R4: Asymptotics, Take 1

Yaroslav Mukhin

Mar 1, 2015

Abstract. We set out to derive the asymptotic distribution of the indirect least squares estimator of the IV model with one instrument, one endogenous variable and some covariates. The objective here is to be very explicit and detailed. To that end, we first recall the essential tools of probability theory necessary to complete the job.

1. Stochastic convergence

In order to derive the asymptotic distribution of the indirect least squares estimator, we will need the following tools: LLN, CLT, CMT, Slutsky's lemma, Delta method. Rather than giving short but unmotivated statements of these results, I would like to put things in some perspective.

1.1 **Probability space.** $(\Omega, \mathcal{F}, P)$ is a "black box" that controls/models random outcomes.

Ω -set of points that represent all possible configurations of random outcomes, or equivalently, all possible states of nature.

$\mathcal{F}$ -collection of subsets of Ω, called events. Events are those combinations of states of nature, to which a probability statement can be assigned. For technical reasons, not all subsets of Ω are events.

P -probability function that assigns probabilities to events in $\mathcal{F}$. That is it!

1.2 **Random variables.** Are measurable maps from the black box $(\Omega, \mathcal{F}, P)$ into the real world:

$$Y : (\Omega, \mathcal{F}, P) \to \mathbb{R} \quad \text{or} \quad X : (\Omega, \mathcal{F}, P) \to \mathbb{R}^d.$$

Measurability simply means that events in $\mathbb{R}$ (e.g. intervals) traced back to to the probability space via Y or X correspond to events there. This allows us to make probability statements about Y and X.

1.3 **Pointwise convergence of random variables.** Classical probability is about large n behavior of sequences of random variables. At least three different notions of convergence are relevant in econometrics. We start with the simpler two convergence concepts. In what follows we work with a fixed sequence of random variables or vectors $X_n : (\Omega, \mathcal{F}, P) \to \mathbb{R}$ or $\mathbb{R}^d$.

1.3.1 A.s. convergence

If we fix a particular realization ω, we are then looking at a sequence real numbers or vectors $\{X_n(\omega)\}_n$. A.s. convergence requires that these sequences of numbers converge for sufficiently many realizations ω:

$$X_n \overset{\text{a.s.}}{\to} X \quad \text{iff} \quad \forall \omega \in \Omega \setminus N, \quad X_n(\omega) \xrightarrow[n\to\infty]{} X(\omega),$$

where N must be an event with probability zero. This is simply pointwise convergence of functions.

1.3.2 Convergence in P

Instead of fixing a realization ω and looking at the sequence of outcomes, we can fix a particular element of the sequence X_n and ask for the proportion of outcomes for which X_n is far from the limit to be small:

$$X_n \xrightarrow{P} X \quad \text{iff} \quad P\underbrace{\{\omega \,;\, |X_n(\omega) - X(\omega)| > \epsilon\}}_{=:A_n} \xrightarrow[n\to\infty]{} 0.$$

This is a weaker notion of pointwise convergence: any given configuration ω may enter and leave the sequence of bad evens A_n infinitely often as we move along the sequence, but the total mass of bad configurations must tend to zero as we move along the sequence.

These convergence notions are simply convergence of functions that we learned in calculus/mathematical analysis. The next concept is not about functions at all.

1.4 **Weak convergence.** From the above sequence of random variables $X_n : (\Omega, \mathcal{F}, P) \to \mathbb{R}$, we can construct a sequence of probability measures on the real world $\mathbb{R}$ that correspond to the following CDFs:

$$F_n(x) := P\{X_n \leq x\}.$$

It may seem strange at first, but we can defined a notion of distance/closeness for pairs of probability measures. In words, two probability measures are close if they assign approximately the same mass to approximately the same locations. This intuitive description can be made formal to define a topology that corresponds to the notion of weak convergence. Here are some examples:

1. Point masses at location $\frac{1}{n}$ converge weakly to a point mass at 0, denoted $\delta_{1/n} \quad \delta_0$.
2. $\left[\frac{n}{n+1}\delta_{1/n} + \frac{1}{n+1}\delta_5\right] \quad \delta_0$.
3. $N(\mu_n, \sigma_n^2) \quad N(\mu, \sigma^2)$ iff $\begin{bmatrix} \mu_n \\ \sigma_n^2 \end{bmatrix} \to \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix}$.
4. For i.i.d. centered[1] sequence Y_i, let $X_n := \sqrt{n}\,\mathbb{E}_n Y_i$ and $F_n(x) := P\{X_n \leq x\}$. Then $F_n \quad N(0, \mathrm{E}Y_1^2)$.

The point worth emphasizing here is that weak convergence and the CLT we apply to our estimators have nothing to do with the random variables as maps from Ω. Contrary to what shorthand notation may suggest, the objects under investigation here are the probability distributions that the estimators induce on $\mathbb{R}$. These summarize their 'global' behavior rather pointwise behavior as in Section 1.3. Important: $X_n \quad X$ does not imply that $X_n(\omega)$ converges at all for any ω. Example: take $Y \sim N(0,1)$, and construct $X_n := (-1)^n \mathrm{Y}$. But at the same time $X_n \quad X$ does imply that the sequence X_n remains bounded, so in particular $X_n/\sqrt{n} \xrightarrow{P} 0$.

How exactly do we define weak convergence? There are many equivalent ways, which is always a good thing!, as it gives as multiple tools for proofs:

- via distribution functions (convergence at continuity points)
- via probabilities attached to open/closed/nice sets
- via integrating nice "test functions"

This characterization is called Portmanteau lemma.

[1] $\mathrm{E}Y_1 = 0$

1.5 **Limit theorems.** Classical probability is all about product measures, which concentrate in the "middle" of the space as dimension grows. In the limit, the infinite product measure concentrates at a single point, but under appropriate rescaling looks like a Gaussian distribution. The framework of limits of statistical experiments makes a very nice use of this phenomena. Modern statistics is about studying the rate of this convergence to singularity in and interplay with the dimensionality of each coordinate, rather than looking only at the limit for fixed d.

It is amazing that the most demanding results mentioned in this note take only two lines to state:

LLN If $X_n : (\Omega, \mathcal{F}, P) \to \mathbb{R}$ or $\mathbb{R}^d$ are i.i.d. then

$$\mathbb{E}_n X_i \xrightarrow{\text{a.s.}} \mathrm{E}X_1 \quad \text{iff} \quad \mathrm{E}|X_1|_1 < \infty.$$

CLT If in addition $\mathrm{E}|X_1|_2^2 < \infty$ then

$$\sqrt{n}\mathbb{E}_n(X_i - \mathrm{E}X_1) \qquad N\Big(0, \mathrm{E}(X_1 - \mathrm{E}X_1)(X_1 - \mathrm{E}X_1)^T\Big).$$

1.6 **Continuous mapping theorem.** (CMT) All three notions of convergence mentioned above are *preserved* under smooth transformations. That is, if $g : \mathbb{R} \to \mathbb{R}$ or $\mathbb{R}^d \to \mathbb{R}^k$ is continuous, then

$X_n \xrightarrow{\text{a.s.}} X$ implies $g(X_n) \xrightarrow{\text{a.s.}} g(X)$;

$X_n \xrightarrow{P} X$ implies $g(X_n) \xrightarrow{P} g(X)$;

$X_n \qquad X$ implies $g(X_n) \qquad g(X)$;

1.7 **Slutsky's lemma.** If $X_n \qquad X$ and $Y_n \xrightarrow{P} c$ a constant, then we have *joint* weak convergence $(X_n, Y_n) \qquad (X, Y)$.

Remark. The result may seem trivial. It is not. Unlike pointwise (in ω) convergence where coordinatewise convergence is equivalent to joint (as whole vector) convergence, this is very different for joint probability measures. The source of the difference is in the fact that while coordinates determine the vector, marginal distributions *do not* determine the joint law. Examples:

1. Take $X, Y \sim N(0, 1)$ and compare (X, Y) and $(X, -Y)$.
2. Fix $Z_i \overset{\text{i.i.d.}}{\sim} N(0, 1)$. Two sequences $X_n := Z_n$ and $Y_n := (-1)^n Z_n$ both converge weakly (to what?), but the vector (X_n, Y_n) does not converge in any sense (why?).

Slutsky's lemma indicated that we must be careful when making statements about joint convergence of two estimators. The reason the lemma is true, is due to the hypothesis that Y_n convergence to a degenerate distribution concentrated at a single point. With on of the marginals being degenerate, there is only one way to construct the joint law.

Corollary Under conditions of Slutsky's lemma, we have that from CMT

$$\begin{aligned} X_n + Y_n &\qquad X + c \\ Y_n \cdot X_n &\qquad c \cdot X \end{aligned}$$

we use both of these every time we derive an asy distribution of an estimator. The last statement is used particularly often: under the conditions for LLN and CLT,

$$\mathbb{E}_n Y_i \cdot \sqrt{n}\mathbb{E}_n[X_i - \mathrm{E}X_i] \qquad \mathrm{E}Y_1 \cdot N(0, \mathrm{Var}(X_1)).$$

1.8 **Delta method.** Suppose $\sqrt{n}(X_n - x_0) \quad N(0,\Sigma)$.

1. If $g(x) = g(x_0) + \nabla g_{x_0}(x - x_0)$ is a linear map, then we get from CMT that
$$\sqrt{n}(g(X_n) - g(x_0)) = \nabla g_{x_0}\left[\sqrt{n}(X_n - x_0)\right] \quad \nabla g \cdot N(0,\Sigma).$$

2. For g that is differentiable at x_0 but possibly nonlinear the same is also true.

Intuition: If $\sqrt{n}(X_n - x_0) \quad N(0,\Sigma)$, then $\sqrt{n}(X_n - x_0)$ is a bounded sequence, and therefore X_n converges to x_0 with the rate of $1/\sqrt{n}$ tending to zero. Now, the smoothness of g at x_0 implies that it is locally linear at x_0, so we are practically back to the linear case.

1.9 **Stochastic o and O notation.** It is very convenient to have the following short-hand notation that we can use in proofs. We say that a sequence X_n of random variables is $o_P(1)$ if $X_n \xrightarrow{P} 0$. We think of $o_P(1)$ as a sequence that gets arbitrarily small eventually, but without any control on the rate of this convergence.

We say that X_n is $O_P(1)$ if it is bounded in the following sense: for any probability tolerance level $\epsilon > 0$ there is a sufficiently large uniform deterministic bound M on the sequence so that
$$P\{\omega\ ;\ |X_n(\omega)| > M\} \le \epsilon \quad \text{for all } n.$$
This property is called uniform tightness. We think of $O_P(1)$ as a sequence that is bounded. Equivalently, uniform tightness is a compactness property of the sequence of distributions of X_n. Some examples:

1. If Y_i is an i.i.d. sequence obeying LLN, then $\mathbb{E}_n(Y_i - \mathrm{E}Y_1)$ is o_P.

2. If Y_i is an i.i.d. sequence obeying CLT, then $\sqrt{n}\,\mathbb{E}_n(Y_i - \mathrm{E}Y_1)$ is O_P.

3. Combining previous two examples with Slutksy's lemma, we have in this specific case
$$\mathbb{E}_n(Y_i - \mathrm{E}Y_1) \cdot \sqrt{n}\,\mathbb{E}_n(Y_i - \mathrm{E}Y_1) = o_P(1) \cdot O_P(1) = o_P(1).$$

Rules like the one above are extremely helpful in econometrics, we list some here:
$$\begin{aligned} o_P(1) + o_P(1) &= o_P(1) \\ o_P(1) + O_P(1) &= O_P(1) \\ O_P(1) + O_P(1) &= O_P(1) \\ o_P(1) \cdot O_P(1) &= o_P(1) \\ o_P(1) \cdot o_P(1) &= o_P(1) \\ (\text{positive} + o_P(1))^{-1} &= O_P(1). \end{aligned}$$
It is a good exercise to prove these properties by combining definitions of o_p, O_P and convergence in probability, uniform tightness.

With this excellent notation, we can restate results of Slutsky's lemma: if $X_n \quad X$, then
$$\begin{array}{rcl} X_n + o_P(1) \quad X & \text{is} & O_P(1) \\ o_P(1) \cdot X_n \xrightarrow{P} 0 & \text{is} & o_P(1). \end{array}$$

We are now fully equipped to study the asymptotic distribution of many estimators in econometrics.

2. Example: Asymptotic distribution of Indirect LS

Recall our first structural model

$$\begin{aligned} Y &= \alpha_1 D + \alpha_2 W + U, \quad U \perp Z, W \\ D &= \beta_1 Z + \beta_2 W + V, \quad V \perp Z, W \end{aligned} \tag{IVM}$$

In order to focus on the structural coefficient of interest α_1 we partial out the controls W:

$$\begin{array}{ccrccc} Y & = & \gamma_{YW} & W & + & \widetilde{Y} \\ D & = & \gamma_{DW} & W & + & \widetilde{D} \\ Z & = & \gamma_{ZW} & W & + & \widetilde{Z} \\ W & = & 1 & W & + & 0 \\ U & = & 0 & W & + & U \\ V & = & 0 & W & + & V \end{array} \tag{ortho}$$

is the orthogonal decomposition of all variables w.r.t. span(W). Upon applying the linear operation of partialling out to eq. (IVM) we obtain

$$\begin{aligned} \widetilde{Y} &= \alpha_1 \widetilde{D} + U, \quad U \perp \widetilde{Z} \\ \widetilde{D} &= \beta_1 \widetilde{Z} + V, \quad V \perp \widetilde{Z}. \end{aligned} \tag{toyIVM}$$

Indirect least squares approach is based on the observation that orthogonality condition in eq. (toyIVM) imply that β_1 is a projection coefficient of $\widetilde{D}$ on $\widetilde{Z}$, the so called first-stage. We push this idea further: upon substituting the first stage into the main structural equation, we obtain the following reduced form:

$$\widetilde{Y} = \underbrace{\alpha_1 \beta_1}_{=:\gamma} \widetilde{Z} + \underbrace{U + \alpha_1 V}_{=:\epsilon}$$

and observe again that the orthogonality conditions of eq. (toyIVM) imply that γ *is the projection* coefficient of $\widetilde{Y}$ onto $\widetilde{Z}$. Thus, assuming that there is a non-trivial first stage, $\beta_1 \neq 0$, we can identify the structural coefficient α_1 as

$$\alpha_1 = \frac{\gamma}{\beta_1}. \tag{1}$$

Our goal here is to characterize the asymptotic distribution of an estimator based on eq. (1). The first step is to wright down an explicit formula for the projection coefficients in eq. (1). It will be necessary to analyze these coefficients jointly, so I set up a vector:

$$\theta := \begin{bmatrix} \beta_1 \\ \gamma \end{bmatrix} = \begin{bmatrix} (\mathrm{E}\widetilde{Z}^2)^{-1}(\mathrm{E}\widetilde{Z}\widetilde{D}) \\ (\mathrm{E}\widetilde{Z}^2)^{-1}(\mathrm{E}\widetilde{Z}\widetilde{Y}) \end{bmatrix},$$

where I used the formula for orthogonal projection coefficient. We obtain an estimator $\widetilde{\theta}$ by replacing expectations (which we don't know how to compute because we don't know the underlying P) with empirical expectation (which works asymptotically as good as P by LLN):

$$\widetilde{\theta} := \begin{bmatrix} (\mathbb{E}_n \widetilde{Z}_i^2)^{-1}(\mathbb{E}_n \widetilde{Z}_i \widetilde{D}_i) \\ (\mathbb{E}_n \widetilde{Z}_i^2)^{-1}(\mathbb{E}_n \widetilde{Z}_i \widetilde{Y}_i) \end{bmatrix}. \tag{2}$$

In order to compute $\widetilde{\theta}$, a sample of $\{\widetilde{Z}_i, \widetilde{D}_i, \widetilde{Y}_i\}$ is required. But we only have a sample of $\{Z_i, D_i, Y_i, W_i\}$, since we do no know the projection coefficients in eq. (ortho) we cannot construct a sample of tilded variables. But we can estimate all the projection coefficients in eq. (ortho) and construct estimates of the tilded variables, which we will denote with checks:

$$\begin{aligned} \breve{Y} &:= Y - \widehat{\gamma}_{YW} W \\ \breve{D} &:= D - \widehat{\gamma}_{DW} W \\ \breve{Z} &:= Z - \widehat{\gamma}_{ZW} W \end{aligned} \tag{ortho$\breve{\ }$}$$

Here $\widehat{\gamma}$'s are OLS estimates; we will use the fact that OLS estimators are consistent in the what follows. We define an estimator based on checks:

$$\widehat{\theta} := \begin{bmatrix} (\mathbb{E}_n \breve{Z}_i^2)^{-1} (\mathbb{E}_n \breve{Z}_i \breve{D}_i) \\ (\mathbb{E}_n \breve{Z}_i^2)^{-1} (\mathbb{E}_n \breve{Z}_i \breve{Y}_i) \end{bmatrix}.$$

The plan is to establish asymptotic properties of $\widetilde{\theta}$ and then show that $\widehat{\theta}$ is equivalent to $\widetilde{\theta}$ in the sense that

$$\sqrt{n}(\widehat{\theta} - \theta) = \sqrt{n}(\widetilde{\theta} - \theta) + o_P(1), \tag{3}$$

which would imply that $\widetilde{\theta}$ and $\widehat{\theta}$ have the same asymptotic distribution, see Section 1.9.

First we make the simple observation that $\widetilde{\theta}$ is consistent, so long as the LLN works, as can be see directly from eq. (2) via CMT. A sufficient condition for the LLN here is $\mathrm{E}\|Z\|_2^2, \mathrm{E}\|D\|_2^2, \mathrm{E}\|Y\|_2^2 < \infty$.

Next we substitute the first stage equation for $\widetilde{D}$ and the reduced form equation for $\widetilde{Y}$ into formula (2)

$$\widetilde{\theta} := (\mathbb{E}_n \widetilde{Z}_i^2)^{-1} \begin{bmatrix} (\mathbb{E}_n \widetilde{Z}_i \widetilde{D}_i) \\ (\mathbb{E}_n \widetilde{Z}_i \widetilde{Y}_i) \end{bmatrix} = (\mathbb{E}_n \widetilde{Z}_i^2)^{-1} \begin{bmatrix} (\mathbb{E}_n \widetilde{Z}_i \beta_1 \widetilde{Z}_i + \widetilde{Z}_i V_i) \\ (\mathbb{E}_n \widetilde{Z}_i \gamma \widetilde{Z}_i + \widetilde{Z}_i \epsilon_i) \end{bmatrix}$$

after regrouping and rescaling we obtain

$$\sqrt{n}(\widetilde{\theta} - \theta) = (\mathbb{E}_n \widetilde{Z}_i^2)^{-1} \sqrt{n} \mathbb{E}_n \begin{bmatrix} (\widetilde{Z}_i V_i) \\ (\widetilde{Z}_i \epsilon_i) \end{bmatrix} \tag{4}$$

From the last display we read off that $\sqrt{n}(\widetilde{\theta} - \theta) \quad (\mathrm{E}\widetilde{Z}^2)^{-2} N(0, V_\theta)$ by LLN, CLT, CMT and Slutsky's lemma, where $V_\theta = \begin{bmatrix} \mathrm{E}(\widetilde{Z}V)^2 & \mathrm{E}(\epsilon V \widetilde{Z}^2) \\ & \mathrm{E}(\widetilde{Z}\epsilon)^2 \end{bmatrix}$. A sufficient condition for the CLT here is $\mathrm{E}\|Z\|_4^4, \mathrm{E}\|D\|_4^4, \mathrm{E}\|Y\|_4^4 < \infty$.

Finally we can use the Delta method to get the asymptotic distribution of $\widetilde{\alpha} = \phi(\widetilde{\theta})$, for $\phi(x_1, x_2) = \frac{x_2}{x_1}$, with $\nabla \phi_x = [-\frac{x_2}{x_1^2} \; \frac{1}{x_1}]$:

$$\sqrt{n}(\widetilde{\alpha} - \alpha) = \sqrt{n}(\phi(\widetilde{\theta}) - \phi(\theta)) \quad [-\frac{\gamma}{\beta_1^2} \; \frac{1}{\beta_1}](\mathrm{E}\widetilde{Z}^2)^{-2} N(0, V_\theta).$$ [2]

We now argue eq. (3). Per discussion in Section 1.7, it is enough to show this for each component of θ. Since the analysis for each component is identical, we will only show that $\sqrt{n}(\widehat{\beta} - \widetilde{\beta}) = o_P(1)$. Here we again begin by looking at the exact formula for each estimator:

$$\begin{aligned} \sqrt{n}(\widehat{\beta} - \widetilde{\beta}) &= \sqrt{n}\Big[(\mathbb{E}_n \breve{Z}_i^2)^{-1} \mathbb{E}_n \breve{Z}_i \breve{D}_i - (\mathbb{E}_n \widetilde{Z}_i^2)^{-1} \mathbb{E}_n \widetilde{Z}_i \widetilde{D}_i\Big] \\ &= \sqrt{n}\Big[(\mathbb{E}_n \breve{Z}_i^2)^{-1} \mathbb{E}_n \breve{Z}_i \breve{D}_i - (\mathbb{E}_n \breve{Z}_i^2)^{-1} \mathbb{E}_n \widetilde{Z}_i \widetilde{D}_i\Big] + \sqrt{n}\Big[(\mathbb{E}_n \breve{Z}_i^2)^{-1} \mathbb{E}_n \widetilde{Z}_i \widetilde{D}_i - (\mathbb{E}_n \widetilde{Z}_i^2)^{-1} \mathbb{E}_n \widetilde{Z}_i \widetilde{D}_i\Big] \\ &= \underbrace{(\mathbb{E}_n \breve{Z}_i^2)^{-1}}_{\text{chunk1}} \sqrt{n} \mathbb{E}_n \Big[\underbrace{\breve{Z}_i (\breve{D}_i - \widetilde{D}_i)}_{\text{chunk2}} + \underbrace{(\breve{Z}_i - \widetilde{Z}_i) \widetilde{D}_i}_{\text{chunk3}}\Big] + \underbrace{\qquad \ldots \qquad}_{\text{Term2}} \end{aligned}$$

It is good to break the problem into small chunks: We work with the two large terms in the last

[2] compare this to Theorem 2 of L2

screen separately, and break each of the terms into three small chunks. To compare checks and tildas, we combine eq. (ortho) and eq. (ortȟo) to obtain:

$$\check{D}_i - \widetilde{D}_i = (\gamma_{DW} - \widehat{\gamma}_{DW})W_i \tag{5}$$

$$\check{Z}_i - \widetilde{Z}_i = (\gamma_{ZW} - \widehat{\gamma}_{ZW})W_i \tag{6}$$

Chunk1 From eq. (6)

$$\begin{aligned}
\mathbb{E}_n\check{Z}_i^2 &= \mathbb{E}_n\widetilde{Z}_i^2 + 2(\gamma_{ZW} - \widehat{\gamma}_{ZW})\mathbb{E}_n\widetilde{Z}_iW_i + (\gamma_{ZW} - \widehat{\gamma}_{ZW})^2\mathbb{E}_nW_i^2 \\
&= \mathbb{E}_n\widetilde{Z}_i^2 + o_P(1)O_P(1) + o_P(1)O_P(1) \\
&= \mathrm{E}\widetilde{Z}_i^2 + o_P(1).
\end{aligned}$$

Chunk2 From eq. (5) and eq. (6)

$$\begin{aligned}
\sqrt{n}\mathbb{E}_n\check{Z}_i(\check{D}_i - \widetilde{D}_i) &= (\gamma_{DW} - \widehat{\gamma}_{DW})\sqrt{n}\mathbb{E}_n\check{Z}_iW_i = o_P(1)\sqrt{n}\mathbb{E}_n\check{Z}_iW_i \\
&= o_P(1)\left[\sqrt{n}\mathbb{E}_n\widetilde{Z}_iW_i + o_P(1)\sqrt{n}\mathbb{E}_nW_i^2\right] \\
&= o_P(1)O_P(1) = o_P(1).
\end{aligned}$$

Chunk3 From eq. (5)

$$\begin{aligned}
\sqrt{n}\mathbb{E}_n(\check{Z}_i - \widetilde{Z}_i)\widetilde{D}_i &= o_P(1)\sqrt{n}\mathbb{E}_n\widehat{D}_iW_i \\
&= o_P(1)O_P(1) = o_P(1).
\end{aligned}$$

Putting these observations together, we conclude that the first term is

$$\begin{aligned}
\mathrm{Term1} &= (\mathrm{positive} + o_P(1))^{-1}\left[o_P(1) + o_P(1)\right] \\
&= O_P(1)o_P(1) = o_P(1).
\end{aligned}$$

Analysis for Term2 is similar.

Remark. It is difficult to keep track of so many different empirical sums of various random variables as on the screen at bottom of previous page. Notation $o_P(1)$ and $O_P(1)$ is extremely helpful here, allowing us to focus on the size of each bit and parse the whole formula by applying simple rules stated in Section 1.9.

References

[1] S. Resnick. "A Probability Path" (1998).

[2] A. W. Van der Vaart. *Asymptotic statistics*. Vol. 3. Cambridge university press, 2000.

A deeper overview of results mentioned in Section 1 can be found in Vaart [2]. A rigorous and very explicit development of probability theory from ground zero with lots of excellent exercises and links to statistics can be found in Resnick [1].