Approximate Distribution Theory

Moment Generating Functions

Definition: The moment generating function of a real valued X is

$$M_X(t) = E(e^{tX})$$

defined for those real t for which the expected value is finite.

Definition: The moment generating function of $X \in R^p$ is

$$M_X(u) = E[\exp u^t X]$$

defined for those vectors u for which the expected value is finite.

The moment generating function has the following formal connection to moments:

$$M_X(t) = \sum_{k=0}^{\infty} E[(tX)^k]/k!$$

$$= \sum_{k=0}^{\infty} \mu'_k t^k/k!$$

It is thus sometimes possible to find the power series expansion of M_X and read off the moments of X from the coefficients of the powers $t^k/k!$. (An analogous multivariate version is available.)

Theorem 1 *If M is finite for all $t \in [-\epsilon, \epsilon]$ for some $\epsilon > 0$ then*

1. *Every moment of X is finite.*
2. *M is C^∞ (in fact M is analytic).*
3. *$\mu'_k = \frac{d^k}{dt^k} M_X(0)$.*

The proof, and many other facts about moment generating functions, rely on techniques of complex variables.

Moment Generating Functions and Sums

If $X_1, \ldots, X_p$ are independent and $Y = \sum X_i$ then the moment generating function of Y is the product of those of the individual X_i:

$$E(e^{tY}) = \prod_i E(e^{tX_i})$$

or $M_Y = \prod M_{X_i}$.

However this formula makes the power series expansion of M_Y not a particularly nice function of the expansions of the individual M_{X_i}. In fact this is related to the following observation. The first 3 moments (meaning μ, σ^2 and μ_3) of Y are just the sums of those of the X_i but this doesn't work for the fourth or higher moment.

$$\begin{aligned} E(Y) &= \sum E(X_i) \\ \mathrm{Var}(Y) &= \sum \mathrm{Var}(X_i) \\ E[(Y - E(Y))^3] &= \sum E[(X_i - E(X_i))^3] \end{aligned}$$

but

$$\begin{aligned} E[(Y - E(Y))^4] = &\sum \{E[(X_i - E(X_i))^4] - E^2[(X_i - E(X_i))^2]\} \\ &+ \left\{\sum E[(X_i - E(X_i))^2]\right\}^2 \end{aligned}$$

Example: If $X_1, \ldots, X_p$ are independent and X_i has a $N(\mu_i, \sigma_i^2)$ distribution then

$$\begin{aligned} M_{X_i}(t) &= \int_{-\infty}^{\infty} e^{tx} e^{-(x-\mu_i)/\sigma_i^2} dx/(\sqrt{2\pi}\sigma_i) \\ &= \int_{-\infty}^{\infty} e^{t(\sigma_i z + \mu_i)} e^{-z^2/2} dz/\sqrt{2\pi} \\ &= e^{t\mu_i} \int_{-\infty}^{\infty} e^{-(z-t\sigma_i)^2/2 + t^2\sigma_i^2/2} dz/\sqrt{2\pi} \\ &= e^{\sigma_i^2 t^2/2 + t\mu_i} \end{aligned}$$

Example: I am having you derive the moment and cumulant generating function and all the moments of a Gamma rv. Suppose that $Z_1, \ldots, Z_\nu$ are

independent $N(0,1)$ rvs. Then we have defined $S_\nu = \sum_1^\nu Z_i^2$ to have a χ^2 distribution. It is easy to check $S_1 = Z_1^2$ has density

$$(u/2)^{-1/2} e^{-u/2}/(2\sqrt{\pi})$$

and then the mgf of S_1 is

$$(1-2t)^{-1/2}$$

It follows that

$$M_{S_\nu}(t) = (1-2t)^{-\nu/2};$$

you will show in homework that this is the mgf of a Gamma$(\nu/2, 2)$ rv. This shows that the χ^2_ν distribution has the Gamma$(\nu/2, 2)$ density which is

$$(u/2)^{(\nu-2)/2} e^{-u/2}/(2\Gamma(\nu/2))\,.$$

Example: The Cauchy density is

$$\frac{1}{\pi(1+x^2)}$$

and the corresponding moment generating function is

$$M(t) = \int_{-\infty}^{\infty} \frac{e^{tx}}{\pi(1+x^2)} dx$$

which is $+\infty$ except for $t = 0$ where we get 1. This mgf is exactly the mgf of *every* t distribution so it is not much use for distinguishing such distributions. The problem is that these distributions do not have infinitely many finite moments.

This observation has led to the development of a substitute for the mgf which is defined for every distribution, namely, the characteristic function.

Characteristic Functions

Definition: The characteristic function of a real rv X is

$$\phi_X(t) = E(e^{itX})$$

where $i = \sqrt{-1}$ is the imaginary unit.
Aside on complex arithmetic.

The complex numbers are the things you get if you add $i = \sqrt{-1}$ to the real numbers and require that all the usual rules of algebra work. In particular if i and any real numbers a and b are to be complex numbers then so must be $a + bi$. If we multiply a complex number $a + bi$ with a and b real by another such number, say $c + di$ then the usual rules of arithmetic (associative, commutative and distributive laws) require

$$\begin{aligned}(a + bi)(c + di) =& ac + adi + bci + bdi^2 \\ =& ac + bd(-1) + (ad + bc)i \\ =& (ac - bd) + (ad + bc)i\end{aligned}$$

so this is precisely how we define multiplication. Addition is simply (again by following the usual rules)

$$(a + bi) + (c + di) = (a + b) + (c + d)i$$

Notice that the usual rules of arithmetic then don't require any more numbers than things of the form

$$x + yi$$

where x and y are real. We can identify a single such number $x + yi$ with the corresponding point (x, y) in the plane. It often helps to picture the complex numbers as forming a plane.

Now look at transcendental functions. For real x we know $e^x = \sum x^k/k!$ so our insistence on the usual rules working means

$$e^{x+iy} = e^x e^{iy}$$

and we need to know how to compute e^{iy}. Remember in what follows that $i^2 = -1$ so $i^3 = -i$, $i^4 = 1$ $i^5 = i^1 = i$ and so on. Then

$$\begin{aligned}e^{iy} =& \sum_0^\infty \frac{(iy)^k}{k!} \\ =& 1 + iy + (iy)^2/2 + (iy)^3/6 + \cdots \\ =& 1 - y^2/2 + y^4/4! - y^6 + \cdots \\ & + iy - iy^3/3! + iy^5/5! + \cdots \\ =& \cos(y) + i\sin(y)\end{aligned}$$

We can thus write

$$e^{x+iy} = e^x(\cos(y) + i\sin(y))$$

Now every point in the plane can be written in polar co-ordinates as $(r\cos\theta, r\sin\theta)$ and comparing this with our formula for the exponential we see we can write

$$x + iy = \sqrt{x^2 + y^2}e^{i\theta}$$

for an angle $\theta \in [0, 2\pi)$.

We will need from time to time a couple of other definitions:

Definition: The **modulus** of the complex number $x + iy$ is

$$|x + iy| = \sqrt{x^2 + y^2}$$

Definition: The **complex conjugate** of $x + iy$ is $\overline{x + iy} = x - iy$.

Notes on calculus with complex variables. Essentially the usual rules apply so, for example,

$$\frac{d}{dt}e^{it} = ie^{it}$$

We will (mostly) be doing only integrals over the real line; the theory of integrals along paths in the complex plane is a very important part of mathematics, however.

End of Aside

Since

$$e^{itX} = \cos(tX) + i\sin(tX)$$

we find that

$$\phi_X(t) = E(\cos(tX)) + iE(\sin(tX))$$

Since the trigonometric functions are bounded by 1 the expected values must be finite for all t and this is precisely the reason for using characteristic rather than moment generating functions in probability theory courses.

Theorem 2 *For any two real rvs X and Y the following are equivalent:*

1. *X and Y have the same distribution, that is, for any (Borel) set A we have*
$$P(X \in A) = P(Y \in A)$$
2. *$F_X(t) = F_Y(t)$ for all t.*
3. *$\phi_X = E(e^{itX}) = E(e^{itY}) = \phi_Y(t)$ for all real t.*

Moreover, all of these are implied if there is a positive ϵ such that for all $|t| \leq \epsilon$

$$M_X(t) = M_Y(t) < \infty \,.$$

Inversion

The previous theorem is a non-constructive characterization. It does not show us how to get from ϕ_X to F_X or f_X. For CDFs or densities with reasonable properties, however, there are effective ways to compute F or f from ϕ. In homework I am asking you to prove the following basic **inversion** formula:
If X is a random variable taking only integer values then for each integer k

$$\begin{aligned} P(X = k) &= \frac{1}{2\pi} \int_0^{2\pi} \phi_X(t) e^{-itk} dt \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \phi_X(t) e^{-itk} dt \,. \end{aligned}$$

The proof proceeds from the formula

$$\phi_X(t) = \sum_k e^{ikt} P(X = k) \,.$$

Now suppose that X has a continuous bounded density f. Define

$$X_n = [nX]/n$$

where $[a]$ denotes the integer part (rounding down to the next smallest integer). We have

$$\begin{aligned} P(k/n \leq X < (k+1)/n) &= P([nX] = k) \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \phi_{[nX]}(t) e^{-itk} dt \,. \end{aligned}$$

Make the substitution $t = u/n$, and get

$$nP(k/n \leq X < (k+1)/n) = \frac{1}{2\pi} \int_{-n\pi}^{n\pi} \phi_{[nX]}(u/n) e^{iuk/n} du$$

Now, as $n \to \infty$ we have

$$\phi_{[nX]}(u/n) = E(e^{iu[nX]/n}) \to E(e^{iuX})$$

(by the dominated convergence theorem – the dominating random variable is just the constant 1). The range of integration converges to the whole real line and if $k/n \to x$ we see that the left hand side converges to the density $f(x)$ while the right hand side converges to

$$\frac{1}{2\pi}\int_{-\infty}^{\infty}\phi_X(u)e^{-iux}du$$

which gives the inversion formula

$$f_X(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\phi_X(u)e^{-iux}du$$

Many other such formulas are available to compute things like $F(b) - F(a)$ and so on.

All such formulas are sometimes referred to as Fourier inversion formulas; the characteristic function itself is sometimes called the Fourier transform of the distribution or CDF or density of X.

Inversion of the Moment Generating Function

The moment generating function and the characteristic function are related formally by

$$M_X(it) = \phi_X(t)$$

When M_X exists this relationship is not merely formal; the methods of complex variables mean there is a "nice" (analytic) function which is $E(e^{zX})$ for any complex $z = x + iy$ for which $M_X(x)$ is finite. All this means that there is an inversion formula for M_X. This formula requires a complex *contour integral*. In general if z_1 and z_2 are two points in the complex plane and C a path between these two points we can define the path integral

$$\int_C f(z)dz$$

by the methods of line integration. When it comes to doing algebra with such integrals the usual theorems of calculus still work. The Fourier inversion formula was

$$2\pi f(x) = \int_{-\infty}^{\infty}\phi(t)e^{-itx}dt$$

so replacing ϕ by M we get

$$2\pi f(x) = \int_{-\infty}^{\infty} M(it)e^{-itx}dt$$

If we just substitute $z = it$ then we find

$$2\pi i f(x) = \int_C M(z)e^{-zx}dz$$

where the path C is the imaginary axis. This formula becomes of use by the methods of complex integration which permit us to replace the path C by any other path which starts and ends at the same place. It is possible, in some cases, to choose this path to make it easy to do the integral approximately; this is what saddlepoint approximations are. This inversion formula is called the inverse Laplace transform; the mgf is also called the Laplace transform of the distribution or of the CDF or of the density.

Applications of Inversion

1): Numerical calculations

Example: Many statistics have a distribution which is approximately that of

$$T = \sum \lambda_j Z_j^2$$

where the Z_j are iid $N(0,1)$. In this case

$$\begin{aligned} E(e^{itT}) &= \prod E(e^{it\lambda_j Z_j^2}) \\ &= \prod (1 - 2it\lambda_j)^{-1/2}\,. \end{aligned}$$

Imhof (*Biometrika*, 1961) gives a simplification of the Fourier inversion formula for

$$F_T(x) - F_T(0)$$

which can be evaluated numerically.

2): The central limit theorem (in some versions) can be deduced from the Fourier inversion formula: if $X_1, \ldots, X_n$ are iid with mean 0 and variance 1

and $T = n^{1/2}\bar{X}$ then with ϕ denoting the characteristic function of a single X we have

$$\begin{aligned} E(e^{itT} &= E(e^{in^{-1/2}t\sum X_j}) \\ &= \left[\phi(n^{-1/2}t)\right]^n \\ &\approx \left[\phi(0) + n^{-1/2}t\phi'(0) + n^{-1}t^2\phi''(0)/2 + o(n^{-1}\right]^n \end{aligned}$$

But now $\phi(0) = 1$ and

$$\phi'(t) = \frac{d}{dt}E(e^{itX_1}) = iE(X_1e^{itX_1})$$

So $\phi'(0) = E(X_1) = 0$. Similarly

$$\phi''(t) = i^2E(X_1^2e^{itX_1})$$

so that

$$\phi''(0) = -E(X_1^2) = -1$$

It now follows that

$$\begin{aligned} E(e^{itT}) &\approx [1 - t^2/(2n) + o(1/n)]^n \\ &\to e^{-t^2/2} \end{aligned}$$

With care we can then apply the Fourier inversion formula and get

$$\begin{aligned} f_T(x) &= \frac{1}{2\pi i}\int_{-\infty}^{\infty} e^{-itx}[\phi(tn^{-1/2})]^n dt \\ &\to \frac{1}{2\pi i}\int_{-\infty}^{\infty} e^{-itx}e^{-t^2/2}dt \\ &= \frac{1}{\sqrt{2\pi}}\phi_Z(-x) \end{aligned}$$

where ϕ_Z is the characteristic function of a standard normal variable Z. Doing the integral we find

$$\phi_Z(x) = \phi_Z(-x) = e^{-x^2/2}$$

so that

$$f_T(x) \to \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$$

which is a standard normal random variable.

This proof of the central limit theorem is not terribly general since it requires T to have a bounded continuous density. The central limit theorem itself is a statement about CDFs not densities and is

$$P(T \leq t) \to P(Z \leq t)$$

Last time derived the Fourier inversion formula

$$f_X(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \phi_X(u) e^{-iux} du$$

and the moment generating function inversion formula

$$2\pi i f(x) = \int_{-i\infty}^{i\infty} M(z) e^{-zx} dz$$

(where the limits of integration indicate a contour integral up the imaginary axis.) The methods of complex variables permit this path to be replaced by any contour running up a line like $Re(z) = c$. ($Re(Z)$ denotes the real part of z, that is, x when $z = x + iy$ with x and y real.) The value of c has to be one for which $M(c) < \infty$. Rewrite the inversion formula using the cumulant generating function $K(t) = \log(M(t))$ to get

$$2\pi i f(x) = \int_{c-i\infty}^{c+i\infty} \exp(K(z) - zx) dz \,.$$

Along the contour in question we have $z = c + iy$ so we can think of the integral as being

$$i \int_{-\infty}^{\infty} \exp(K(c + iy) - (c + iy)x) dy$$

Now do a Taylor expansion of the exponent:

$$K(c + iy) - (c + iy)x = K(c) - cx + iy(K'(c) - x) - y^2 K''(c)/2 + \cdots$$

Ignore the higher order terms and select a c so that the first derivative

$$K'(c) - x$$

vanishes. Such a c is a saddlepoint. We get the formula

$$2\pi f(x) \approx \exp(K(c) - cx) \int_{-\infty}^{\infty} \exp(-y^2 K''(c)/2) dy$$

The integral is just a normal density calculation and gives $\sqrt{2\pi/K''(c)}$. The saddlepoint approximation is

$$f(x) = \frac{\exp(K(c) - cx)}{\sqrt{2\pi K''(c)}}$$

Essentially the same idea lies at the heart of the proof of Sterling's approximation to the factorial function:

$$n! = \int_0^{\infty} \exp(n \log(x) - x) dx$$

The exponent is maximized when $x = n$. For n large we approximate $f(x) = n \log(x) - x$ by

$$f(x) \approx f(x_0) + (x - x_0) f'(x_0) + (x - x_0)^2 f''(x_0)/2$$

and choose $x_0 = n$ to make $f'(x_0) = 0$. Then

$$n! \approx \int_0^{\infty} \exp[n \log(n) - n - (x - n)^2/(2n)] dx$$

Substitute $y = (x - n)/\sqrt{n}$ to get the approximation

$$n! \approx n^{1/2} n^n e^{-n} \int_{-\infty}^{\infty} e^{-y^2/2} dy$$

or

$$n! \approx \sqrt{2\pi} n^{n+1/2} e^{-n}$$

This tactic is called Laplace's method. Notice that I am very sloppy about the limits of integration. To make the foregoing rigorous you must show that the contribution to the integral from x not sufficiently close to n is negligible.

Convergence in Distribution

In undergraduate courses we often teach the central limit theorem: if $X_1, \ldots, X_n$ are iid from a population with mean μ and standard deviation σ then $n^{1/2}(\bar{X} - \mu)/\sigma$ has approximately a normal distribution. We also say that a Binomial(n, p) random variable has approximately a $N(np, np(1-p))$ distribution.

To make precise sense of these assertions we need to assign a meaning to statements like "X and Y have approximately the same distribution". The meaning we want to give is that X and Y have nearly the same CDF but even here we need some care. If n is a large number is the $N(0, 1/n)$ distribution close to the distribution of $X \equiv 0$? Is it close to the $N(1/n, 1/n)$ distribution? Is it close to the $N(1/\sqrt{n}, 1/n)$ distribution? If $X_n \equiv 2^{-n}$ is the distribution of X_n close to that of $X \equiv 0$?

The answer to these questions depends in part on how close close needs to be so it's a matter of definition. In practice the usual sort of approximation we want to make is to say that some random variable X, say, has nearly some continuous distribution, like $N(0, 1)$. In this case we must want to calculate probabilities like $P(X > x)$ and know that this is nearly $P(N(0, 1) > x)$. The real difficulty arises in the case of discrete random variables; in this course we will not actually need to approximate a distribution by a discrete distribution.

When mathematicians say two things are close together they either can provide an upper bound on the distance between the two things or they are talking about taking a limit. In this course we do the latter.

Definition: A sequence of random variables X_n converges in distribution to a random variable X if

$$E(g(X_n)) \to E(g(X))$$

for every bounded continuous function g.

Theorem: The following are equivalent:

1. X_n converges in distribution to X.
2. $P(X_n \le x) \to P(X \le x)$ for each x such that $P(X = x) = 0$
3. The characteristic functions of X_n converge to that of X:

$$E(e^{itX_n}) \to E(e^{itX})$$

for every real x.

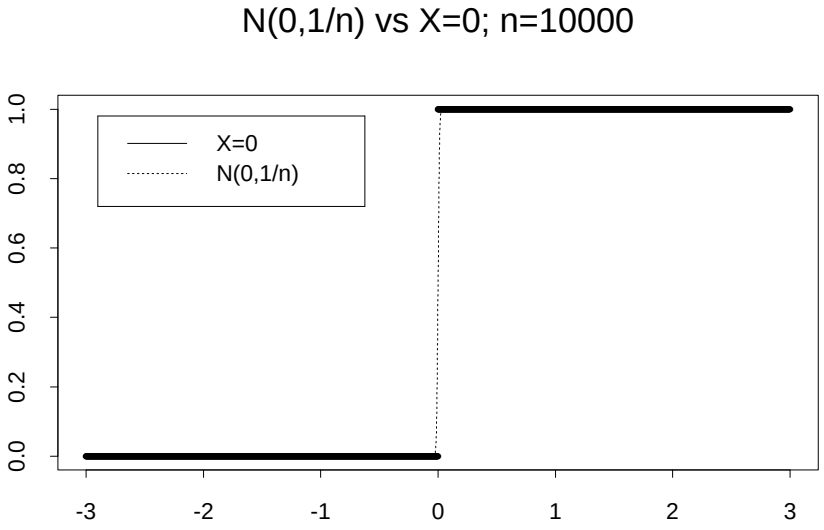
These are all implied by

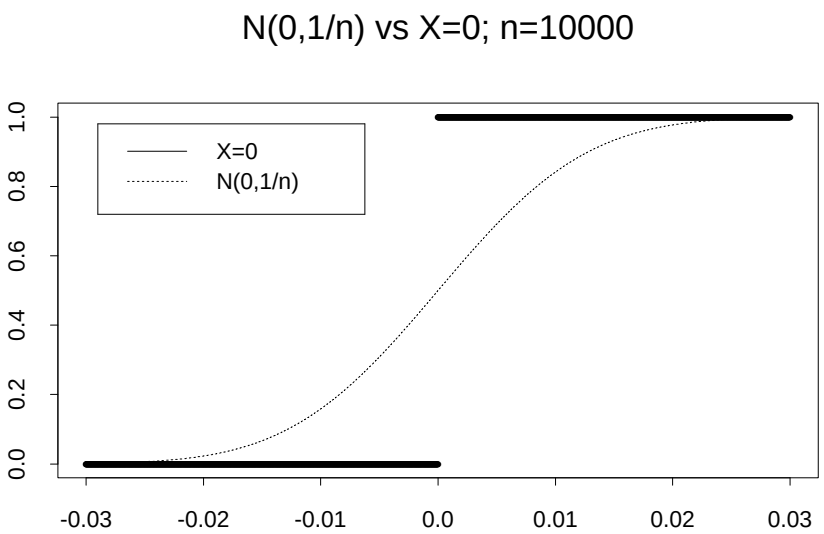
$$M_{X_n}(t) \to M_X(t) < \infty$$

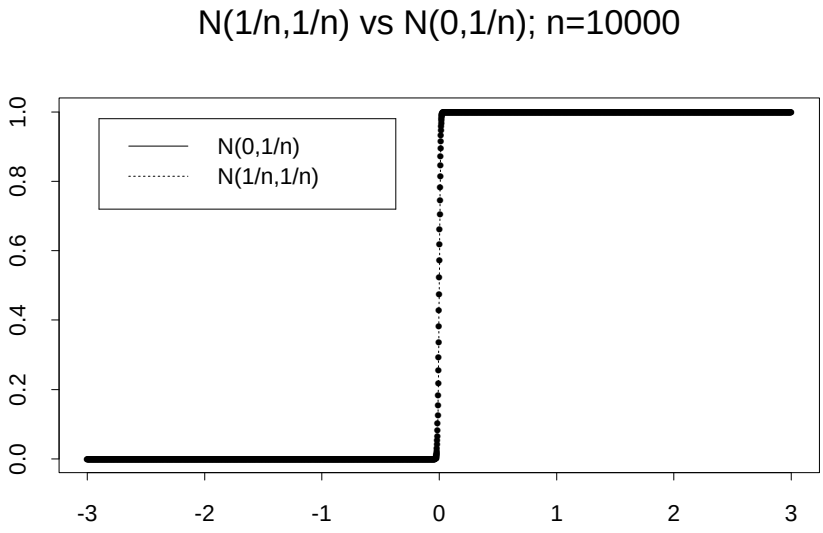
for all $|t| \le \epsilon$ for some positive ϵ.

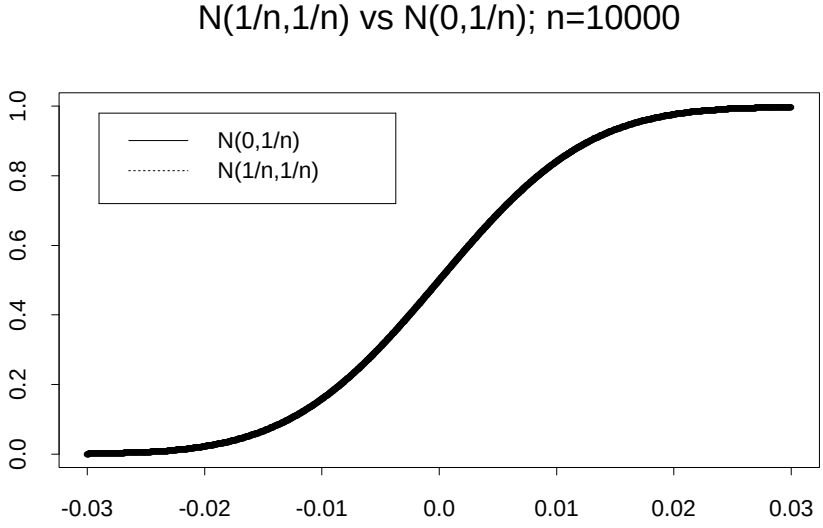
Now let's go back to the questions I asked:

- $X_n \sim N(0, 1/n)$ and $X = 0$.
- Then
$$P(X_n \le x) \to \begin{cases} 1 & x > 0 \\ 0 & x < 0 \\ 1/2 & x = 0 \end{cases}$$
- Now the limit is the CDF of $X = 0$ except for $x = 0$ and the CDF of X is not continuous at $x = 0$ so yes, X_n converges to X in distribution.
- I asked if $X_n \sim N(1/n, 1/n)$ had a distribution close to that of $Y_n \sim N(0, 1/n)$. The definition I gave really requires me to answer by finding a limit X and proving that both X_n and Y_n converge to X in distribution. Take $X = 0$. Then
$$E(e^{tX_n}) = e^{t/n + t^2/(2n)} \to 1 = E(e^{tX})$$
and
$$E(e^{tY_n}) = e^{t^2/(2n)} \to 1$$
so that both X_n and Y_n have the same limit in distribution.
- Now multiply both X_n and Y_n by $n^{1/2}$ and let $X \sim N(0, 1)$. Then $\sqrt{n}X_n \sim N(n^{-1/2}, 1)$ and $\sqrt{n}Y_n \sim N(0, 1)$. You can use characteristic functions to prove that both $\sqrt{n}X_n$ and $\sqrt{n}Y_n$ converge to $N(0, 1)$ in distribution.


N(0,1/n) vs X=0; n=10000
X=0
N(0,1/n)
0.0
0.2
0.4
0.6
0.8
1.0
-3
-2
-1
0
1
2
3

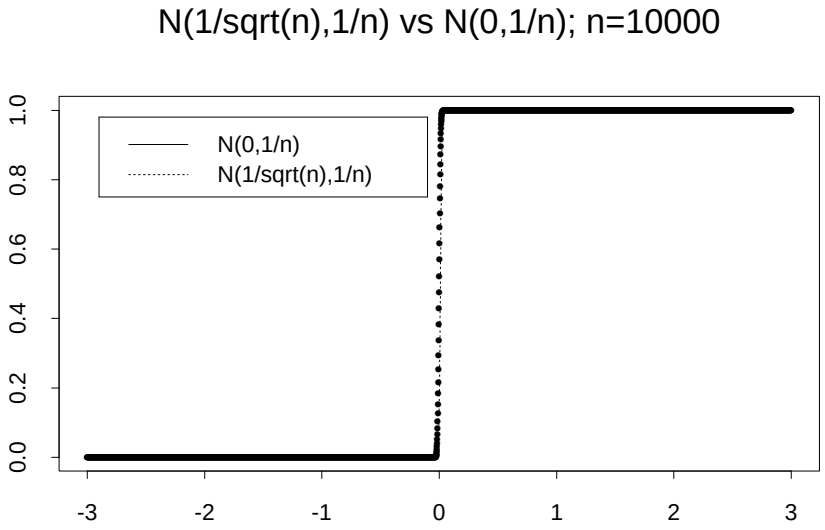

N(0,1/n) vs X=0; n=10000
X=0
N(0,1/n)
0.0
0.2
0.4
0.6
0.8
1.0
-0.03
-0.02
-0.01
0.0
0.01
0.02
0.03

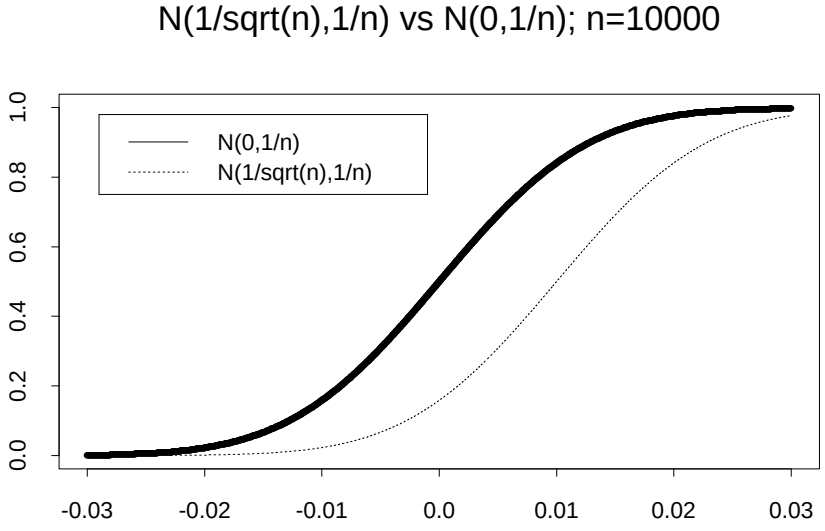




- If you now let $X_n \sim N(n^{-1/2}, 1/n)$ and $Y_n \sim N(0, 1/n)$ then again both X_n and Y_n converge to 0 in distribution.
- If you multiply these X_n and Y_n by $n^{1/2}$ then $n^{1/2}X_n \sim N(1, 1)$ and

$n^{1/2}Y_n \sim N(0,1)$ so that $n^{1/2}X_n$ and $n^{1/2}Y_n$ are **not** close together in distribution.





- You can check that $2^{-n} \to 0$ in distribution.

Here is the message you are supposed to take away from this discussion. You do distributional approximations by showing that a sequence of random variables X_n converges to some X. The limit distribution should be non-trivial, like say $N(0,1)$. We don't say X_n is approximately $N(1/n, 1/n)$ but that $n^{1/2}X_n$ converges to $N(0,1)$ in distribution.

The Central Limit Theorem

If $X_1, X_2, \cdots$ are iid with mean 0 and variance 1 then $n^{1/2}\bar{X}$ converges in distribution to $N(0,1)$. That is,

$$P(n^{1/2}\bar{X} \le x) \to \frac{1}{2\pi}\int_{-\infty}^{x} e^{-y^2/2}dy\,.$$

Proof: As before

$$E(e^{itn^{1/2}\bar{X}}) \to e^{-t^2/2}$$

This is the characteristic function of a $N(0,1)$ random variable so we are done by our theorem.

Edgeworth expansions

In fact if $\gamma = E(X^3)$ then

$$\phi(t) \approx 1 - t^2/2 - i\gamma t^3/6 + \cdots$$

keeping one more term. Then

$$\log(\phi(t)) = \log(1+u)$$

where

$$u = -t^2/2 - i\gamma t^3/6 + \cdots$$

Use $\log(1+u) = u - u^2/2 + \cdots$ to get

$$\log(\phi(t)) \approx -[t^2/2 - i\gamma t^3/6 + cdots] - [\ldots]^2/2 + \cdots$$

which rearranged is

$$\log(\phi(t)) \approx -t^2/2 + i\gamma t^3/6 + \cdots$$

Now apply this calculation to

$$\log(\phi_T(t)) \approx -t^2/2 + iE(T^3)t^3/6 + \cdots$$

Remember $E(T^3) = \gamma/\sqrt{n}$ and exponentiate to get

$$\phi_T(t) \approx e^{-t^2/2} \exp\{i\gamma t^3/(6\sqrt{n}) + \cdots\}$$

You can do a Taylor expansion of the second exponential around 0 because of the square root of n and get

$$\phi_T(t) \approx e^{-t^2/2}(1 - i\gamma t^3/(6\sqrt{n}))$$

neglecting higher order terms. This approximation to the characteristic function of T can be inverted to get an **Edgeworth** approximation to the density (or distribution) of T which looks like

$$f_T(x) \approx \frac{1}{\sqrt{2\pi}} e^{-x^2/2}[1 - \gamma(x^3 - 3x)/(6\sqrt{n}) + \cdots]$$

Remarks:

1. The error using the central limit theorem to approximate a density or a probability is proportional to $n^{-1/2}$
2. This is improved to n^{-1} for symmetric densities for which $\gamma = 0$.
3. The expansions are **asymptotic**. This means that the series indicated by $\cdots$ usually does **not** converge. When $n = 25$ it may help to take the second term but get worse if you include the third or fourth or more.
4. You can integrate the expansion above for the density to get an approximation for the CDF.

Multivariate convergence in distribution

Definition: $X_n \in R^p$ converges in distribution to $X \in R^p$ if

$$E(g(X_n)) \to E(g(X))$$

for each bounded continuous real valued function g on R^p.

This is equivalent to either of

Cramér Wold Device: $a^t X_n$ converges in distribution to $a^t X$ for each $a \in R^p$

or

Convergence of Characteristic Functions:

$$E(e^{ia^t X_n}) \to E(e^{ia^t X})$$

for each $a \in R^p$.

Extensions of the CLT

1. If $Y_1, Y_2, \cdots$ are iid in R^p with mean μ and variance covariance Σ then $n^{1/2}(\bar{Y} - \mu)$ converges in distribution to $MVN(0, \Sigma)$.

2. If for each n we have a set of independent mean 0 random variables $X_{n1}, \ldots, X_{nn}$ and $E(X_{ni}) = 0$ and $Var(\sum_i X_{ni}) = 1$ and

$$\sum E(|X_{ni}|^3) \to 0$$

then $\sum_i X_{ni}$ converges in distribution to $N(0, 1)$. This is the Lyapunov central limit theorem.

3. As in the Lyapunov central CLT but replace the third moment condition with

$$\sum E(X_{ni}^2 1(|X_{ni}| > \epsilon)) \to 0$$

for each $\epsilon > 0$ then again $\sum_i X_{ni}$ converges in distribution to $N(0, 1)$. This is the Lindeberg central limit theorem. (Lyapunov's condition implies Lindeberg's.)

4. There are extensions to random variables which are not independent. Examples include the m-dependent central limit theorem, the martingale central limit theorem, the central limit theorem for mixing processes.

5. Many important random variables are not sums of independent random variables. We handle these with Slutsky's theorem and the δ method.

Slutsky's Theorem: If X_n converges in distribution to X and Y_n converges in distribution (or in probability) to c, a constant, then $X_n + Y_n$ converges in distribution to $X + c$.

Warning: the hypothesis that the limit of Y_n be constant is essential.

The delta method: Suppose a sequence Y_n of random variables converges to some y a constant and that if we define $X_n = a_n(Y_n - y)$ then X_n converges in distribution to some random variable X. Suppose that f is a differentiable function on the range of Y_n. Then $a_n(f(Y_n) - f(y))$ converges in distribution to $f'(y)X$. If X_n is in R^p and f maps R^p to R^q then f' is the $q \times p$ matrix of first derivatives of components of f.

Example: Suppose $X_1, \ldots, X_n$ are a sample from a population with mean μ, variance σ^2, and third and fourth central moments μ_3 and μ_4. Then

$$n^{1/2}(s^2 - \sigma^2) \Rightarrow N(0, \mu_4 - \sigma^4)$$

where $\Rightarrow$ is notation for convergence in distribution. For simplicity I define $s^2 = \overline{X^2} - \bar{X}^2$.

We take Y_n to be the vector with components $(\overline{X^2}, \bar{X})$. Then Y_n converges to $y = (\mu^2 + \sigma^2, \mu)$. Take $a_n = n^{1/2}$. Then

$$n^{1/2}(Y_n - y)$$

converges in distribution to $MVN(0, \Sigma)$ with

$$\Sigma = \begin{bmatrix} \mu_4 - \sigma^4 & \mu_3 - \mu(\mu^2 + \sigma^2) \\ \mu_3 - \mu(\mu^2 + \sigma^2) & \sigma^2 \end{bmatrix}$$

Define $f(x_1, x_2) = x_1 - x_2^2$. Then $s^2 = f(Y_n)$ and the gradient of f has components $(1, -2x_2)$. This leads to

$$n^{1/2}(s^2 - \sigma^2) \approx n^{1/2}(1, -2\mu) \begin{bmatrix} \overline{X^2} - (\mu^2 + \sigma^2) \\ \bar{X} - \mu \end{bmatrix}$$

which converges in distribution to the law of $(1, -2\mu)Y$ which is $N(0, a^t \Sigma a)$ where $a = (1, -2\mu)^t$. This boils down to $N(0, \mu_4 - \sigma^2)$.

Remark: In this sort of problem it is best to learn to recognize that the sample variance is unaffected by subtracting μ from each X. Thus there is no loss in assuming $\mu = 0$ which simplifies Σ and a.

Special case: if the observations are $N(\mu, \sigma^2)$ then $\mu_3 = 0$ and $\mu_4 = 3\sigma^4$. Our calculation has

$$n^{1/2}(s^2 - \sigma^2) \Rightarrow N(0, 2\sigma^4)$$

You can divide through by σ^2 and get

$$n^{1/2}(\frac{s^2}{\sigma^2} - 1) \Rightarrow N(0, 2)$$

In fact $(n-1)s^2/\sigma^2$ has a χ^2_{n-1} distribution and so the usual central limit theorem shows that

$$(n-1)^{1/2}[(n-1)s^2/\sigma^2 - (n-1)] \Rightarrow N(0, 2)$$

(using mean of χ^2_1 is 1 and variance is 2). Factoring out $n - 1$ gives the assertion that

$$(n-1)^{1/2}(s^2/\sigma^2 - 1) \Rightarrow N(0, 2)$$

which is our δ method calculation except for using $n-1$ instead of n. This difference is unimportant as can be checked using Slutsky's theorem.

Monte Carlo

The last method of distribution theory that I will review is Monte Carlo simulation. Suppose you have some random variables $X_1, \ldots, X_n$ whose joint distribution is specified and a statistic $T(X_1, \ldots, X_n)$ whose distribution you want to know. To compute something like $P(T > t)$ for some specific value of t we appeal to the limiting relative frequency interpretation of probability: $P(T > t)$ is the limit of the proportion of trials in a long sequence of trials in which T occurs. We use a (pseudo) random number generator to generate a sample $X_1, \ldots, X_n$ and then calculate the statistic getting T_1. Then we generate a new sample (independently of our first, say) and calculate T_2. We repeat this a large number of times say N and just count up how many of the T_k are larger than t. If there are M such T_k we estimate that $P(T > t) = M/N$.

The quantity M has a Binomial$(N, p = P(T > t))$ distribution. The standard error of M/N is then $p(1-p)/N$ which is estimated by $M(N-M)/N^3$. This permits us to guess the accuracy of our study.

Notice that the standard deviation of M/N is $\sqrt{p(1-p)}/\sqrt{N}$ so that to improve the accuracy by a factor of 2 requires 4 times as many samples. This makes Monte Carlo a relatively time consuming method of calculation. There are a number of tricks to make the method more accurate (though they only change the constant of proportionality – the SE is still inversely proportional to the square root of the sample size).

Generating the Sample

Most computer languages have a facility for generating pseudo uniform random numbers, that is, variables U which have (approximately of course) a Uniform$[0, 1]$ distribution. Other distributions are generated by transformation:

Exponential: $X = -\log U$ has an exponential distribution:

$$P(X > x) = P(-\log(U) > x) = P(U \leq e^{-x}) = e^{-x}$$

Random uniforms generated on the computer sometimes have only 6 or 7 digits or so of detail. This can make the tail of your distribution grainy. If U were actually a multiple of 10^{-6} for instance then the largest possible value of X is $6\log(10)$. This problem can be ameliorated by the following algorithm:

- Generate U a Uniform[0,1] variable.
- Pick a small ϵ like 10^{-3} say. If $U > \epsilon$ take $Y = -\log(U)$.
- If $U \leq \epsilon$ remember that the conditional distribution of $Y - y$ given $Y > y$ is exponential. You use this by generating a new U' and computing $Y' = -\log(U')$. Then take $Y = Y' - \log(\epsilon)$. The resulting Y has an exponential distribution. You should check this by computing $P(Y > y)$.

Normal: In general if F is a continuous CDF and U is Uniform[0,1] then $Y = F^{-1}(U)$ has CDF F because

$$P(Y \leq y) = P(F^{-1}(U) \leq y) = P(U \leq F(y)) = F(y)$$

This is almost the technique in the exponential distribution. For the normal distribution $F = \Phi$ (Φ is a common notation for the standard normal CDF) there is no closed form for F^{-1}. You could use a numerical algorithm to compute F^{-1} or you could use the following Box Müller trick. Generate U_1, U_2 two independent Uniform[0,1] variables. Define $Y_1 = \sqrt{-2\log(U_1)}\cos(2\pi U_2)$ and $Y_2 = \sqrt{-2\log(U_1)}\sin(2\pi U_2)$. Then you can check using the change of variables formula that Y_1 and Y_2 are independent $N(0,1)$ variables.

Acceptance Rejection

If you can't easily calculate F^{-1} but you know f you can try the acceptance rejection method. Find a density g and a constant c such that $f(x) \leq cg(x)$ for each x and G^{-1} is computable or you otherwise know how to generate observations $W_1, W_2, \ldots$ independently from g. Generate W_1. Compute $p = f(W_1)/(cg(W_1)) \leq 1$. Generate a uniform[0,1] random variable U_1 independent of all the Ws and let $Y = W_1$ if $U_1 \leq p$. Otherwise get a new W and a new U and repeat until you find a $U_i \leq f(W_i)/(cg(W_i))$. You make Y be the last W you generated. This Y has density f.

Markov Chain Monte Carlo

In the last 10 years the following tactic has become popular, particularly for generating multivariate observations. If $W_1, W_2, \ldots$ is an (ergodic) Markov chain with stationary transitions and the stationary initial distribution of W has density f then you can get random variables which have the marginal density f by starting off the Markov chain and letting it run for a long time. The marginal distributions of the W_i converge to f. So you can estimate things like $\int_A f(x)dx$ by computing the fraction of the W_i which land in A.

There are now many versions of this technique. Examples include Gibbs Sampling and the Metropolis-Hastings algorithm. (The technique was invented in the 1950s by physicists: Metropolis et al. One of the authors of the paper was Edward Teller "father of the hydrogen bomb".)

Importance Sampling

If you want to compute

$$\theta \equiv E(T(X)) = \int T(x)f(x)dx$$

you can generate observations from a different density g and then compute

$$\hat{\theta} = n^{-1}\sum T(X_i)f(X_i)/g(X_i)$$

Then

$$\begin{aligned} E(\hat{\theta}) &= n^{-1}\sum E(T(X_i)f(X_i)/g(X_i) \\ &= \int [T(x)f(x)/g(x)]g(x)dx \\ &= \int T(x)f(x)dx \\ &= \theta \end{aligned}$$

Variance reduction

Consider the problem of estimating the distribution of the sample mean for a Cauchy random variable. The Cauchy density is

$$f(x) = \frac{1}{\pi(1+x^2)}$$

We generate $U_1, \ldots, U_n$ uniforms and then define $X_i = \tan^{-1}(\pi(U_i - 1/2))$. Then we compute $T = \bar{X}$. Now to estimate $p = P(T > t)$ we would use

$$\hat{p} = \sum_{i=1}^{N} 1(T_i > t)/N$$

after generating N samples of size n. This estimate is unbiased and has standard error $\sqrt{p(1-p)/N}$.

We can improve this estimate by remembering that $-X_i$ also has Cauchy distribution. Take $S_i = -T_i$. Remember that S_i has the same distribution as T_i. Then we try (for $t > 0$)

$$\tilde{p} = [\sum_{i=1}^{N} 1(T_i > t) + \sum_{i=1}^{N} 1(S_i > t)]/(2N)$$

which is the average of two estimates like $\hat{p}$. The variance of $\tilde{p}$ is

$$(4N)^{-1} Var(1(T_i > t) + 1(S_i > t)) = (4N)^{-1} Var(1(|T| > t))$$

which is

$$\frac{2p(1-2p)}{4N} = \frac{p(1-2p)}{2N}$$

Notice that the variance has an extra 2 in the denominator and that the numerator is also smaller – particularly for p near $1/2$. So this method of variance reduction has resulted in a need for only half the sample size to get the same accuracy.

Regression estimates

Suppose we want to compute

$$\theta = E(|Z|)$$

where Z is standard normal. We generate N iid $N(0,1)$ variables $Z_1, \ldots, Z_N$ and compute $\hat{\theta} = \sum |Z_i|/N$. But we know that $E(Z_i^2) = 1$ and can see easily that $\hat{\theta}$ is positively correlated with $\sum Z_i^2/N$. So we consider using

$$\tilde{\theta} = \hat{\theta} - c(\sum Z_i^2/N - 1)$$

Notice that $E(\tilde{\theta}) = \theta$ and

$$Var(\tilde{\theta}) = Var(\hat{\theta}) - 2cCov(\hat{\theta}, \sum Z_i^2/n) + c^2 Var(\sum Z_i^2/N)$$

The value of c which minimizes this is

$$c = \frac{Cov(\hat{\theta}, \sum Z_i^2/n)}{Var(\sum Z_i^2/N)}$$

and this value can be estimated by regressing the $|Z_i|$ on the Z_i^2!