Asymptotic Analysis

Paolo Zacchia

Probability and Statistics

Lecture 6

Why asymptotics?

- This lecture characterizes probabilistic results for **samples** in **asymptotic** settings: when the sample size N is large.
- The focus is on **convergence** results for selected statistics: their value and distribution for large N.
- These results greatly facilitate estimation & inference when *exact* results on sampling distributions are hard to obtain.
- The main objective is to characterize the behavior of known **estimators** (MM and MLE) in asymptotic environments.
- This is achieved via the analysis of two fundamental results about the sample mean $\bar{\boldsymbol{x}}$: the **law of large numbers** and the **central limit theorem**.

Random sequences

To characterize asymptotic results it is necessary to adopt a notation that helps express the dependence of these results on the sample size.

Definition 1

Random sequence. Any random vector expressed as an N-indexed sequence, write it as $\boldsymbol{x}_N = (X_{1N}, \ldots, X_{KN})^{\mathrm{T}}$, is a random *sequence*. In the univariate context ($K = 1$), one can write it simply as X_N.

The definition can also apply to sequences of *random matrices* having dimension $J \times K$, which combine J vectorial sequences $\boldsymbol{x}_{jN}$ of length K for $j = 1, \ldots, J$. One such matrix is denoted for example as follows.

$$\boldsymbol{X}_N = \begin{bmatrix} \boldsymbol{x}_{1N} & \boldsymbol{x}_{2N} & \ldots & \boldsymbol{x}_{JN} \end{bmatrix}^{\mathrm{T}}$$

Example: both the sample mean and the sample variance-covariance:

$$\bar{\boldsymbol{x}}_N = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{x}_i \quad \text{and} \quad \boldsymbol{S}_N = \frac{1}{N-1} \sum_{i=1}^{N} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_N)(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_N)^{\mathrm{T}}$$

are two random sequences, as they are statistics that depend upon the sample size N. Their univariate versions are written as $\bar{X}_N$ and S_N^2.

Boundedness in probability

The first asymptotic concept related to the idea of "convergence," as applied to random sequences, is defined next.

Definition 2

Boundedness in Probability. A sequence $\boldsymbol{x}_N$ of random vectors is *bounded in probability* if and only if, for any $\varepsilon > 0$, there exists some number $\delta_\varepsilon < \infty$ and an integer N_ε such that

$$\mathbb{P}\left(\|\boldsymbol{x}_N\| \geq \delta_\varepsilon\right) < \varepsilon \quad \forall N \geq N_\varepsilon$$

which is also written as $\boldsymbol{x}_N = \mathcal{O}_p\left(1\right)$ and read as "$\boldsymbol{x}_N$ is big p-oh one."

This is a desirable properties of random sequences, however it is still not fully satisfactory, as it allows the distribution of random vectors to remain "dense" around a certain interval.

Convergence in probability

The following "convergence" concept is stronger than the previous.

Definition 3

Convergence in Probability. A sequence $\boldsymbol{x}_N$ of random vectors converges in probability to a constant vector $\mathbf{c}$ if

$$\lim_{N\to\infty} \mathbb{P}\left(\|\boldsymbol{x}_N - \mathbf{c}\| > \delta\right) = 0$$

for *any* positive real number $\delta > 0$.

This definition formalizes the idea that as the sample size N grows increasingly larger, the probability distribution of $\boldsymbol{x}_N$ concentrates within an increasingly smaller neighborhood of $\mathbf{c}$.

Convergence in probability is usually denoted in one of two ways.

$$\boldsymbol{x}_N \xrightarrow{p} \mathbf{c}$$
$$\operatorname{plim} \boldsymbol{x}_N = \mathbf{c}$$

Here the first type of notation (using $\xrightarrow{p}$) is preferred.

Convergence implies boundedness in probability

Theorem 1

Convergent Random Sequences are also Bounded. *If some sequence* $\boldsymbol{x}_N$ *of random vectors converges in probability to a constant* $\mathbf{c}$, *that is* $\boldsymbol{x}_N \xrightarrow{p} \mathbf{c}$, *then it is also bounded:* $\boldsymbol{x}_N = \mathcal{O}_p(1)$.

Proof.

By the definition of convergence in probability, for any $\varepsilon > 0$ there is always an integer N_ε such that

$$\mathbb{P}\left(\|\boldsymbol{x}_N - \mathbf{c}\| > \delta\right) < \varepsilon \quad \forall N \geq N_\varepsilon$$

thus by setting $\delta_\varepsilon = \delta + \|\boldsymbol{x}_{N_\varepsilon}\| - \|\boldsymbol{x}_{N_\varepsilon} - \mathbf{c}\|$ one gets $\boldsymbol{x}_N = \mathcal{O}_p(1)$. □

Hence, while "boundedness" is valid for a specific constant δ_ϵ so long as N large enough, "convergence" must be true for any δ.

Convergence of random to real sequances

If convergence in probability holds for $\mathbf{c} = \mathbf{0}$, one can also write:

$$\boldsymbol{x}_N = o_p(1)$$

which is read as "$\boldsymbol{x}_N$ is little p-oh one." This notation helps develop the following concept.

Definition 4

Convergence of Random Sequences to Real Sequences. Consider a random sequence $\boldsymbol{x}_N$, and a *non*-random sequence $\mathbf{a}_N$ of the same dimension K as $\boldsymbol{x}_N$. Moreover, define the random sequence $\boldsymbol{z}_N = (Z_{1N}, \ldots, Z_{KN})^{\mathrm{T}}$ where $Z_{kn} = X_{kn}/a_{kn}$ for $k = 1, \ldots, K$ and for $n = 1, 2, \ldots$ to infinity.

1. If $\boldsymbol{z}_N = \mathcal{O}_p(1)$, then $\boldsymbol{x}_N$ is said to be bounded in probability by $\mathbf{a}_N$, which one can write as $\boldsymbol{x}_N = \mathcal{O}_p(\mathbf{a}_N)$.
2. If $\boldsymbol{z}_N = o_p(1)$, then $\boldsymbol{x}_N$ is said to converge in probability to $\mathbf{a}_N$, which one can write as $\boldsymbol{x}_N = o_p(\mathbf{a}_N)$.

Convergence in r-th Mean

The following asymptotic concept is even stronger than convergence in probability.

Definition 5

Convergence in r-th Mean. A random sequence $\boldsymbol{x}_N$ is said to converge in r-th mean to a constant vector $\mathbf{c}$ if the following holds.

$$\lim_{N\to\infty} \mathbb{E}\left[\|\boldsymbol{x}_N - \mathbf{c}\|^r\right] = 0$$

In the special case with $r = 2$, this concept is known as **Convergence in Quadratic Mean** and is also expressed as follows.

$$\boldsymbol{x}_N \overset{qm}{\to} \mathbf{c}$$

This particular kind of convergence is not as general as convergence in probability, but it may be more convenient to work with, given that it is based upon moments.

Convergence in lower means

Intuitively, if a random sequence converges in the r-th mean it shall also converge to means of order lowern than r.

Theorem 2

Convergence in Lower Means. *A random sequence* $\boldsymbol{x}_N$ *that converges in* r*-th mean to some constant vector* $\mathbf{c}$ *also converges in* s*-th mean to* $\mathbf{c}$ *for* $s < r$*.*

Proof.

The proof is based on Jensen's Inequality:

$$\begin{aligned}\lim_{N\to\infty} \mathbb{E}\left[\|\boldsymbol{x}_N - \mathbf{c}\|^s\right] &= \lim_{N\to\infty} \mathbb{E}\left[\left(\|\boldsymbol{x}_N - \mathbf{c}\|^r\right)^{\frac{s}{r}}\right] \\ &\leq \lim_{N\to\infty} \left\{\mathbb{E}\left[\|\boldsymbol{x}_N - \mathbf{c}\|^r\right]\right\}^{\frac{s}{r}} \\ &= 0\end{aligned}$$

since $\lim_{N\to\infty} \mathbb{E}\left[\|\boldsymbol{x}_N - \mathbf{c}\|^r\right] = 0$. □

Convergence in quadratic mean (1/2)

Theorem 3

Convergence in Quadratic Mean and Probability. *If a random sequence $\boldsymbol{x}_N$ converges in r-th mean to a constant vector $\mathbf{c}$ for $r \geq 2$ (that is, at least $\boldsymbol{x}_N \overset{qm}{\to} \mathbf{c}$), then it also converges in probability to $\mathbf{c}$.*

Proof.

Define the (one-dimensional) *nonnegative* random sequence Q_N as:

$$Q_N = \|\boldsymbol{x}_N - \mathbf{c}\| = \sqrt{(\boldsymbol{x}_N - \mathbf{c})^{\mathrm{T}} (\boldsymbol{x}_N - \mathbf{c})} \in \mathbb{R}_+$$

and notice that by Theorem 2 it must converge in *first* mean.

$$\lim_{N\to\infty} \mathbb{E}\left[Q_N\right] = \lim_{N\to\infty} \mathbb{E}\left[\|\boldsymbol{x}_N - \mathbf{c}\|\right] = 0$$

In addition, quadratic mean convergence implies the following.

$$\lim_{N\to\infty} \mathbb{V}\mathrm{ar}\left[Q_N\right] = \lim_{N\to\infty} \mathbb{E}\left[Q_N^2\right] = \lim_{N\to\infty} \mathbb{E}\left[\|\boldsymbol{x}_N - \mathbf{c}\|^2\right] = 0$$

(Continues...)

Convergence in quadratic mean (2/2)

Theorem 3

Proof.

(Continued.) At the same time, by Čebyšëv's Inequality:

$$\mathbb{P}\left(|Q_N - \mathbb{E}\left[Q_N\right]| > \delta\right) \leq \frac{\mathbb{V}\mathrm{ar}\left[Q_N\right]}{\delta^2}$$

therefore, taking limits on both sides gives:

$$\begin{aligned}\lim_{N\to\infty} \mathbb{P}\left(\|\boldsymbol{x}_N - \mathbf{c}\| > \delta\right) &= \lim_{N\to\infty} \mathbb{P}\left(|Q_N - \mathbb{E}\left[Q_N\right]| > \delta\right) \\ &\leq \lim_{N\to\infty} \frac{\mathbb{V}\mathrm{ar}\left[Q_N\right]}{\delta^2} \\ &= 0\end{aligned}$$

implying convergence in probability: $\boldsymbol{x}_N \xrightarrow{p} \mathbf{c}$. □

This result is useful to verify that in random samples drawn from a random vector $\boldsymbol{x}$ with finite variance $\mathbb{V}\mathrm{ar}\left[\boldsymbol{x}\right] < \infty$, the sample mean $\boldsymbol{x}_N$ converges in probability to the mean of the population, $\mathbb{E}\left[\boldsymbol{x}\right]$.

Convergence in probability of the sample mean

In a random sample drawn from some random variable X:

$$\lim_{N\to\infty} \mathbb{E}\left[\bar{X}_N\right] = \lim_{N\to\infty} \frac{N}{N}\, \mathbb{E}\left[X\right] = \mathbb{E}\left[X\right]$$

and in addition, if $\mathbb{V}\mathrm{ar}\left[X\right] < \infty$:

$$\lim_{N\to\infty} \mathbb{E}\left[\left(\bar{X}_N - \mathbb{E}\left[\bar{X}_N\right]\right)^2\right] = \lim_{N\to\infty} \mathbb{V}\mathrm{ar}\left[\bar{X}_N\right] = \lim_{N\to\infty} \frac{\mathbb{V}\mathrm{ar}\left[X\right]}{N} = 0$$

and therefore, $\bar{X}_N \overset{qm}{\to} \mathbb{E}\left[X\right]$ which also implies $\bar{X}_N \overset{p}{\to} \mathbb{E}\left[X\right]$.

This generalizes to a multivariate environment: given a random sample drawn from a random vector $\boldsymbol{x}$ with $\mathbb{V}\mathrm{ar}\left[\boldsymbol{x}\right] < \infty$, it is:

$$\bar{\boldsymbol{x}}_N = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i \overset{qm}{\to} \mathbb{E}\left[\boldsymbol{x}\right]$$

which also implies convergence in probability, $\bar{\boldsymbol{x}}_N \overset{p}{\to} \mathbb{E}\left[\boldsymbol{x}\right]$

Almost Sure Convergence

There is yet another, stronger notion of convergence.

Definition 6

Almost Sure Convergence. A sequence $\boldsymbol{x}_N$ of random vectors converges *almost surely*, or *with probability one* to a constant vector $\mathbf{c}$ if it holds that:

$$\mathbb{P}\left(\lim_{N\to\infty} \boldsymbol{x}_N = \mathbf{c}\right) = 1$$

where $\lim_{N\to\infty} \boldsymbol{x}_N$ is a random vector. This is also writes as $\boldsymbol{x}_N \overset{a.s.}{\to} \mathbf{c}$.

One can prove that "almost sure" convergence implies convergence in probability, but occasionally the converse is not true.

Convergence of random matrix sequences

- All concepts and results discussed until here also apply to random sequences that are *matrix-valued*.

- A matrix-valued random sequence $\boldsymbol{X}_N$ is said to converge in probability to some matrix $\mathbf{C}$ if:

$$\lim_{N\to\infty} \mathbb{P}\left(\|\boldsymbol{X}_N - \mathbf{C}\| > \delta\right) = 0$$

- (. . . where, for any matrix $\mathbf{B}$, it is as follows).

$$\|\mathbf{B}\| = \sqrt{\mathrm{tr}\left(\mathbf{B}^{\mathrm{T}}\mathbf{B}\right)}$$

- Convergence in probability of random matrix sequences can be written as follows.

$$\boldsymbol{X}_N \xrightarrow{p} \mathbf{C}$$

Continuous mapping theorem (1/2)

What follows is an extremely useful result.

Theorem 4

Continuous Mapping Theorem. *Consider a vector-valued random sequence* $\boldsymbol{x}_N \in \mathbb{X}$*, a vector* $\mathbf{c} \in \mathbb{X}$ *with the same length as* $\boldsymbol{x}_N$*, as well as a vector-valued continuous function* $\boldsymbol{g}(\cdot)$ *with a set of discontinuity points* $\mathbb{D}_{\boldsymbol{g}}$ *such that:*

$$\mathbb{P}(\mathbf{x} \in \mathbb{D}_{\boldsymbol{g}}) = 0$$

(the probability mass at the discontinuities is zero). It follows that:

$$\boldsymbol{x}_N \xrightarrow{p} \mathbf{c} \Rightarrow \boldsymbol{g}(\boldsymbol{x}_N) \xrightarrow{p} \boldsymbol{g}(\mathbf{c})$$
$$\boldsymbol{x}_N \xrightarrow{a.s.} \mathbf{c} \Rightarrow \boldsymbol{g}(\boldsymbol{x}_N) \xrightarrow{a.s.} \boldsymbol{g}(\mathbf{c})$$

thus, convergence in probability and almost sure convergence are preserved when functions are applied to random sequences.

Proof.

(Sketched.) Only the case about convergence in probability is proved here, for the sake of illustration. **(Continues...)**

Continuous mapping theorem (2/2)

Theorem 4

Proof.

(Continued.) For a given positive number $\delta > 0$, define the set:

$$\mathbb{G}_\delta = \{\mathbf{x} \in \mathbb{X} | \, \mathbf{x} \notin \mathbb{D}_{\boldsymbol{g}} : \exists \mathbf{y} \in \mathbb{X} : \|\mathbf{x} - \mathbf{y}\| < \delta, \|\boldsymbol{g}(\mathbf{x}) - \boldsymbol{g}(\mathbf{y})\| > \varepsilon\}$$

that is, the set of points in $\mathbb{X}$ where $\boldsymbol{g}(\cdot)$ "amplifies" the distance with some other point $\mathbf{y}$ beyond a small neighborhood of ε. In light of this definition:

$$\mathbb{P}(\|\boldsymbol{g}(\boldsymbol{x}_N) - \boldsymbol{g}(\mathbf{c})\| > \varepsilon) \leq \mathbb{P}(\|\boldsymbol{x}_N - \mathbf{c}\| \geq \delta) + \mathbb{P}(\mathbf{c} \in \mathbb{G}_\delta) + \mathbb{P}(\mathbf{c} \in \mathbb{D}_{\boldsymbol{g}})$$

and note that upon taking the limit of the right-hand side as $N \to \infty$, the second term vanishes by definition of a continuous function, while the third term is zero by hypothesis. Therefore:

$$\lim_{N\to\infty} \mathbb{P}(\|\boldsymbol{g}(\boldsymbol{x}_N) - \boldsymbol{g}(\mathbf{c})\| > \varepsilon) \leq \lim_{N\to\infty} \mathbb{P}(\|\boldsymbol{x}_N - \mathbf{c}\| \geq \delta)$$

which proves the result on convergence in probability. □

Uses of the continuous mapping theorem (1/3)

- This result also applies to scalar-valued and matrix-valued sequences.
- The theorem already showcases the convenience of working in an asymptotic environment.
- Note that in general, one cannot derive the expected value of *some function* $g(\widehat{\mu}_N)$ of some given unbiased estimator $\widehat{\mu}_N$ such that $\mathbb{E}[\widehat{\mu}_N] = \mu_0$ for some μ_0.
- (The best one can do is to derive *approximations* based on Jensen's Inequality.)
- If $\widehat{\mu}_N$ converges in probability to μ_0 though, the continuous mapping theorem ensures that in large samples $g(\widehat{\mu}_N)$ also converges in probability to $g(\mu_0)$.

Uses of the continuous mapping theorem (2/3)

A list of 'properties' of random sequences, which can be derived from the continuous mapping theorem, follows suit.

1. **Scalars.** Given two scalar random sequences $X_N \xrightarrow{p} x$ and $Y_N \xrightarrow{p} y$, the following holds.

$$(X_N + Y_N) \xrightarrow{p} x + y$$
$$X_N Y_N \xrightarrow{p} xy$$
$$X_N / Y_N \xrightarrow{p} x/y \quad \text{if } y \neq 0$$

2. **Vectors.** Given two vector random sequences $\boldsymbol{x}_N \xrightarrow{p} \mathbf{x}$ and $\boldsymbol{y}_N \xrightarrow{p} \mathbf{y}$ of equal length, the following holds.

$$\boldsymbol{x}_N^{\mathrm{T}} \boldsymbol{y}_N \xrightarrow{p} \mathbf{x}^{\mathrm{T}} \mathbf{y}$$
$$\boldsymbol{x}_N \boldsymbol{y}_N^{\mathrm{T}} \xrightarrow{p} \mathbf{x}\mathbf{y}^{\mathrm{T}}$$

Uses of the continuous mapping theorem (3/3)

3. **Matrices.** Given two matrix random sequences $\boldsymbol{X}_N \xrightarrow{p} \mathbf{X}$ and $\boldsymbol{Y}_N \xrightarrow{p} \mathbf{Y}$ of appropriate dimension it holds that:

$$\boldsymbol{X}_N \boldsymbol{Y}_N \xrightarrow{p} \mathbf{X}\mathbf{Y}$$

while for sequences of random square matrices of full rank $\boldsymbol{Z}_N \xrightarrow{p} \mathbf{Z}$, it is as follows.

$$\boldsymbol{Z}_N^{-1} \xrightarrow{p} \mathbf{Z}^{-1}$$

4. **Combinations of the above.** Consider the three random sequences X_N, $\boldsymbol{x}_N$ and $\boldsymbol{X}_N$ as above, and suppose that the column dimension of $\boldsymbol{X}_N$ corresponds to the row dimension of $\boldsymbol{x}_N$. Then, the following holds.

$$X_N \boldsymbol{X}_N \boldsymbol{x}_N \xrightarrow{p} x\mathbf{X}\mathbf{x}$$

Laws of Large Numbers

- Endowed with these convergence concepts, it is possible to state and prove the **Laws of Large Numbers**.

- These fundamental results in asymptotic analysis show how **sample means converge to population means** in ways that depend on the assumptions that one makes.

- There are two distinct kinds of Law: **weak** (for convergence in probability) and **strong** (for almost sure convergence).

- While both are stated next, only the weak law is proved. A full-fledged proof would use characteristic functions; for the sake of simplicity, a slightly less general proof that is based on moment-generating functions is given.

- The weak law resembles the result following from quadratic mean convergence, but it does not impose finite variances.

Weak Law of Large Numbers (1/3)

The simplest version of the Law is introduced next.

Theorem 5

Weak Law of Large Numbers (Khinčin's). *The sample mean associated with a random (i.i.d.) sample drawn from the distribution of a random vector* $\boldsymbol{x}$ *with finite mean* $\mathbb{E}[\boldsymbol{x}] < \infty$ *converges in probability to such population mean of* $\boldsymbol{x}$*.*

$$\bar{\boldsymbol{x}}_N = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i \xrightarrow{p} \mathbb{E}[\boldsymbol{x}]$$

Proof.

(Sketched.) The proof is restricted to random vectors $\boldsymbol{x}$ for which the m.g.f. $M_{\boldsymbol{x}}(\mathbf{t})$ is actually defined. A more general analysis, which also allows for random vectors that lack an m.g.f., would use characteristic functions $\phi_{\boldsymbol{x}}(\mathbf{t})$ instead. **(Continues...)**

Weak Law of Large Numbers (2/3)

Theorem 5

Proof.

(Continued.) The m.g.f. of the sample mean $\bar{\boldsymbol{x}}_N$ is, *for a given* N:

$$\begin{aligned} M_{\bar{\boldsymbol{x}}_N}(\mathbf{t}) &= \mathbb{E}\left[\exp\left(\mathbf{t}^{\mathrm{T}}\bar{\boldsymbol{x}}_N\right)\right] \\ &= \mathbb{E}\left[\exp\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{t}^{\mathrm{T}}\boldsymbol{x}_i\right)\right] \\ &= \prod_{i=1}^{N}\mathbb{E}\left[\exp\left(\frac{1}{N}\mathbf{t}^{\mathrm{T}}\boldsymbol{x}_i\right)\right] \\ &= \left[M_{\boldsymbol{x}}\left(\frac{1}{N}\mathbf{t}\right)\right]^N \end{aligned}$$

where the third line follows from independence between observations, and the fourth line relies on observations being identically distributed (so that they have the same m.g.f.); basically, this is an application of the theorem on the m.g.f. of linear combinations. **(Continues. . .)**

Weak Law of Large Numbers (3/3)

Theorem 5

Proof.

(Continued.) From a Taylor expansion around $\mathbf{t}_0 = \mathbf{0}$:

$$M_{\bar{\boldsymbol{x}}_N}(\mathbf{t}) = \left[1 + \frac{\mathbf{t}^{\mathrm{T}}\,\mathbb{E}[\boldsymbol{x}]}{N} + o\left(\frac{\mathbf{t}^{\mathrm{T}}\boldsymbol{\iota}}{N}\right)\right]^N$$

hence, taking the limit gives the following result.

$$\lim_{N\to\infty} M_{\bar{\boldsymbol{x}}_N}(\mathbf{t}) = \exp\left(\mathbf{t}^{\mathrm{T}}\,\mathbb{E}[\boldsymbol{x}]\right)$$

This is a trivial m.g.f.: the one of a *degenerate* discrete random vector where the entire probability mass is concentrated in $\mathbb{E}[\boldsymbol{x}]$! Therefore, exploiting the fact that m.g.f. uniquely characterize distributions, one can conclude that the sample mean indeed converges in probability to its mean as N grows larger. □

Strong Laws of Large Numbers (1/2)

Under appropriate restrictions about the variance of the population, one can also establish almost sure convergence.

Theorem 6

Strong Law of Large Numbers (Kolmogorov's). *If in a random (i.i.d.) sample drawn from the distribution of some random vector* $\boldsymbol{x}$ *it simultaneously holds that:*

i. $\mathbb{E}[\boldsymbol{x}] < \infty$,

ii. $\mathbb{V}\text{ar}[\boldsymbol{x}] < \infty$,

iii. $\sum_{n=1}^{\infty} n^{-2}\,\mathbb{V}\text{ar}[\boldsymbol{x}_n] < \infty$,

then the sample mean $\bar{\boldsymbol{x}}_N = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i$ *converges almost surely to its population mean.*

$$\bar{\boldsymbol{x}}_N = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i \overset{a.s.}{\to} \mathbb{E}[\boldsymbol{x}]$$

Strong Laws of Large Numbers (2/2)

The following result applies to *non-identically distributed* observations.

Theorem 7

Strong Law of Large Numbers (Markov's). *Consider a sample with independent, non identically distributed observations (i.n.i.d.) so that the random vectors $\boldsymbol{x}_i$ that generate it have possibly heterogeneous moments $\mathbb{E}[\boldsymbol{x}_i]$ and $\mathbb{V}\text{ar}[\boldsymbol{x}_i]$. If for some $\delta > 0$ it holds that:*

$$\lim_{N\to\infty} \sum_{i=1}^{\infty} \frac{1}{i^{1+\delta}} \mathbb{E}\left[\left|\boldsymbol{x}_i - \mathbb{E}[\boldsymbol{x}_i]\right|^{1+\delta}\right] < \infty$$

then the following almost sure convergence result holds.

$$\frac{1}{N}\sum_{i=1}^{N} (\boldsymbol{x}_i - \mathbb{E}[\boldsymbol{x}_i]) \xrightarrow{a.s.} \mathbf{0}$$

More complex results that apply to *non-independent* observations also exist. These are widely applied in econometrics.

LLN: Illustrative simulations (1/5)

- To illustrate the functioning of the Theorem in practice, a **simulation** is presented next.

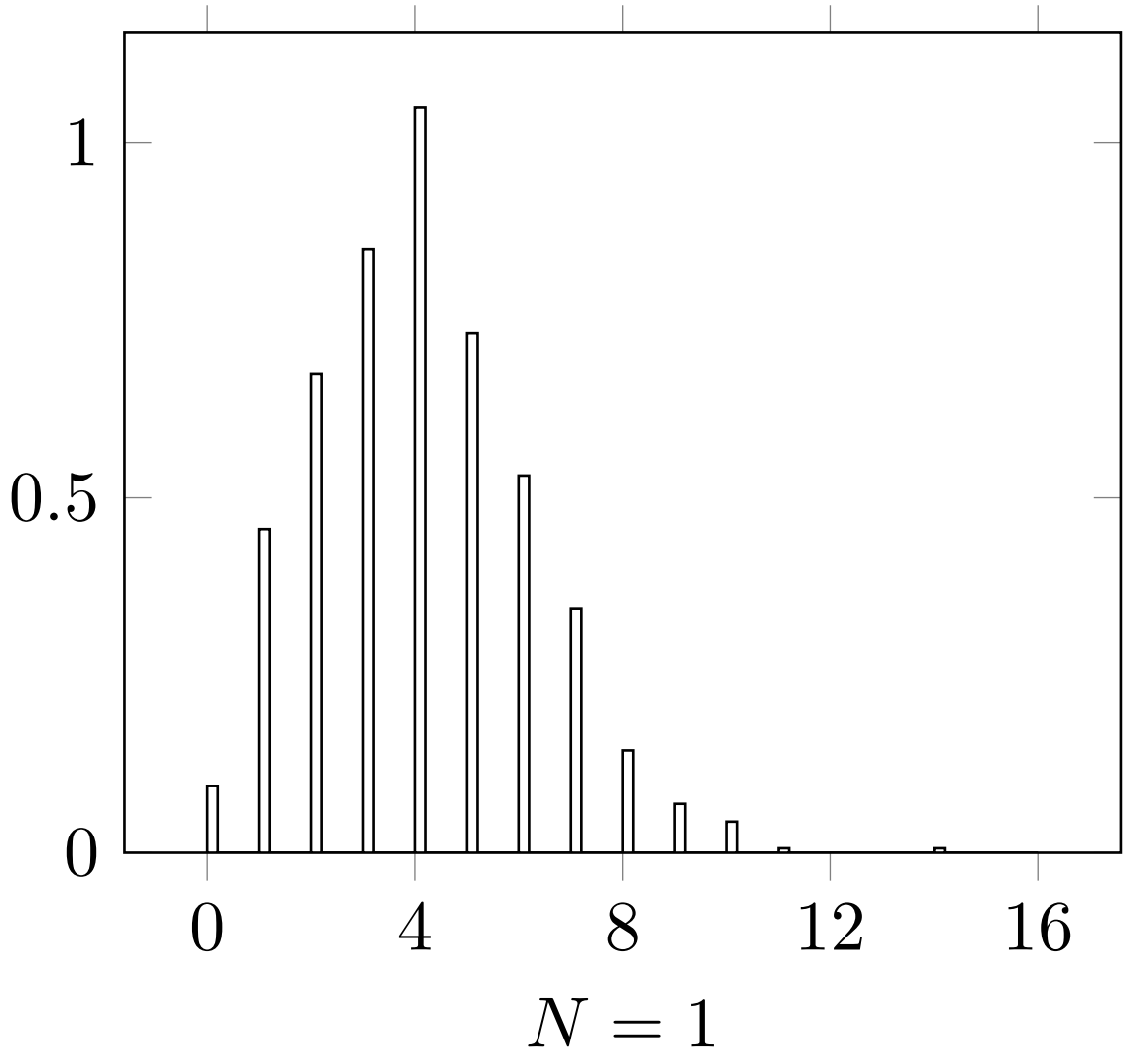
- The simulation is based off random draws from the Poisson distribution with parameter $\lambda = 4$.

- Four **histograms** are presented. All of them bin the values of 800 sample means from 800 simulated samples.

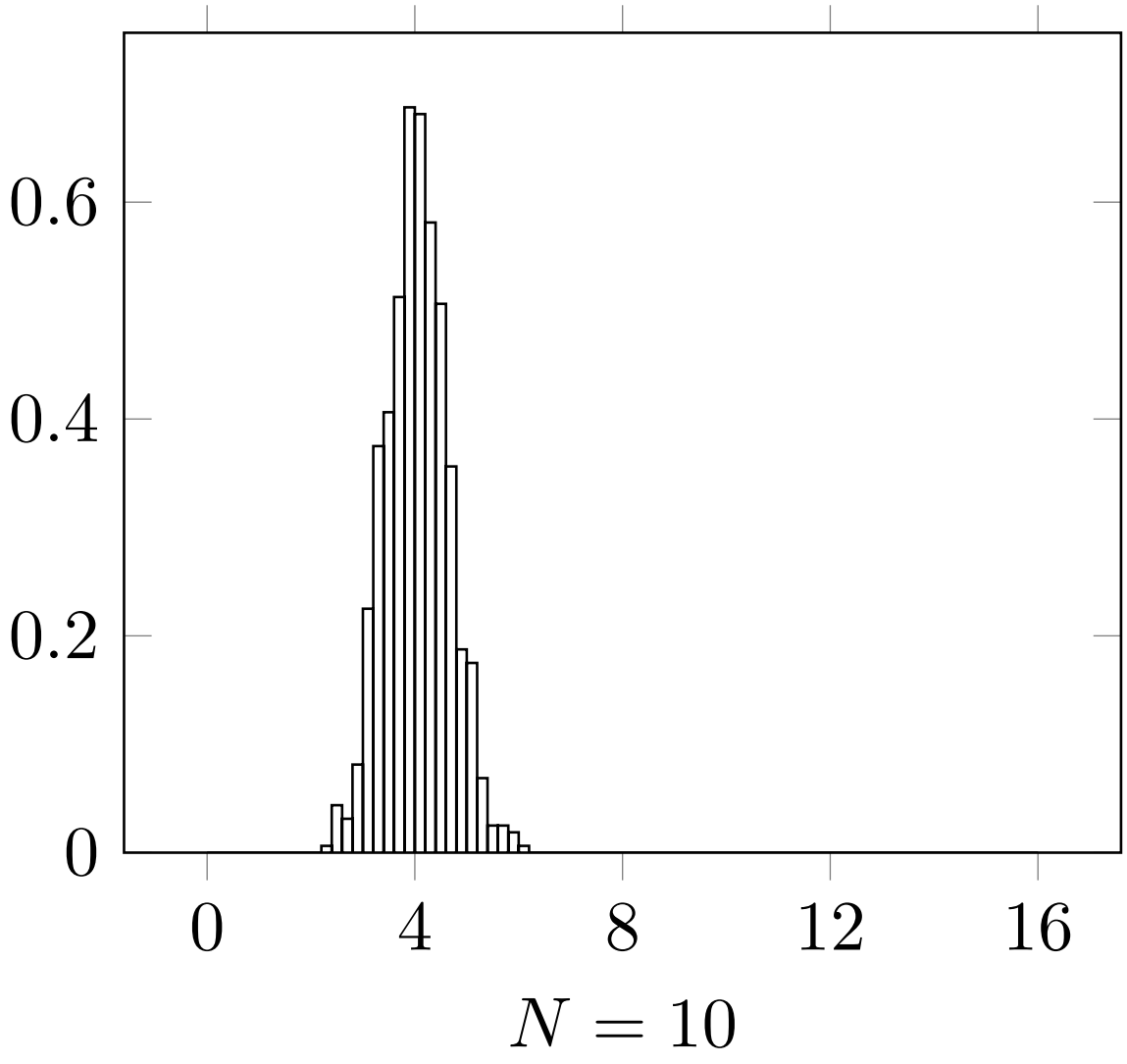
- Across histograms the size of the sample N varies. Thus, if $N = 1$, 800 observations are grouped across 800 samples; if $N = 10$, 8,000 observations are grouped across 800 samples, and so on.

- This helps showing how the empirical sampling distribution of the 800 sample means becomes increasingly concentrated around 4 as N becomes larger.

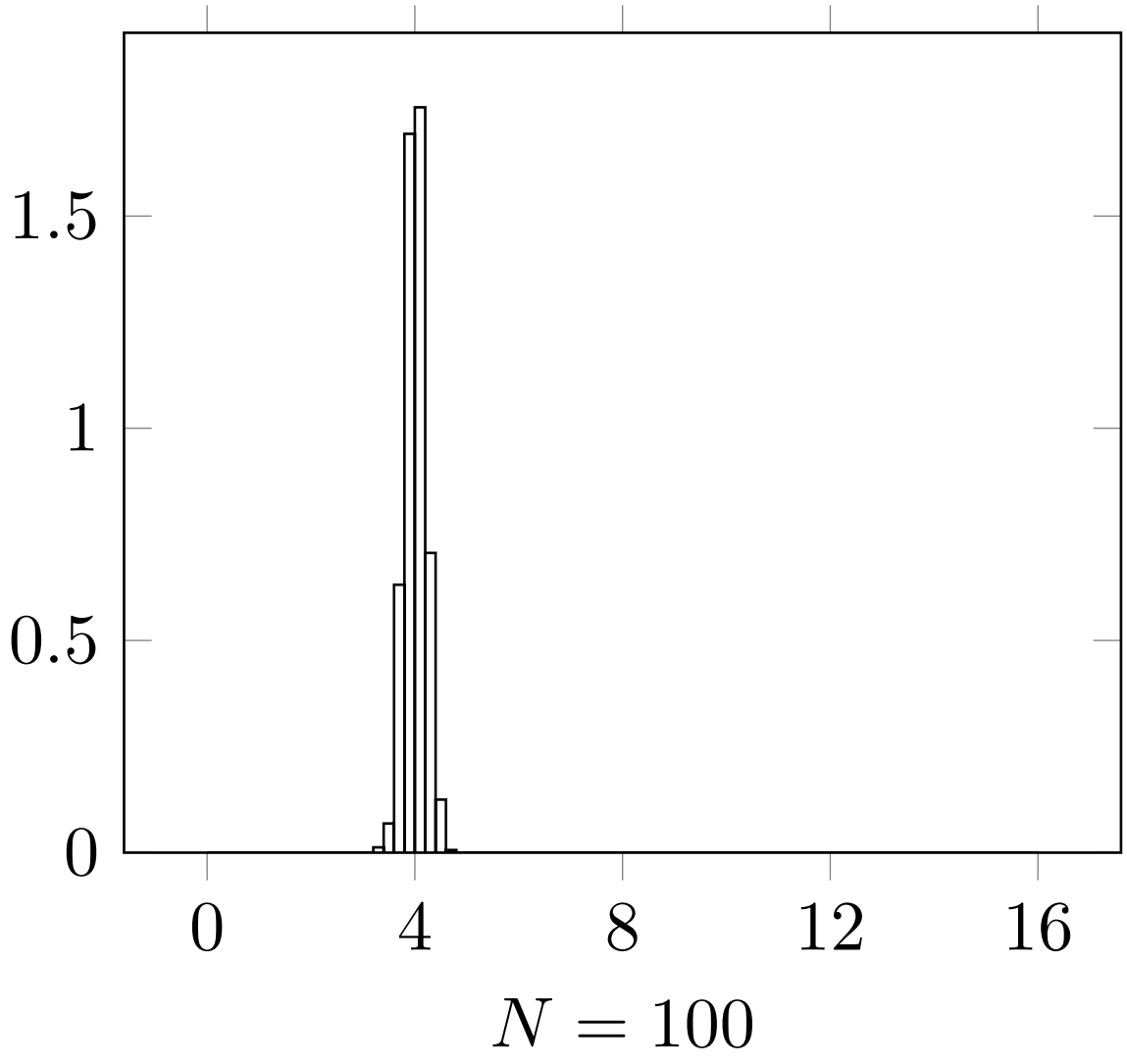
LLN: Illustrative simulations (2/5)



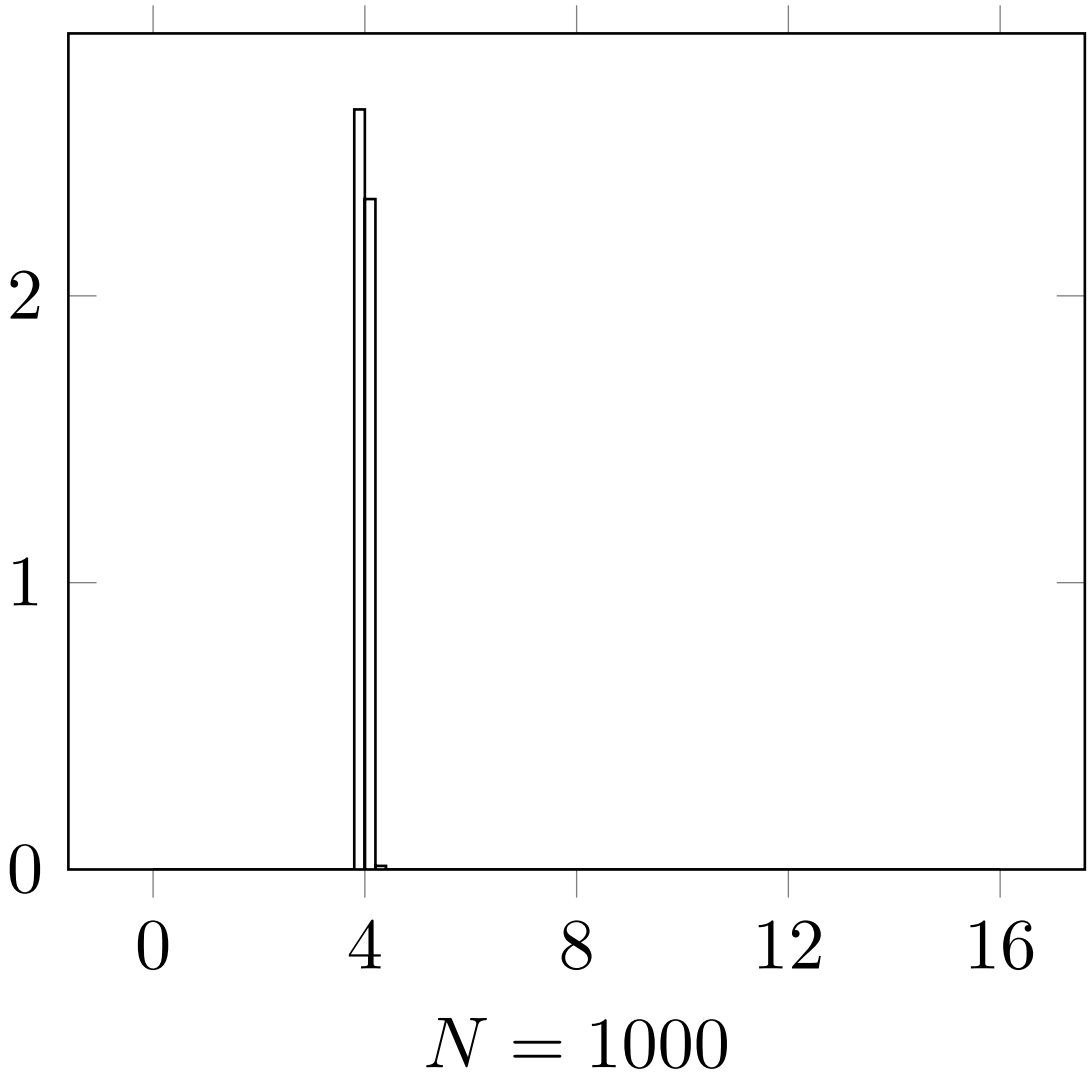
LLN: Illustrative simulations (3/5)



LLN: Illustrative simulations (4/5)



LLN: Illustrative simulations (5/5)



Consistent estimators

The Laws of Large Numbers can be exploited to show that selected estimators converge in probability to the parameters of interest for estimation.

Definition 7

Consistent Estimators. An estimator $\widehat{\theta}_N$ is called *consistent* if it converges in probability to the *true* population parameters θ_0 which it is meant to estimate.

$$\widehat{\theta}_N \xrightarrow{p} \theta_0$$

Note: from now on, the subscript 0 – as in θ_0 – is used to denote the "true" value of the parameters of interest, the ones that characterize the distribution which generates the data.

Before proving consistency of both MM and MLE estimators (under some loose conditions), it is useful to provide an example about the bivariate linear regression model, showing that the estimator(s) that are associated with it are consistent.

Bivariate regression and consistency (1/2)

Consider the bivariate linear regression model from Lecture 3. The MM estimator of the *true* slope parameter β_1 is given by (Lecture 4) as:

$$\widehat{\beta}_{1,MM} = \frac{\sum_{i=1}^{N}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2}$$

where $\bar{X} = N^{-1}\sum_{i=1}^{N} X_i$ and $\bar{Y} = N^{-1}\sum_{i=1}^{N} Y_i$. This estimator can also be obtained via MLE under certain assumptions.

Observe that after dividing both sides of the above ratio by N, the numerator and the denominator are, respectively:

- the *sample covariance* between X_i and Y_i, and
- the *sample variance* of X_i,

(both multiplied by the $(N-1)/N$ factor).

Bivariate regression and consistency (2/2)

Since the two sides of the ratio are (generalized) sample means, by the Weak Law of Large Numbers:

$$\frac{1}{N}\sum_{i=1}^{N}\left(X_i-\bar{X}\right)\left(Y_i-\bar{Y}\right)\xrightarrow{p}\mathbb{C}\mathrm{ov}\left[X_i,Y_i\right]$$

$$\frac{1}{N}\sum_{i=1}^{N}\left(X_i-\bar{X}\right)^2\xrightarrow{p}\mathbb{V}\mathrm{ar}\left[X_i\right]$$

and thanks to the Continuous Mapping Theorem:

$$\widehat{\beta}_{1,MM}\xrightarrow{p}\beta_1$$

this estimator of the slope parameter β_1 is consistent!

An extended analysis shows that the MM estimator of β_0:

$$\widehat{\beta}_{0,MM}=\bar{Y}-\widehat{\beta}_{1,MM}\cdot\bar{X}$$

is also consistent: $\widehat{\beta}_{0,MM}\xrightarrow{p}\beta_0$, again thanks to the Continuous Mapping Theorem.

Consistency of the Method of Moments

Theorem 8

Consistency of the Method of Moments. *An estimator defined as the solution* $\widehat{\boldsymbol{\theta}}_{MM}$ *of a set of sample moments*

$$\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{m}\left(\boldsymbol{x}_i;\widehat{\boldsymbol{\theta}}_{MM}\right)=\mathbf{0}$$

is consistent for the set of parameters $\boldsymbol{\theta}_0$ *that solves the corresponding population moments* $\mathbb{E}\left[\boldsymbol{m}\left(\boldsymbol{x}_i;\boldsymbol{\theta}\right)\right]=\mathbf{0}$*, if such a solution exists (i.e. if the estimation problem is well defined).*

Proof.

(Heuristic.) By some applicable Law of Large Numbers:

$$\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{m}\left(\boldsymbol{x}_i;\widehat{\boldsymbol{\theta}}_{MM}\right)\xrightarrow{p}\mathbb{E}\left[\boldsymbol{m}\left(\boldsymbol{x}_i;\widehat{\boldsymbol{\theta}}_{MM}\right)\right]=\mathbf{0}$$

where the equality to $\mathbf{0}$ follows by definition of MM estimator, which is maintained throughout the sequence as $N\to\infty$. As by hypothesis the zero moment conditions have only one admissible solution, at the probability limit it is $\operatorname{plim}\widehat{\boldsymbol{\theta}}_{MM}=\boldsymbol{\theta}_0$. □

Consistency of Maximum Likelihood (1/3)

Theorem 9

Consistency of Maximum Likelihood Estimators. *In a random sample, an estimator* $\widehat{\theta}_{MLE}$ *which is defined as the maximizer of a log-likelihood function as per*

$$\widehat{\theta}_{MLE} = \arg\max_{\theta \in \Theta} \sum_{i=1}^{N} \log f_{\boldsymbol{x}_i} \left(\theta | \, \mathbf{x}_i \right)$$

converges in probability to that parameter set θ_0 *that maximizes the corresponding population moment function.*

$$\theta_0 = \arg\max_{\theta \in \Theta} \mathbb{E} \left[\log f_{\boldsymbol{x}} \left(\boldsymbol{x}; \theta \right) \right]$$

If such a maximum exists, by the likelihood principle it corresponds to the true parameter of the distribution under analysis.

Proof.

(Continues...)

Consistency of Maximum Likelihood (2/3)

Theorem 9

Proof.

(Continued.) *(Heuristic.)* By the Weak Law of Large Numbers, for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ including $\widehat{\boldsymbol{\theta}}_{MLE}$ and $\boldsymbol{\theta}_0$:

$$\frac{1}{N}\sum_{i=1}^{N} \log f_{\boldsymbol{x}}\left(\boldsymbol{x}_i; \boldsymbol{\theta}\right) \xrightarrow{p} \mathbb{E}\left[\log f_{\boldsymbol{x}}\left(\boldsymbol{x}; \boldsymbol{\theta}\right)\right]$$

while by the definition of MLE the following holds for all $N \in \mathbb{N}$.

$$\frac{1}{N}\sum_{i=1}^{N} \log f_{\boldsymbol{x}}\left(\boldsymbol{x}_i; \widehat{\boldsymbol{\theta}}_{MLE}\right) \geq \frac{1}{N}\sum_{i=1}^{N} \log f_{\boldsymbol{x}}\left(\boldsymbol{x}_i; \boldsymbol{\theta}_0\right) \xrightarrow{p} \mathbb{E}\left[\log f_{\boldsymbol{x}}\left(\boldsymbol{x}; \boldsymbol{\theta}_0\right)\right]$$

Moreover, since $\boldsymbol{\theta}_0$ maximizes the expected logarithmic p.d.f. or p.m.f. *in the population*, the following holds too.

$$\lim_{N\to\infty} \mathbb{P}\left(\mathbb{E}\left[\log f_{\boldsymbol{x}}\left(\boldsymbol{x}; \boldsymbol{\theta}_0\right)\right] \geq \mathbb{E}\left[\log f_{\boldsymbol{x}}\left(\boldsymbol{x}; \widehat{\boldsymbol{\theta}}_{MLE}\right)\right]\right) = 1$$

(Continues. . .)

Consistency of Maximum Likelihood (3/3)

Theorem 9

Proof.

(Continued.) All these facts can be simultaneously reconciled only if, *at the limit*:

$$\frac{1}{N}\sum_{i=1}^{N}\log f_{\boldsymbol{x}}\left(\boldsymbol{x}_i;\widehat{\theta}_{MLE}\right)\xrightarrow{p}\mathbb{E}\left[\log f_{\boldsymbol{x}}\left(\boldsymbol{x};\widehat{\theta}_{MLE}\right)\right]$$

while, at the same time:

$$\mathbb{E}\left[\log f_{\boldsymbol{x}}\left(\boldsymbol{x};\widehat{\theta}_{MLE}\right)\right]=\mathbb{E}\left[\log f_{\boldsymbol{x}}\left(\boldsymbol{x};\theta_0\right)\right]$$

hence, *at the limit* it follows that $\operatorname{plim}\widehat{\theta}_{MLE}=\theta_0$ by the Continuous Mapping Theorem. □

Convergence to random vectors?

- All convergence concepts that were discuss so far concern convergence to a point or interval in $\mathbb{R}^K$.

- What if interest falls on convergence *to a random vector*?

- Consider the expression:

$$\boldsymbol{x}_N \xrightarrow{p} \boldsymbol{x}$$

  which can be read as: "the random sequence $\boldsymbol{x}_N$ converges in probability to the random vector $\boldsymbol{x}$" as follows.

$$\lim_{N\to\infty} \mathbb{P}\left(\|\boldsymbol{x}_N - \boldsymbol{x}\| > \delta\right) = 0$$

- Is that enough so to guarantee that at the probability limit, the distribution (and moments) of $\boldsymbol{x}_N$ and $\boldsymbol{x}$ coincide?

Convergence in distribution

The answer to the question is "no:" such a result is only obtained when the following stronger concept can be applied.

Definition 8

Convergence in Distribution. Consider:

- a sequence of random vectors $\boldsymbol{x}_N$, whose each element has a cumulative distribution function $F_{\boldsymbol{x}_N}(\mathbf{x}_N)$,
- and a random vector $\boldsymbol{x}$ with cumulative distribution function $F_{\boldsymbol{x}}(\mathbf{x})$.

The random sequence $\boldsymbol{x}_N$ is said to converge *in distribution* to $\boldsymbol{x}$ if:

$$\lim_{N\to\infty} |F_{\boldsymbol{x}_N}(\mathbf{x}_N) - F_{\boldsymbol{x}}(\mathbf{x})| = 0$$

at all *continuity* points $\mathbf{x} \in \mathbb{X}$ belonging to the support of $\boldsymbol{x}$. This is usually expressed with the following formalism.

$$\boldsymbol{x}_N \xrightarrow{d} \boldsymbol{x}$$

Limiting distribution, and discussion

The distribution $F_{\boldsymbol{x}}(\mathbf{x})$ from the definition takes the following name.

Definition 9

Limiting Distribution. If $\boldsymbol{x}_N \xrightarrow{d} \boldsymbol{x}$, that is some random sequence $\boldsymbol{x}_N$ converges in distribution to a random vector $\boldsymbol{x}$, $F_{\boldsymbol{x}}(\mathbf{x})$ is said to be the *limiting* distribution of $\boldsymbol{x}_N$.

Observe the following.

- The definition of convergence in distribution indicates that the probabilistic behavior of $\boldsymbol{x}_N$ and $\boldsymbol{x}$ becomes increasingly closer as N grows, eventually it shall coincide;

- This is a stronger concept than is convergence in probability to a random vector, which implies that $\boldsymbol{x}_N$ and $\boldsymbol{x}$ are going to deliver increasingly close realizations as N grows, even if not necessarily with the same probabilities on the support.

Student's t convergence to the normal (1/2)

Observation 1

Asymptotics of Student's t-distribution. Consider a random variable that follows the Student's t-distribution with parameter ν, that is $X \sim \mathcal{T}(\nu)$. As $\nu \to \infty$, the probability distribution of X tends to that of the standard normal distribution, i.e. $\lim_{\nu \to \infty} X = Z \sim \mathcal{N}(0, 1)$.

Proof.

Taking the limit of the probability density function of the Student's t-distribution as $\nu \to \infty$:

$$\lim_{\nu \to \infty} \frac{1}{\mathrm{B}\left(\frac{1}{2}, \frac{\nu}{2}\right)} \frac{1}{\sqrt{\nu}} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

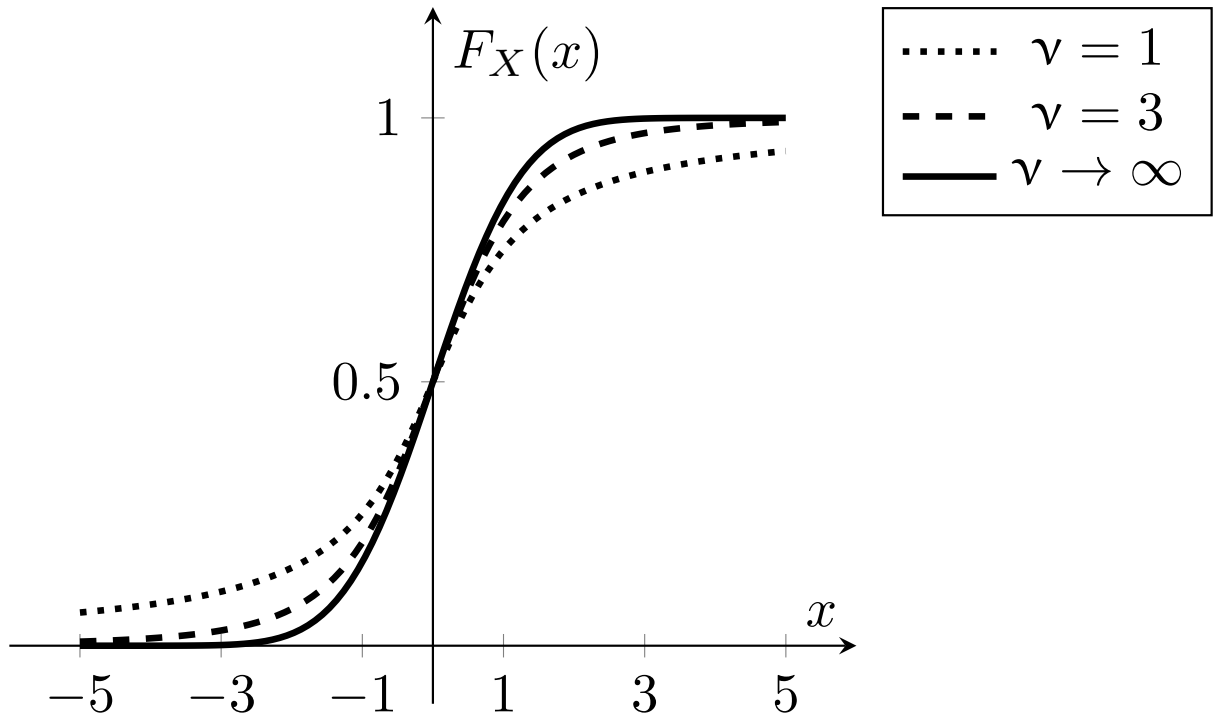
as $\lim_{\nu \to \infty} \sqrt{\nu}\, \mathrm{B}\left(\frac{1}{2}, \frac{\nu}{2}\right) = \sqrt{2\pi}$ by the properties of the Beta function; while:

$$\lim_{\nu \to \infty} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{(\nu+1)}{2}} = \exp\left(-\frac{x^2}{2}\right)$$

by more standard arguments. □

Student's t convergence to the normal (2/2)

This result was already anticipated in Lecture 2. It is worthwhile to replicate the graphical intuition in a different graphical form.



Convergence of *t*-statistics

- Consider a random sample which is drawn from a normally distributed random variable $X \sim \mathcal{N}(\mu, \sigma^2)$.

- As it has been analyzed in Lecture 4, a t-statistic follows a t-distribution with $N-1$ degrees of freedom.

$$\mathrm{t}_N = \sqrt{N}\,\frac{\bar{X}_N - \mu}{S_N} \sim \mathcal{T}_{(N-1)}$$

- If seen as a random sequence, the t-statistic thus converges in distribution to the standard normal.

$$\mathrm{t}_N \xrightarrow{d} \mathcal{N}(0,1)$$

- Hence, with large N one can very reliably perform inference using t-statistics evaluated against the standard normal.

Snedecor's F convergence to the chi-squared

Observation 2

Asymptotics of Snedecor's F-distribution. Consider a random variable that follows Snedecor's F-distribution having parameters ν_1 and ν_2, $X \sim \mathcal{F}(\nu_1, \nu_2)$. As $\nu_2 \to \infty$, the probability distribution of $W = \nu_1 X$ tends to that of a chi-squared distribution with parameter ν_1, i.e. $\lim_{\nu_2 \to \infty} \nu_1 X = W \sim \chi^2(\nu_1)$.

Proof.

After deriving the p.d.f. of $W = \nu_1 X$, take its limit as $\nu_2 \to \infty$:

$$\begin{aligned}
\lim_{\nu_2 \to \infty} f_W(w) &= \lim_{\nu_2 \to \infty} \frac{1}{\mathrm{B}\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)} \left(\frac{1}{\nu_2}\right)^{\frac{\nu_1}{2}} w^{\frac{\nu_1}{2}-1} \left(1 + \frac{w}{\nu_2}\right)^{-\frac{\nu_1+\nu_2}{2}} \\
&= \lim_{\nu_2 \to \infty} \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_2}{2}\right)} (\nu_2 + w)^{-\frac{\nu_1}{2}} \frac{w^{\frac{\nu_1}{2}-1}}{\Gamma\left(\frac{\nu_1}{2}\right)} \left(1 + \frac{w}{\nu_2}\right)^{-\frac{\nu_2}{2}} \\
&= \frac{1}{\Gamma\left(\frac{\nu_1}{2}\right) \cdot 2^{\frac{\nu_1}{2}}} w^{\frac{\nu_1}{2}-1} \exp\left(-\frac{w}{2}\right)
\end{aligned}$$

where $\lim_{\nu_2 \to \infty} \left[\Gamma\left(\frac{\nu_1+\nu_2}{2}\right) / \Gamma\left(\frac{\nu_2}{2}\right)\right] (\nu_2 + w)^{-\frac{\nu_1}{2}} = 2^{-\frac{\nu_1}{2}}$ derives from the properties of the Gamma function. □

Convergence of Hotelling's *t*-squared statistics

- Recall Hotelling's *rescaled* t-squared statistic.

$$\frac{N-K}{K\,(N-1)}\,\mathrm{t}_N^2 =$$

$$= \frac{(N-K)\,N}{K\,(N-1)}\,(\bar{\boldsymbol{x}}-\boldsymbol{\mu})^{\mathrm{T}}\,\boldsymbol{S}^{-1}\,(\bar{\boldsymbol{x}}-\boldsymbol{\mu}) \sim \mathcal{F}_{K,N-K}$$

  For a given N, this statistic follows the F-distribution with paired degrees of freedom K and $N-K$.

- Per Observation 2, Hotelling's t-squared statistic converges in distribution to a chi-squared distribution with K degrees of freedom:

$$\mathrm{t}_N^2 \xrightarrow{d} \chi_K^2$$

  note that the term $(N-K)\,/\,(N-1)$ vanishes as $N \to \infty$.

- Like in the univariate case, this result facilitates statistical inference in multivariate settings.

Gamma convergence to the normal

Observation 3

Asymptotics of the Gamma distribution. Consider a random variable that follows the Gamma distribution with parameters α and β, $X \sim \Gamma(\alpha, \beta)$. Let $\mu = \alpha/\beta$ as well as $\sigma^2 = \alpha/\beta^2$. As $\alpha \to \infty$, the probability distribution of X tends to that of a normal distribution with parameters μ and σ^2, i.e. $\lim_{\alpha\to\infty} X \sim \mathcal{N}(\mu, \sigma^2)$.

Proof.

Define the random variable $Z = (X - \mu)/\sigma = \left(\beta/\sqrt{\alpha}\right) X - \sqrt{\alpha}$; by the properties of m.g.f.s it is:

$$M_Z(t) = \exp\left(-\sqrt{\alpha}t\right) \cdot M_X\left(\frac{\beta}{\sqrt{\alpha}}\, t\right) = \exp\left(-\sqrt{\alpha}t\right)\left(1 - \frac{t}{\sqrt{\alpha}}\right)^{-\alpha}$$

and after some manipulation, the limit as $\alpha \to \infty$ gives:

$$\lim_{\alpha\to\infty} M_Z(t) = \lim_{\alpha\to\infty} \exp\left(-\sqrt{\alpha}t\right)\left(1 - \frac{t}{\sqrt{\alpha}}\right)^{-\alpha} = \exp\left(\frac{t^2}{2}\right)$$

showing that *at the limit*, $Z \sim \mathcal{N}(0, 1)$ and thus $X \sim \mathcal{N}(\mu, \sigma^2)$. □

Mean convergence in exponential samples

- Recall that in a random sample drawn from $X \sim \text{Exp}(\lambda)$ the sample mean is Gamma-distributed, $\bar{X} \sim \Gamma(N, N/\lambda)$.

- Thus, by Observation 3, the following holds.

$$\sqrt{N}\left(\bar{X}_N - \lambda\right) \xrightarrow{d} \mathcal{N}\left(0, \lambda^2\right)$$

- This statement is interpreted in the sense that *for a fixed value of* N:

$$\bar{X}_N \overset{A}{\sim} \mathcal{N}\left(\lambda, \frac{\lambda^2}{N}\right)$$

  where A there stands for "asymptotic" (observe that by the definition of convergence in distribution, N cannot show up in the expression of a limiting distribution).

- This is a particular case of the Central Limit Theorem, one that can help inference about the exponential distribution.

Continuous Mapping Theorem, continued

The Continuous Mapping Theorem also applies to the concept of convergence in distribution.

Theorem 10

Continuous Mapping Theorem (convergence in distribution). *Under the hypotheses of Theorem 4:*

$$\boldsymbol{x}_N \xrightarrow{d} \boldsymbol{x} \Rightarrow \boldsymbol{g}(\boldsymbol{x}_N) \xrightarrow{d} \boldsymbol{g}(\boldsymbol{x})$$

that is, a random sequence which is obtained from the application of a transformation $\boldsymbol{g}(\cdot)$ to some original random sequence $\boldsymbol{x}_N$, converges in distribution to the distribution resulting from applying the transformation $\boldsymbol{g}(\cdot)$ to the random vector $\boldsymbol{x}$ associated with the limiting distribution of $\boldsymbol{x}_N$.

The proof of this statement is omitted: it involves advanced measure theory. This version of the continuous mapping theorem is important, as it allows to prove some properties of random sequences – which are exploited in statistics and econometrics – that are presented next.

Slutskij's Theorem (1/2)

Theorem 11

Slutskij's Theorem. *Consider any two (scalar) random sequences X_N and Y_N such that:*

$$X_N \xrightarrow{d} X$$
$$Y_N \xrightarrow{p} c$$

that is, X_N converges in distribution to that of the random variable X, while Y_N converges in probability to a constant c. Then, the following holds.

$$(X_N + Y_N) \xrightarrow{d} X + c$$
$$X_N Y_N \xrightarrow{d} cX$$
$$X_N / Y_N \xrightarrow{d} X/c \quad \text{if } c \neq 0$$

Proof.

(Continues...)

Slutskij's Theorem (2/2)

Theorem 11

Proof.

(Continued.) Recognize that, as $Y_N \xrightarrow{p} c$, Y_N has a degenerate limiting distribution, and the (vector-valued) random sequence (X_N, Y_N) converges in distribution to that of the random vector (X, c). All the results above follow, therefore, from applying the Continuous Mapping Theorem to three given continuous functions of X_N and Y_N. □

Corollary: Cramér-Wold Device. *Given a random sequence* $\boldsymbol{x}_N$ *and a constant vector* $\mathbf{a}$ *of the same dimension:*

$$\boldsymbol{x}_N \xrightarrow{d} \boldsymbol{x} \Rightarrow \mathbf{a}^{\mathrm{T}}\boldsymbol{x}_N \xrightarrow{d} \mathbf{a}^{\mathrm{T}}\boldsymbol{x}$$

that is, if a vectorial random sequence has a limiting distribution, any linear combination of its elements will converge in distribution to the distribution of the corresponding "limiting" linear combination.

The Extreme Value Theorem (1/4)

It is worth to briefly sketch here the central result of **extreme value theory**: that is, the asymptotic theory of order statistics.

Theorem 12

Extreme Value Theorem (Fisher-Tippett-Gnedenko). *Given a random (i.i.d.) sample* $(X_1, \ldots, X_N)$*, if a convergence in distribution result of the kind*

$$\frac{X_{(N)} - b_N}{a_N} \xrightarrow{d} W$$

can be established – where $X_{(N)}$ *is the maximum order statistic while* $a_N > 0$ *and* b_N *are sequences of real constants – then:*

$$W \sim \text{GEV}\,(0, 1, \xi)$$

for some real ξ*. That is, the limiting distribution of the "normalized" maximum is some standardized type of the Generalized Extreme Value distribution.*

Proof.

(Outline.) The extended proof is quite elaborate. **(Continues...)**

The Extreme Value Theorem (2/4)

Theorem 12

Proof.

(Continued.) The objective is to show that, given a random variable X from which the random sample is drawn, for all the points $x \in \mathbb{X}$ in its support where the distribution $F_X(x)$ is continuous:

$$\lim_{N \to \infty} \left[F_X \left(a_N x - b_N \right) \right]^N = \exp\left(- \left(1 + \xi x \right)^{-\frac{1}{\xi}} \right)$$

where the left-hand side is the limit of the cumulative distribution of the standardized maximum, and the right-hand side is the expression of the cumulative standardized GEV distribution.

By taking the logarithm of this expression, the above is:

$$\lim_{N \to \infty} N \log F_X \left(a_N x - b_N \right) = - \left(1 + \xi x \right)^{\frac{1}{\xi}}$$

showing that $F_X \left(a_N x - b_N \right) \to 1$ as $N \to \infty$. **(Continues...)**

The Extreme Value Theorem (3/4)

Theorem 12

Proof.

(Continued.) Since $-\log(x) \approx 1 - x$ for any given x is close to 1, the above expression approximates the following.

$$\lim_{N\to\infty} \frac{1}{N\left[1 - F_X\left(a_N x - b_N\right)\right]} = \frac{1}{\left(1 + \xi x\right)^{\frac{1}{\xi}}}$$

The rest of the proof is mathematically involved, and it proceeds to:

i. show that the right-hand side of the above expression on is the only admissible limit; and

ii. establish conditions under which $\xi = 0$ (Type I GEV, Gumbel), $\xi > 0$ (Type II GEV, Fréchet) and $\xi < 0$ (Type III GEV, reverse Weibull).

In this context, $\xi = 0$ is interpreted as a limit case (see Lecture 2). □

The Extreme Value Theorem (4/4)

The Extreme Value Theorem has the following implications.

1. A standardized maximum does *not* necessarily *always* converge to a GEV distribution; the Theorem states that *if* it converges, the limiting distribution is GEV.

2. By defining $Y = -X$, for every N it clearly is:

$$Y_{(1)} = -X_{(N)}$$

   which helps identify the distribution of the minimum if the maximum's is known (e.g. reverse vs. traditional Weibull).

3. The technical conditions in the proof that help identify the GEV Type are often useful. For example, one can establish that in sampling from the normal distribution, maxima are Gumbel-distributed.

Central Limit Theorems

- Convergence in distribution is a useful concept, but it is of limited practical use in inference if the limiting distribution of a statistic cannot be derived.

- In this regard, **Central Limit Theorems** are paramount: they prove that some specific functions of **sample means** converge in distribution to the (multivariate) **normal**.

- This is even more important as the result does **not** depend upon the underlying distribution that generates the sample.

- This results helps conduct inference in a variety of settings, including – as discussed later – estimation results from MM and MLE frameworks alike.

- Once again (as in the Law of Large Numbers case), various versions of the result exist, for different sets of assumptions.

Classic Central Limit Theorem (1/5)

Theorem 13

Central Limit Theorem (Lindeberg and Lévy's). *The sample mean $\bar{\boldsymbol{x}}_N$ associated with a random (i.i.d.) sample drawn from the distribution of a random vector $\boldsymbol{x}$ with mean and variance that are both finite: $\mathbb{E}[\boldsymbol{x}] < \infty$ and $\mathbb{V}\mathrm{ar}[\boldsymbol{x}] < \infty$, is such that the random sequence defined as a centered sample mean multiplied by $\sqrt{N}$ converges in distribution to a multivariate normal distribution.*

$$\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i - \mathbb{E}[\boldsymbol{x}]\right) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbb{V}\mathrm{ar}[\boldsymbol{x}])$$

Proof.

(Sketched.) Like in earlier proof for the Weak Law of Large Numbers (Theorem 5) this one will make use of moment-generating functions, but in order to be general enough, characteristic functions should be used instead. **(Continues...)**

Classic Central Limit Theorem (2/5)

Theorem 13

Proof.
(Continued.) Consider the *standardized* random vector

$$\boldsymbol{z} = \left[\mathbb{V}\mathrm{ar}\left[\boldsymbol{x}\right]\right]^{-\frac{1}{2}} \left(\boldsymbol{x} - \mathbb{E}\left[\boldsymbol{x}\right]\right)$$

where the matrix $\left[\mathbb{V}\mathrm{ar}\left[\boldsymbol{x}\right]\right]^{-\frac{1}{2}}$ and its inverse $\left[\mathbb{V}\mathrm{ar}\left[\boldsymbol{x}\right]\right]^{\frac{1}{2}}$ satisfies:

$$\left[\mathbb{V}\mathrm{ar}\left[\boldsymbol{x}\right]\right]^{-\frac{1}{2}} \mathbb{V}\mathrm{ar}\left[\boldsymbol{x}\right] \left[\mathbb{V}\mathrm{ar}\left[\boldsymbol{x}\right]\right]^{-\frac{1}{2}} = \mathbf{I}$$

as well as the following.

$$\left[\mathbb{V}\mathrm{ar}\left[\boldsymbol{x}\right]\right]^{\frac{1}{2}} \left[\mathbb{V}\mathrm{ar}\left[\boldsymbol{x}\right]\right]^{\frac{1}{2}} = \mathbb{V}\mathrm{ar}\left[\boldsymbol{x}\right]$$

Such a matrix can always be constructed because variance-covariance matrices are positive semi-definite. **(Continues...)**

Classic Central Limit Theorem (3/5)

Theorem 13

Proof.

(Continued.) The objective of the proof is to show that:

$$\bar{\bar{\boldsymbol{z}}}_N \equiv \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \boldsymbol{z}_i \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I})$$

that is, the random sequence $\bar{\bar{\boldsymbol{z}}}_N$ defined above converges in distribution to a *standard* multivariate normal distribution. If this holds, the main result would also follow thanks to the (linear) properties of the multivariate normal distribution, per the following relationship.

$$\begin{aligned}
\sqrt{N}\left(\bar{\boldsymbol{x}}_N - \mathbb{E}[\boldsymbol{x}]\right) &= \sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i - \mathbb{E}[\boldsymbol{x}]\right) \\
&= \left[\mathbb{V}\mathrm{ar}[\boldsymbol{x}]\right]^{\frac{1}{2}} \left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} \boldsymbol{z}_i\right)
\end{aligned}$$

(Continues...)

Classic Central Limit Theorem (4/5)

Theorem 13

Proof.

(Continued.) To show this, express the m.g.f. of $\bar{\bar{\boldsymbol{z}}}_N$, for fixed N, as:

$$\begin{aligned} M_{\bar{\bar{\boldsymbol{z}}}_N}(\mathbf{t}) &= \mathbb{E}\left[\exp\left(\mathbf{t}^{\mathrm{T}}\bar{\bar{\boldsymbol{z}}}_N\right)\right] \\ &= \mathbb{E}\left[\exp\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{t}^{\mathrm{T}}\boldsymbol{z}_i\right)\right] \\ &= \prod_{i=1}^{N}\mathbb{E}\left[\exp\left(\frac{1}{\sqrt{N}}\mathbf{t}^{\mathrm{T}}\boldsymbol{z}\right)\right] \\ &= \left[M_{\boldsymbol{z}}\left(\frac{1}{\sqrt{N}}\mathbf{t}\right)\right]^N \end{aligned}$$

by a derivation analogous to the one in the proof of the Weak Law of Large Numbers. **(Continues. . .)**

Classic Central Limit Theorem (5/5)

Theorem 13

Proof.

(Continued.) Just like in that proof, apply here a Taylor expansion of the above expression around $\mathbf{t}_0 = \mathbf{0}$, but now of *second degree*:

$$M_{\bar{\boldsymbol{z}}_N}(\mathbf{t}) = \left[1 + \frac{\mathbf{t}^{\mathrm{T}}\,\mathbb{E}[\boldsymbol{z}]}{\sqrt{N}} + \frac{\mathbf{t}^{\mathrm{T}}\,\mathbb{E}\left[\boldsymbol{z}\boldsymbol{z}^{\mathrm{T}}\right]\mathbf{t}}{2N} + o\left(\frac{\mathbf{t}^{\mathrm{T}}\mathbf{t}}{2N}\right)\right]^N$$
$$= \left[1 + \frac{\mathbf{t}^{\mathrm{T}}\mathbf{t}}{2N} + o\left(\frac{\mathbf{t}^{\mathrm{T}}\mathbf{t}}{2N}\right)\right]^N$$

where the second line exploits the fact that $\mathbb{E}[\boldsymbol{z}] = \mathbf{0}$ and $\mathbb{E}\left[\boldsymbol{z}\boldsymbol{z}^{\mathrm{T}}\right] = \mathbf{I}$ by construction of $\boldsymbol{z}$. Taking the limit for $N \to \infty$ now gives:

$$\lim_{N\to\infty} M_{\bar{\boldsymbol{z}}_N}(\mathbf{t}) = \exp\left(\frac{\mathbf{t}^{\mathrm{T}}\mathbf{t}}{2}\right)$$

which is precisely the m.g.f. of the standard multivariate normal, as it was postulated. □

Use of the Central Limit Theorem

- How to "use" the Central Limit Theorem? Note that for a *given* N, the result can be restated as follows.

$$\bar{\boldsymbol{x}}_N = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i \overset{A}{\sim} \mathcal{N}\left(\mathbb{E}\left[\boldsymbol{x}\right], \frac{1}{N}\mathbb{V}\mathrm{ar}\left[\boldsymbol{x}\right]\right)$$

- The sample mean is "approximately" normally distributed with a variance-covariance decreasing in the sample size.

- The notation $\overset{A}{\sim}$ indicates here that the normal distribution in question, which is called the **asymptotic distribution**, is approximate and is valid for a fixed N, instead of being a "limiting" distribution.

- Recall that limiting distributions cannot be functions of N.

CLT: Illustrative simulations (1/5)

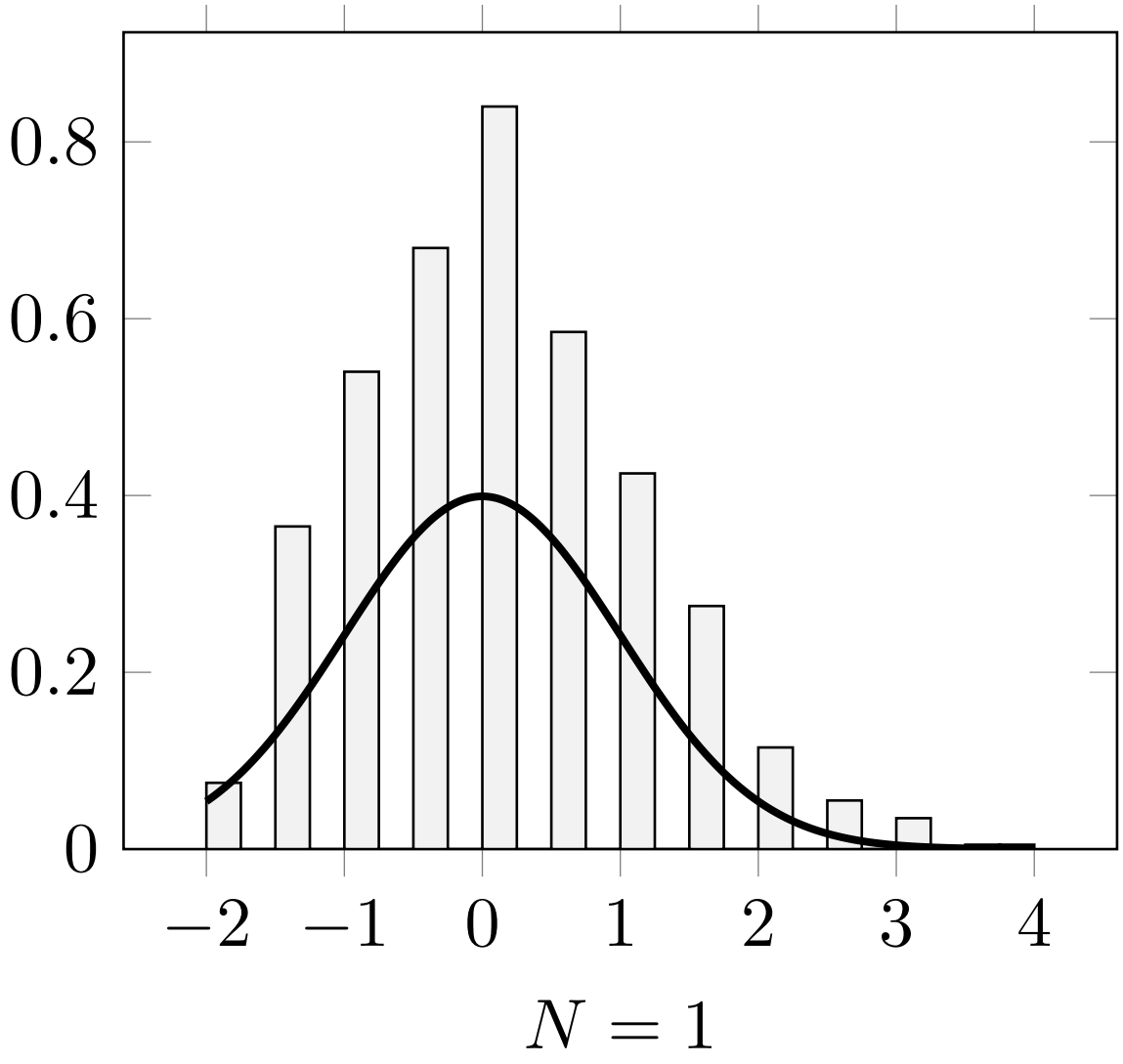
- Once again, **simulations** reveal themselves useful.
- These are based on exactly the same random draws from the Poisson distribution with parameter $\lambda = 4$.
- The four **histograms** now bin 800 values calculated as:

$$\bar{\bar{z}}_N = \sqrt{N} \cdot \frac{\bar{x}_N - 4}{2}$$

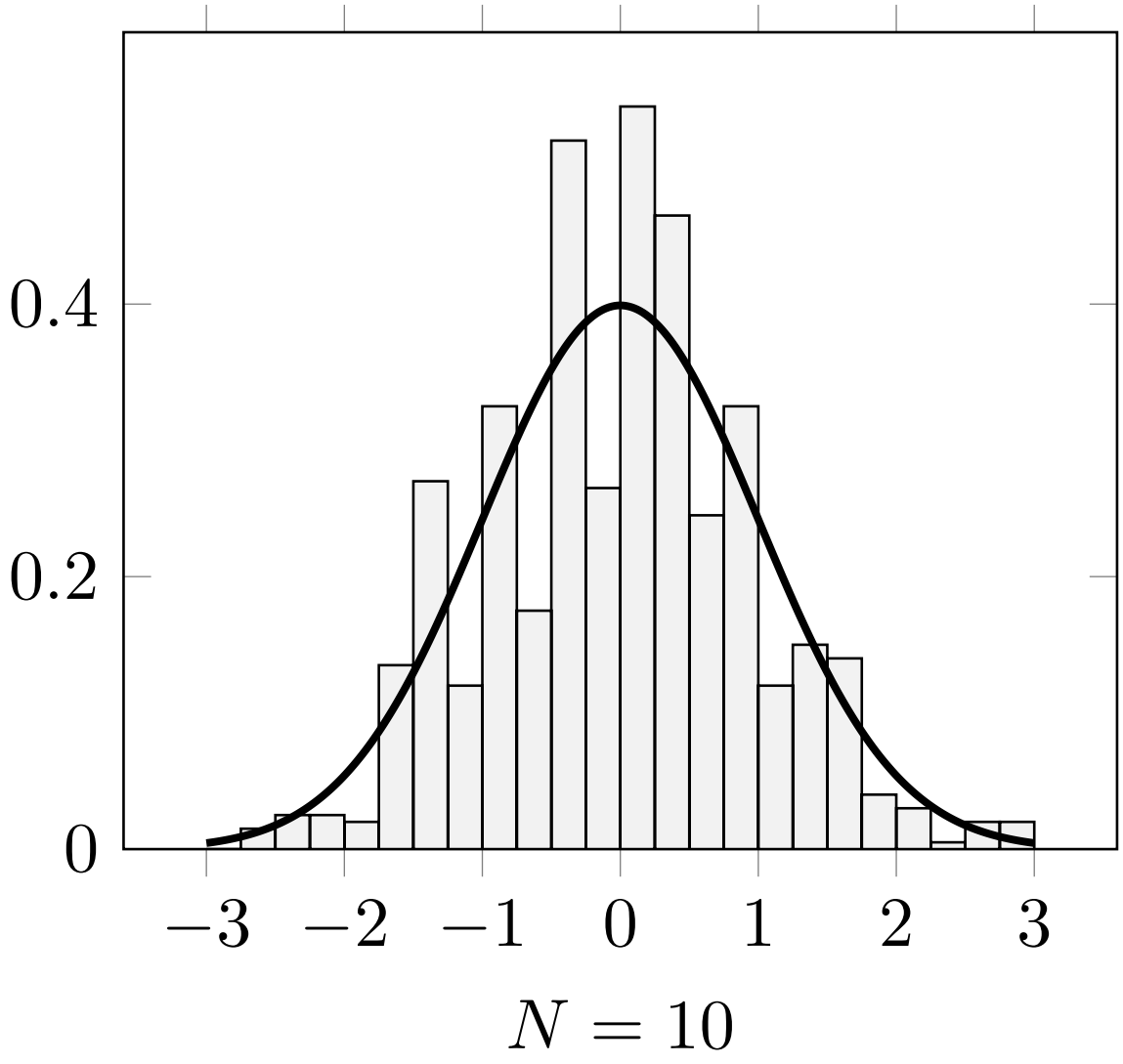
  where $\bar{x}_N$ is a realized mean from the previous simulation. Across histograms the size of the sample N varies as before.

- Note the standardization in the construction of $\bar{\bar{z}}_N$: in this Poisson distribution, both the mean and variance equal 4.
- An overlayed standard normal p.d.f. helps showing how the sampling distribution of this statistic resembles the normal increasingly better as N increases.

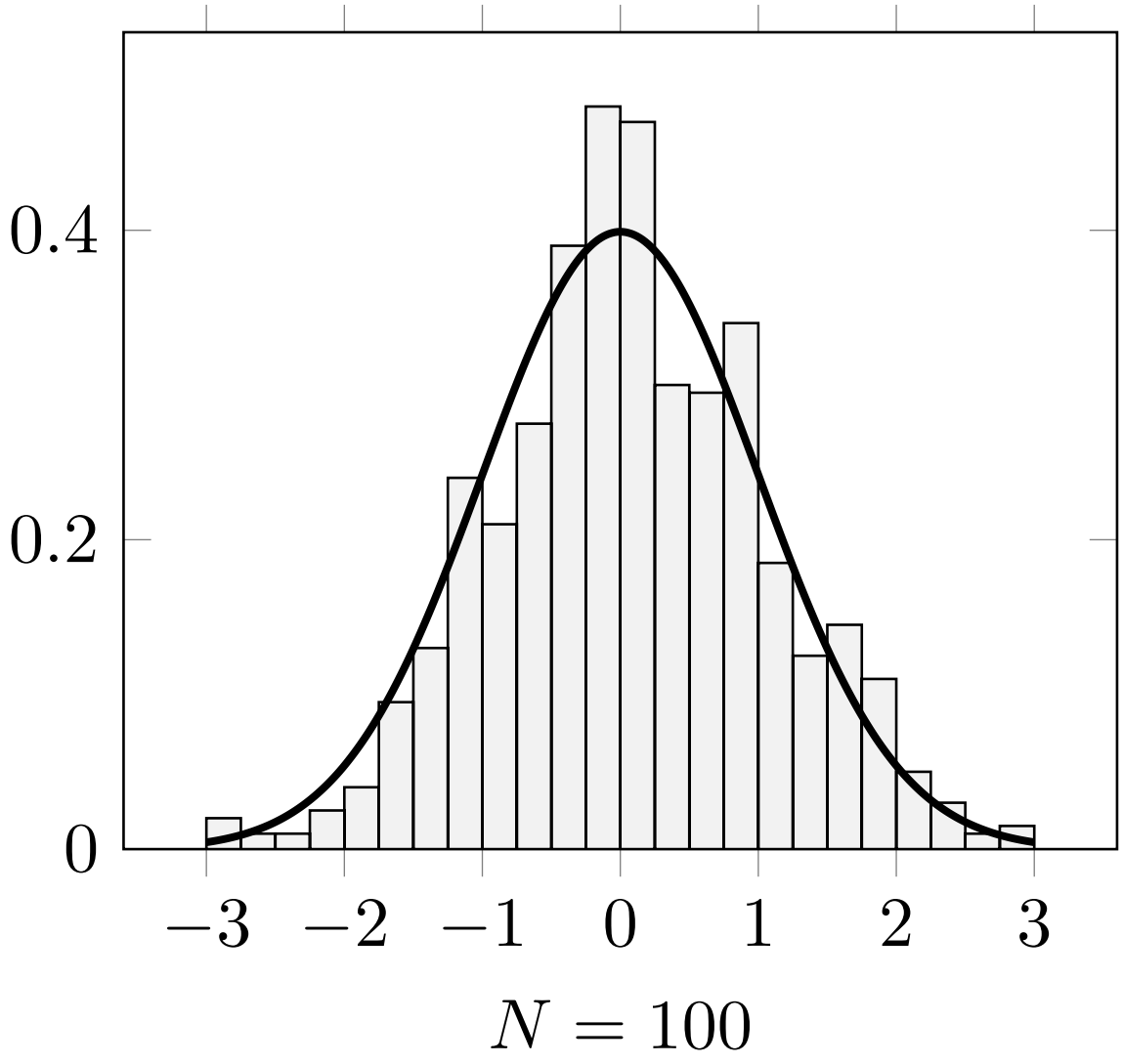
CLT: Illustrative simulations (2/5)



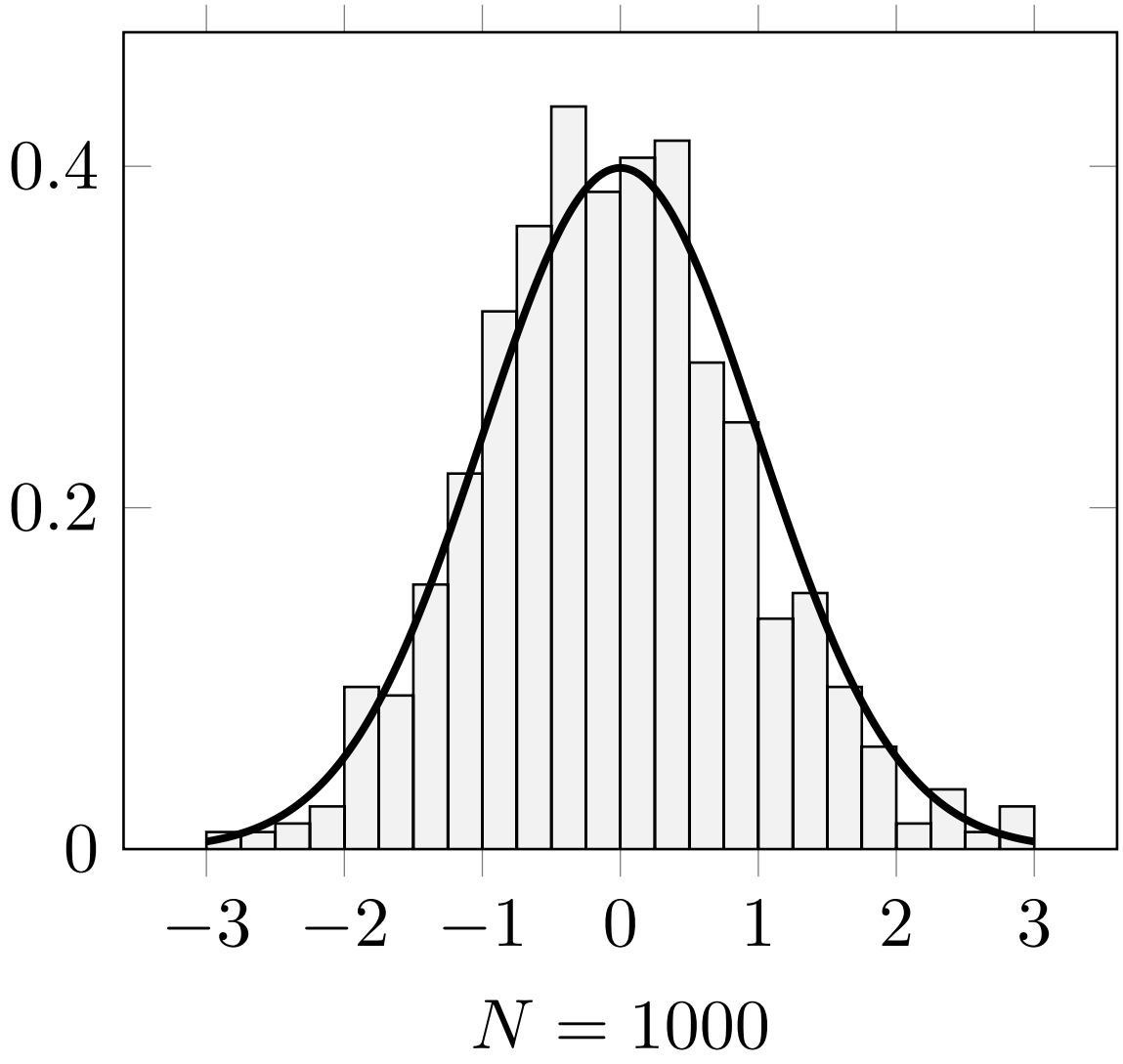
CLT: Illustrative simulations (3/5)



CLT: Illustrative simulations (4/5)



CLT: Illustrative simulations (5/5)



More general Central Limit Theorems

- As with the Laws of Large Numbers, more general Central Limit Theorems with less restrictive assumptions exist.

- Two famous versions, which both allow for i.n.i.d. data, are presented next without proof.

- Of these two, the one that is named after A. Ljapunov is of particular interest, as it is based on a condition which often shows up in some technical econometric papers.

- The so-called *Ljapunov condition* requires that in a sample, at least some cross-osservation moments of order "slightly" higher than two (as detailed later) are finite.

- Even in this case, some more general versions that allow for *weakly dependent* observations also exist.

Lindeberg-Feller Central Limit Theorem

Theorem 14

Central Limit Theorem (Lindeberg and Feller's). *Consider a non-random (i.n.i.d.) sample where the random vectors* $\boldsymbol{x}_i$ *that generate it have possibly heterogeneous finite means* $\mathbb{E}[\boldsymbol{x}_i] < \infty$*, variances* $\mathbb{V}\text{ar}[\boldsymbol{x}_i] < \infty$*, and all mixed third moments are finite too. If:*

$$\lim_{N\to\infty} \left(\sum_{i=1}^{N} \mathbb{V}\text{ar}[\boldsymbol{x}_i] \right)^{-1} \mathbb{V}\text{ar}[\boldsymbol{x}_i] = \mathbf{0}$$

then it holds that:

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} (\boldsymbol{x}_i - \mathbb{E}[\boldsymbol{x}_i]) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbb{V}\text{ar}[\boldsymbol{x}])$$

where:

$$\frac{1}{N} \sum_{i=1}^{N} \mathbb{V}\text{ar}[\boldsymbol{x}_i] \xrightarrow{p} \mathbb{V}\text{ar}[\boldsymbol{x}]$$

that is, the positive semi-definite matrix $\mathbb{V}\text{ar}[\boldsymbol{x}]$ *is the probability limit of the observations' variances.*

Ljapunov's Central Limit Theorem

Theorem 15

Central Limit Theorem (Ljapunov's). *Consider a non-random (i.n.i.d.) sample where the random vectors* $\boldsymbol{x}_i$ *that generate it have possibly heterogeneous finite moments* $\mathbb{E}[\boldsymbol{x}_i] < \infty$, $\mathbb{V}\text{ar}[\boldsymbol{x}_i] < \infty$. *If:*

$$\lim_{N\to\infty} \left(\sum_{i=1}^{N} \mathbb{V}\text{ar}[\boldsymbol{x}_i] \right)^{-\left(1+\frac{\delta}{2}\right)} \sum_{i=1}^{N} \mathbb{E}\left[\left| \boldsymbol{x}_i - \mathbb{E}[\boldsymbol{x}_i] \right|^{2+\delta} \right] = \mathbf{0}$$

for some $\delta > 0$, *then:*

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} (\boldsymbol{x}_i - \mathbb{E}[\boldsymbol{x}_i]) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbb{V}\text{ar}[\boldsymbol{x}])$$

where $\mathbb{V}\text{ar}[\boldsymbol{x}]$ *is the variances' probability limit as in Theorem 14.*

Note: in econometric applications with $\mathbb{E}[\boldsymbol{x}_i] = 0$ for $i = 1, \ldots, N$, the "Ljapunov condition" specializes, for some $\delta > 0$, to:

$$\mathbb{E}\left[\left| X_{ik} X_{i\ell} \right|^{1+\delta} \right] < \infty$$

for any two elements $k, \ell = 1, \ldots, K$ of $\boldsymbol{x}$ and for all observations i.

Asymptotic normality & linear regression (1/5)

To show how the Central Limit can help statistical inference in practice, consider again the estimator of the slope parameter in the bivariate regression model. Rework it as follows.

$$\begin{aligned}
\widehat{\beta}_{1,MM} &= \frac{\sum_{i=1}^{N}\left(X_i-\bar{X}\right)Y_i}{\sum_{i=1}^{N}\left(X_i-\bar{X}\right)^2} \\
&= \beta_1\frac{\sum_{i=1}^{N}\left(X_i-\bar{X}\right)X_i}{\sum_{i=1}^{N}\left(X_i-\bar{X}\right)^2}+\frac{\sum_{i=1}^{N}\left(X_i-\bar{X}\right)\varepsilon_i}{\sum_{i=1}^{N}\left(X_i-\bar{X}\right)^2} \\
&= \beta_1+\frac{\frac{1}{N}\sum_{i=1}^{N}\left(X_i-\bar{X}\right)\varepsilon_i}{\frac{1}{N}\sum_{i=1}^{N}\left(X_i-\bar{X}\right)^2}
\end{aligned}$$

where

$$\varepsilon_i \equiv Y_i-\beta_0-\beta_1X_i$$

is the **error term** of the model: the deviation between Y_i and the linear CEF, $\mathbb{E}\left[Y_i|\,X_i\right]=\beta_0+\beta_1X_i$. Note that $\mathbb{E}\left[\varepsilon_i\right]=0$.

Asymptotic normality & linear regression (2/5)

Recall that in the bivariate linear regression model, the Law of Iterated Expectations gives $\mathbb{E}[X_i \varepsilon_i] = 0$. This provides another avenue to demonstrate consistency of the MM estimator for β_1. In fact, by the Continuous Mapping Theorem:

$$\frac{1}{N} \sum_{i=1}^{N} \left(X_i - \bar{X} \right) \varepsilon_i \xrightarrow{p} \underbrace{\mathbb{E}[X_i \varepsilon_i]}_{=0} - \mathbb{E}[X_i] \underbrace{\mathbb{E}[\varepsilon_i]}_{=0} = 0$$

implying $\widehat{\beta}_{1,MM} \xrightarrow{p} \beta_1$.

As the expression on the left-hand side above is a sample mean, under adequate assumptions about the sample some applicable Central Limit Theorem would imply the following.

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left(X_i - \bar{X} \right) \varepsilon_i \xrightarrow{d} \mathcal{N}\left(0, \mathbb{E}\left[\varepsilon_i^2 \left(X_i - \mathbb{E}[X_i] \right)^2 \right] \right)$$

Here the limiting variance takes this form because $\bar{X} \xrightarrow{p} \mathbb{E}[X_i]$ at the probability limit.

Asymptotic normality & linear regression (3/5)

The limiting variance obtains as:

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left[\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(X_i-\mathbb{E}\left[X_i\right]\right)\varepsilon_i\right] &= \frac{1}{N}\sum_{i=1}^{N}\mathbb{V}\mathrm{ar}\left[\left(X_i-\mathbb{E}\left[X_i\right]\right)\varepsilon_i\right] \\ &= \mathbb{E}\left[\varepsilon_i^2\left(X_i-\mathbb{E}\left[X_i\right]\right)^2\right]\end{aligned}$$

while in the more specialized case where the squared deviations of X_i and ε_i from their means are mutually independent, it is:

$$\mathbb{E}\left[\varepsilon_i^2\left(X_i-\mathbb{E}\left[X_i\right]\right)^2\right] = \mathbb{E}\left[\varepsilon_i^2\right]\mathbb{E}\left[\left(X_i-\mathbb{E}\left[X_i\right]\right)^2\right] = \sigma_\varepsilon^2\cdot\mathbb{V}\mathrm{ar}\left[X_i\right]$$

where $\sigma_\varepsilon^2 \equiv \mathbb{E}\left[\varepsilon_i^2\right]$.

This latter case is that where the conditional variance function of ε_i given X_i is actually a constant – a scenario usually called *homoscedasticity* (as opposed to *heteroscedasticity*, the general case). This is typical terminology in regression parlance.

Asymptotic normality & linear regression (4/5)

By the Cramér-Wold device and the following implication of the Continuous Mapping Theorem:

$$\left[\frac{1}{N}\sum_{i=1}^{N}\left(X_i-\bar{X}\right)^2\right]^{-1} \xrightarrow{p} \left[\mathbb{V}\mathrm{ar}\left[X_i\right]\right]^{-1}$$

these results allow, together, to obtain the limiting distribution of the MM estimator as:

$$\sqrt{N}\left(\widehat{\beta}_{1,MM}-\beta_1\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\mathbb{E}\left[\varepsilon_i^2\left(X_i-\mathbb{E}\left[X_i\right]\right)^2\right]}{\left(\mathbb{V}\mathrm{ar}\left[X_i\right]\right)^2}\right)$$

and for some given N, its asymptotic distribution as follows.

$$\widehat{\beta}_{1,MM} \overset{A}{\sim} \mathcal{N}\left(\beta_1, \frac{1}{N}\frac{\mathbb{E}\left[\varepsilon_i^2\left(X_i-\mathbb{E}\left[X_i\right]\right)^2\right]}{\left(\mathbb{V}\mathrm{ar}\left[X_i\right]\right)^2}\right)$$

Asymptotic normality & linear regression (5/5)

In the specialized homoscedastic case, the limiting distribution of the estimator is:

$$\sqrt{N}\left(\widehat{\beta}_{1,MM} - \beta_1\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma_\varepsilon^2}{\mathbb{V}\mathrm{ar}\left[X_i\right]}\right)$$

and for some given N, its asymptotic distribution as follows.

$$\widehat{\beta}_{1,MM} \overset{A}{\sim} \mathcal{N}\left(\beta_1, \frac{1}{N}\frac{\sigma_\varepsilon^2}{\mathbb{V}\mathrm{ar}\left[X_i\right]}\right)$$

For them to be used in statistical inference, the results for both the heteroscedastic and homoscedastic cases require knowledge of the various components of the limiting variances. In general, these are unknown by researchers and must be **estimated**.

This is best discussed later after reviewing the application of the Central Limit Theorem to general MM and MLE estimators.

The Delta Method (1/2)

Theorem 16

Delta Method. *Suppose that some random sequence of dimension K – call it $\boldsymbol{x}_N$ – is asymptotically normal:*

$$\sqrt{N}\left(\boldsymbol{x}_N - \mathbf{c}\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Upsilon}\right)$$

for some $K \times 1$ vector $\mathbf{c}$ and for some $K \times K$ matrix $\boldsymbol{\Upsilon}$. In addition, consider some vector-valued function $\boldsymbol{d}\left(\boldsymbol{x}\right) : \mathbb{R}^K \to \mathbb{R}^J$. If the latter is continuously differentiable at $\mathbf{c}$ and the $J \times K$ Jacobian matrix

$$\boldsymbol{\Delta} \equiv \frac{\partial}{\partial \boldsymbol{x}^{\mathrm{T}}} \boldsymbol{d}\left(\mathbf{c}\right)$$

has full row rank J, the limiting distribution of $\boldsymbol{d}\left(\boldsymbol{x}_N\right)$ is as follows.

$$\sqrt{N}\left(\boldsymbol{d}\left(\boldsymbol{x}_N\right) - \boldsymbol{d}\left(\mathbf{c}\right)\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Delta}\boldsymbol{\Upsilon}\boldsymbol{\Delta}^{\mathrm{T}}\right)$$

Proof.

(Continues...)

The Delta Method (2/2)

Theorem 16

Proof.

(Continued.) From the mean value theorem it is:

$$\boldsymbol{d}\left(\boldsymbol{x}_N\right) = \boldsymbol{d}\left(\mathbf{c}\right) + \frac{\partial}{\partial \boldsymbol{x}^{\mathrm{T}}} \boldsymbol{d}\left(\widetilde{\boldsymbol{x}}_N\right)\left(\boldsymbol{x}_N - \mathbf{c}\right)$$

where $\widetilde{\boldsymbol{x}}_N$ is a convex combination of $\boldsymbol{x}_N$ and $\mathbf{c}$. However, as $\boldsymbol{x}_N \xrightarrow{p} \mathbf{c}$:

$$\frac{\partial}{\partial \boldsymbol{x}^{\mathrm{T}}} \boldsymbol{d}\left(\widetilde{\boldsymbol{x}}_N\right) \xrightarrow{p} \frac{\partial}{\partial \boldsymbol{x}^{\mathrm{T}}} \boldsymbol{d}\left(\mathbf{c}\right) = \boldsymbol{\Delta}$$

hence, at the probability limit:

$$\sqrt{N}\left(\boldsymbol{d}\left(\boldsymbol{x}_N\right) - \boldsymbol{d}\left(\mathbf{c}\right)\right) \xrightarrow{p} \boldsymbol{\Delta} \cdot \sqrt{N}\left(\boldsymbol{x}_N - \mathbf{c}\right)$$

which, by the given hypotheses, implies the result. □

This result is extremely useful to derive the asymptotic distribution of estimators that relate with sample means, but are not sample means.

Method of moments asymptotic normality (1/4)

Theorem 17

Asymptotically, MM estimators are normally distributed. *An estimator* $\widehat{\theta}_{MM}$ *defined as the solution of a set of sample moments*

$$\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{m}\left(\boldsymbol{x}_i;\widehat{\theta}_{MM}\right)=\mathbf{0}$$

is asymptotically normal. If the sample is random and the moment conditions are differentiable the limiting distribution is:

$$\sqrt{N}\left(\widehat{\theta}_{MM}-\theta_0\right)\xrightarrow{d}\mathcal{N}\left(\mathbf{0},\mathbf{M}_0\boldsymbol{\Upsilon}_0\mathbf{M}_0^{\mathrm{T}}\right)$$

so long as the following matrices exist, are finite and nonsingular.

$$\boldsymbol{\Upsilon}_0=\mathbb{V}\mathrm{ar}\left[\boldsymbol{m}\left(\boldsymbol{x}_i;\theta_0\right)\right]\qquad\mathbf{M}_0\equiv\left[\mathbb{E}\left[\frac{\partial}{\partial\theta^{\mathrm{T}}}\boldsymbol{m}\left(\boldsymbol{x}_i;\theta_0\right)\right]\right]^{-1}$$

Proof.

(Continues...)

Method of moments asymptotic normality (2/4)

Theorem 17

Proof.

(Continued.) The proof applies the same logic as the Delta Method. By the mean value theorem, the sample moment conditions become:

$$\begin{aligned}\mathbf{0} &= \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{m}\left(\boldsymbol{x}_i; \widehat{\theta}_{MM}\right) = \\ &= \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{m}\left(\boldsymbol{x}_i; \theta_0\right) + \left[\frac{1}{N}\sum_{i=1}^{N} \frac{\partial}{\partial \theta^{\mathrm{T}}} \boldsymbol{m}\left(\boldsymbol{x}_i; \widetilde{\theta}_N\right)\right]\left(\widehat{\theta}_{MM} - \theta_0\right)\end{aligned}$$

where the first expression in the first line equals to zero by construction of all MM estimators. After multiplying both sides by $\sqrt{N}$ and some manipulation the above expression is rendered as follows.

$$\sqrt{N}\left(\widehat{\theta}_{MM} - \theta_0\right) = -\left[\frac{1}{N}\sum_{i=1}^{N} \frac{\partial}{\partial \theta^{\mathrm{T}}} \boldsymbol{m}\left(\boldsymbol{x}_i; \widetilde{\theta}_N\right)\right]^{-1} \frac{1}{\sqrt{N}}\sum_{i=1}^{N} \boldsymbol{m}\left(\boldsymbol{x}_i; \theta_0\right)$$

(Continues...)

Method of moments asymptotic normality (3/4)

Theorem 17

Proof.

(Continued.) Since this is a random sample:

1. by a suitable Central Limit Theorem:

$$-\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\boldsymbol{m}\left(\boldsymbol{x}_i;\boldsymbol{\theta}_0\right)\xrightarrow{d}\mathcal{N}\left(\mathbf{0},\mathbb{V}\mathrm{ar}\left[\boldsymbol{m}\left(\boldsymbol{x}_i;\boldsymbol{\theta}_0\right)\right]\right)$$

   given that $\mathbb{E}\left[\boldsymbol{m}\left(\boldsymbol{x}_i;\boldsymbol{\theta}_0\right)\right]=\mathbf{0}$ by hypothesis;

2. while by the Weak Law of Large Numbers:

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial\boldsymbol{\theta}^{\mathrm{T}}}\boldsymbol{m}\left(\boldsymbol{x}_i;\widetilde{\boldsymbol{\theta}}_N\right)\xrightarrow{p}\mathbb{E}\left[\frac{\partial}{\partial\boldsymbol{\theta}^{\mathrm{T}}}\boldsymbol{m}\left(\boldsymbol{x}_i;\boldsymbol{\theta}_0\right)\right]$$

   since $\widetilde{\boldsymbol{\theta}}_N\xrightarrow{p}\boldsymbol{\theta}_0$ by consistency of the estimator (at the limit, $\widetilde{\boldsymbol{\theta}}_N$, $\widehat{\boldsymbol{\theta}}_{MM}$ and $\boldsymbol{\theta}_0$ all coincide).

(Continues...)

Method of moments asymptotic normality (4/4)

Theorem 17

Proof.

(Continued.) These intermediate results are together combined via the Continuous Mapping Theorem, Slutskij's Theorem as well as the Cramér-Wold device so to imply the statement. Therefore, *for a fixed* N the asymptotic distribution is:

$$\widehat{\boldsymbol{\theta}}_{MM} \overset{A}{\sim} \mathcal{N}\left(\boldsymbol{\theta}_0, \frac{1}{N}\mathbf{M}_0\boldsymbol{\Upsilon}_0\mathbf{M}_0^{\mathrm{T}}\right)$$

which concludes the proof. □

This expression of the asymptotic variance-covariance is typically unknown and must be thus *estimated*. The general approach to address this issue is shown later alongside the MLE case.

Maximum likelihood asymptotic normality (1/5)

Theorem 18

Asymptotically, ML estimators are normally distributed and they attain the Cramér-Rao bound. *An estimator* $\widehat{\theta}_{MLE}$ *defined as the maximizer of a log-likelihood function as per*

$$\widehat{\theta}_{MLE} = \arg\max_{\theta \in \boldsymbol{\Theta}} \sum_{i=1}^{N} \log f_{\boldsymbol{x}_i} \left(\theta | \, \mathbf{x}_i \right)$$

is asymptotically normal. Define the following 'regularity conditions:'

i. *the problem is well defined, i.e.* θ_0 *is the maximizer of the population expression* $\mathbb{E}\left[\log f_{\boldsymbol{x}} \left(\boldsymbol{x}_i ; \theta \right)\right]$ *– where* $f_{\boldsymbol{x}} \left(\boldsymbol{x}_i ; \theta \right)$ *is the probability mass or density function that generates the data;*

ii. $f_{\boldsymbol{x}} \left(\boldsymbol{x}_i ; \theta \right)$ *is three times continuously differentiable and its derivatives are bounded in absolute value;*

iii. *the support of* $\boldsymbol{x}_i$ *does not depend on* θ*, so that derivatives for* θ *can pass at least twice through an integral defined over* $f_{\boldsymbol{x}} \left(\boldsymbol{x}_i ; \theta \right)$*.*

(Continues...)

Maximum likelihood asymptotic normality (2/5)

Theorem 18

(Continued.) *If the sample is random and the regularity conditions hold, then the limiting distribution is expressible as:*

$$\sqrt{N}\left(\widehat{\theta}_{MLE} - \theta_0\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \left[\boldsymbol{I}\left(\theta_0\right)\right]^{-1}\right)$$

where $\boldsymbol{I}\left(\theta_0\right)$ – written without the N subscript – is the expression for the following "single-observation" information matrix evaluated at θ_0.

$$\begin{aligned}\boldsymbol{I}\left(\theta_0\right) &\equiv \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f_{\boldsymbol{x}}\left(\boldsymbol{x}_i;\theta_0\right)\right)\left(\frac{\partial}{\partial\theta}\log f_{\boldsymbol{x}}\left(\boldsymbol{x}_i;\theta_0\right)\right)^{\mathrm{T}}\right] \\ &= -\,\mathbb{E}\left[\frac{\partial^2}{\partial\theta\partial\theta^{\mathrm{T}}}\log f_{\boldsymbol{x}}\left(\boldsymbol{x}_i;\theta_0\right)\right]\end{aligned}$$

Consequently, $\widehat{\theta}_{MLE}$ asymptotically attains the Cramér-Rao bound.

Proof.

(Continues...)

Maximum likelihood asymptotic normality (3/5)

Theorem 18

Proof.

(Continued.) The proof proceeds similarly to the MM case. By the mean value theorem, the MLE First Order Conditions can write as:

$$\mathbf{0} = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial \boldsymbol{\theta}}\log f_{\boldsymbol{x}}\left(\boldsymbol{x}_i;\widehat{\boldsymbol{\theta}}_{MLE}\right) = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial \boldsymbol{\theta}}\log f_{\boldsymbol{x}}\left(\boldsymbol{x}_i;\boldsymbol{\theta}_0\right) + \\ + \left[\frac{1}{N}\sum_{i=1}^{N}\frac{\partial^2}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^{\mathrm{T}}}\log f_{\boldsymbol{x}}\left(\boldsymbol{x}_i;\widetilde{\boldsymbol{\theta}}_N\right)\right]\left(\widehat{\boldsymbol{\theta}}_{MLE}-\boldsymbol{\theta}_0\right)$$

where the entire expression is zero by definition of MLE. Once again:

$$\sqrt{N}\left(\widehat{\boldsymbol{\theta}}_{MLE}-\boldsymbol{\theta}_0\right) = \\ = -\left[\frac{1}{N}\sum_{i=1}^{N}\frac{\partial^2}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^{\mathrm{T}}}\log f_{\boldsymbol{x}}\left(\boldsymbol{x}_i;\widetilde{\boldsymbol{\theta}}_N\right)\right]^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{\partial}{\partial \boldsymbol{\theta}}\log f_{\boldsymbol{x}}\left(\boldsymbol{x}_i;\boldsymbol{\theta}_0\right)$$

but here additional simplifications are possible. **(Continues...)**

Maximum likelihood asymptotic normality (4/5)

Theorem 18

Proof.

(Continued.) Thanks to the Information Matrix Equality, under the regularity conditions the following holds.

1. A suitable Central Limit Theorem implies that:

$$-\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{x}}\left(\boldsymbol{x}_i;\widehat{\boldsymbol{\theta}}_{MLE}\right)\xrightarrow{d}\mathcal{N}\left(\mathbf{0},\boldsymbol{I}\left(\boldsymbol{\theta}_0\right)\right)$$

   as $\boldsymbol{\theta}_0$ maximizes $\mathbb{E}\left[\log f_{\boldsymbol{x}}\left(\boldsymbol{x}_i;\boldsymbol{\theta}_0\right)\right]$, so: $\mathbb{E}\left[\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{x}}\left(\boldsymbol{x}_i;\boldsymbol{\theta}_0\right)\right]=\mathbf{0}$;

2. while by the Weak Law of Large Numbers:

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}}\log f_{\boldsymbol{x}}\left(\boldsymbol{x}_i;\widetilde{\boldsymbol{\theta}}_N\right)\xrightarrow{p}-\boldsymbol{I}\left(\boldsymbol{\theta}_0\right)$$

   again since $\widetilde{\boldsymbol{\theta}}_N\xrightarrow{p}\boldsymbol{\theta}_0$ by consistency of MLE as per Theorem 9.

(Continues...)

Maximum likelihood asymptotic normality (5/5)

Theorem 18

Proof.

(Continued.) Here, the application of the Delta Method results in a simplified expression of the limiting variance – given in the statement of the Theorem. Collecting terms, *for a fixed* N the asymptotic distribution is:

$$\widehat{\theta}_{MLE} \overset{A}{\sim} \mathcal{N}\left(\theta_0, \left[\boldsymbol{I}_N\left(\theta_0\right)\right]^{-1}\right)$$

where $\boldsymbol{I}_N\left(\theta_0\right)$ is the grand (sample) information matrix for a fixed N. Since the MLE is asymptotically consistent, at the probability limit its bias is zero, hence the estimator attains the Cramér-Rao bound. □

Some comments are in order here.

- Asymptotic attainment of the Cramér-Rao bound is a desirable property of MLE (alongside invariance – see Lecture 5).
- Yet it hinges on correctly assuming the underlying distribution. If this is incorrect, the MLE can fail utterly (be inconsistent).
- By contrast, the MM is more robust: there is a trade-off here!

Estimating asymptotic variance-covariances

- The above results develop expressions for both *limiting* and *asymptotic* variance-covariances of MM and ML estimators.
- However, the elements inside such expressions, like $\mathbf{M}_0$, $\boldsymbol{\Upsilon}_0$ and $\boldsymbol{I}(\boldsymbol{\theta}_0)$, are generally unknown *ex-ante*.
- To use these results in statistical inference it is necessary to **estimate** these quantities.
- By the analogy principle, one could use **sample analogues** as consistent estimators of population variance-covariances.
- Example: if $\sqrt{N}\left(\bar{X}-\mu\right) \xrightarrow{d} \mathcal{N}\left(0,\sigma^2\right)$, then $\frac{N-1}{N}S^2 \xrightarrow{p} \sigma^2$.
- A more elaborate example on the bivariate linear regression model is developed next.
- General estimators for the MM and MLE cases then follow.

Asymptotic inference in linear regression (1/2)

Suppose one wants to perform a two-sided test of hypothesis on the bivariate linear regression slope parameter β_1.

$$\mathrm{H}_0 : \ \beta_1 = C \qquad\qquad \mathrm{H}_1 : \ \beta_1 \neq C$$

If $C = 0$, this is a so-called *significance test* of the regression: a test whether the explanatory variable X_i affects the mean of Y_i in a conditional (CEF) sense.

In small samples this test may be problematic and require extra assumptions. In an asymptotic environment, the earlier analysis of the model allows to establish the following property.

$$\mathrm{t}_N = \sqrt{N}\,\frac{\widehat{\beta}_{1,MM} - C}{S_{\beta_1}} \xrightarrow{d} \mathcal{N}(0,1)$$

Here t_N is a *t*-statistic and S_{β_1} is a suitable consistent estimator of the limiting standard deviation of $\sqrt{N}\left(\widehat{\beta}_{1,MM} - \beta_1\right)$.

Asymptotic inference in linear regression (2/2)

The expression of S_{β_1} differs across assumptions. In the general heteroscedastic case, its squared version is:

$$S^2_{\beta_1} = N \frac{\sum_{i=1}^N \left(Y_i - \widehat{\beta}_{0,MM} - \widehat{\beta}_{1,MM} X_i\right)^2 \left(X_i - \bar{X}\right)^2}{\left[\sum_{i=1}^N \left(X_i - \bar{X}\right)^2\right]^2}$$

while in the more restricted homoscedastic case $S^2_{\beta_1}$ is as follows.

$$S^2_{\beta_1} = \frac{\sum_{i=1}^N \left(Y_i - \widehat{\beta}_{0,MM} - \widehat{\beta}_{1,MM} X_i\right)^2}{\sum_{i=1}^N \left(X_i - \bar{X}\right)^2}$$

The quantity $S_{\beta_1}/\sqrt{N}$ is called the **standard error** of $\widehat{\beta}_{1,MM}$. A proper confidence interval for $\widehat{\beta}_{1,MM}$ would be as follows.

$$\beta_1 \in \left[\widehat{\beta}_{1,MM} - z^*_{\alpha/2} \frac{S_{\beta_1}}{\sqrt{N}}, \widehat{\beta}_{1,MM} + z^*_{\alpha/2} \frac{S_{\beta_1}}{\sqrt{N}}\right]$$

Estimating MM asymptotic variance-covariances

In the MM case, $\widehat{\mathbf{M}}_N \widehat{\boldsymbol{\Upsilon}}_N \widehat{\mathbf{M}}_N^{\mathrm{T}}/N$ is a consistent estimator of the asymptotic variance-covariance in random samples, where:

$$\widehat{\mathbf{M}}_N \equiv \left[\frac{1}{N} \sum_{i=1}^{N} \frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}} \boldsymbol{m}\left(\boldsymbol{x}_i; \widehat{\boldsymbol{\theta}}_{MM} \right) \right]^{-1} \xrightarrow{p} \mathbf{M}_0$$

is a consistent estimator of $\mathbf{M}_0$ (by some Law of Large Numbers and the Continuous Mapping Theorem), while

$$\widehat{\boldsymbol{\Upsilon}}_N \equiv \frac{1}{N} \sum_{i=1}^{N} \left[\boldsymbol{m}\left(\boldsymbol{x}_i; \widehat{\boldsymbol{\theta}}_{MM} \right) \right] \left[\boldsymbol{m}\left(\boldsymbol{x}_i; \widehat{\boldsymbol{\theta}}_{MM} \right) \right]^{\mathrm{T}} \xrightarrow{p} \boldsymbol{\Upsilon}_0$$

is also a consistent estimator of the variance of the zero moment conditions by some applicable Law of Large Numbers, since in a random sample the following holds.

$$\boldsymbol{\Upsilon}_0 = \mathbb{V}\mathrm{ar}\left[\boldsymbol{m}\left(\boldsymbol{x}_i; \boldsymbol{\theta}_0 \right) \right] = \mathbb{E}\left[\left(\boldsymbol{m}\left(\boldsymbol{x}_i; \boldsymbol{\theta}_0 \right) \right) \left(\boldsymbol{m}\left(\boldsymbol{x}_i; \boldsymbol{\theta}_0 \right) \right)^{\mathrm{T}} \right]$$

These estimators also work under general i.n.i.d. assumptions.

Estimating ML asymptotic variance-covariances

In MLE, there are two ways to estimate the information matrix, corresponding to both sides of the Information Matrix Equality. The first option is based on the Hessian of the p.m.f. or p.d.f.:

$$\widehat{\mathbf{H}}_N \equiv -\frac{1}{N}\sum_{i=1}^{N}\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}}\log f_{\boldsymbol{x}}\left(\boldsymbol{x}_i;\widehat{\boldsymbol{\theta}}_{MLE}\right)\xrightarrow{p}\boldsymbol{I}\left(\boldsymbol{\theta}_0\right)$$

while the second option exploits the "squared" score:

$$\widehat{\mathbf{J}}_N \equiv \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{x}}\left(\boldsymbol{x}_i;\widehat{\boldsymbol{\theta}}_{MLE}\right)\right)\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{x}}\left(\boldsymbol{x}_i;\widehat{\boldsymbol{\theta}}_{MLE}\right)\right)^{\mathrm{T}}$$

with

$$\widehat{\mathbf{J}}_N \xrightarrow{p}\boldsymbol{I}\left(\boldsymbol{\theta}_0\right).$$

The choice between $\widehat{\mathbf{H}}_N$ and $\widehat{\mathbf{J}}_N$ is based on convenience and is context-dependent. Observe how all these estimators (both MM and MLE) are evaluated at the consistent parameter estimates.