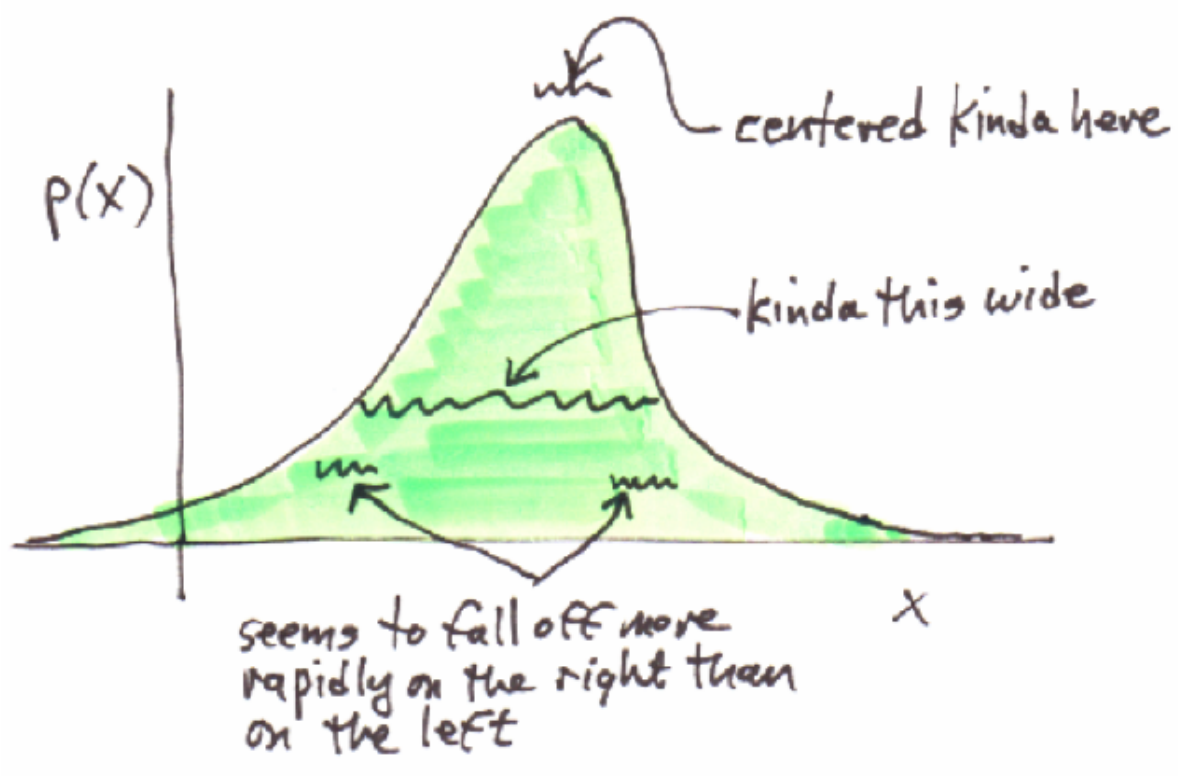
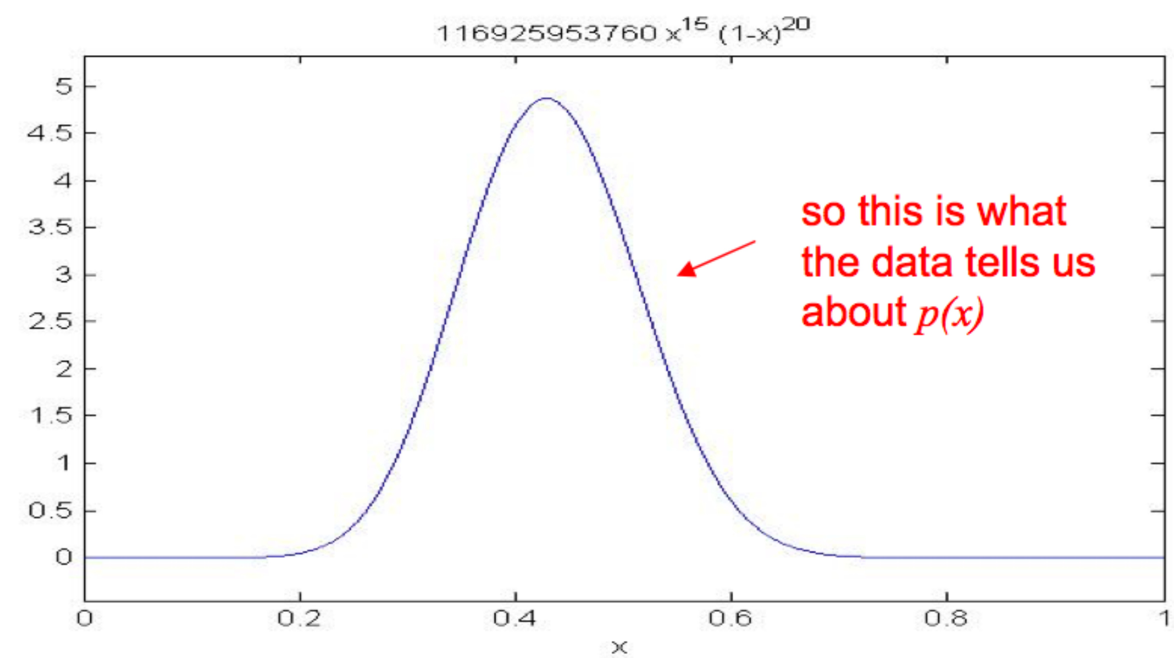
Lectures 5: central limit theorem

distribution functions:

We are often interested in distributions that have some kind of localization (because why would we be interested if they didn't?)





distribution on [0,1] interval

Suppose we want to <u>summarize</u> *p(x)* by a single number *a*, its "value". Let's find the value *a* that <u>minimizes the mean-square discrepancy of the "typical" value</u> *x*:

distribution functions:

Recall expectation notation:

$$\langle \text{anything} \rangle \equiv \int_x (\text{anything})\, p(x) dx$$

i.e., the weighted average of “anything”, weighted by the probable values of x.
Expectation is linear over “anything” (sums, constants times, etc.).

$$\begin{aligned} \text{minimize:} \quad \Delta^2 &\equiv \left\langle (x-a)^2 \right\rangle = \left\langle x^2 - 2ax + a^2 \right\rangle \\ &= \left(\left\langle x^2 \right\rangle - \langle x \rangle^2 \right) + \left(\langle x \rangle - a \right)^2 \end{aligned}$$

This is the variance Var(x),
but all we care about here is
that it doesn’t depend on a.

(in physics this is called the “parallel axis theorem”)

The minimum is obviously $a = \langle x \rangle$. (Take derivative wrt a and set to zero if you like mechanical calculations.)

distribution functions:

Why mean-square? Why not mean-absolute? Try it!

$$\Delta = \langle |x-a| \rangle = \int_{-\infty}^{\infty} |x-a|\, p(x)dx$$

$$= \int_{-\infty}^{a} (a-x)\, p(x)dx + \int_{a}^{\infty} (x-a)\, p(x)dx$$

So,

$$0 = \frac{d\Delta}{da} = \int_{-\infty}^{a} p(x)dx + 0 - \int_{a}^{\infty} p(x)dx + 0$$

Integrand at a

$$\Rightarrow$$

$$\int_{-\infty}^{a} p(x)dx = \int_{a}^{\infty} p(x)dx = \tfrac{1}{2}$$

$\Rightarrow a$ is the median value

Mean and median are both “measures of central tendency”.

distribution functions:

Higher moments, centered moments are conventionally defined by

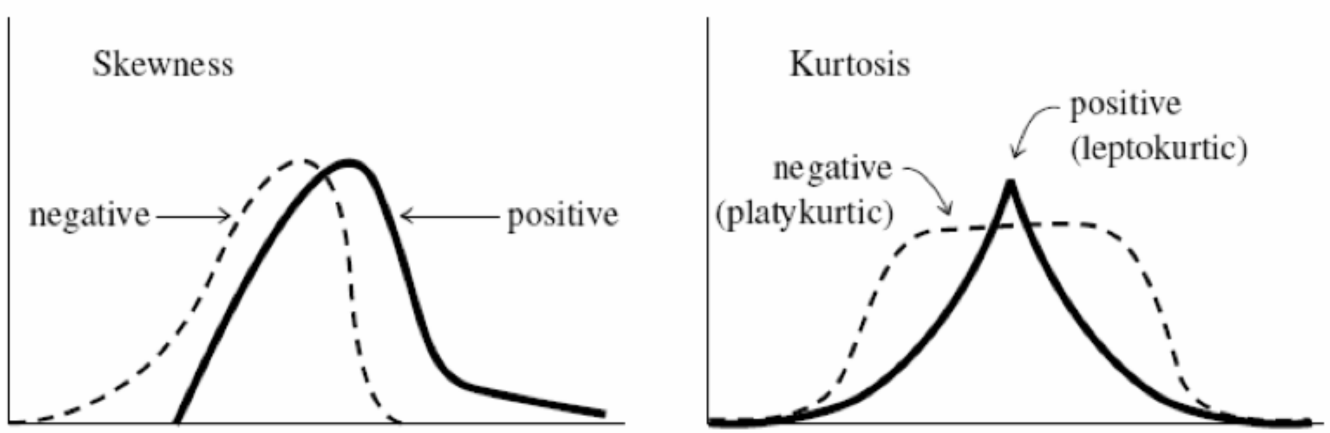
$$\mu_i \equiv \langle x^i \rangle = \int x^i \, p(x) dx$$
$$M_i \equiv \langle (x - \langle x \rangle)^i \rangle = \int (x - \langle x \rangle)^i \, p(x) dx$$

The centered second moment M_2, the variance, is by far most useful

$$M_2 \equiv \mathrm{Var}(x) \equiv \langle (x - \langle x \rangle)^2 \rangle = \langle x^2 \rangle - \langle x \rangle^2$$
$$\sigma(x) \equiv \sqrt{\mathrm{Var}(x)}$$

"standard deviation" summarizes a distribution's half-width (r.m.s. deviation from the mean)

Third and fourth moments also have "names"

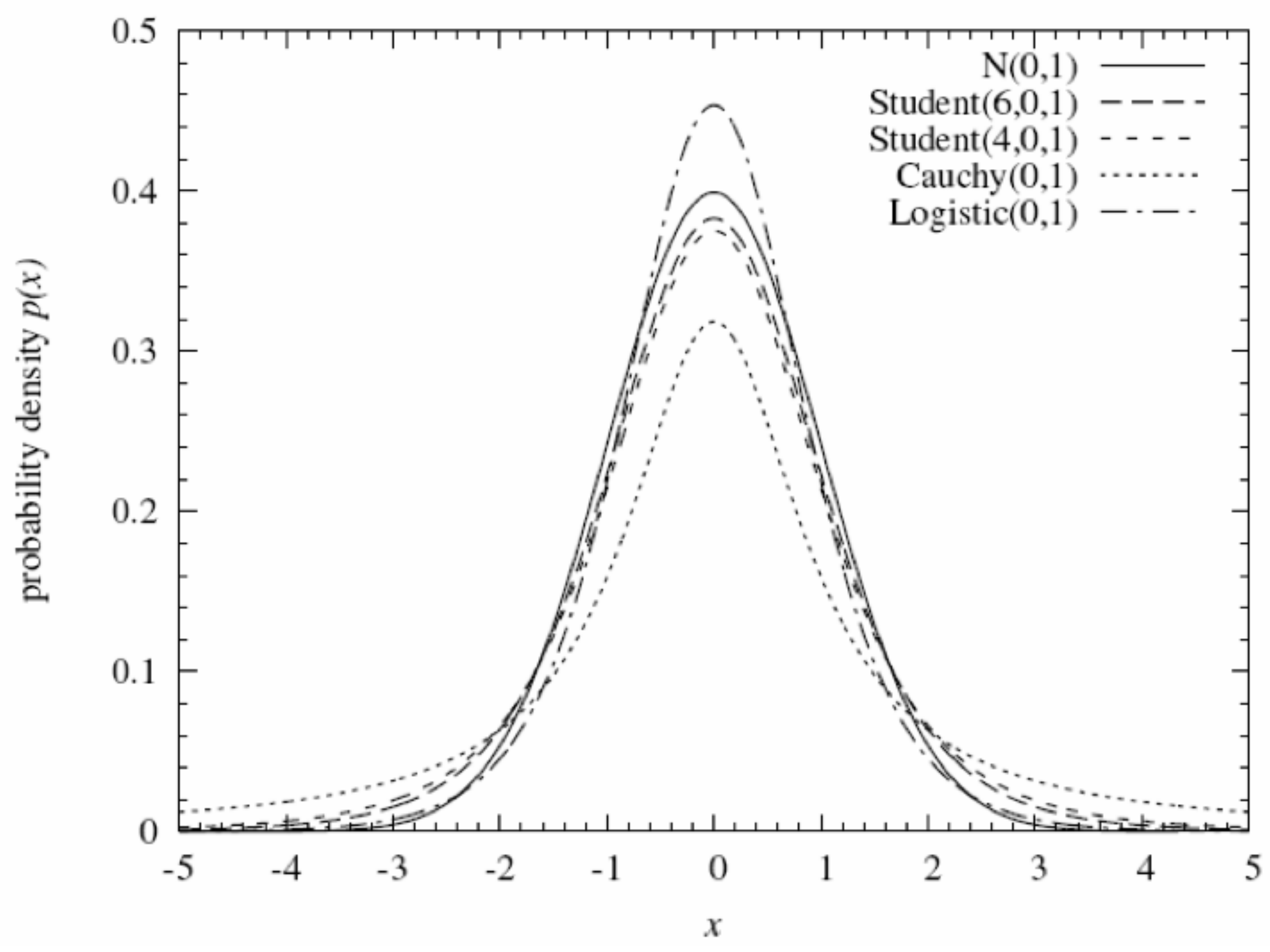


But generally wise to be cautious about using high moments. Otherwise perfectly good distributions don't have them at all (divergent). And (related) it can take a lot of data to measure them accurately.

distribution functions:

Let us review some standard (i.e., frequently occurring) distributions:

The “bell shaped” ones differ qualitatively by their tail behaviors:



distribution functions:

Normal (Gaussian) has the fastest falling tails:

$$x \sim \mathrm{N}(\mu, \sigma), \qquad \sigma > 0$$
$$p(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{1}{2}\left[\frac{x-\mu}{\sigma}\right]^2\right)$$

Cauchy (aka Lorentzian) has the slowest falling tails:

$$x \sim \mathrm{Cauchy}(\mu, \sigma), \qquad \sigma > 0$$
$$p(x) = \frac{1}{\pi\sigma}\left(1 + \left[\frac{x-\mu}{\sigma}\right]^2\right)^{-1}$$

Cauchy has area=1 (zeroth moment), but no defined mean or variance (1st and 2nd moments divergent).

characteristic function:

The Central Limit Theorem is the reason that the Normal (Gaussian) distribution is uniquely important. We need to understand where it does and doesn't apply.

The characteristic function of a distribution is its Fourier transform.

$$\phi_X(t) \equiv \int_{-\infty}^{\infty} e^{itx} p_X(x) dx$$

(Statisticians often use notational convention that *X* is a random variable, *x* its value, $p_X(x)$ its distribution.)

$$\phi_X(0) = 1$$

$$\phi_X'(0) = \int ix p_X(x) dx = i\mu$$

$$-\phi_X''(0) = \int x^2 p_X(x) dx = \sigma^2 + \mu^2$$

So, the coefficients of the Taylor series expansion of the characteristic function are the (uncentered) moments.

$$\phi_{\text{Normal}}(t) = e^{i\mu t - \frac{1}{2}\sigma^2 t^2}$$

characteristic function:

Addition of independent r.v.'s:

$$\text{let } S = X + Y$$

$$p_S(s) = \int p_X(u) p_Y(s-u) du$$

$$\phi_S(t) = \phi_X(t) \phi_Y(t)$$

Last line follows immediately from the Fourier convolution theorem. (In fact, it <u>is</u> the Fourier convolution theorem!)

distribution functions:

Proof of convolution theorem:

$$\left.\begin{aligned}
\phi_X(t) &\equiv \int_{-\infty}^{\infty} e^{itx} p_X(x)dx \\
p_X(x) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \phi_X(t)e^{-itx}dt
\end{aligned}\right\} \text{Fourier transform pair}$$

$$\begin{aligned}
p_S(s) &= \int_{-\infty}^{\infty} p_X(u)p_Y(s-u)du \\
&= \int_{-\infty}^{\infty} p_X(u)\left[\frac{1}{2\pi}\int_{-\infty}^{\infty} \phi_Y(t)e^{-it(s-u)}dt\right]du \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty} \phi_Y(t)e^{-its}\left[\int_{-\infty}^{\infty} p_X(u)e^{itu}du\right]dt \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty} \phi_Y(t)\phi_X(t)e^{-its}dt
\end{aligned}$$

So, $\quad \phi_S(t) = \phi_Y(t)\phi_X(t)$

distribution functions:

Mean and variance are additive over independent random variables:

$$\overline{(x+y)} = \bar{x} + \bar{y} \qquad \mathrm{Var}(x+y) = \mathrm{Var}(x) + \mathrm{Var}(y)$$

note “bar” notation, equivalent to < >

Certain combinations of higher moments are also additive. These are called semi-invariants.

$$I_2 = M_2 \qquad I_3 = M_3 \qquad I_4 = M_4 - 3M_2^2$$
$$I_5 = M_5 - 10M_2M_3 \qquad I_6 = M_6 - 15M_2M_4 - 10M_3^2 + 30M_2^3$$

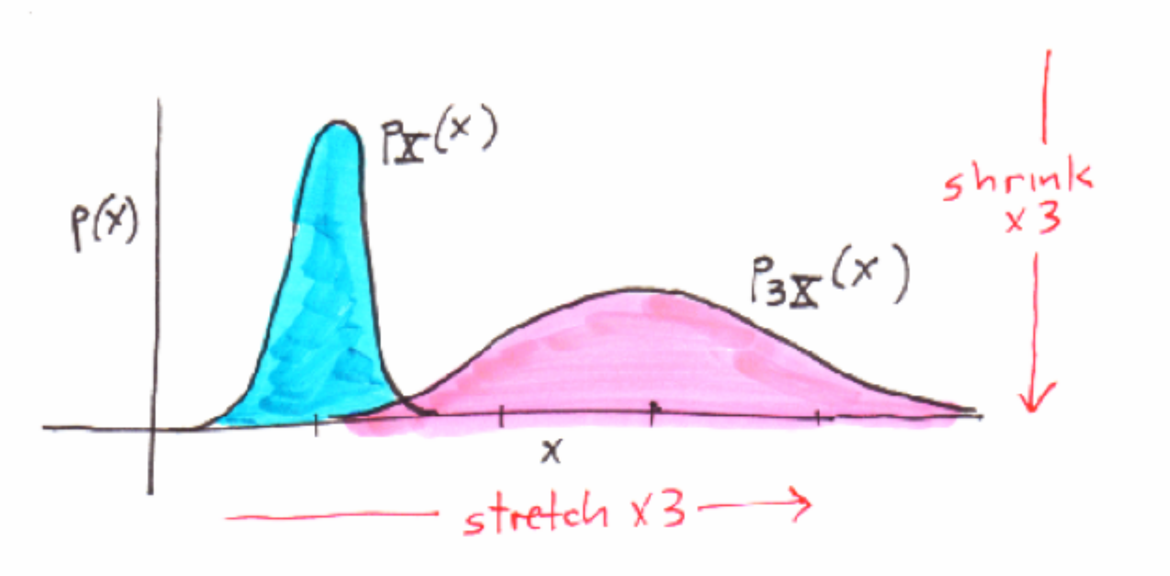
Skew and kurtosis are dimensionless combinations of semi-invariants

$$\mathrm{Skew}(x) = I_3/I_2^{3/2} \qquad \mathrm{Kurt}(x) = I_4/I_2^2$$

A Gaussian has all of its semi-invariants higher than I_2 equal to zero.
A Poisson distribution has all of its semi-invariants equal to its mean.

characteristic function:

Scaling law for r.v.'s:



Scaling law for characteristic functions:

$$
\begin{aligned}
\phi_{aX}(t) &= \int e^{itx} \underline{p_{aX}(x)} dx \\
&= \int e^{itx} \underline{\frac{1}{a} p_X \left(\frac{x}{a}\right)} dx \\
&= \int e^{i(at)(x/a)} p_X \left(\frac{x}{a}\right) \frac{dx}{a} \\
&= \phi_X(at)
\end{aligned}
$$

central limit theorem:

Let $S = \frac{1}{N}\sum X_i = \sum \frac{X_i}{N}$ with $\langle X_i \rangle \equiv 0$

Can always subtract off the means, then add back later.

Then

$$\phi_S(t) = \prod_i \phi_{X_i/N}(t) = \prod_i \phi_{X_i}\left(\frac{t}{N}\right)$$

$$= \prod_i \left(1 - \tfrac{1}{2}\sigma_i^2 \frac{t^2}{N^2} + \cdots\right)$$

Whoa! It better have a convergent Taylor series around zero! (Cauchy doesn't, e.g.)

$$= \exp\left[\sum_i \ln\left(1 - \tfrac{1}{2}\sigma_i^2 \frac{t^2}{N^2} + \cdots\right)\right]$$

These terms decrease with N, but how fast?

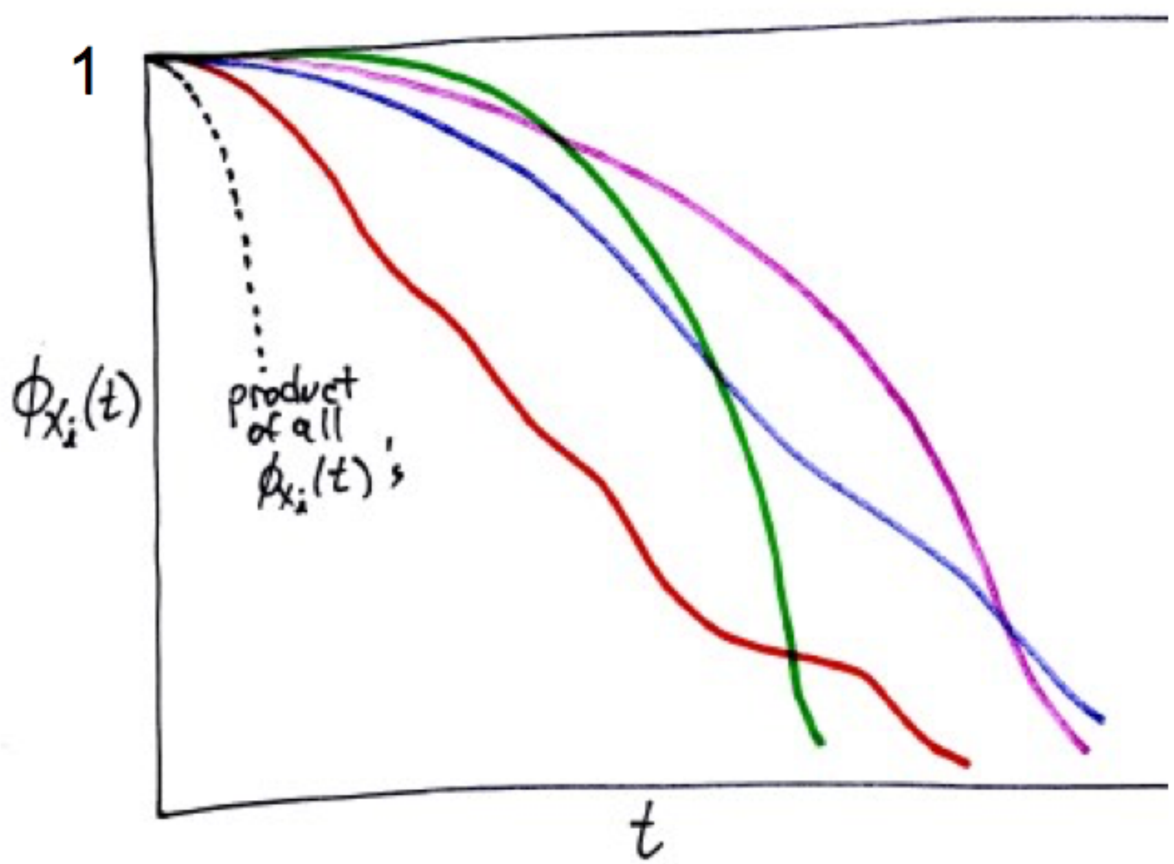
$$\approx \exp\left[-\frac{1}{2}\left(\frac{1}{N^2}\sum_i \sigma_i^2\right) t^2 + \cdots\right]$$

So, S is normally distributed

$$p_S(\cdot) \sim \text{Normal}(0, \tfrac{1}{N^2}\sum \sigma_i^2)$$

central limit theorem:

Intuitively, the product of a lot of arbitrary functions that all start at 1 and have zero derivative looks like this:



Because the product falls off so fast, it loses all memory of the details of its factors except the starting value 1 and fact of zero derivative. In characteristic function space that's basically the CLT.

central limit theorem:

CLT is usually stated about the sum of RVs, not the average, so

$$p_S(\cdot) \sim \mathrm{Normal}(0, \tfrac{1}{N^2} \sum \sigma_i^2)$$

Now, since

$$NS = \sum X_i \quad \text{and} \quad \mathrm{Var}(NS) = N^2 \mathrm{Var}(S)$$

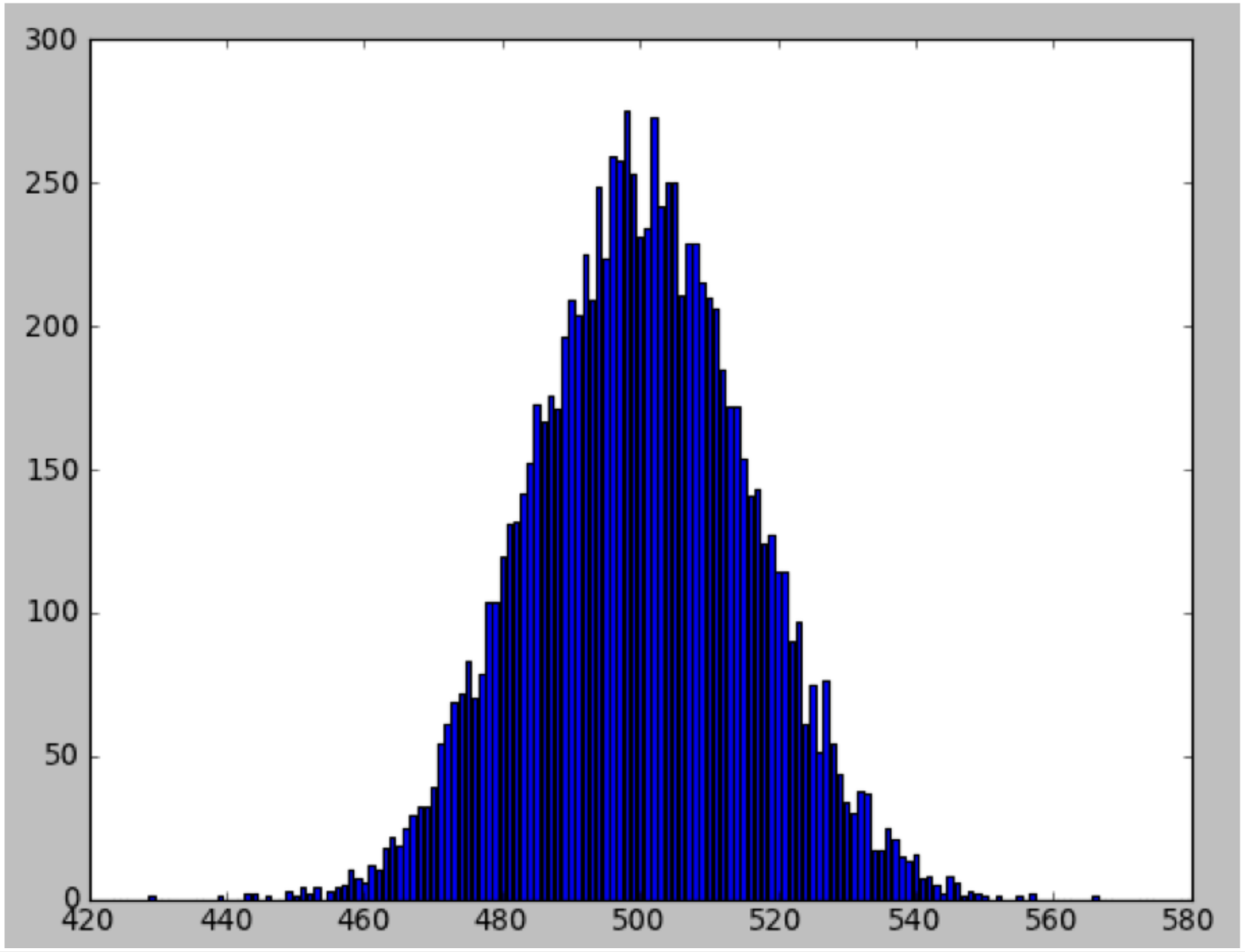
it follows that the simple sum of a large number of r.v.'s is normally distributed, with variance equal to the sum of the variances:

$$p_{\sum x_i}(\cdot) \sim \mathrm{Normal}(0, \sum \sigma_i^2)$$

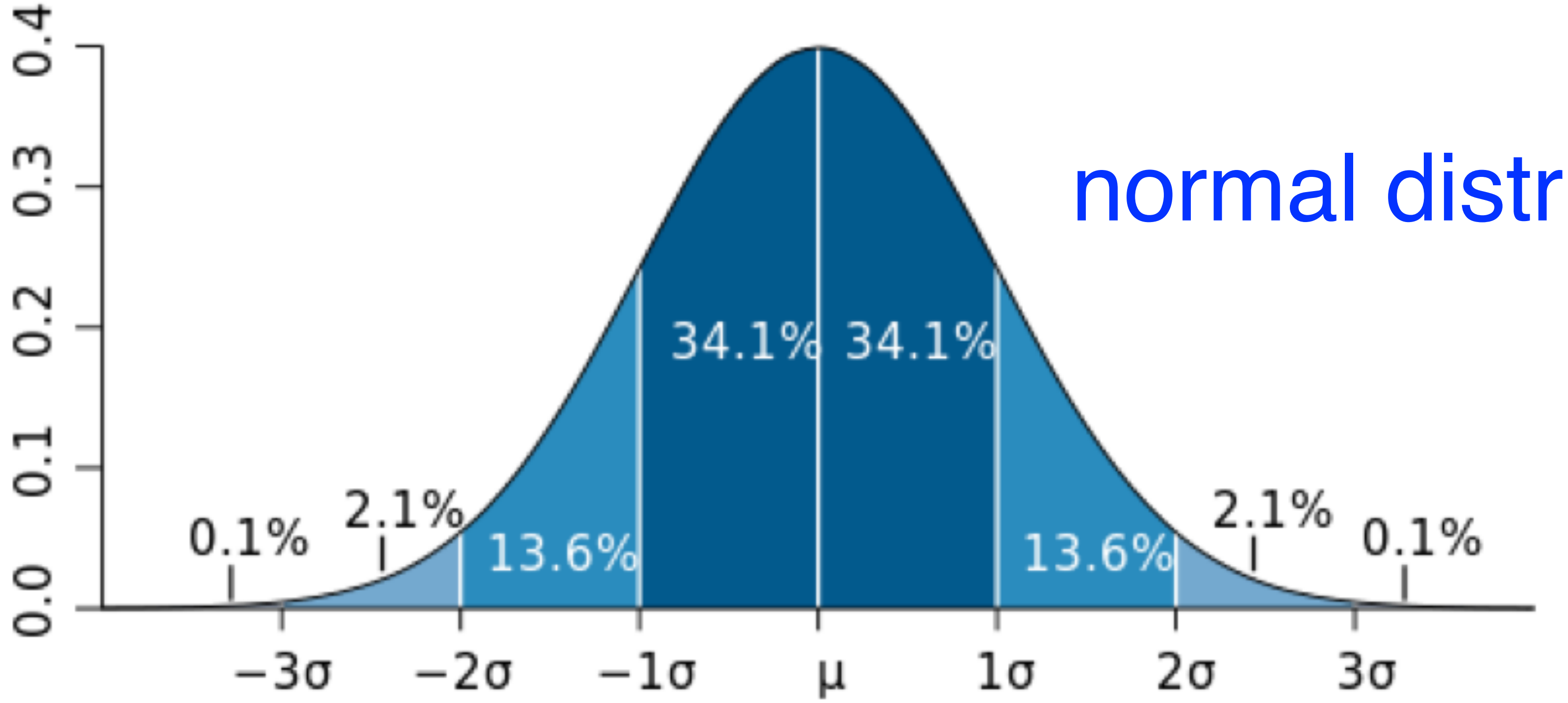
If N is large enough, and if the higher moments are well-enough behaved, and if the Taylor series expansion exists!

Also beware of borderline cases where the assumptions technically hold, but convergence to Normal is slow and/or highly nonuniform. (This can affect p-values for tail tests, as we will soon see.)

central limit theorem:



10,000 trials of 1,000 tosses



normal distribution (bell curve)

central limit theorem:

Since Gaussians are so universal, let's learn estimate the parameters μ and σ of a Gaussian from a set of points drawn from it:

For now, we'll just find the maximum of the posterior distribution of (μ,σ), given some data, for a uniform prior. This is called "maximum a posteriori (MAP)" by Bayesians, and "maximum likelihood (MLE)" by frequentists.

The data is: $x_i,\ i = 1, \ldots, N$

The statistical model is: $$P(\mathbf{x}|\mu,\sigma) = \prod_i \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}}$$

The posterior estimate is: $$P(\mu,\sigma|\mathbf{x}) \propto \frac{1}{\sqrt{2\pi}\sigma^N} e^{-\frac{1}{2\sigma^2}\sum_i (x_i-\mu)^2} \times \cancel{P(\mu,\sigma)}$$ uniform

Now find the MAP (MLE):

$$0 = \frac{\partial P}{\partial \mu} = \frac{P}{\sigma^3}\Big(\sum_i x_i - N\mu\Big) \Rightarrow \mu = \frac{1}{N}\sum_i x_i$$

Ha! The MAP mean is the sample mean, the MAP variance is the sample variance!

$$0 = \frac{\partial P}{\partial \sigma} = \frac{P}{\sigma^4}\Big[-N\sigma^2 + \sum_i (x_i-\mu)^2\Big] \Rightarrow \sigma^2 = \frac{1}{N}\sum_i (x_i-\mu)^2$$