Laws of Large Numbers

When the extremely strong assumptions of the classical normal linear model do not hold, we often use **asymptotic theory** to test hypotheses and form confidence intervals

Asymptotic theory depends on two types of fundamental result, laws of large numbers and central limit theorems.

A **law of large numbers**, or **LLN**, may apply to anything that can be written as an average of N random variables.

Let x_i, $i = 1, \ldots, N$, be independent random variables, each with bounded finite variance σ_i^2 and with a common mean μ. The sample mean of the x_i is

$$\bar{x} \equiv \frac{1}{N} \sum_{i=1}^{N} x_i. \tag{1}$$

A fairly simple LLN says that, as $N \to \infty$, $\bar{x}$ tends to μ.

An LLN also lets us prove the **Fundamental Theorem of Statistics**.

It is concerned with the **empirical distribution function**, or **EDF**, of a random sample.

We obtain a random sample of size N, with typical element x_i, from the distribution $F(x)$, where the x_i are independent.

An **empirical distribution** is a discrete distribution that gives weight $1/N$ to each x_i for $i = 1, \ldots, N$.

The EDF is the CDF of the empirical distribution:

$$\hat{F}(x) \equiv \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(x_i \leq x), \tag{2}$$

where $\mathbb{I}(\cdot)$ is the **indicator function**.

For a given x, the sum on the r.h.s. of (2) counts the number of realizations x_i that are smaller than or equal to x.

The EDF has the form of a step function.

The height of each step is $1/N$, and the width is the difference between two successive values of x_i.

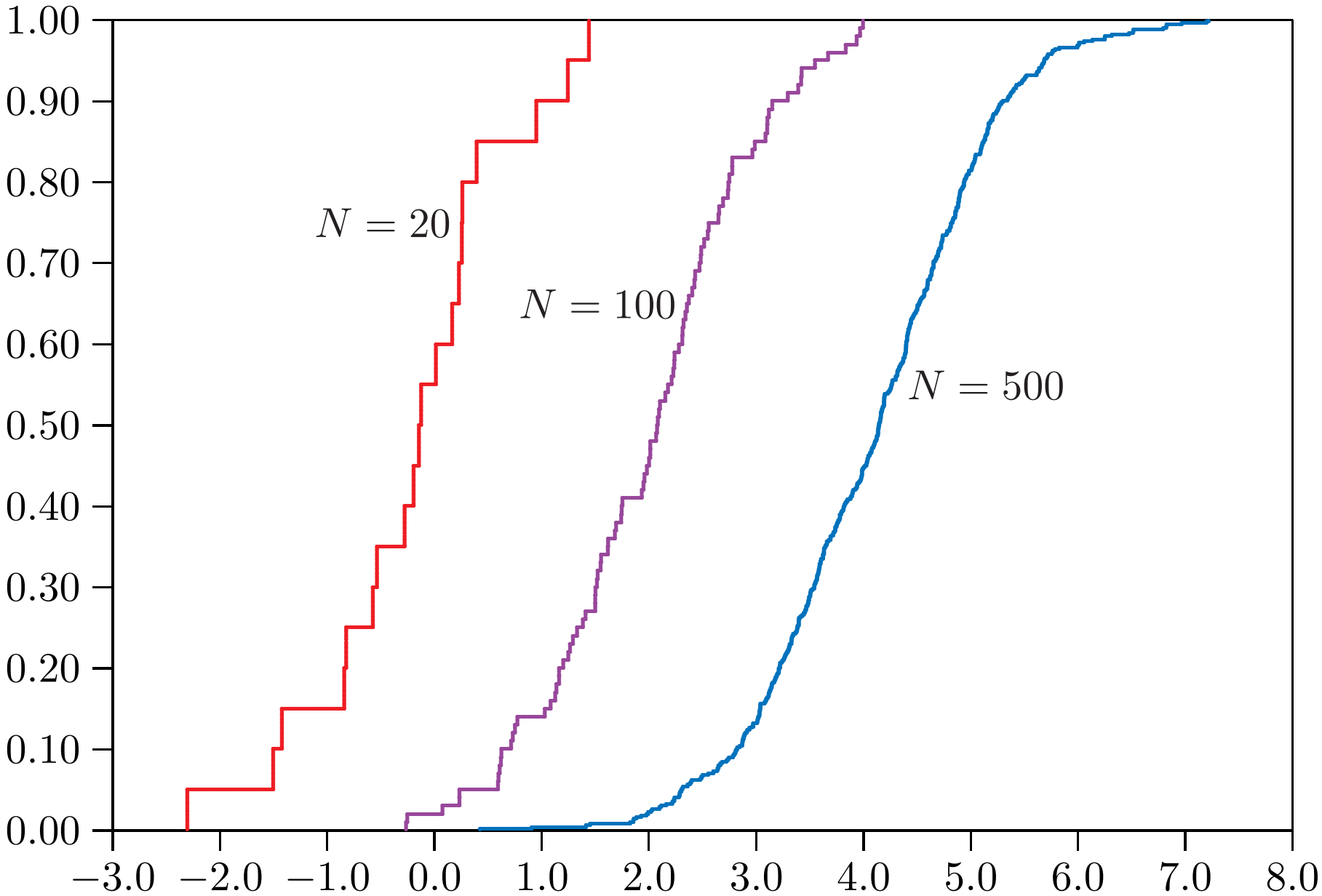
The next figure shows the EDFs for three samples of sizes 20, 100, and 500 drawn from three normal distributions, each with variance 1 and with means 0, 2, and 4.

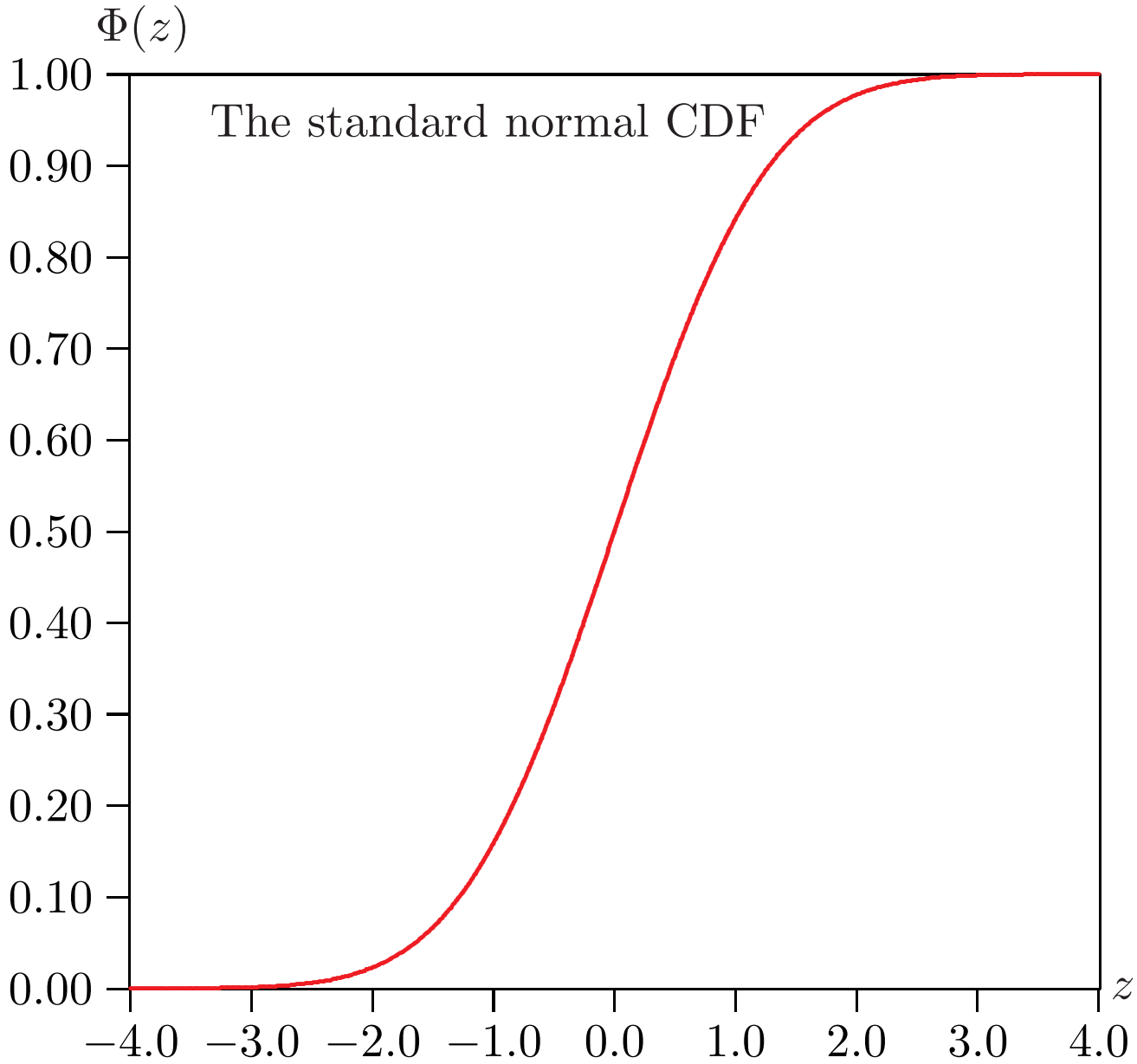
Compare these with the CDF of the standard normal distribution in the figure that follows.

The Fundamental Theorem of Statistics tells us that the EDF for a sample, x_i, $i = 1, \ldots, N$, consistently estimates the CDF of the random variable X:

$$\operatorname*{plim}_{N \to \infty} \hat{F}(x) = F(x). \tag{3}$$

This is true for every admissible value x.


1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
$N = 20$
$N = 100$
$N = 500$
−3.0
−2.0
−1.0
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0


$\Phi(z)$
The standard normal CDF
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
z
−4.0
−3.0
−2.0
−1.0
0.0
1.0
2.0
3.0
4.0

Proof of the Fundamental Theorem of Statistics

For any x, each term in the sum that defines $\hat{F}(x)$ depends only on x_i.

$\mathbb{I}(x_i \leq x)$ can take on only two values, 1 and 0.

The expectation of $\mathbb{I}(x_i \leq x)$ is

$$\mathbb{E}\left(\mathbb{I}(x_i \leq x)\right) = 0 \cdot \Pr\left(\mathbb{I}(x_i \leq x) = 0\right) + 1 \cdot \Pr\left(\mathbb{I}(x_i \leq x) = 1\right) \tag{4}$$

$$= \Pr\left(\mathbb{I}(x_i \leq x) = 1\right) = \Pr(x_i \leq x) = F(x). \tag{5}$$

The x_i are independent and follow the same distribution, so the $\mathbb{I}(x_i \leq x)$ must do so as well.

Thus $\hat{F}(x)$ is the sample mean of N IID random terms, each with finite expectation. A very simple LLN (due to Khinchin) applies to it.

We conclude that, for every x, $\hat{F}(x)$ is a consistent estimator of $F(x)$.

More commonly, we apply an LLN directly to averages of functions of the x_i, like $\bar{x}$.

There are many different LLNs. Some do not require the x_i to have a common mean or to be independent, although the amount of dependence must be limited.

- In general, the weaker the conditions on dependence that we impose, the stronger the conditions on moments have to be.
- For example, Khinchin's LLN only requires that the first moment exists, but it requires identical distributions and no dependence.
- LLNs that allow for dependence and heterogeneity may require that four (or more) moments exist.
- If we can apply an LLN to a random quantity, we can treat it as nonrandom for the purpose of asymptotic analysis.
- In many cases, we must divide the quantity by N.
  - $\boldsymbol{X}^\top \boldsymbol{X}$ generally does not converge to anything as $N \to \infty$.
  - But $N^{-1}\boldsymbol{X}^\top \boldsymbol{X}$, under many plausible assumptions, tends to a non-stochastic limiting matrix $\boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}$.

Central Limit Theorems

The second type of fundamental result on which asymptotic theory is based is called a **central limit theorem**, or **CLT**.

$1/\sqrt{N}$ times the sum of N centered random variables approximately follows a normal distribution when N is sufficiently large.

Let the random variables x_i, $i = 1, \ldots, N$, be independently and identically distributed with mean μ and variance σ^2.

According to the Lindeberg-Lévy CLT,

$$z_N \equiv \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{x_i - \mu}{\sigma} \tag{6}$$

is **asymptotically distributed** as $\mathrm{N}(0, 1)$.

We can write this result compactly as $z_N \xrightarrow{d} \mathrm{N}(0, 1)$.

It is essential to divide by $\sqrt{N}$ in (6). Why?

Because the x_i are independent, $\text{Var}(z_N)$ is just a sum of variances:

$$\text{Var}(z_N) = N \,\text{Var}\left(\frac{1}{\sqrt{N}}\frac{x_i - \mu}{\sigma}\right) = \frac{N}{N} = 1. \tag{7}$$

If we had divided by N, we would have obtained a random variable with a plim of 0, by an LLN, instead of one with a limiting standard normal distribution.

We must always ensure that a factor of $N^{-1/2} = 1/\sqrt{N}$ is present.

There are many different CLTs.

- We can relax the assumption that the x_i are identically distributed and the assumption that they are independent.
- However, with either too much dependence or too much heterogeneity, a CLT may not apply.

In many cases of interest to us, a CLT says that, for a sequence of uncorrelated random variables x_i, $i = 1, \ldots, N$, with $\mathrm{E}(x_i) = 0$,

$$\mathrm{N}^{-1/2} \sum_{i=1}^{N} x_i = x_N^0 \xrightarrow{d} \mathrm{N}\Big(0, \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \mathrm{Var}(x_i)\Big). \tag{8}$$

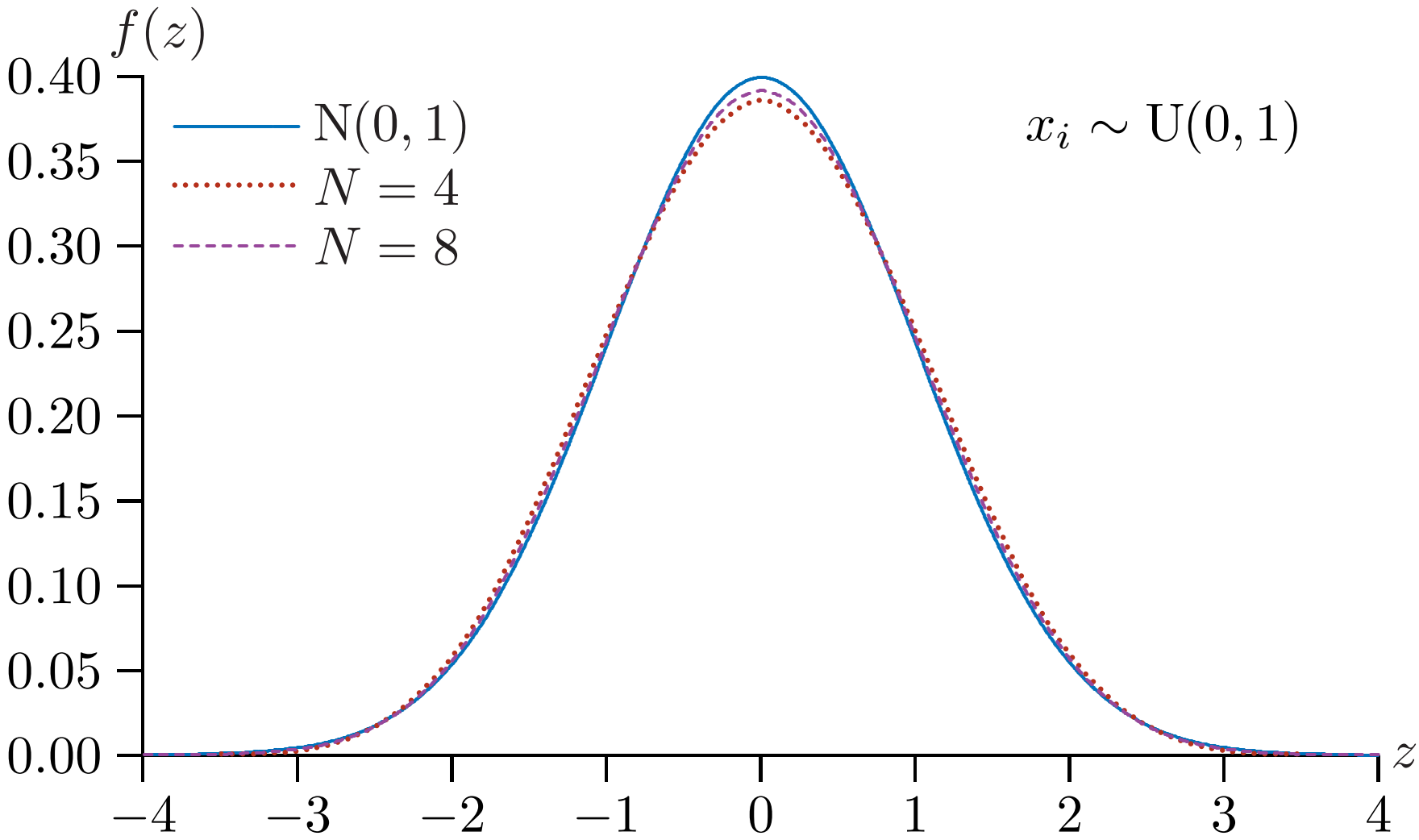
We sometimes need vector, or **multivariate**, CLTs.

Suppose that we have a sequence of uncorrelated random m-vectors $\boldsymbol{x}_i$, for some fixed m, with $\mathrm{E}(\boldsymbol{x}_i) = \mathbf{0}$. Then

$$\mathrm{N}^{-1/2} \sum_{i=1}^{N} \boldsymbol{x}_i = \boldsymbol{x}_N^0 \xrightarrow{d} \mathrm{N}\Big(\mathbf{0}, \lim_{N\to\infty} \mathrm{N}^{-1} \sum_{i=1}^{N} \mathrm{Var}(\boldsymbol{x}_i)\Big), \tag{9}$$

where $\boldsymbol{x}_N^0$ is multivariate normal, and each $\mathrm{Var}(\boldsymbol{x}_i)$ is an $m \times m$ matrix.

CLTs often provide good approximations even when N is not very large; see the next figure. Here the x_i are uniformly distributed.


$f(z)$
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
N(0, 1)
$N = 4$
$N = 8$
$x_i \sim \mathrm{U}(0, 1)$
z
−4
−3
−2
−1
0
1
2
3
4

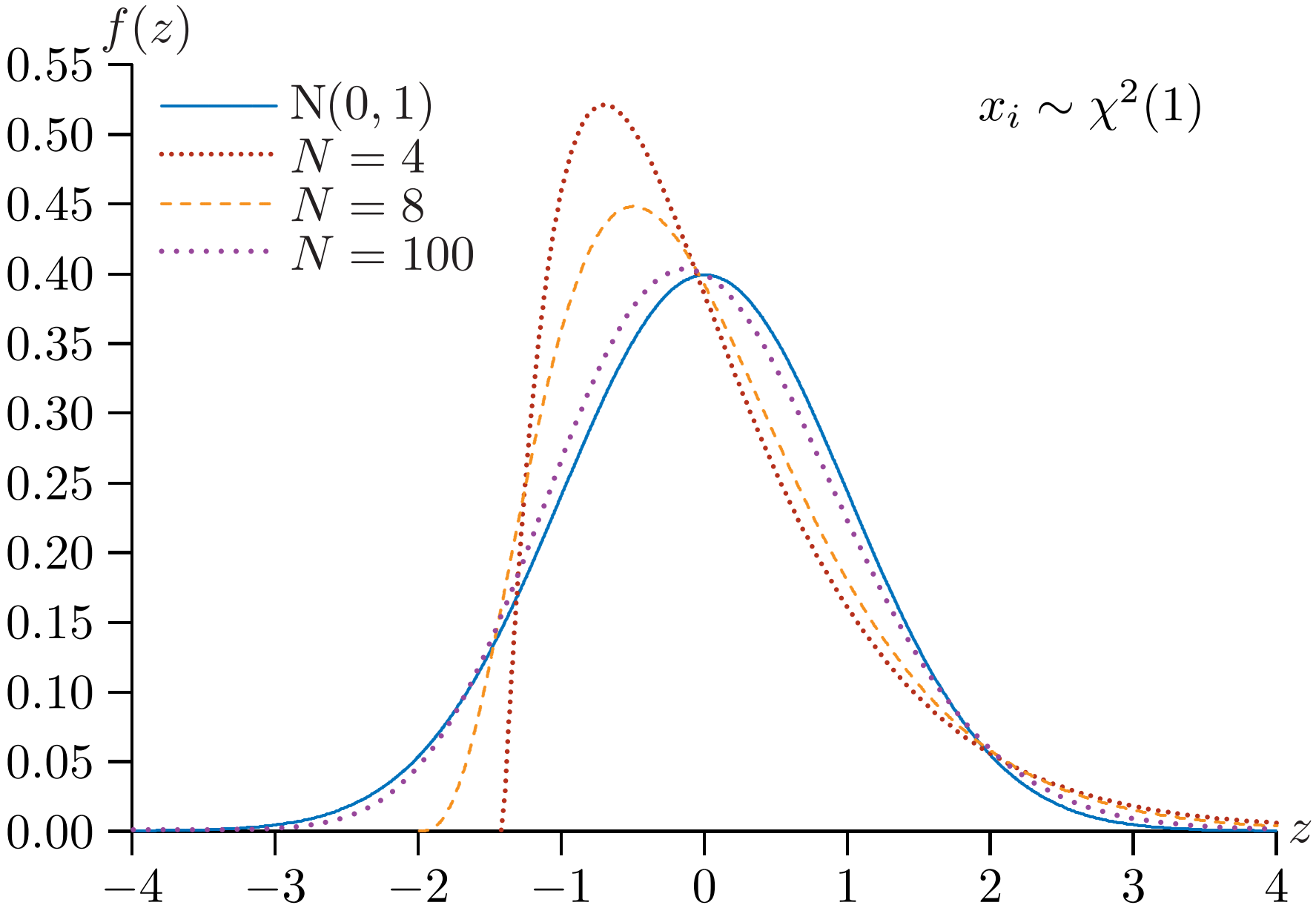
At one time, some routines to generate $\mathrm{N}(0,1)$ pseudo-random numbers actually generated 12 $\mathrm{U}[0,1]$ random numbers!

But CLTs do not always provide good approximations.

In the next figure, the x_i follow the $\chi^2(1)$ distribution, which is extremely right skewed.

- The mode is 0, there are no values less than 0, and there is a very long right-hand tail.
- For $N = 4$ and $N = 8$, the standard normal provides a poor approximation to the actual distribution of z_N.
- The approximation is not bad for $N = 100$, but it is far from perfect. The mode is negative, and it is skewed to the right.
- Thus it can be very dangerous to rely on a central limit theorem!

Asymptotic theory may work poorly when the LLNs or CLTs that it relies on do not provide good approximations.


$f(z)$
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
N(0, 1)
$N = 4$
$N = 8$
$N = 100$
$x_i \sim \chi^2(1)$
−4
−3
−2
−1
0
1
2
3
4
z

Asymptotic Normality and Root-N Consistency

Suppose the data are generated by the DGP

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{u}, \quad \boldsymbol{u} \sim \text{IID}(\boldsymbol{0}, \sigma_0^2 \mathbf{I}), \tag{10}$$

instead of the classical normal linear model.

The disturbances have mean 0 and variance σ_0^2.

If we are using time-series (or panel) data, $\boldsymbol{X}_i$ may contain lagged dependent variables, so the exogeneity assumption is replaced by

$$\text{E}(u_i \mid \Omega_i) = 0 \quad \text{and} \quad \text{E}(u_i^2 \mid \Omega_i) = \sigma_0^2. \tag{11}$$

Here Ω_i denotes the information set, which includes all current explanatory variables and all lagged variables.

Equations (11) say that the disturbances are **innovations**, which implies that the explanatory variables $\boldsymbol{X}_i$ are **predetermined**.

The second equation in (11) just says that the conditional variance of u_i is constant, so that they are homoskedastic. We will relax this assumption later on.

We also assume that the DGP for the explanatory variables is such that

$$\operatorname*{plim}_{N\to\infty} \frac{1}{N} \boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}, \tag{12}$$

where $\boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}$ is a finite, deterministic, positive definite matrix. We have seen it several times before.

Condition (12) would not hold if one of the columns of the $\boldsymbol{X}$ matrix were a linear time trend, because $\sum_{t=1}^{T} t^2$ grows at a rate faster than T.

We wish to apply a multivariate CLT to the k-vector

$$\boldsymbol{v} \equiv N^{-1/2} \boldsymbol{X}^\top \boldsymbol{u} = N^{-1/2} \sum_{i=1}^{N} u_i \boldsymbol{X}_i{}^\top. \tag{13}$$

By assumption, $\mathrm{E}(u_i \,|\, \boldsymbol{X}_i) = 0$. This implies that $\mathrm{E}(u_i \boldsymbol{X}_i^\top) = \mathbf{0}$, as required for the CLT.

Thus, assuming that the $u_i \boldsymbol{X}_i^\top$ satisfy technical assumptions, we have

$$\begin{aligned} \boldsymbol{v} &\xrightarrow{d} \mathrm{N}\Big(\mathbf{0}, \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \mathrm{Var}(u_i \boldsymbol{X}_i^\top)\Big) \\ &= \mathrm{N}\Big(\mathbf{0}, \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \mathrm{E}(u_i^2 \boldsymbol{X}_i^\top \boldsymbol{X}_i)\Big). \end{aligned} \tag{14}$$

Because $\boldsymbol{X}_i$ is a $1 \times k$ row vector, the covariance matrix here is $k \times k$, as it must be.

The assumption that the disturbances are homoskedastic allows us to simplify the asymptotic covariance matrix.

$$\mathrm{Var}(v) \overset{a}{=} \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \mathrm{E}(u_i^2 \boldsymbol{X}_i^\top \boldsymbol{X}_i) \tag{15}$$

$$= \lim_{N\to\infty} \sigma_0^2 \frac{1}{N} \sum_{i=1}^{N} \mathrm{E}(\boldsymbol{X}_i^\top \boldsymbol{X}_i) \tag{16}$$

$$= \sigma_0^2 \operatorname*{plim}_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{X}_i^\top \boldsymbol{X}_i = \sigma_0^2 \operatorname*{plim}_{N\to\infty} \frac{1}{N} \boldsymbol{X}^\top \boldsymbol{X} \tag{17}$$

$$= \sigma_0^2 \boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}. \tag{18}$$

The last equality follows from assumption (12).

Thus we conclude that

$$v \xrightarrow{d} \mathrm{N}(\boldsymbol{0}, \sigma_0^2 \boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}). \tag{19}$$

The estimation error of the vector of OLS estimates is

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{u}. \tag{20}$$

Since $\hat{\boldsymbol{\beta}}$ is consistent, $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$ must tend to a limit of $\mathbf{0}$ as $N \to \infty$.

- Under standard assumptions, sums of random variables with nonzero means, like the elements of $\boldsymbol{X}^\top \boldsymbol{X}$, are $O_p(N)$.
- Weighted sums of random variables with zero means, like the elements of the vector $\boldsymbol{X}^\top \boldsymbol{u}$, are $O_p(N^{1/2})$.
- We need to multiply the former by N^{-1} and the latter by $N^{-1/2}$ in order to obtain quantities that are $O_p(1)$.

We can rewrite (20) as

$$N^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = (N^{-1}\boldsymbol{X}^\top \boldsymbol{X})^{-1} N^{-1/2} \boldsymbol{X}^\top \boldsymbol{u} = (N^{-1}\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{v}. \tag{21}$$

Asymptotically, $N^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = \boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}^{-1} \boldsymbol{v}$. We have to blow up $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$ here so that (21) is $O_p(1)$.

Because $S_{X^\top X}$ is deterministic, we find that, asymptotically,

$$\text{Var}\left(N^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\right) = \sigma_0^2 \boldsymbol{S}_{X^\top X}^{-1} \boldsymbol{S}_{X^\top X} \boldsymbol{S}_{X^\top X}^{-1} = \sigma_0^2 \boldsymbol{S}_{X^\top X}^{-1}. \tag{22}$$

Moreover, because v is asymptotically normally distributed,

$$\mathrm{N}^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \sigma_0^2 \boldsymbol{S}_{X^\top X}^{-1}). \tag{23}$$

Informally, we say that the vector $\hat{\boldsymbol{\beta}}$ is **asymptotically normal**.

Because s^2 estimates σ^2 consistently, we also conclude that

$$\operatorname*{plim}_{N\to\infty} s^2 (N^{-1}\boldsymbol{X}^\top \boldsymbol{X})^{-1} = \sigma_0^2 \boldsymbol{S}_{X^\top X}^{-1}. \tag{24}$$

In ETM2, the results (23) and (24) are collected into Theorem 4.3.

(23) allows us to pretend that $\hat{\boldsymbol{\beta}}$ is normally distributed with mean $\boldsymbol{\beta}_0$.

(24) allows us to use $s^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ to estimate $\text{Var}(\hat{\boldsymbol{\beta}})$.

These are both just approximations. **Asymptotic inference** based on these approximations is not necessarily reliable.

$\hat{\boldsymbol{\beta}}$ is said to be **root-N consistent**, because its **rate of convergence** to $\boldsymbol{\beta}_0$ is $N^{-1/2} = 1/N^{1/2}$.

The factors of N in (23) and (24) are only there for purposes of asymptotic theory.

In practice, we pretend that

$$\hat{\boldsymbol{\beta}} \sim \text{N}\big(\boldsymbol{\beta}_0, s^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\big). \tag{25}$$

But this involves two approximations, which may be good or bad.