# Computational Statistics with Application to Bioinformatics

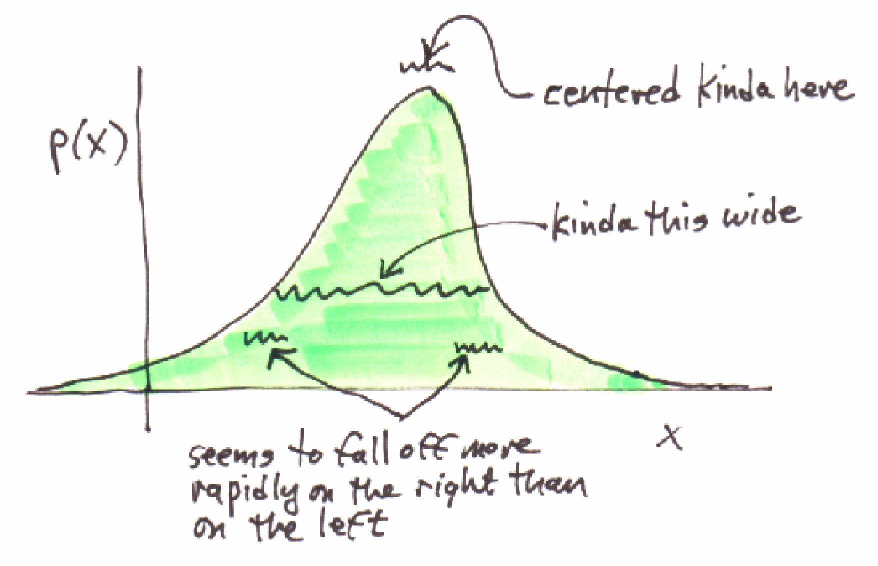
Prof. William H. Press
Spring Term, 2008
The University of Texas at Austin

## Unit 2: Univariate Distributions and the Central Limit Theorem

## Unit 2: Univariate Distributions and the Central Limit Theorem (Summary)

- Two “measures of central tendency”
  - mean (minimize square deviation)
  - median (minimize absolute deviation)
  - discuss higher moments and semi-invariants
- Some common distributions on the real line
  - Normal (Gaussian)
  - Student
- Some common distributions on the positive real line
  - Lognormal
  - Exponential
  - Gamma
  - Chi-square
- Example of an NR3 class for a distribution function
  - PDF, CDF, inverse CDF
- Properties of any distribution’s characteristic function
  - moments appear in Taylor series
  - product for sum of random variables
  - scaling laws
- Compute the characteristic function for the Normal and Cauchy distributions
  - Cauchy has divergent moments and (related) has nonanalytic CF
- Prove the Central Limit Theorem
  - and see how various assumptions come into it

Many, though not all, common distributions look sort-of like this:




We just saw the beta distribution with $\alpha, \beta > 0$ as an example on the interval [0,1]. We'll see more examples in a few minutes.

Suppose we want to summarize $p(x)$ by a single number $a$, its "center". Let's find the value $a$ that minimizes the mean-square distance of the "typical" value $x$:

expectation notation:
$\langle \text{anything} \rangle \equiv \int_x (\text{anything})\, p(x) dx$
expectation is linear, etc.

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

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.)

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

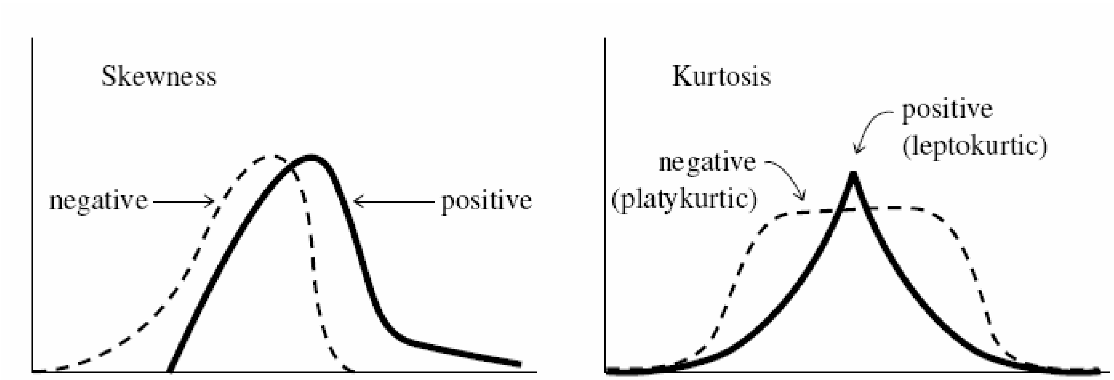
$$\Rightarrow$$
$$\int_{-\infty}^{a} p(x) dx = \int_{a}^{\infty} p(x) dx = \tfrac{1}{2}$$
$$\Rightarrow a \text{ is the median value}$$

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

Higher moments, centered moments defined by

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

Semi-invariants are combinations of higher moments that are additive, like the variance ($M_2$)




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.

## 14.1.1 Semi-Invariants

The mean and variance of independent random variables are additive: If $x$ and $y$ are drawn independently from two, possibly different, probability distributions, then

$$\overline{(x+y)} = \overline{x} + \overline{y} \qquad \text{Var}(x+y) = \text{Var}(x) + \text{Var}(x) \tag{14.1.9}$$

Higher moments are not, in general, additive. However, certain combinations of them, called *semi-invariants*, are in fact additive. If the centered moments of a distribution are denoted $M_k$,

$$M_k \equiv \left\langle (x_i - \overline{x})^k \right\rangle \tag{14.1.10}$$

so that, e.g., $M_2 = \text{Var}(x)$, then the first few semi-invariants, denoted $I_k$, are given by

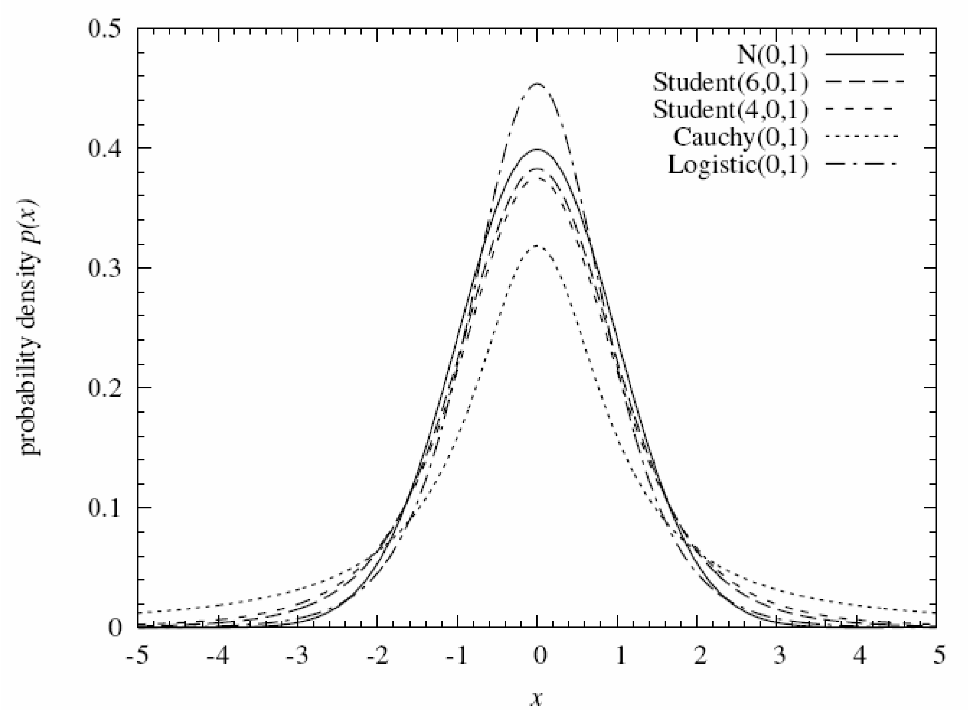
$$\begin{gathered} I_2 = M_2 \qquad I_3 = M_3 \qquad I_4 = M_4 - 3M_2^2 \\ I_5 = M_5 - 10M_2M_3 \qquad I_6 = M_6 - 15M_2M_4 - 10M_3^2 + 30M_2^3 \end{gathered} \tag{14.1.11}$$

Notice that the skewness and kurtosis, equations (14.1.5) and (14.1.6), are simple powers of the semi-invariants,

$$\text{Skew}(x) = I_3/I_2^{3/2} \qquad \text{Kurt}(x) = I_4/I_2^2 \tag{14.1.12}$$

A Gaussian distribution has all its semi-invariants higher than $I_2$ equal to zero. A Poisson distribution has all of its semi-invariants equal to its mean. For more details, see [2].

Univariate distributions to know and love on the real line:




Normal (Gaussian):

$$x \sim \mathrm{N}(\mu, \sigma), \qquad \sigma > 0$$

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

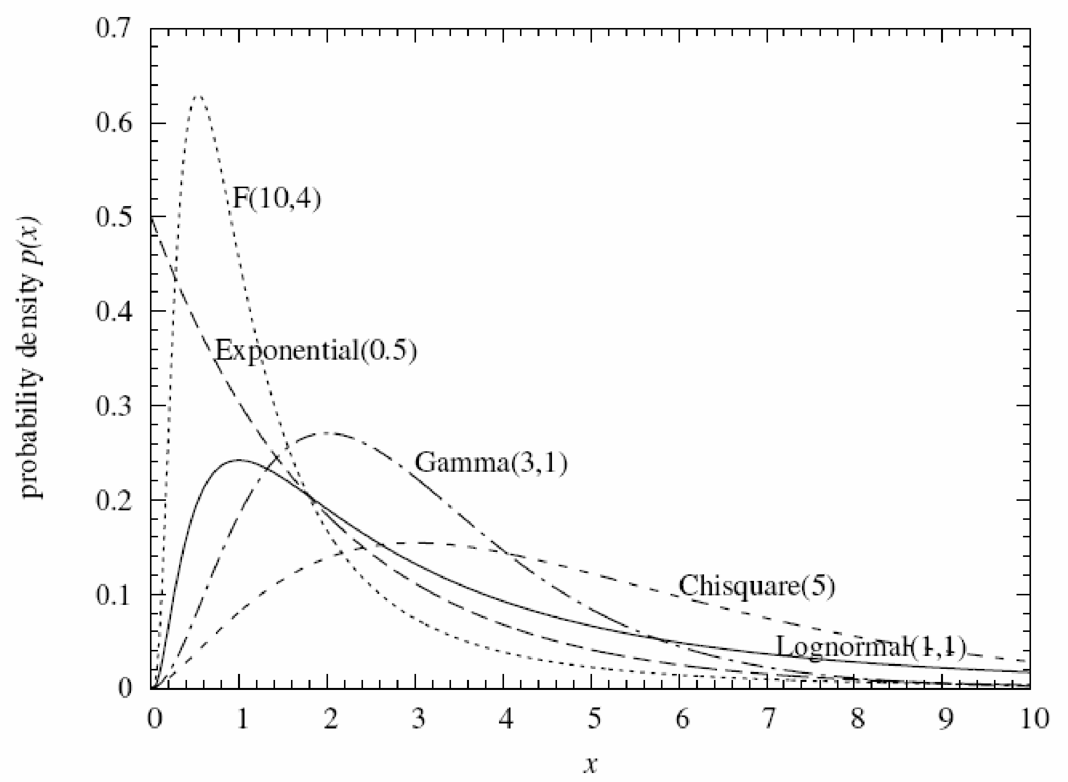
Student:

$$t \sim \text{Student}(\nu, \mu, \sigma), \qquad \nu > 0,\ \sigma > 0$$

$$p(t) = \frac{\Gamma(\frac{1}{2}[\nu + 1])}{\Gamma(\frac{1}{2}\nu)\sqrt{\nu\pi}\sigma} \left(1 + \frac{1}{\nu}\left[\frac{t-\mu}{\sigma}\right]^2\right)^{-\frac{1}{2}(\nu+1)}$$

“bell shaped” but you get to specify the power with which the tails fall off. Normal and Cauchy are limiting cases. (Also occurs in some statistical tests.)

Common distributions on positive real line:




Exponential:

$$x \sim \text{Exponential}(\beta), \qquad \beta > 0$$
$$p(x) = \beta \exp(-\beta x), \qquad x > 0$$

Lognormal:

$$x \sim \text{Lognormal}(\mu, \sigma), \qquad \sigma > 0$$

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma x} \exp\left(-\frac{1}{2}\left[\frac{\log(x)-\mu}{\sigma}\right]^2\right), \qquad x > 0 \tag{6.14.31}$$

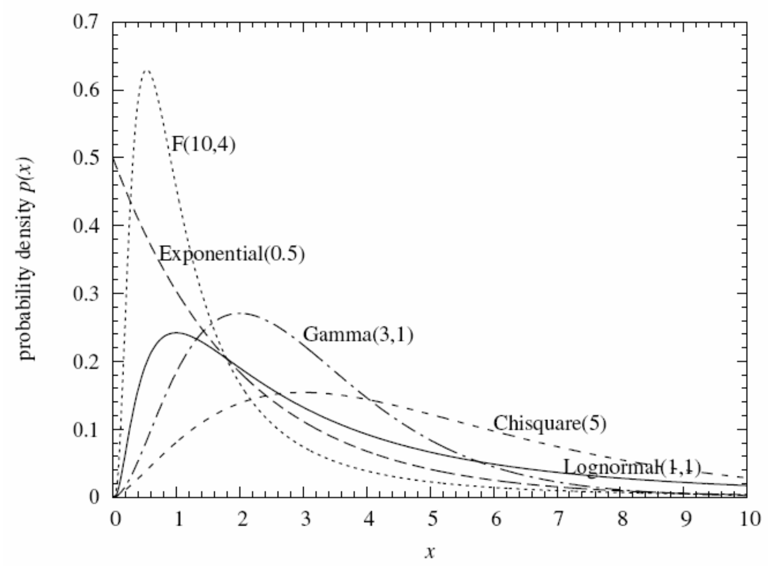
Note the required extra factor of $x^{-1}$ in front of the exponential: The density that is "normal" is $p(\log x) d \log x$.

While $\mu$ and $\sigma$ are the mean and standard deviation in $\log x$ space, they are *not* so in $x$ space. Rather,

$$\begin{aligned} \text{Mean}\{\text{Lognormal}(\mu, \sigma)\} &= e^{\mu + \frac{1}{2}\sigma^2} \\ \text{Var}\{\text{Lognormal}(\mu, \sigma)\} &= e^{2\mu} e^{\sigma^2} (e^{\sigma^2} - 1) \end{aligned} \tag{6.14.32}$$

```
syms mu sig positive
syms x
Pi = sym(pi);
p = (1/sqrt(2*Pi))*(1/(sig*x))*exp(-(1/2)*(log(x)-mu)^2/sig^2)
p =
1/2*2^(1/2)/pi^(1/2)/sig/x*exp(-1/2*(log(x)-mu)^2/sig^2)
norm = int(p,0,inf)
norm =
1
mean = int(x*p,0,inf)
mean =
exp(mu+1/2*sig^2)
sig = simplify(sqrt(int(x^2*p,0,inf)-mean^2))
sig =
exp(mu+1/2*sig^2)*(exp(sig^2)-1)^(1/2)
```

Gamma distribution:




$$x \sim \text{Gamma}(\alpha, \beta), \qquad \alpha > 0, \ \beta > 0$$

$$p(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \qquad x > 0$$

$$\text{Mean}\{\text{Gamma}(\alpha, \beta)\} = \alpha/\beta$$

$$\text{Var}\{\text{Gamma}(\alpha, \beta)\} = \alpha/\beta^2$$

When $\alpha \geq 1$ there is a single mode at $x = (\alpha - 1)/\beta$

- Gamma and Lognormal are both commonly used as convenient 2-parameter fitting functions for “peak with tail” positive distributions.
- Both have parameters for peak location and width.
- Neither has a separate parameter for how the tail decays.
  - Gamma: exponential decay
  - Lognormal: long-tailed (exponential of square of log)

Chi-square distribution (we'll use this a lot!)

Has only one parameter $\nu$ that determines both peak location and width.
$\nu$ is often an integer, called "number of degrees of freedom" or "DF"

the independent variable is $\chi^2$, not $\chi$

$$\chi^2 \sim \text{Chisquare}(\nu), \qquad \nu > 0$$

$$p(\chi^2)d\chi^2 = \frac{1}{2^{\frac{1}{2}\nu}\Gamma(\frac{1}{2}\nu)}(\chi^2)^{\frac{1}{2}\nu-1}\exp\left(-\tfrac{1}{2}\chi^2\right)d\chi^2, \qquad \chi^2 > 0$$

$$\text{Mean}\{\text{Chisquare}(\nu)\} = \nu$$

$$\text{Var}\{\text{Chisquare}(\nu)\} = 2\nu$$

When $\nu \geq 2$ there is a single mode at $\chi^2 = \nu - 2$

It's actually just a special case of Gamma, namely Gamma($\nu$/2,1/2)

For a univariate distribution, one wants efficient computational methods for all of:

- PDF $p(x)$
- CDF $P(x)$
- Inverse of CDF $x(P)$
- Random deviates drawn from it (we'll get to soon)

NR3 has classes for many common distributions, with algorithms for p, cdf, and inverse cdf.

```
struct Normaldist : Erf {
Normal distribution, derived from the error function Erf.
    Doub mu, sig;
    Normaldist(Doub mmu = 0., Doub ssig = 1.) : mu(mmu), sig(ssig) {
    Constructor. Initialize with μ and σ. The default with no arguments is N(0, 1).
        if (sig <= 0.) throw("bad sig in Normaldist");
    }
    Doub p(Doub x) {
    Return probability density function.
        return (0.398942280401432678/sig)*exp(-0.5*SQR((x-mu)/sig));
    }
    Doub cdf(Doub x) {
    Return cumulative distribution function.
        return 0.5*erfc(-0.707106781186547524*(x-mu)/sig);
    }
    Doub invcdf(Doub p) {
    Return inverse cumulative distribution function.
        if (p <= 0. || p >= 1.) throw("bad p in Normaldist");
        return -1.41421356237309505*sig*inverfc(2.*p)+mu;
    }
};
```

Matlab and Mathematica both have many distributions, e.g.,

```
chi2pdf(x,v)
chi2cdf(x,v)
chi2inv(p,v)
```

Let's understand better the Central Limit Theorem and where it might <u>or might not</u> 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.

Addition of independent r.v.'s:

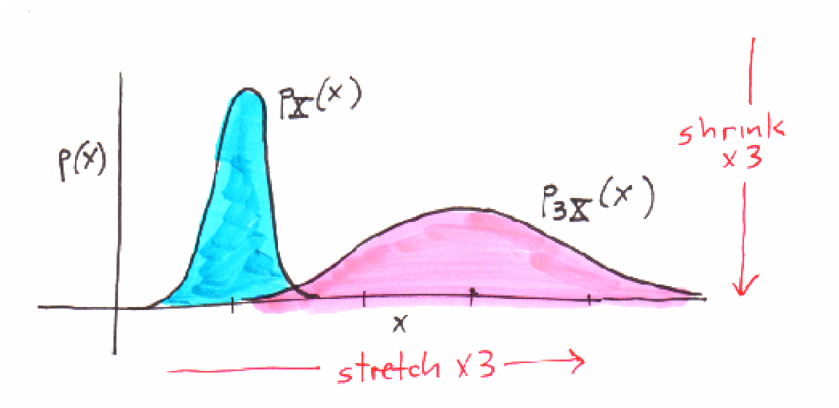
$$\text{let } S = X + Y$$

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

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

(Fourier convolution theorem.)

Scaling law for r.v.'s:




Scaling law for characteristic functions:

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

What's the characteristic function of a Gaussian?

```
syms x mu pi t sigma
p = exp(-(x-mu)^2 / (2*sigma^2)) / (sqrt(2*pi)*sigma)
p =
1/2*exp(-1/2*(x-mu)^2/sigma^2)*2^(1/2)/pi^(1/2)/sigma
norm = int(p,x,-Inf,Inf)
norm =
1
cf = simplify(int(p*exp(i*t*x),x,-Inf,Inf))
cf =
exp(1/2*i*t*(2*mu+i*t*sigma^2))
```

```
In[14]:= $Assumptions = $Assumptions && (sig > 0)

In[15]:=
         p = (1 / (Sqrt[2 Pi] sig)) Exp[-(1 / 2) ((x - mu) / sig) ^2]
Out[15]=
```

$$\frac{e^{-\frac{(-mu+x)^2}{2\,sig^2}}}{\sqrt{2\pi}\,sig}$$

Tell Mathematica that sig is positive. Otherwise it gives “cases” when taking the square root of sig^2

```
In[16]:= Integrate[p, {x, -Infinity, Infinity}]
Out[16]=
         1

In[17]:= Integrate[p Exp[I t x], {x, -Infinity, Infinity}]
Out[17]=
```

$$e^{i\,mu\,t-\frac{sig^2\,t^2}{2}}$$

Cauchy distribution has ill-defined mean and infinite variance, but it has a perfectly good characteristic function:

$$x \sim \text{Cauchy}(\mu, \sigma), \qquad \sigma > 0$$

$$p(x) = \frac{1}{\pi\sigma}\left(1 + \left[\frac{x-\mu}{\sigma}\right]^2\right)^{-1}$$

Matlab and Mathematica both sadly fails at computing the characteristic function of the Cauchy distribution, but you can use fancier methods* and get:

$$\phi_{\text{Cauchy}}(t) = e^{i\mu t - \sigma|t|}$$

note non-analytic at t=0

*If t>0, close the contour in the upper 1/2-plane with a big semi-circle, which adds nothing. So the integral is just the residue at the pole $(x-\mu)/\sigma=i$, which gives $\exp(-\sigma t)$. Similarly, close the contour in the lower 1/2-plane for t<0, giving $\exp(\sigma t)$. So answer is $\exp(-|\sigma t|)$. The factor $\exp(i\mu t)$ comes from the change of x variable to $x-\mu$.

## 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

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

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

Moreover, 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:

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

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.)