Introduction. Probability and distributions

Joe Felsenstein

Genome 560, Spring 2011

Probabilities

We will assume you know that

- Probabilities of mutually exclusive events are summed:

$$\begin{aligned}&\text{Prob (a die comes up 2 or 3)}\\ &= \text{Prob (comes up 2)} + \text{Prob (comes up 3)}\\ &= (1/6) + (1/6) \;=\; 1/3\end{aligned}$$

- Probabilities of independent events are multiplied to get the joint probability:

$$\begin{aligned}&\text{Prob (one die comes up 2 and the other one comes up 3)}\\ &= \text{Prob (first one comes up 2)} \times \text{Prob (second one comes up 3)}\\ &= (1/6) \times (1/6) \;=\; 1/36\end{aligned}$$

- Conditional probabilities are the joint probability divided by the probability of the event that they are conditioned on:

$$\begin{aligned}&\text{Prob (a die comes up 2 given that it comes up even)}\\ &= \text{Prob (comes up 2)}/\text{Prob (it comes up even)}\\ &= (1/6)/(1/2) = 1/3\end{aligned}$$

Stochastic processes (i.e., random processes)

We will think of some simple, fairly easily understood, random processes and analogize biological processes to them. The ones we will use are

1. Tossing of coins (from which we get the binomial distribution, the geometric distribution, and the negative binomial distribution)
2. Tossing a die (singular of dice – the multinomial distribution)
3. Randomly drawing balls of different colors out of an urn (the hypergeometric distribution)
4. Telephone calls coming in on a line at random times (the uniform distribution, the Poisson distribution, the exponential distribution, and the Gamma distribution)
   - The uniform distribution is, when only one call comes in, when that is.
   - The Poisson distribution is the number of calls that come in during a fixed amount of time
   - The exponential distribution is the time until the first call (when you are willing to wait long enough to get it).
   - The Gamma distribution is the time until the k-th call.
5. Change due to Brownian motion in a given time (the Normal distribution and the lognormal distribution), or distribution of a quantity that is a sum of a very large number of very small random quantities (or the product of a very large number of quantities each close to 1).

Tossing coins and the binomial distribution

If we have a coin with Heads probability 0.4 tossed independently 100 times, what is the probability that we get 48 Heads?
The probability of k particular tosses coming up Heads out of n tosses (say TTHTHHTTT...HT) is

$$(1-p)(1-p)p(1-p)pp(1-p)(1-p)(1-p)\dots p(1-p) = p^k(1-p)^{n-k}$$

There are $\binom{n}{k} = n!/(k!(n-k)!)$ different ways to choose k coins to be Heads out of n tosses. So the noting that each of these choices is mutually exclusive, we add up the above probability that many times, so the total probability of all ways of getting k Heads out of n tosses is

$$\binom{n}{k}p^k(1-p)^{n-k} = \frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}$$

Our coin example

which in our numerical example is

$$\binom{100}{48}(0.4)^{48}(1-0.4)^{100-48} = \frac{100!}{48!\ 52!}\ 0.4^{48}\ 0.6^{52}$$

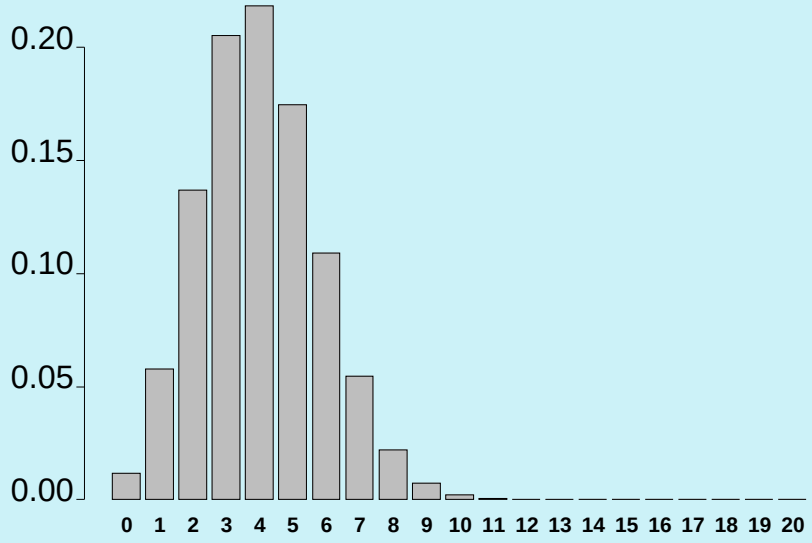
$$= 93,206,558,875,049,876,949,581,681,100$$

$$\times 7.92282 \times 10^{-20} \times 2.90981 \times 10^{-12} = 0.0214878$$

(which really is best done with a computer and/or logarithms).

The histogram of a binomial distribution

This is for $n = 20$ and $p = 0.2$:



A *histogram* is a bar graph whose bars show the probabilities of being at each value on an axis. They are made for discrete variables.

- The bars can be thin enough that there are gaps between them
- ... or they can be fat enough to touch.
- They can be a solid color
- ... or they can have a filler color and a different border.
- The vertical scale can be either number of occurrences of each value found
- ... or they can be the fraction found, or the fraction expected.
- Histograms are often used to group (or “bin”) values from a continuous scale into discrete classes. In doing so they are an approximate reflection of the data.

The Poisson distribution

When there are a vast number of tosses, and each has a tiny probability of Heads, we can have a situation where the expected number of Heads is small but not tiny. The expected number of Heads is $\lambda = n\,p$. The binomial probability becomes, as we consider cases with p smaller and smaller and n larger and larger, holding their product λ constant, finally

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

This is a good model for events such as

- the number of bacterial colonies on a Petri plate (where there are many “places”, each with a small probability of being the center of a colony, but the average (expected) number of colonies λ is modest.
- the number of occurrences of a rare sequence motif such as a restriction site in a stretch of genome (note that this ignores the nonindependence of events in the unlikely case where the prospective restriction sites would overlap).
- the number of radioactive decays in a substance in a given amount of time, where there are vastly many “instants” of time, each with a very small and independent probability that a decay is observed in it.

Proof of the Poisson distribution

(not to be covered in class)

It turns out that because $p = \lambda/n$, the probability of getting k Heads becomes from the binomial distribution formula

$$\frac{n(n-1)(n-2)\ldots(3)(2)(1)}{(n-k)(n-k-1)\ldots(3)(2)(1)\ k!}\ \left(\frac{\lambda}{n}\right)^k\ \left(1-\frac{\lambda}{n}\right)^{n-k}$$

since $(1-\lambda/n) \approx \exp(-\lambda/n)$ which is a good approximation since λ/n is tiny we have

$$\left(1-\frac{\lambda}{n}\right)^{n-k} = \left(1-\frac{\lambda}{n}\right)^{n}\left(1-\frac{\lambda}{n}\right)^{-k} \rightarrow e^{-\lambda}\times 1$$

$$\frac{n}{n}\frac{(n-1)}{n}\frac{(n-2)}{n}\cdots\frac{n-k+1}{n}\frac{1}{k!}\ \lambda^k\ e^{-\lambda}$$

and as n gets large the fraction in front goes to $1/k!$ so that.

$$\frac{1}{k!}\ \lambda^k\ e^{-\lambda}$$

The hypergeometric distribution

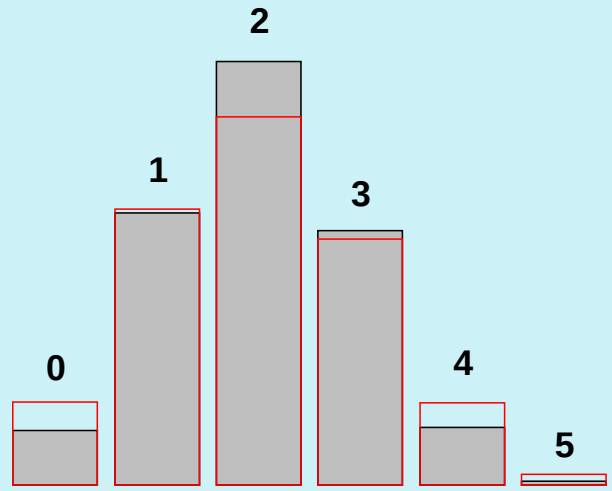
If we have an urn with N balls in it, M of which are red (the rest are white), if we draw n balls out of it, *without replacement* what is the probability that m of those are red?

It turns out to be the fraction, out of all the ways we could choose n balls out of N, in which there are m of them red and $n-m$ of them white:

$$\frac{\binom{M}{m} \times \binom{N-M}{n-m}}{\binom{N}{n}} = \frac{M!(N-M)!n!(N-n)!}{N!m!(M-m)!(n-m)!(N-(n-m))!}$$

Here are histograms when there are few balls in the urn, versus many balls in the urn, each compared with the binomial distribution which in effect chooses *with* replacement:

gray boxes are the hypergeometric distribution, (5 draws out of 8+12), red outlines are the corresponding binomial distribution



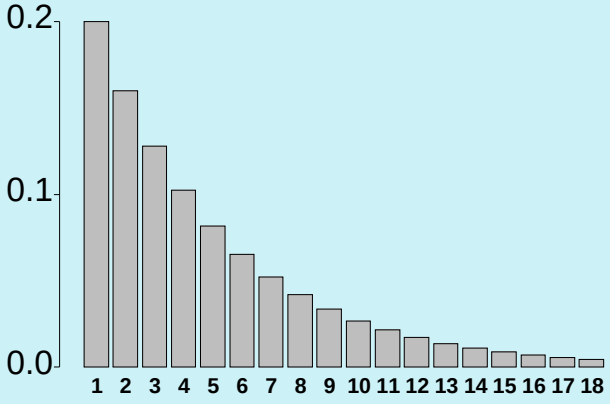
The geometric distribution

Suppose we toss a coin until Heads comes up (for the first time) and count how many tosses it takes. This random variable has a *geometric distribution*.
The probability of k tosses is

$$\underbrace{(1-p)(1-p)(1-p)\ldots(1-p)}_{(k-1)\text{ times}} \quad p$$

$$= (1-p)^{k-1}\, p$$

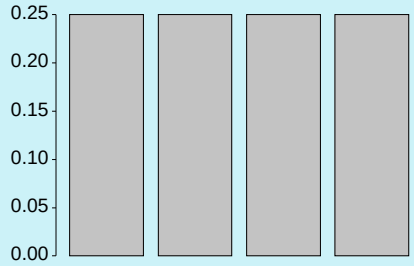
It looks like this:



with each histogram class $(1-p)$ times as high as the one before.

The uniform distribution

Suppose we have a telephone line and we listen to it for one hour. A call comes in at a random time in that hour. How can we make a histogram of when the call comes in? If we divide it into four 15-minute periods the histogram of probabilities of being in them is:

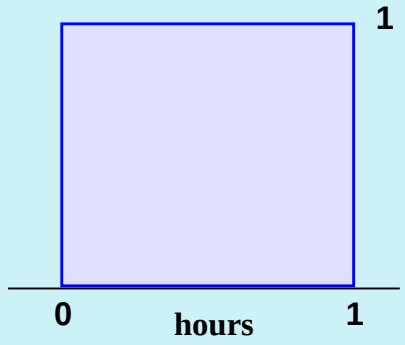


but if we divide it into 60 1-minute blocks we get:

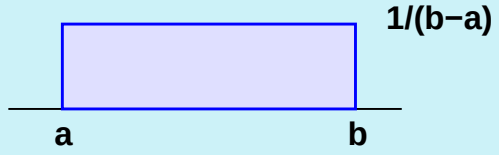
The paradox is that the more finely we record the data (and divide the histogram) the lower is the probability of each class. If we record exact times, we end up with zillions of bars, all of height zero!

Continuous distributions

The solution is to not record probabilities of being “at” a value, but a *density function* which shows the relative probabilities of being in different regions, scaled so that the area under it is 1. Then we can use it to compute the probability of being in any interval, by integrating the function between those bounds.



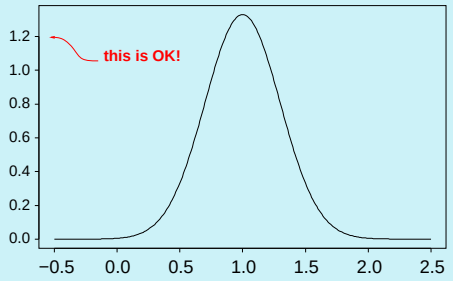
and here is that distribution more generally, the *uniform distribution* (or *uniform density*) between a and b:



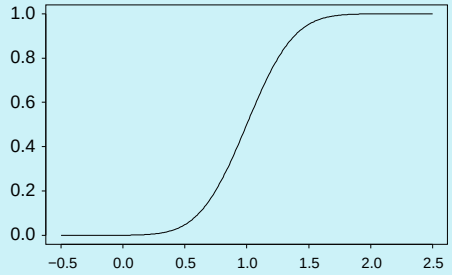
Cumulative distribution functions

We can also plot the total fraction (or total number) of values that are less than or equal to a value. This gives a *cumulative distribution function* that rises from (e.g.) 0 to 1 as you go left to right.

Here is a density function

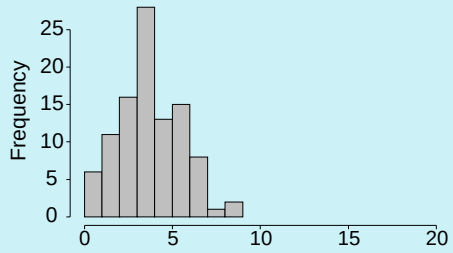


and its cumulative distribution function

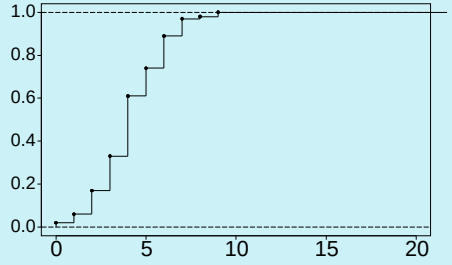


They can also be used for histograms and even for mixtures of variables that are partly discrete and partly continuous (not shown here).

Here is an empirical histogram



and its cumulative distribution function



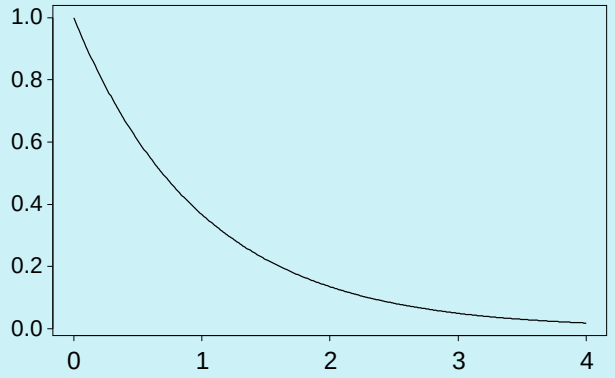
The exponential distribution

If we have a time axis sliced into very fine segments, and have a geometric distribution, in the limit as the slices get finer and finer (with the same probability of Heads per unit time) we get an *exponential distribution*.

Its density function is (when λ is the rate at which events occur per unit time):

$$\lambda e^{-\lambda t}$$

The density function of the exponential distribution is very similar to a histogram of the geometric distribution in that it decays exponentially. It starts at 0 and goes out toward ∞.

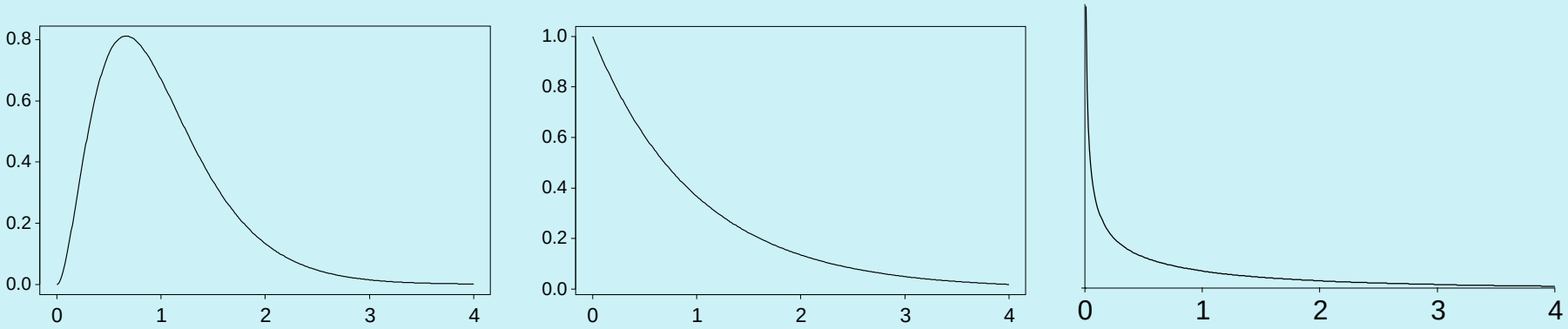


All exponential distributions are the same shape, just stretched out horizontally to different extents.

The Gamma distribution

The Gamma distribution is the waiting time to the k -th telephone call. (It is related to the Poisson in this way: if we wait a fixed amount of time, each little chunk of time is like the toss of a coin with a very small probability of Heads, and we receive a Poisson number of telephone calls; but if instead we wait until the k-th call comes, the waiting time is Gamma-distributed).

Three Gamma densities for different parameters:

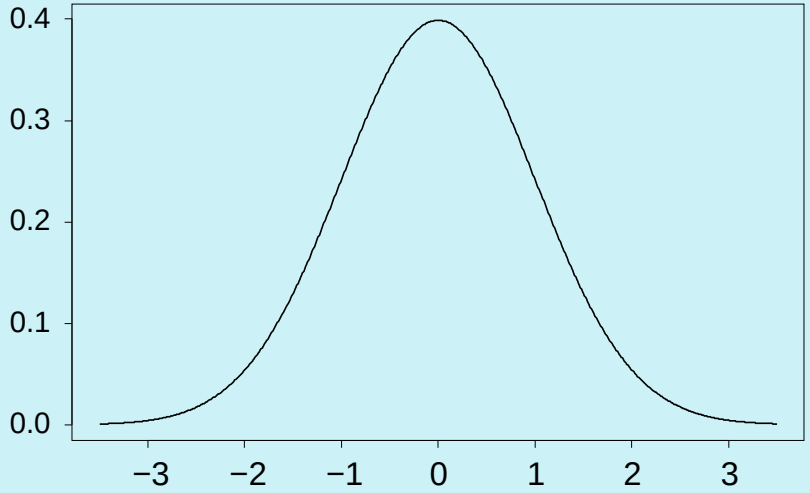


The Gamma has two parameters, a scale parameter and a shape parameter. The rate of phone calls expected and the number of the call you are waiting until set these. But the parameters are continuous and so also you can get a Gamma density for fractional phone calls too, in effect.

The Normal distribution

When a quantity is the sum of a very large number of independent random variables (such as displacements in a given time in Brownian motion), no one of which has a very large effect, there are mathematical proofs (Central Limit Theorems) that the sum is distributed in a *Normal Distribution* or *Gaussian Distribution*

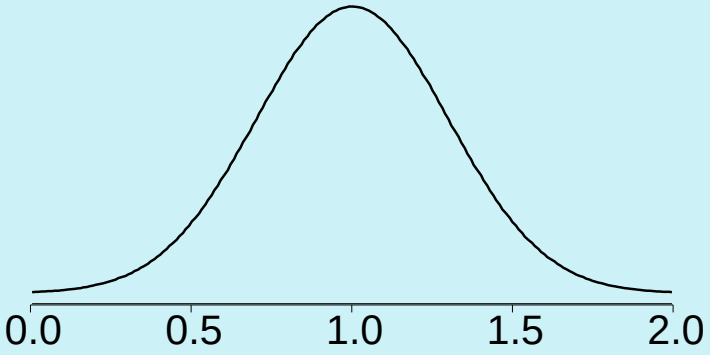
Here is the density function of a Normal distribution with mean 0 and standard deviation 1 *(more on that later)*. They can be scaled to have mean (center) anywhere and any width (except of course the height has to adjust so that the area under the density is 1).



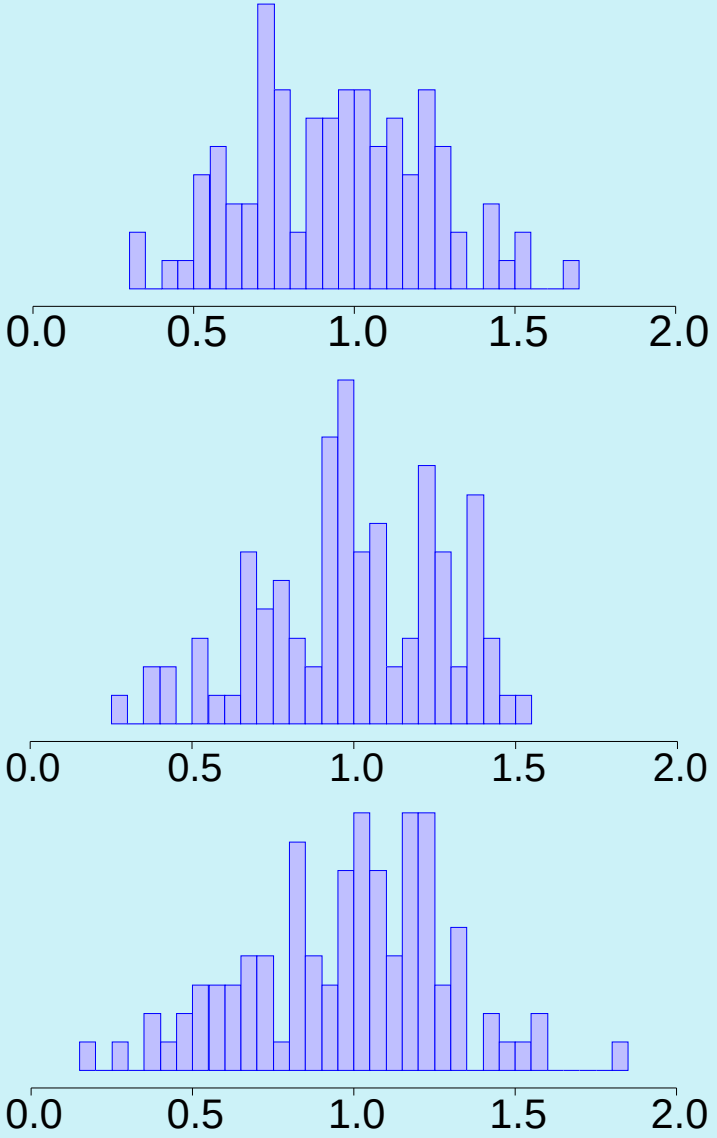
(It is said that biologists believe in the Normal distribution, thinking that mathematicians have proven it to be the correct one, while mathematicians think biologists have shown empirically that it is the correct one).

Probability vs. statistics

The true $\mathrm{N}(1, 0.09)$ distribution



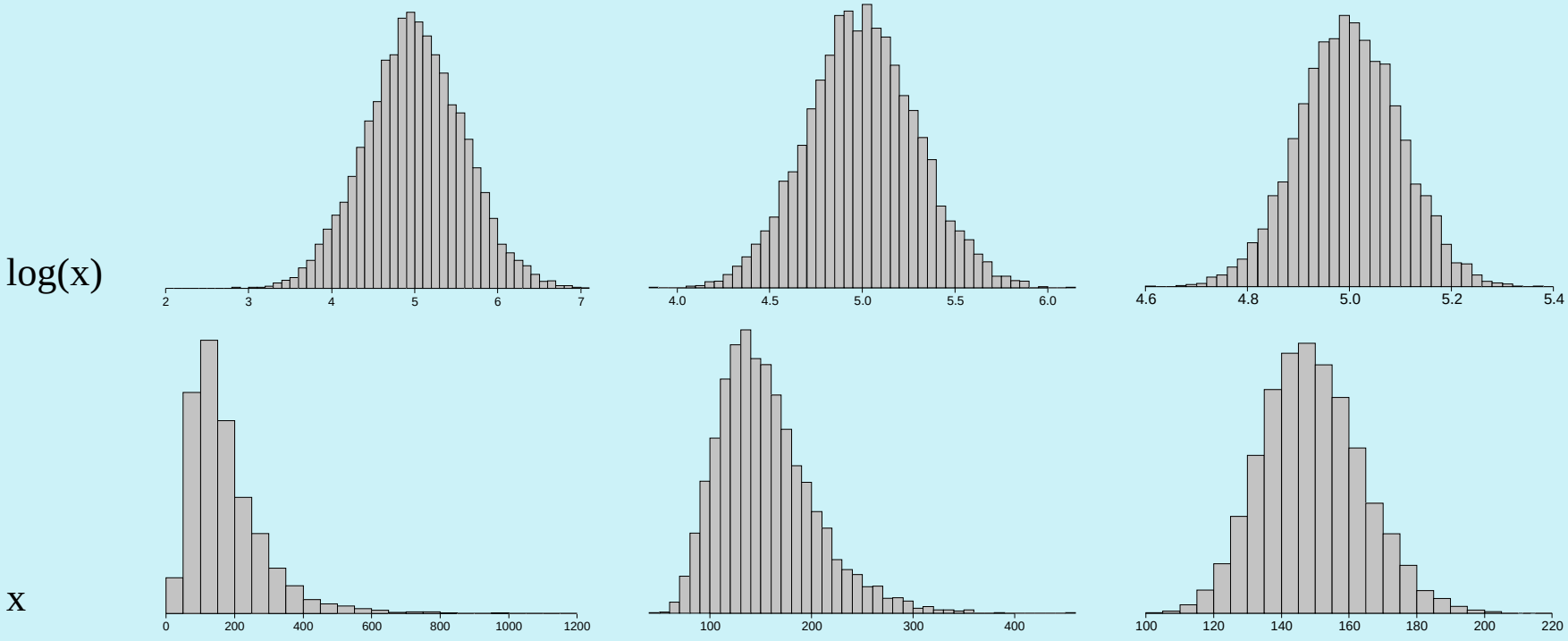
three samples of 100 points



The lognormal distribution

Plainly and simply, the logarithm of the quantity is normally distributed. (Which log, natural or common? As they're multiples of each other, both).

This is a natural distribution for measurements that might be a product of random quantities, rather than a sum. It cannot be negative. Here are three lognormal distributions, each on both scales:



Note that as the spread on the $\log(x)$ scale gets less, the asymmetry on the x scale becomes less. Many quantities in nature are better approximated by lognormal distributions than by normal distributions.