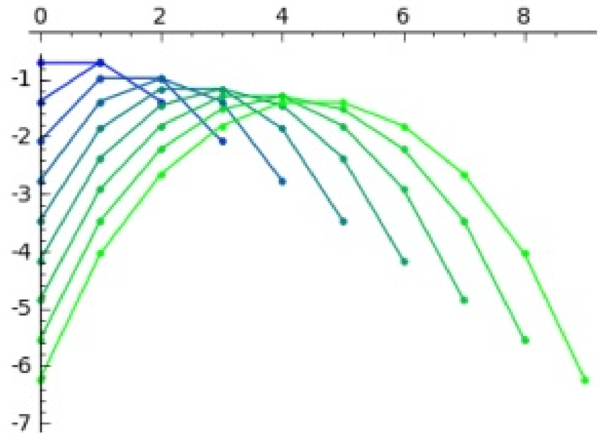
## Menagerie



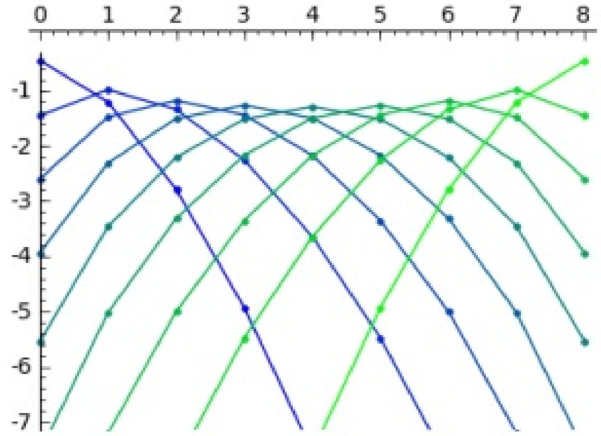
Log of Beta Distributions with the same mean



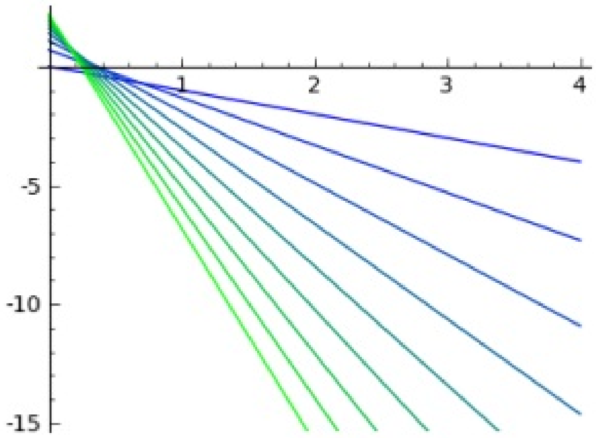
Log of Beta Distributions with the same Variance



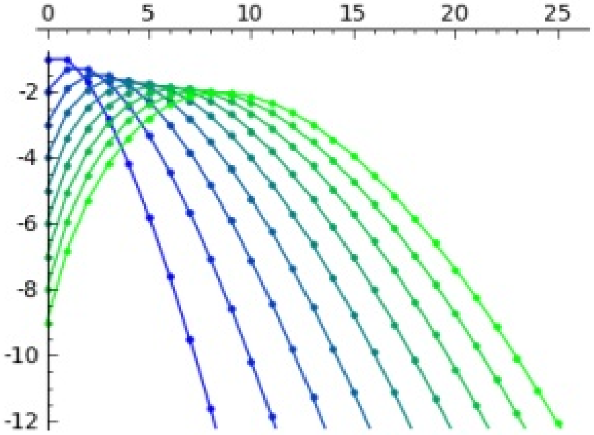
Log of Binomial Distributions with the same p and varying n



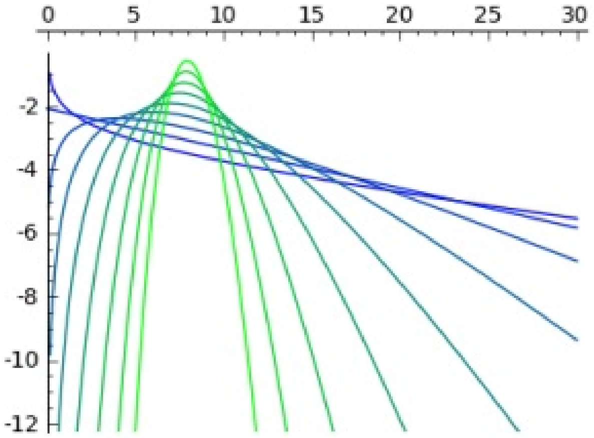
Log of Binomial Distributions with the same n and varying p



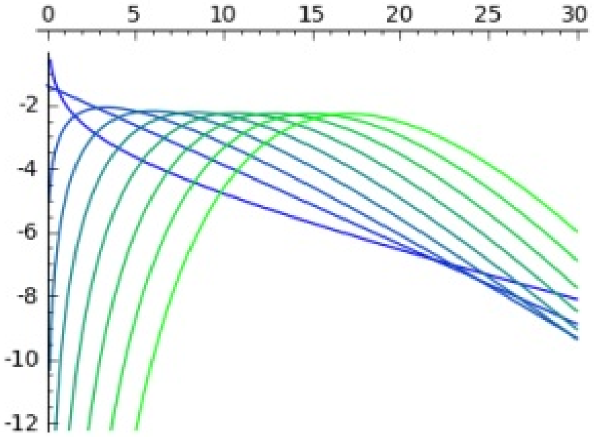
Log of Exponential Distributions



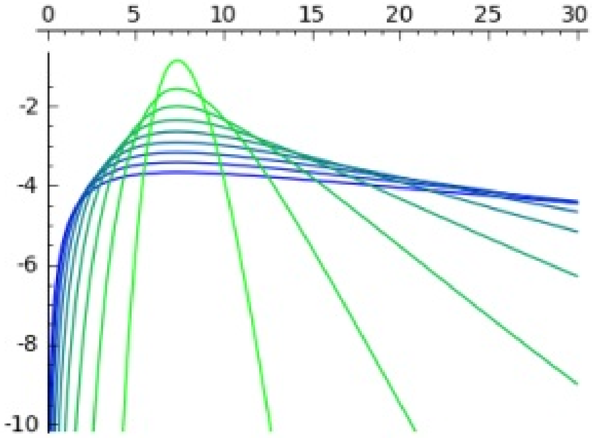
Log of Poisson Distributions



Log of Gamma Distributions with the same mean



Log of Gamma Distributions with the same variance



Log of Log-Normal Distributions with the same mode