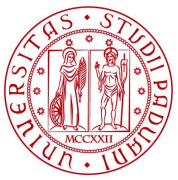


UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Università degli Studi di Padova

DIPARTIMENTO DI SCIENZE STATISTICHE

Corso di Laurea Triennale in
Statistica per le Tecnologie e le Scienze

ERROR BOUNDS FOR THE CENTRAL LIMIT THEOREM
via the Berry–Esseen Theorem and Stein's method

Relatore: Prof. Alberto Chiarini
Dipartimento di Matematica "Tullio Levi–Civita"

Laureando: Alessandro Rizzi
Matricola: 2036017

Anno Accademico 2024/2025

17 luglio 2025

Contents

Chapter 0

Introduction

Throughout the many hours of lectures during my three-year undergraduate studies in Statistics, I have heard countless times about the Central Limit Theorem, its big role in Statistics, its usefulness, and so on. I cannot remember how many times a professor remarked its importance, or used it to draw some other conclusions. It is so simple in its formulation, and it is mainly derived from the Strong Law of Large Numbers, which is even easier to grasp. More so than that, it is extremely versatile, and that's exactly why it is that largely used in probabilistic results. Yet looking closely at its formulation, one can notice the theorem states an asymptotic result, a convergence at infinity, and does not specify any quantitative measure. Therefore, one may ask themselves: "How much error in the approximation do I have to expect from a fixed sample size?". On the other hand, what is the minimum sample size needed to obtain a sufficiently small error? An overly large error is likely to lead to produce results substantially different from reality, the moment it is wrongly assumed the Central Limit Theorem proves a suitable convergence to the normal distribution on a sample of too few observations. From Klenke (2013)

Theorem 1 (Central Limit theorem)**.** *Let $X_1, X_2, \ldots, X_n$ be i.i.d. real random variables with $\mu := \mathbb{E}[X_1] \in \mathbb{R}$ and $\sigma^2 := Var(X_1) \in (0, \infty)$. For $n \in \mathbb{N}$, let $\overline{X}_n := (1/\sqrt{\sigma^2 n}) \sum_{i=1}^{n} (X_i - \mu)$. Then*

$$\overline{X}_n \xrightarrow{d} N(0,1).$$

Some methods to obtain a quite precise measure of the error do exist. They usually consist of The most classical one is the Berry–Esseen theorem.

Theorem 2 (Berry–Esseen)**.** *Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with $\mathbb{E}|X_1|^3 < \infty, \mathbb{E}X_1 = 0$, and $Var(X_1) = 1$. If Φ denotes the cumulative distribution function of a*

standard normal distribution and $W_n = \sqrt{n}\,\overline{X}_n = n^{-\frac{1}{2}} \sum X_i$, *then*

$$|\mathbb{P}(W_n \leq x) - \Phi(x)| \leq 0.4785\, \frac{\mathbb{E}|X_1|^3}{\sqrt{n}}.$$

This is a nice upper bound, however it requires estimating the Berry– constant, which is not easy. Instead, in this thesis, I will go through how to get bounds using a very powerful method called Stein's method. This method is very powerful and it's scope far exceeds what we want to achieve here; in particular it is useful to bound the distance from two distributions, and therefore we can use it to obtain bounds for the error of approximation in the Central Limit Theorem. The key is first to define a metric to measure distance between probability distributions, then convert the problem of measuring the error when approximating a distribution with a known one with a problem of bounding the expectation of a certain functional of the variable of interest, and finally bound the aforementioned expectation. Stein's method, as we will see, returns a bound which shares some similarities with the one obtained through Berry–Esseen. However, the result we will obtain through Stein's method cannot be directly compared to the one in the Berry–Esseen theorem as they use different metrics to measure distance between probability distributions.

A great amount of this thesis refers to Ross (2011) and the work of Charles Stein (Stein 1972 and Stein 1986).

Chapter 1

Stein's method

This chapter is dedicated to explaining step by step Stein's method and the process to obtain error bounds for the Central Limit Theorem using it. We begin by defining a class of distances that can be used to measure the difference in the probability distributions of random variables. This is followed by the definition of three particular distances and some related properties.

1.1 Distances between probability distributions

Let μ, ν be the two probability measures of random variables U, V, and $\mathcal{H}$ a family of 'test' functions, we introduce a class of probability metrics defined as:

$$d_{\mathcal{H}}(\mu,\nu) := \sup_{h\in\mathcal{H}} \left| \int h(x)d\mu(x) - \int h(x)d\nu(x) \right| = \sup_{h\in\mathcal{H}} \left| \mathbb{E}_\mu h(U) - \mathbb{E}_\nu h(V) \right|. \tag{1.1}$$

From now on, it will be written simply as $d_{\mathcal{H}}(U,V)$, ignoring the little abuse of notation.

There are three important families of functions that define three important metrics:

1. Taking $\mathcal{H} = \{\mathbb{1}_{(-\infty,x]} : x \in \mathbb{R}\}$ we get the Kolmogorov metric, which will be written as d_K. It is easy to understand this metric as, $\forall x \in \mathbb{R}$,

   $$\mathbb{E}[\mathbb{1}_{(-\infty,x]}(U)] = \mathbb{P}(U \leq x).$$

   Then $d_K(U,V) = \sup_{x\in\mathbb{R}} |\mathbb{P}(U \leq x) - \mathbb{P}(V \leq x)|$, which is just the maximum difference between distribution functions. Therefore, as $n \longrightarrow \infty$,

   $$d_K(X_n, X) \to 0 \implies X_n \xrightarrow{d} X \qquad \text{(weak convergence)}.$$

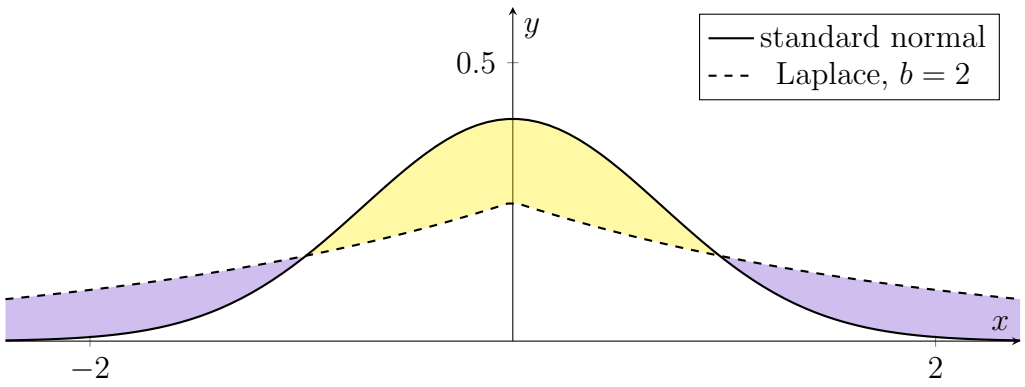


Figure 1.1: An example of the total variation metric, measuring the distance between two continuous distributions (standard normal and Laplace with median absolute deviation 2). d_{TV} is either the area in yellow or the area in violet; we will see that there is actually no difference between them (keep in mind that the y axis in this image is 'stretched', thus it isn't a perfect representation of the real plot).

2. With $\mathcal{H} = \{f : \mathbb{R} \longrightarrow \mathbb{R} \ : |f(x) - f(y)| \leq |x - y|\}$ we get d_W, the Wasserstein metric. It is less intuitive than the Kolmogorov metric, but it will be useful when working with continuous distributions, because of some properties that we will see later.

3. Finally, $\mathcal{H} = \{\mathbb{1}_{\mathbb{A}} : \mathbb{A} \in \text{Borel}(\mathbb{R})\}$ defines d_{TV}, the total variation metric. For visualization it is useful to remember that $\mathbb{E}[\mathbb{1}_{[a,b]}(X)] = \mathbb{P}(a \leq X \leq b)$, with $a, b \in \mathbb{R}, a \leq b$. The set $\mathbb{A}$ which gives the maximum absolute difference between the two expected values is then either the set containing all the x where $\mathbb{P}(U = x) > \mathbb{P}(V = x)$, or all the y where $\mathbb{P}(U = y) < \mathbb{P}(V = y)$, in the case we are trying to compute $d_{TV}(U, V)$, for U, V random variables. In Figure 1.1 there is a representation of the distance between two continuous distributions in the total variation metric. In that example $\mathbb{A}$ is either $(-1, 1)$ or the complement. Although it can be used with continuous distributions switching probabilities with densities, it is with discrete ones that we can really understand its usefulness, and we will see why in the next proposition.

We will now prove three propositions that underscore the importance of the three particular distances defined above using the notion of a $d_{\mathcal{H}}$ metric. The Kolmogorov metric is by far the simplest one to understand, so naturally we would prefer to choose it amongst the three. However, it is as simple to understand as it is difficult to compute it, and that is why the other distances will prove to be more practical.

Proposition 1. *Retaining the notation above for the Kolmogorov metric d_K and the total*

variation metric d_{TV}, let W, Z be random variables, then

$$d_K(W, Z) \leq d_{TV}(W, Z).$$

Proof. This comes directly after the fact that every $h \in \mathcal{H}_K$ is, by definition, an indicator function over an interval $(-\infty, x),\ x \in \mathbb{R}$, which is Borel as it is equal to $\mathbb{R} \setminus (x, \infty)$. It follows that $\mathcal{H}_K \subset \mathcal{H}_{TV}$, therefore $d_{TV}(W, Z) \geq d_K(W, Z)$ because it is a sup over a larger set. □

In the case of discrete distributions, the total variation distance has the following useful representation.

Proposition 2. *Let W, Z be discrete random variables with sample space Ω,*

$$d_{TV}(W, Z) = \frac{1}{2} \sum_{\omega \in \Omega} |\mathbb{P}(W = \omega) - \mathbb{P}(Z = \omega)|.$$

Proof. Let's begin by proving that taking $\mathbb{1}_{\mathbb{A}_i} \in \mathcal{H}_{TV}$ with $\mathbb{A}_1 = \{\omega \in \Omega : \mathbb{P}(W = \omega) > \mathbb{P}(Z = \omega)\}$ or with $\mathbb{A}_2 = \{\omega \in \Omega : \mathbb{P}(Z = \omega) > \mathbb{P}(W = \omega)\}$ is the same (this is analogous with continuous distributions, using densities instead of probabilities, see Figure 1.1). Obviously the set $\mathbb{A}_3 = \{\omega \in \Omega : \mathbb{P}(W = \omega) = \mathbb{P}(Z = \omega)\}$ can be ignored, as its contribution is zero. Considering that $\mathbb{P}(W \in \mathbb{A}_1 \cap \mathbb{A}_2 \cap \mathbb{A}_3) = \mathbb{P}(Z \in \mathbb{A}_1 \cap \mathbb{A}_2 \cap \mathbb{A}_3) = \mathbb{P}(\Omega) = 1$, and that the 3 sets are disjoint, then it comes with ease that

$$\begin{aligned}
\mathbb{P}_W(\mathbb{A}_3) &= \mathbb{P}_Z(\mathbb{A}_3) \\
\mathbb{P}_W(\mathbb{A}_1 \cap \mathbb{A}_3) - \mathbb{P}_Z(\mathbb{A}_1 \cap \mathbb{A}_3) &= \mathbb{P}_W(\mathbb{A}_1) + \cancel{\mathbb{P}_W(\mathbb{A}_3)} - \mathbb{P}_Z(\mathbb{A}_1) + \cancel{\mathbb{P}_Z(\mathbb{A}_3)} \\
&= \mathbb{P}_W(\mathbb{A}_1) - \mathbb{P}_Z(\mathbb{A}_1) \\
\mathbb{P}_Z(\mathbb{A}_2 \cap \mathbb{A}_3) - \mathbb{P}_W(\mathbb{A}_2 \cap \mathbb{A}_3) &= \big[1 - \mathbb{P}_Z(\mathbb{A}_1)\big] - \big[1 - \mathbb{P}_W(\mathbb{A}_1)\big] \\
&= \mathbb{P}_W(\mathbb{A}_1) - \mathbb{P}_Z(\mathbb{A}_1)
\end{aligned}$$

$$\begin{aligned}
&\implies |\mathbb{P}_W(\mathbb{A}_1) - \mathbb{P}_Z(\mathbb{A}_1)| + |\mathbb{P}_W(\mathbb{A}_2) - \mathbb{P}_Z(\mathbb{A}_2)| + |\mathbb{P}_W(\mathbb{A}_3) - \mathbb{P}_Z(\mathbb{A}_3)| = \\
&= \sum_{\omega \in \Omega} |\mathbb{P}(W = \omega) - \mathbb{P}(Z = \omega)| = 2|\mathbb{P}_W(\mathbb{A}_1) - \mathbb{P}_Z(\mathbb{A}_1)| = 2d_{TV}(W, Z).
\end{aligned}$$

□

Finally, this next proposition highlights a useful relation between the Kolmogorov and the Wasserstein distance when comparing two continuous distributions, with the only

condition that one of them has a bounded density. The main idea is to approximate indicator functions with Lipschitz continuous functions determined by a 'smoothness' factor ε, effectively linking the two distances.

Proposition 3. *Let Z be a random variable with Lebesgue density bounded by C, then for any random variable W:*

$$d_K(W, Z) \leq \sqrt{2C d_W(W, Z)}.$$

Proof. To prove this we have to consider $h_x(w)$ as the functions $\mathbb{1}_{w \leq x}$, and the $h_{x,\varepsilon}(w)$ ($\varepsilon > 0$) a 'smoothed' version (see Figure 1.2) which is linear between $(x, h_x(x)) = (x, 1)$ and $(x + \varepsilon, h_x(x + \varepsilon)) = (x + \varepsilon, 0)$, hence

$$h_{x,\varepsilon}(w) := \begin{cases} 1 & w \leq x \\ \frac{x+\varepsilon-w}{\varepsilon} & x < w \leq x + \varepsilon \\ 0 & w > x + \varepsilon. \end{cases} \tag{1.2}$$

We can clearly see $h_{x,\varepsilon}(w) = h_x(w) + g_{x,\varepsilon}(w)$, where $g_{x,\varepsilon}(w) \geq 0 \; \forall w \in \mathbb{R}$. Then $\mathbb{E}h_{x,\varepsilon}(W) = \mathbb{E}[h_x(W) + g_{x,\varepsilon}(W)] \geq \mathbb{E}h_x(W) \; \forall \varepsilon > 0, x \in \mathbb{R}$, because $\mathbb{E}g_{x,\varepsilon}(W)$ is always non-negative[1], and we have

$$\begin{aligned} \mathbb{E}h_x(W) - \mathbb{E}h_x(Z) &= \mathbb{E}h_x(W) - \mathbb{E}h_{x,\varepsilon}(Z) + \mathbb{E}h_{x,\varepsilon}(Z) - \mathbb{E}h_x(Z) && (1.3) \\ &\leq \mathbb{E}h_{x,\varepsilon}(W) - \mathbb{E}h_{x,\varepsilon}(Z) + [\mathbb{E}h_{x,\varepsilon}(Z) - \mathbb{E}h_x(Z)] && (1.4) \\ &\leq \mathbb{E}h_{x,\varepsilon}(W) - \mathbb{E}h_{x,\varepsilon}(Z) + \frac{1}{2}C\varepsilon && (1.5) \\ &\leq \frac{1}{\varepsilon} d_W(W, Z) + \frac{1}{2}C\varepsilon. && (1.6) \end{aligned}$$

The step (1.5) is justified by first taking again $\mathbb{E}h_{x,\varepsilon}(Z) = \mathbb{E}h_x(Z) + \mathbb{E}g_{x,\varepsilon}(Z)$, then it is helpful to see that the area under $g_{x,\varepsilon}(w)$ is a triangle, exactly half of a rectangle with the same width and height (Figure 1.2). Knowing that the expected value of $g_{x,\varepsilon}(Z)$ is the integral over $\mathbb{R}$ of $g_{x,\varepsilon}(z)d(z)$, where $d(z)$ is the density of Z, and remembering that this is bounded by C, we find that $\mathbb{E}g_{x,\varepsilon}(W) \leq \int_{\mathbb{R}} C g_{x,\varepsilon}(z)\, dz$. The area under $Cg_{x,\varepsilon}(z)$ is just half of the rectangle of height $1 \cdot C$ and width ε, whereby we concluded the proof for this step.

[1] To be more precise, $0 \leq \mathbb{E}g_{x,\varepsilon}(W) \leq \mathbb{P}(x < W < x + \varepsilon) \leq 1$.

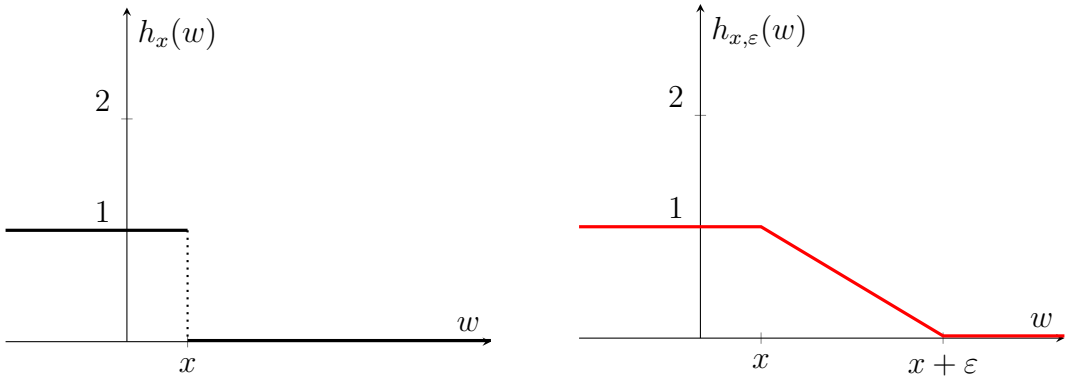


Figure 1.2: On the left $h_x(w)$, on the right $h_{x,\varepsilon}(w)$, with $x = 0.5, \varepsilon = 1.5$.

For the step Equation 1.6 it suffices to prove that a modified version of $h_{x,\varepsilon}(w)$ is contained in $\mathcal{H}_W$. Since $h_{x,\varepsilon}(w_1) - h_{x,\varepsilon}(w_2) = 0$ for every $w_1, w_2 \leq x \vee w_1, w_2 > x + \varepsilon$ (as the function is constant), the critical part is the linear one where, as defined in (1.2), $|h_{x,\varepsilon}(w_1) - h_{x,\varepsilon}(w_2)| = |\frac{1}{\varepsilon}(w_2 - w_1)|$, thus $\varepsilon h_{x,\varepsilon}(w) \in \mathcal{H}_W$ (as the constant parts multiplied by ε are still constant, and if $w = x$ or $w = x + \varepsilon$ the function $\varepsilon h_{x,\varepsilon}(w)$ is unambiguously ε or 0, so it is still continuous). It follows that $d_W(W, Z) \geq \varepsilon |\mathbb{E}h_{x,\varepsilon}(W) - \mathbb{E}h_{x,\varepsilon}(Z)|$.

We want to find the ε which gives the minimum possible boundary, or rather the solution of $\min_\varepsilon \frac{1}{\varepsilon} d_W(W, Z) + \frac{1}{2} C\varepsilon$.

$$\frac{d}{d\varepsilon}\left(d_W(W, Z)\, \varepsilon^{-1} + \frac{C}{2}\,\varepsilon\right) = \frac{C}{2} - d_W(W, Z)\, \varepsilon^{-2},$$

and we check where it is equal to zero:

$$\begin{aligned} \frac{C}{2} - d_W(W, Z)\, \varepsilon^{-2} &= 0 \\ d_W(W, Z)\, \varepsilon^{-2} &= \frac{C}{2} \\ \varepsilon^2 &= \frac{2 d_W(W, Z)}{C} \\ \varepsilon &= \pm\sqrt{\frac{2 d_W(W, Z)}{C}}. \end{aligned}$$

Computing the second derivative we get

$$\frac{d^2}{d\varepsilon^2}\left(d_W(W, Z)\, \varepsilon^{-1} + \frac{C}{2}\,\varepsilon\right) = 2 d_W(W, Z)\, \varepsilon^{-3}$$

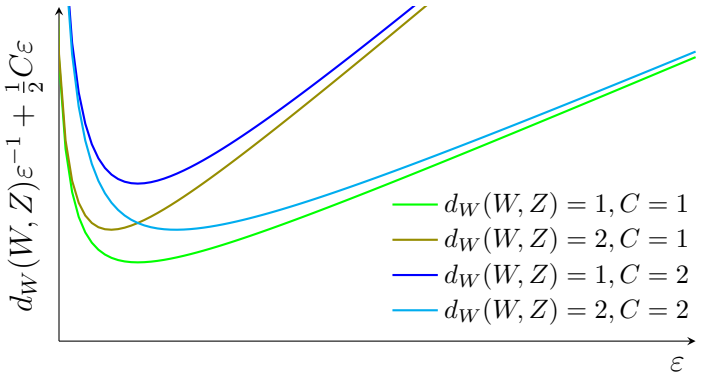


Figure 1.3: Plots of $d_W(W,Z)\varepsilon^{-1} + \frac{1}{2}C\varepsilon$ for varying values of $d_K(W,Z)$ and C, restricted to $\varepsilon \in \mathbb{R}^+$. One can clearly observe the presence of a unique minimum in the functions.

which, at $\varepsilon = \pm\sqrt{2d_W(W,Z)C^{-1}}$, is

$$\pm\sqrt{\frac{C^3}{2d_W(W,Z)}}.$$

Since d_W and C are always non-negative, the square root is real, and we take the positive solution to get the local minimum. $\varepsilon = \sqrt{2d_W(W,Z)C^{-1}}$ is also the global minimum, because the limit of (1.6) as $\varepsilon \to 0$ and the limit of (1.6) as $\varepsilon \to \infty$ are both equal to $+\infty$, and the function is continuous in $(0,\infty)$ (Figure 1.3).

As we assumed $\varepsilon > 0$, the right hand side of the inequality is always positive. Since, by the definition of Kolmogorov distance, $\sup_x|\mathbb{E}h_x(W) - \mathbb{E}h_x(Z)| = d_K(W,Z)$, then

$$\begin{aligned}\mathbb{E}h_x(W) - \mathbb{E}h_x(Z) &\in [-d_K(W,Z), d_K(W,Z)],\\ |\mathbb{E}h_x(W) - \mathbb{E}h_x(Z)| &\in [0, d_K(W,Z)],\\ [0, d_K(W,Z)] &\subset [-d_K(W,Z), d_K(W,Z)],\end{aligned}$$

so taking the absolute value of the left-hand side preserves the inequality. Finally, taking the supremum on the left-hand side and letting $\varepsilon = \sqrt{2d_W(W,Z)C^{-1}}$ shows the desired inequality. □

1.2 At the core of Stein's method

In this section we see some theory which will be useful to understand the first theorem of the next section, regarding the Wasserstein distance restricted to the case where one of the random variables has the standard normal distribution. Before going too deep with

the Wasserstein distance, we will look at Stein's Lemma, the core of the Stein's method, which establish the relation we need between the distance of a probability distribution from the standard normal and a functional of this variable, and apply it on the intuitive Kolmogorov distance. In particular the focus will be on the characterizing operator of the standard normal distribution applied to a piecewise differentiable function. To be precise this statement is not entirely correct, as Stein's Lemma actually specifies that the function should be *absolutely continuous* (Ross 2011). To avoid diving too deep into complex measure theoretic concepts I opted for a piecewise C^1 function, which is sufficient for the purposes of this exposition. At the end, in subsection 1.2.1, we will briefly see a generalisation of these results that will help switching to the Wasserstein distance in the next section.

A great part of this section will be spent on discussing whether the first derivative of the function is bounded or not. This is important because it guarantees that the expectation of the absolute value of $f'(W)$ is finite, which lets us make use of Fubini's theorem to prove part of Stein's Lemma. Ultimately, the bounds for f and f' will be helpful in the next section, when bounding the expectation obtained with Stein's Lemma. Furthermore I want to mention I decided not to rely too much on known theorems in the analytical study of the function, as I personally prefer to follow a thought process and gradually visualise the plot in my mind as much as possible when analysing a function. Nonetheless, I supplied some images and counterexamples alongside the mathematical reasoning to illustrate the arguments in practice.

Finally, we introduce Stein's Lemma, which we will rely on throughout the remainder of this thesis.

Lemma 1 (Stein's Lemma)**.** *Define a functional operator $\mathcal{A}$ by*

$$\mathcal{A}f(x) := f'(x) - xf(x).$$

1. *If the random variable Z has the standard normal distribution, then $\mathbb{E}\mathcal{A}f(Z) = 0$ for all piecewise differentiable functions f with $\mathbb{E}|f'(Z)| < \infty$.*

2. *If for some random variable W, $\mathbb{E}\mathcal{A}f(W) = 0$ for all piecewise differentiable functions f with $\|f'(Z)\| < \infty$, then* ***W has the standard normal distribution****.*

The operator $\mathcal{A}$ is referred to as the ***characterizing operator of the standard normal distribution****.*

Proof of Lemma 1 (1). Let Z be a standard normal random variable, f a piecewise differentiable function such that $\mathbb{E}|f'(Z)| < \infty$, Φ the standard normal cumulative distribution function and ϕ the standard normal density function. Then, as $\mathbb{E}|f'(Z)| < \infty$, we write the expectation as a double integral and by Fubini's theorem we can change the order of integration:

$$\begin{aligned}
\mathbb{E}f'(Z) &= \int_{-\infty}^{0} f'(t)\phi(t)\,dt + \int_{0}^{\infty} f'(t)\phi(t)\,dt \\
&= \int_{-\infty}^{0} f'(t) \int_{-\infty}^{t} u\phi(u)\,du\,dt + \int_{0}^{\infty} f'(t) \int_{t}^{\infty} u\phi(t)\,du\,dt \\
&= \int_{-\infty}^{0} u\phi(u) \int_{u}^{0} f'(t)\,dt\,du + \int_{-\infty}^{0} u\phi(u) \int_{0}^{u} f'(t)\,dt\,du \\
&= \mathbb{E}[Zf(Z)].
\end{aligned}$$

□

Before proving Lemma 1 (2), we proceed by looking for a function which satisfies

$$\mathbb{E}\mathcal{A}f_x(W) = \mathbb{P}(W \leq x) - \Phi(x). \tag{1.7}$$

This is useful because the supremum over x of the absolute value of the right-hand side of (1.7) corresponds to the distance of a distribution to the standard normal distribution in the Kolmogorov metric, and it is zero if and only if W has a standard normal distribution, then the left-hand side is also zero if W is standard normal, by Lemma 1 (1). Finding f_x is equivalent to solving the following differential equation:

$$f_x'(w) - wf_x(w) = \mathbb{1}_{w\leq x} - \Phi(x). \tag{1.8}$$

Applying the method of integrating factors shows

$$\begin{aligned}
e^{-\frac{w^2}{2}} f_x'(w) - we^{-\frac{w^2}{2}} f_x(w) &= e^{-\frac{w^2}{2}} \left[\mathbb{1}_{w\leq x} - \Phi(x)\right] \\
\frac{d}{dw}\left(e^{-\frac{w^2}{2}} f_x(w)\right) &= e^{-\frac{w^2}{2}} \left[\mathbb{1}_{w\leq x} - \Phi(x)\right],
\end{aligned}$$

then, integrating and considering the homogeneous solution:

$$f_x(w) = e^{\frac{w^2}{2}} \int_{-\infty}^{w} e^{-\frac{t^2}{2}} \left[\mathbb{1}_{t\leq x} - \Phi(x)\right] dt + Ce^{\frac{w^2}{2}}.$$

But, personally, I prefer to write it in a more explicit form:

$$f_x(w) = e^{\frac{w^2}{2}} \left(\int_{-\infty}^{w} \mathbb{1}_{t \leq x}\, e^{-\frac{t^2}{2}}\, dt - \Phi(x) \int_{-\infty}^{w} e^{-\frac{t^2}{2}}\, dt + C \right)$$
$$f_x(w) = \sqrt{2\pi} e^{\frac{w^2}{2}} \big[\Phi(\min\{w, x\}) - \Phi(x)\Phi(w) \big] + C e^{\frac{w^2}{2}}. \tag{1.9}$$

Is this function actually a solution of the differential equation (1.8)? We check it by computing the first derivative:

$$f_x'(w) = w f_x(w) + \cancel{e^{\frac{w^2}{2}} e^{-\frac{w^2}{2}}} \big[\mathbb{1}_{w \leq x} - \Phi(x) \big],$$

and we verified it really is a solution.

This solution is clearly unbounded when $C \neq 0$, since $e^{\frac{w^2}{2}} \to +\infty$ as $w \to \pm\infty$. Why do we need f to be bounded? Because we have to remember it is crucial that $\mathbb{E}|f'(W)| < \infty$ otherwise we can't prove (1) of Stein's Lemma, and if f_x' was bounded then we would automatically know that expectation is finite. Since $f_x'(w) \leq w f_x(w) + 1$ (see (1.8)), f_x doesn't have to be bounded only, but it should also decrease at the rate of w^{-1}, so f_x' is bounded if and only if $f_x = \mathcal{O}(w^{-1})$ as $w \to \pm\infty$ and $\|f_x\|_\infty < \infty$. It remains to analyse the behaviour of f_x when $C = 0$.

Knowing that $\|\Phi\|_\infty = 1$ and letting $k(w) = \sqrt{2\pi} e^{\frac{w^2}{2}}$:

$$\begin{aligned} w \leq x \implies & f_x(w) = k(w)\Phi(w)[1 - \Phi(x)] \ \wedge \ 1 - \Phi(w) \geq 1 - \Phi(x) \\ \implies & f_x(w) \leq k(w)\Phi(w)[1 - \Phi(w)] \end{aligned} \tag{1.10}$$

$$\begin{aligned} w > x \implies & f_x(w) = k(w)\Phi(x)[1 - \Phi(w)] \ \wedge \ \Phi(w) > \Phi(x) \\ \implies & f_x(w) < k(w)\Phi(w)[1 - \Phi(w)]. \end{aligned} \tag{1.11}$$

Combining (1.10) and (1.11) we get, $\forall w \in \mathbb{R}$,

$$f_x(w) \leq \sqrt{2\pi} e^{\frac{w^2}{2}} \Phi(w)[1 - \Phi(w)]. \tag{1.12}$$

The function on the right-hand side of (1.12) is even, as w^2 is clearly even and $\Phi(w) = 1 - \Phi(-w)$, consequently we can restrict our search for bounds for (1.12) to $\mathbb{R}^+$.

Now it would be useful to rewrite $1 - \Phi(w)$ in order to simplify the term $e^{\frac{w^2}{2}}$ (which is unbounded!). Considering $1 - \Phi(w)$ is the primitive of a constant multiplied by $e^{-\frac{w^2}{2}}$, it is reasonable to think that it also decreases exponentially. Then we want to find an upper bound for $1 - \Phi(w)$ in the subdomain $\mathbb{R}^+$. The critical part now is to notice that,

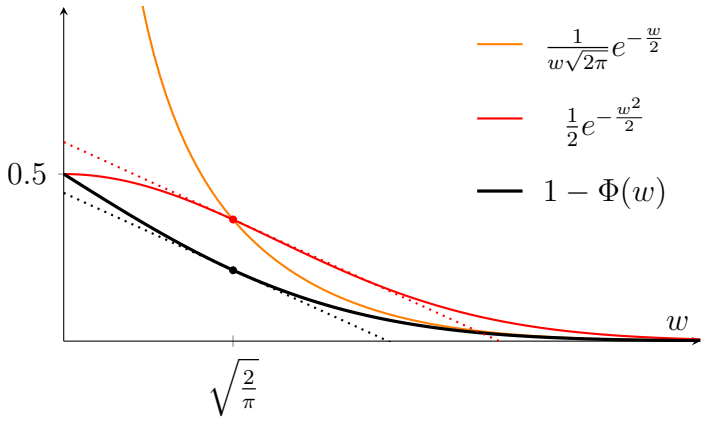


Figure 1.4: This figure displays the plot of two functions that upper-bound $1-\Phi(w)$, and the tangents at $\sqrt{2/\pi}$ marked with dotted lines where the red and the black functions have the same slope.

restricting to $w \geq 0$,

$$1-\Phi(w) \leq \frac{1}{2}e^{-\frac{w^2}{2}}. \tag{1.13}$$

What we would love to do is just solve the inequality to find if it is true for all $w \geq 0$. However this cannot be done, as $\Phi(\cdot)$ doesn't have an analytical form. Therefore it comes necessary to study both functions in unorthodox ways.

$$1-\Phi(0) = \frac{1}{2} = \frac{1}{2}e^{-\frac{0^2}{2}}$$
$$\lim_{w\to\infty} 1-\Phi(0) = 0^+ = \lim_{w\to\infty} \frac{1}{2}e^{-\frac{0^2}{2}},$$

so they both start at $(0, \frac{1}{2})$, and they both tend to zero as $w \to \infty$.

To study these functions more in depth, we differentiate them:

$$\frac{d}{dw}\Big(1-\Phi(w)\Big) \leq \frac{d}{dw}\left(\frac{1}{2}e^{-\frac{w^2}{2}}\right),$$

$$\iff -\frac{1}{\sqrt{2\pi}}e^{-\frac{w^2}{2}} \leq -\frac{1}{2}we^{-\frac{w^2}{2}} \tag{1.14}$$
$$\iff \frac{1}{\sqrt{2\pi}} \geq \frac{1}{2}w$$
$$\iff w \leq \sqrt{\frac{2}{\pi}}. \tag{1.15}$$

(1.14) implies they both decrease monotonically (and we know they are piecewise differ-

entiable) on $\mathbb{R}^+$, while (1.15) implies they get increasingly farther from each other (with $\frac{1}{2}e^{-\frac{w^2}{2}}$ on top) from $w = 0$ to $w = \sqrt{2/\pi}$, then they start getting closer. Again we would like to check at what points they are equal if it was possible, instead we analyse the first derivative of the difference. We know the difference increases when $0 \leq w \leq \sqrt{2/\pi}$ and then decreases, hence the first derivative will be positive before $\sqrt{2/\pi}$ and negative after. The difference tends to zero asymptotically, which implies that the first derivative of the difference also tends to zero.

Suppose the functions cross at some point $w_1 > \sqrt{2/\pi}$, and let $w_2 > w_1$ be either the next point where they cross again or $+\infty$[2]. Consequently the difference between the functions is negative in (w_1, w_2). But since the difference tends to zero at infinity, it follows that the difference function must increase at some point, thus there is a local minimum in the difference function at $w_0 \in (w_1, w_2)$, and its derivative must be zero at that point. An example of this applied to two functions that instead cross each other is illustrated in Figure 1.5. Taking the derivatives calculated at (1.14) we find

$$\frac{d}{dw}\left(\frac{1}{2}e^{-\frac{w^2}{2}} - 1 + \Phi(w)\right) = e^{-\frac{w^2}{2}}\left(\frac{1}{\sqrt{2\pi}} - \frac{1}{2}w\right) \neq 0 \iff w \neq \sqrt{\frac{2}{\pi}},$$

and we conclude the inequality at (1.13) is true for every $w \geq 0$ (Figure 1.4).

Putting together the results (1.9), (1.12) and (1.13) we can now easily find bounds for f_x (with $C = 0$):

$$\begin{aligned}
\|f_x\|_\infty &= \sup_w \left|\sqrt{2\pi}e^{\frac{w^2}{2}}\left[\Phi(\min\{w, x\}) - \Phi(x)\Phi(w)\right]\right| \\
&\leq \sup_w \sqrt{2\pi}e^{\frac{w^2}{2}}\Phi(w)[1 - \Phi(w)] \\
&\leq \sqrt{2\pi} \cdot \sup_w \cancel{e^{\frac{w^2}{2}}}\Phi(w)\frac{1}{2}\cancel{e^{-\frac{w^2}{2}}} \\
&\leq \sqrt{\frac{\pi}{2}}.
\end{aligned}$$

Finally we check if $wf_x(w)$ is bounded, but before that we show that $\forall w > 0$

$$1 - \Phi(w) \leq \frac{1}{w\sqrt{2\pi}}e^{-\frac{w^2}{2}}. \tag{1.16}$$

[2]With 'crossing' I mean actually crossing not just 'touching'. The formal definitions would be

$$w_1 = \inf\left\{w > \sqrt{\frac{2}{\pi}} : 1 - \Phi(w) > \frac{1}{2}e^{-\frac{w^2}{2}}\right\}, \quad w_2 = \sup\left\{w > w_1 : \forall k \in (w_1, w),\ 1 - \Phi(k) > \frac{1}{2}e^{-\frac{k^2}{2}}\right\}.$$

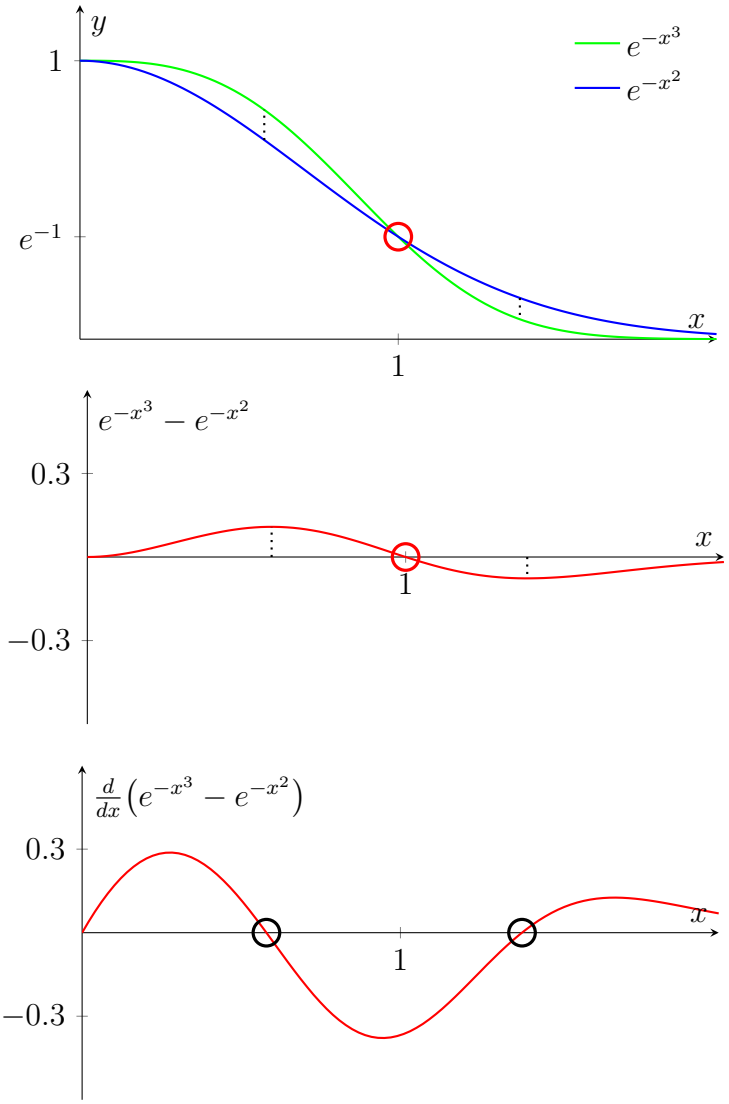


Figure 1.5: This figure illustrates an example of two functions starting at the same point and tending both to zero but where one does not bound the other. The focus in particular is at the difference, showing that if the functions cross through each other than there is at least one local minimum and one local maximum in the difference function.

Once again, this cannot be directly verified. First, we want to check the extremes of the function in $\mathbb{R}^+$:

$$\lim_{w\to 0^+} \frac{1}{w\sqrt{2\pi}} e^{-\frac{w^2}{2}} = +\infty,$$
$$\lim_{w\to \infty} \frac{1}{w\sqrt{2\pi}} e^{-\frac{w^2}{2}} = 0^+.$$

Then we check the first derivative of the difference between the two functions:

$$\begin{aligned}\frac{d}{dw}\left(\frac{1}{w\sqrt{2\pi}} e^{-\frac{w^2}{2}} - 1 + \Phi(w)\right) &= \frac{-w^2 e^{-\frac{w^2}{2}} - e^{-\frac{w^2}{2}}}{w^2\sqrt{2\pi}} + \frac{1}{\sqrt{2\pi}} e^{-\frac{w^2}{2}} \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{w^2}{2}} \left(1 - \frac{w^2+1}{w^2}\right) \\ &= -\frac{1}{w^2\sqrt{2\pi}} e^{-\frac{w^2}{2}},\end{aligned}$$

which is always strictly negative for $w > 0$ and tends to 0 as $w \to \infty$. By this last statement, the fact both functions tend to zero from above as $w \to \infty$ and considering both are always positive in $\mathbb{R}^+$, we conclude the inequality (1.16) is true $\forall w > 0$. Recalling $\|\Phi\| = 1$ and that $\Phi(w)[1-\Phi(w)]$ is even, we can deduce $\Phi(w)[1-\Phi(w)] \leq (|w|\sqrt{2\pi})^{-1} e^{-\frac{w^2}{2}}$ for every $w \neq 0$ (and it is 0.25 at $w = 0$).

At last, using all of the results we gathered until now, we check if f'_x is bounded:

$$\begin{aligned}\|f'_x\|_\infty &= \max\left\{\sup_{w\neq 0} |w f_x(w) + \mathbb{1}_{w\leq x} - \Phi(x)|, |f'_x(0)|\right\} \\ &\leq \max\left\{\sup_{w\neq 0} |w\sqrt{2\pi} e^{\frac{w^2}{2}} \Phi(w)[1-\Phi(w)] + \mathbb{1}_{w\leq x} - \Phi(x)|, |\mathbb{1}_{0\leq x} - \Phi(x)|\right\} \\ &\leq \max\left\{\sup_{w\neq 0} |\cancel{w\sqrt{2\pi}} \cancel{e^{\frac{w^2}{2}}} \frac{1}{|w|\cancel{\sqrt{2\pi}}} \cancel{e^{-\frac{w^2}{2}}} + \mathbb{1}_{w\leq x} - \Phi(x)|, 1 - \Phi(x)\right\} \\ &\leq 2.\end{aligned}$$

The first step follows directly from (1.8), while the second and third rows are due to (1.12) and (1.16).

With the result above we showed $\mathbb{E}|f'_x(W)| \leq 2$, and we are ready to prove the lemma.

Proof of Lemma 1 (2). Assume W is a random variable such that $\mathbb{E}[f'(W) - W f(W)] = 0$ for all bounded, continuous, and piecewise continuously differentiable functions f with $\mathbb{E}|f'(W)| < \infty$. f_x, the solution of the differential equation (1.8), is such a function, so

that for all $x \in \mathbb{R}$,

$$P(W \leq x) - \Phi(x) = \mathbb{E}[f_x'(W) - W f_x(W)] = 0,$$

and this implies W has a standard normal distribution. □

1.2.1 A generalisation of the results in this section

What we developed until now is based on the Kolmogorov metric but, as I said before, the Wasserstein metric will turn out to be more useful when working with continuous distributions.

Let X, Y be random variables with continuous distributions, $\mathcal{H}$ a family of functions, and recall the metric

$$d_H(X, Y) = \sup_{h \in \mathcal{H}} |\mathbb{E}h(X) - \mathbb{E}h(Y)| \tag{1.17}$$

we defined at the beginning of this chapter. Abusing notation, let $\Phi(h) := \mathbb{E}h(Z)$, where Z is a standard normal random variable. For $h \in \mathcal{H}$ let f_h be a function which solve this generalisation of (1.8):

$$f_h'(w) - w f_h(w) = h(w) - \Phi(h).$$

If $h \in \mathcal{H}_K$ we clearly get (1.8) again.

Proposition 4. *Let W a random variable and Z a random variable with the standard normal distribution, then*

$$d_{\mathcal{H}}(W, Z) = \sup_{h \in \mathcal{H}} |\mathbb{E}[f_h'(W) - W f_h(W)]|. \tag{1.18}$$

Proof. This follows directly from the definition of f_h. □

The next lemma contains a generalisation of what

Lemma 2. *Let f_h be the solution of the differential equation*

$$f_h'(w) - w f_h(w) = h(w) - \Phi(h) \tag{1.19}$$

which is given by

$$\begin{aligned} f_h(w) &= e^{\frac{w^2}{2}} \int_w^\infty e^{-\frac{t^2}{2}} [\Phi(h) - h(t)]\, dt \\ &= -e^{\frac{w^2}{2}} \int_{-\infty}^w e^{-\frac{t^2}{2}} [\Phi(h) - h(t)]\, dt. \end{aligned}$$

1. If h is bounded, then

$$\|f_h\|_\infty \leq \sqrt{\frac{\pi}{2}}\|h(\cdot) - \Phi(h)\|_\infty, \quad \textit{and} \quad \|f_h'\|_\infty \leq 2\|h(\cdot) - \Phi(h)\|_\infty.$$

2. If h is piecewise differentiable, then

$$\|f_h\|_\infty \leq 2\|h'\|_\infty, \quad \|f_h'\|_\infty \leq \sqrt{\frac{2}{\pi}}\|h'\|_\infty, \quad \textit{and} \quad \|f_h''\|_\infty \leq 2\|h'\|_\infty.$$

We refer to Chen, Goldstein, and Shao (2010) (Lemma 2.4) for the proof.

1.3 Adopting a different metric

We will now use the theory in 1.2.1 to switch from the Kolmogorov metric to the Wasserstein metric, when working with continuous distributions. Proposition 3 supports this idea: if W, Z are two random variables and Z has a standard normal distribution, calling its density ϕ we know $\phi(0) = 1/\sqrt{2\pi}$ and that it is the global maximum of ϕ, therefore ϕ is bounded by $C = 1/\sqrt{2\pi}$. Then $\sqrt{2C} = \sqrt[4]{2/\pi}$ and

$$d_K(W, Z) \leq \sqrt[4]{\frac{2}{\pi}}\sqrt{d_W(W, Z)}$$

where d_K, the distance between distributions in the Kolmogorov metric, is the maximum difference between distribution functions (as seen in item 1 of section 1.1).

The main reason to prefer the Wasserstein metric over the Kolmogorov metric is that—by definition—if $h \in \mathcal{H}_W$ and $x, y \in \mathbb{R}, x \neq y$ then (item 2, p.7)

$$|h(x) - h(y)| \leq |x - y| \implies \frac{|h(x) - h(y)|}{|x - y|} \leq 1.$$

Taking the limit as $y \to x$, we have

$$\lim_{y \to x}\left|\frac{h(x) - h(y)}{x - y}\right| \leq 1 \implies |h'(x)| \leq 1,$$

in particular h has Lipschitz constant less than or equal to one and $\|h'\|_\infty \leq 1$, so that by Lemma 2 item 2 f_h is a solution for (1.19) which is bounded with two bounded derivatives. Also, by the same Lemma, an h in the Kolmogorov family of functions gives a solution f_h bounded with only one bounded derivative (item 1), as $\forall x \in \mathbb{R}, h_x(w) = \mathbb{1}_{w \leq x}$ is bounded

but not piecewise differentiable.

Finally, we summarise everything achieved in the last sections in this theorem.

Theorem 3. *Let W a random variable and Z a standard normal random variable,*

$$d_W(W,Z) \leq \sup_{f\in\mathcal{F}} |\mathbb{E}[\mathcal{A}f(W)]|, \tag{1.20}$$

where the family of functions $\mathcal{F} := \left\{ f : \|f\|_\infty, \|f''\|_\infty \leq 2, \|f'\|_\infty \leq \sqrt{\frac{2}{\pi}} \right\}$.

Proof. Since $h \in \mathcal{H}_W \implies h$ piecewise differentiable, $\|h'\|_\infty \leq 1$, by Lemma 2

$$\|f_h\|_\infty, \|f_h''\|_\infty \leq 2, \|f_h'\|_\infty \leq \sqrt{\frac{2}{\pi}},$$

then $\mathcal{H}_W \subseteq \mathcal{F}$ and by Proposition 4 $d_W(W,Z) = \sup_{h\in\mathcal{H}} |\mathcal{A}f_h| \leq \sup_{f\in\mathcal{F}} |\mathcal{A}f|$. □

1.4 Bounding the error in the Central Limit Theorem

Finally, we can give bounds for the error in the Central Limit Theorem. Retaining the notation in the theorem, let $X_1, \ldots, X_n$ be independent random variables (it is **not** necessary to assume identical distributions) with $\mathbb{E}X_i = 0, \mathrm{Var}(X_i) = \mathbb{E}X_i^2 = 1, \mathbb{E}|X_i|^4 < \infty$. What we would like is to use what we discovered in the last chapters to quantify the convergence rate to the normal distribution for the sample mean $\overline{X}_n$ and find error bounds for the approximation, or in short we want a way to use Theorem 3 on it. Before writing anything we remind—by the Central Limit Theorem—$\overline{X}_n$ is approximately distributed as a $N\big(\mathbb{E}[X_i], \mathrm{Var}[X_i]\frac{1}{\sqrt{n}}\big)$, thus $W = (\sqrt{n})\overline{X}_n$ converges to a standard normal and we can compute the distance from the standard normal distribution. $Z \sim N(0,1)$, as always.

By Theorem 3 we have

$$d_W(W,Z) \leq \sup_{f\in\mathcal{F}} |\mathbb{E}[f'(W) - Wf(W)]|,$$

$\mathcal{F}$ a family of functions as defined in the same theorem. We now aim to reformulate the expression so that it depends on minimal information about the X_i. Moreover, let

$W_i := \frac{1}{\sqrt{n}} \sum_{j \neq i} X_j = W - \frac{1}{\sqrt{n}} X_i$.

$$\begin{aligned}
\mathbb{E}[Wf(W)] &= \mathbb{E}\left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i f(W)\right] \\
&= \mathbb{E}\left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \Big(f(W) - f(W_i) - (W - W_i) f'(W) \right. \\
&\qquad \left. + f(W_i) + (W - W_i) f'(W) \Big)\right] \\
&= \mathbb{E}\left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \Big(f(W) - f(W_i) - (W - W_i) f'(W) \Big)\right] + \qquad (1.21) \\
&\quad + \mathbb{E}\left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i (W - W_i) f'(W)\right] \qquad (1.22)
\end{aligned}$$

where in (1.22) we have $\mathbb{E}[X_i f(W_i)] = 0$, since each X_i is independent of all $X_j : j \neq i$ and so X_i, W_i are independent $\forall i$. Next, by the Taylor expansion of $f(W)$ at W_i with the remainder in the Lagrange form, we get

$$f(W) = f(W_i) + (W - W_i) f'(W_i) + \frac{1}{2}(W - W_i)^2 f''(\xi),$$

where $\xi \in \mathbb{R}$ is in the closed interval between W and W_i. Then, considering $W - W_i = X_i \sqrt{1/n}$, it follows (1.21) is less or equal than

$$\mathbb{E}\left[\frac{1}{2\sqrt{n}} \sum_{i=1}^{n} X_i (W - W_i)^2 f''(\xi)\right] = \frac{1}{2\sqrt{n^3}} \sum_{i=1}^{n} \mathbb{E}[X_i^3 \cdot f''(\xi)].$$

By the triangle inequality, pushing the absolute value inside the expectation, and considering $|f''(\xi)| \leq \|f''\|_\infty$ by definition of the infinity norm:

$$|\mathbb{E}[f'(W) - Wf(W)]| \leq \frac{\|f''\|_\infty}{2n^{3/2}} \sum_{i=1}^{n} \mathbb{E}|X_i|^3 + \mathbb{E}\left| f'(W) \left(1 - \frac{1}{n} \sum X_i^2\right)\right|,$$

which follows by taking the term $f'(W)$ in $\mathbb{E}\mathcal{A}f(W)$ and the result in (1.22). For the second expectation in the expression above we get, by considering $|\cdot| = \sqrt{\cdot^2}$,

$$\begin{aligned}
\mathbb{E}\left| f'(W) \left(1 - \frac{1}{n} \sum X_i^2\right)\right| &= \mathbb{E}\left| f'(W) \frac{1}{n} \left(n - \sum X_i^2\right)\right| \\
&\leq \frac{\|f'\|_\infty}{n} \mathbb{E}\left| n - \sum X_i^2 \right|,
\end{aligned}$$

$$\frac{\|f'\|_\infty}{n}\,\mathbb{E}\left|n-\sum X_i^2\right| = \frac{\|f'\|_\infty}{n}\,\mathbb{E}\sqrt{\Big(\sum 1-X_i^2\Big)^2}.$$

Recalling $\mathbb{E}X_i^2 = 1$, $\mathrm{Var}(X) = \mathbb{E}[(X-\mathbb{E}X)^2]$, Jensen's inequality and that the square root is a concave function:

$$\begin{aligned}\frac{\|f'\|_\infty}{n}\mathbb{E}\sqrt{\Big(\sum 1-X_i^2\Big)^2} &\le \frac{\|f'\|_\infty}{n}\sqrt{\mathbb{E}\left[\Big(\sum \mathbb{E}X_i^2-X_i^2\Big)^2\right]}\\ &= \frac{\|f'\|_\infty}{n}\sqrt{\mathbb{E}\left[\Big(\mathbb{E}\Big(\sum X_i^2\Big)-\sum X_i^2\Big)^2\right]}\\ &= \frac{\|f'\|_\infty}{n}\sqrt{\mathrm{Var}\Big(\sum X_i^2\Big)}.\end{aligned}$$

Since the X_i are independent, and $\mathrm{Var}(X_i^2) = \mathbb{E}X_i^4 - \mathbb{E}\big[X_i^2\big]^2 = \mathbb{E}X_i^4 - 1$:

$$\mathrm{Var}\Big(\sum X_i^2\Big) = \sum \mathrm{Var}X_i^2 \le \sum \mathbb{E}X_i^4.$$

Summarising everything, and substituting function norms with the corresponding boundaries in Theorem 3, we obtain the following theorem.

Theorem 4. *Let $X_1,\dots,X_n$ be independent random variables with $\mathbb{E}|X_i|^4 < \infty$, $\mathbb{E}X_i = 0$, and $\mathbb{E}X_i^2 = 1$. If $W = \frac{1}{\sqrt{n}}\sum_{i=1}^n X_i$ and Z has the standard normal distribution, then*

$$d_W(W,Z) \le \frac{1}{n^{\frac{3}{2}}}\sum_{i=1}^n \mathbb{E}|X_i|^3 + \frac{1}{n}\sqrt{\frac{2}{\pi}\sum_{i=1}^n \mathbb{E}[X_i^4]}.$$

We'll discuss this theorem in the next chapter.

Chapter 2

Applications and extensions

This chapter is thought to give a little hint to the reader as to how versatile Theorem 4 can be in a lot of applications, as well as showing some extensions to the theorem. We introduce it by discussing about this theorem.

Theorem 4 gives boundaries to the Wasserstein distance of a sum of independent random variables from the standard normal, in a beautiful way. An interesting detail is that, even though the Central Limit Theorem requires identical distribution in the sample, this assumption is dropped with this theorem, making it more powerful in a range of different contexts. This however, comes at the cost of having to know both $\mathbb{E}|X_i|^3$ and $\mathbb{E}[X_i^4]$, for every variable X_i in the sample, or at least an upper bound for these expectations. One usually doesn't have prior knowledge on this type of informations. Moreover, the theorem assumes the mean of every X_i is 0 and the variance is 1, but this can be easily extended, and we'll see how.

There is one thing to say before reading further, namely that the main objective of these theorems is to give a quantitative measure of the 'effectiveness' of the Central Limit Theorem, before everything else. Looking at Theorem 4 with this perspective, one may notice that the convergence rate is of order $n^{-\frac{1}{2}}$ if the random variables are identically distributed, which is the best possible.

We first look at direct generalisations of the last theorem, relaxing different assumptions.

2.1 Bounded random variables

If the X_i are bounded, it is possible to solve all of the unknown quantities and get an expression which depends only on n and this upper bound.

Let $X_1, \ldots, X_n$ be random variables with $\mathbb{E}[X_i] = 0, \mathrm{Var}(X_i) = 1$ and

$$|X_i| \leq M$$

$\forall i \in \{1, \ldots, n\}$. This implies $\mathbb{E}|X_i|^3 = M^3$ and $\mathbb{E}X_i^4 = M^4$, therefore by Theorem 4

$$\begin{aligned} d_W(W, Z) &\leq \frac{1}{n^{\frac{3}{2}}} n M^3 + \frac{1}{n}\sqrt{\frac{2}{\pi}}\sqrt{nM^4} \\ &= \frac{M^3}{\sqrt{n}} + \frac{M^2}{\sqrt{n}}\sqrt{\frac{2}{\pi}} \\ &= \frac{1}{\sqrt{n}}\left(M^3 + M^2\sqrt{\frac{2}{\pi}}\right). \end{aligned} \tag{2.1}$$

2.1.1 Example: Uniform distributions with unknown extremes

Let $U_1, \ldots, U_n$ be random variables with continuous uniform distribution and zero mean. In short $U_i \sim U(-a_i, a_i)$, $a_i > 0$. Then $|X_i| \leq a_i$, $\mathrm{Var}(U_i) = \frac{1}{3}a^2$, and letting $X_i = \frac{\sqrt{3}}{a_i}U_i$ we can apply Theorem 4 on the X_i:

$$d_W(W, Z) \leq \frac{1}{\sqrt{n}}\left(\sqrt{3^3} + 3\sqrt{\frac{2}{\pi}}\right),$$

since $|X_i| = \frac{\sqrt{3}}{a_i}|U_i| \leq \frac{\sqrt{3}}{a_i}a_i = \sqrt{3}$. Computing the actual values we get

$$d_W(W, Z) \leq \frac{7.6}{\sqrt{n}}.$$

Now instead, let $V_i \sim U(b_i, c_i)$, $b_i < c_i$. $Y_i = (V_i - \frac{b_i + c_i}{2})\frac{\sqrt{12}}{c_i - b_i}$ and we can apply Theorem 4 on the Y_i the usual way. The Y_i are centred at zero, therefore we get the same result as above.

2.2 Unknown means and common variance

As we will see, not knowing the expectations of the variables does not significantly change the boundaries, and preserves the convergence rate order. On the other hand, the order will change according to the variance of the variables. We assume common variance for simplicity, and to avoid needing prior knowledge on other informations about the variables.

First we briefly introduce this elegant lemma.

Lemma 3. *Let $a, b \in \mathbb{R}$,*

$$(a+b)^n \leq 2^{n-1}(a^n + b^n).$$

Let $Y_1, \ldots, Y_n$ be independent random variables with $\mathbb{E}|Y_i|^4 < \infty$ and $\mathrm{Var}(Y_i) = \sigma^2$. Let

$$X_i = \frac{Y_i - \mathbb{E}Y_i}{\sigma}, \quad \forall i \in \{1, \ldots, n\},$$

then $\mathbb{E}X_i = 0$, $\mathrm{Var}(X_i) = 1$ and we can apply Theorem 4 on $X_1, \ldots, X_n$ and $W = (\sqrt{n})\overline{X}_n$ to get

$$d_W(W, Z) \leq \frac{1}{n^{\frac{3}{2}}} \sum_{i=1}^{n} \mathbb{E}\left|\frac{Y_i - \mathbb{E}Y_i}{\sigma}\right|^3 + \frac{1}{n}\sqrt{\frac{2}{\pi}}\sqrt{\sum_{i=1}^{n} \mathbb{E}\left[\left(\frac{Y_i - \mathbb{E}Y_i}{\sigma}\right)^4\right]}.$$

To write this expression in a simpler form, we see another lemma which follows directly from Lemma 3.

Lemma 4. *Let Y be a random variable and $n \geq 2$, then*

$$\mathbb{E}|Y - \mathbb{E}Y|^n \leq 2^n \mathbb{E}|Y|^n$$

Proof. $x \mapsto |x|^n$ is convex and by Jensen's inequality, the triangle inequality and Lemma 3:

$$\begin{aligned}
\mathbb{E}|Y_i - \mathbb{E}Y_i|^n &\leq \mathbb{E}[(|Y_i| + |\mathbb{E}Y_i|)^n] \\
&\leq 2^{n-1}(\mathbb{E}|Y_i|^n + |\mathbb{E}Y_i|^n) \\
&\leq 2^n \mathbb{E}|Y_i|^n.
\end{aligned}$$

□

By Lemma 4, and remembering $\sigma > 0$,

$$\mathbb{E}\left|\frac{Y_i - \mathbb{E}Y_i}{\sigma}\right|^3 \leq \left(\frac{2}{\sigma}\right)^3 \mathbb{E}|Y_i|^3$$

and

$$\mathbb{E}\left[\left(\frac{Y_i - \mathbb{E}Y_i}{\sigma}\right)^4\right] \leq \left(\frac{2}{\sigma}\right)^4 \mathbb{E}Y_i^4,$$

obtaining

$$d_W(W, Z) \leq \frac{8}{(\sigma\sqrt{n})^3} \sum_{i=1}^{n} \mathbb{E}|Y_i|^3 + \frac{4}{\sigma^2 n}\sqrt{\frac{2}{\pi}}\sqrt{\sum_{i=1}^{n} \mathbb{E}[Y_i^4]}.$$

This gives an idea of how large n needs to be to have a good approximation with the Central Limit Theorem on known distributions.

2.2.1 Example: Bernoulli variables

Let $Y_1, \dots, Y_n$ be a sample of random variables, with $Y_i \sim Ber(p)$ or, equivalently, $Bin(1, p)$. Then

$$d_W(W, Z) \leq \frac{8\mathbb{E}|Y_1|^3}{\sqrt{n}(p(1-p))^{\frac{3}{2}}} + \frac{4\sqrt{\mathbb{E}[Y_1^4]}}{\sqrt{n}p(1-p)}\sqrt{\frac{2}{\pi}},$$

and knowing $|Y_1| \leq 1$ we can simplify the two expected values in the expression. Moreover, since the k-th moment of a Bernoulli is known, it is possible to simplify this expression even more so it depends only on n and p. Another possible solution would have required noticing that $Y_i - p$ has zero mean and is bounded by $1 - p$.

2.3 Dependency neighbourhoods

In this section we will see a generalisation of Theorem 4 which assumes local dependence between the variables. We look now at the definition of a dependency neighbourhood.

Definition 1. *In a collection of random variables $(X_1, \dots, X_n)$, the **dependency neighbourhood** of X_i is a set $N_i \subseteq \{1, \dots, n\}$ such that*

$$X_i \text{ is independent of } \{X_j\}_{j \notin N_i}$$

For Stein's method applied to sums of random variables with this dependency structure, it is possible to rewrite the expression in a similar way to what we have done with the sums of independent variables, applying some modification to the argument. After that, however, some more steps are necessary to bound the quantity. The interesting part here is that this dependency structure can be viewed as a graph, where every vertex has a 1:1 correspondence with a variable, and two vertex are linked by an edge only if the corresponding variables are dependent on each other, or better if they are both present in their dependency neighbourhoods (clearly every variable is dependent on itself, so $|N_i| \geq 1$ for every i).

Theorem 5. *Let $X_1, X_2, \dots, X_n$ be random variables with dependency neighbourhoods $N_1, N_2, \dots, N_n$, $\mathbb{E}[X_i] = 0$ and $\mathbb{E}[X_i^4] < \infty$. Let $W = \sigma^{-1}\sum_i X_i$, $D = \max|N_i|$ and $\sigma^2 = Var(\sum_i X_i)$. Then*

$$d_W(W, Z) \leq \frac{D^2}{\sigma^3}\sum_{i=1}^{n}\mathbb{E}|X_i|^3 + \frac{D^{\frac{3}{2}}}{\sigma^2}\sqrt{\frac{28}{\pi}\sum_{i=1}^{n}\mathbb{E}[X_i^4]}. \tag{2.2}$$

Proof. Here I show a sketch of the proof, to give an idea of how it is done. The complete version is in Ross (2011).

The key here is constructing W_i independent of X_i, just like in the proof of Theorem 4, and $W_i = \sigma^{-1} \sum_{j \notin N_i} X_j$ is exactly what we want. Theorem 3 is still valid, therefore we consider f as defined in that theorem.

The beginning of the proof is identical:

$$|\mathbb{E}[f'(W) - Wf(W)] \leq \left| \mathbb{E}\left[\frac{1}{\sigma} \sum_{i=1}^{n} X_i \Big(f(W) - f(W_i) - (W - W_i) f'(W_i) \Big) \right] \right| \tag{2.3}$$

$$+ \left| \mathbb{E}\left[f'(W) \left(1 - \frac{1}{\sigma} \sum_{i=1}^{n} X_i (W - W_i) \right) \right] \right|. \tag{2.4}$$

By Taylor expansion, the triangle inequality, and after pushing the absolute value inside the expectation, we find out (2.3) is upper bounded by

$$\frac{\|f''\|_\infty}{2\sigma} \mathbb{E} \left| \sum_{i=1}^{n} X_i (W - W_i)^2 \right| \leq \frac{2}{2\sigma^3} \mathbb{E} \left| \sum_{i=1}^{n} X_i \left(\sum_{j \in N_i} X_j \right)^2 \right|$$

$$\leq \frac{1}{\sigma^3} \sum_{i=1}^{n} \sum_{j,k \in N_i} \mathbb{E}|X_i X_j X_k|. \tag{2.5}$$

By the arithmetic-mean inequality

$$\mathbb{E}|X_i X_j X_k| \leq \frac{1}{3}(\mathbb{E}|X_i|^3 + \mathbb{E}|X_j|^3 + \mathbb{E}|X_k|^3),$$

and therefore (2.5) is bounded above by

$$\frac{1}{\sigma^3} \sum_{i=1}^{n} \sum_{j,k \in N_i} \frac{1}{3}(\mathbb{E}|X_i|^3 + \mathbb{E}|X_j|^3 + \mathbb{E}|X_k|^3) \leq \frac{D^2}{\sigma^3} \sum_{i=1}^{n} \mathbb{E}|X_i|^3,$$

where we used the fact that, since every vertex in the dependency graph doesn't have more than D neighbours and every vertex cannot be present more than once in a neighbourhood, then for every variable X_i there exist maximum D neighbourhoods N_j such that $X_i \in N_j$. This should be straightforward, X_i having D neighbours implies X_i is the neighbour of itself and $D - 1$ other variables. Then, summing over all the variables in the N_i, every

variable cannot be summed more than D times. It follows that

$$\sum_{i=1}^{n}\sum_{j\in N_i}\mathbb{E}|X_j| \le \sum_{i=1}^{n}(D\cdot\mathbb{E}|X_i|)$$

and the first addend of (2.2) is proved.

The second part is more complex and we won't see every step. We first push the absolute value inside the expectation, as always, secondly we explicit $W - W_i$ as we've already done above, and finally we factor out σ^{-2} from the difference to get the left-hand side of (2.6). Then consider rewriting σ^2:

$$\sigma^2 = \mathrm{Var}\left(\sum_i X_i\right) = \mathbb{E}\left[\left(\sum_i X_i\right)^2\right] - \cancel{\left(\sum_i \mathbb{E}[X_i]\right)^2} = \mathbb{E}\left[\sum_{i=1}^{n} X_i \sum_{j\in N_i} X_j\right].$$

where we implicitly excluded every term with $\mathbb{E}[X_iX_j] = 0$ in the last step, which happens only if $X_j \notin N_i$ and conversely $X_i \notin N_j$.

After all of this, we find (2.4) is bounded above by

$$\frac{\|f'\|}{\sigma^2}\mathbb{E}\left|\sigma^2 - \sum_{i=1}^{n} X_i\sum_{j\in N_i} X_j\right| \le \frac{1}{\sigma^2}\sqrt{\frac{2}{\pi}\mathrm{Var}\left(\sum_{i=1}^{n}\sum_{j\in N_i} X_iX_j\right)}. \tag{2.6}$$

The right-hand side follows also from $|\cdot| = \sqrt{\cdot^2}$ and Jensen's inequality.

The remainder of the proof consists of analysis on the right-hand side of (2.6), and it is omitted here. Again, I refer to Ross (2011) for the full explanation. Anyway, at the end we obtain

$$\begin{aligned}\mathrm{Var}\left(\sum_{i=1}^{n}\sum_{j\in N_i} X_iX_j\right) &= \mathbb{E}\left[\left(\sum_{i=1}^{n}\sum_{j\in N_i} X_iX_j\right)^2\right] - \sigma^4 \\ &\le (12D^3 + 2D^2)\sum_{i=1}^{n}\mathbb{E}[X_i^4] \le 14D^3\sum_{i=1}^{n}\mathbb{E}[X_i^4],\end{aligned}$$

which yields the theorem. □

2.3.1 Application: Triangles in Erdős–Rényi random graphs

An Erdős–Rényi random graph is a graph $G = G(n,p)$ with n nodes where the presence of an edge between two vertices is determined by a random variable B_i with the distribution of a $Ber(p)$. Intuitively, the sum of these variables is $S \sim Bin(n,p)$ since the result of a

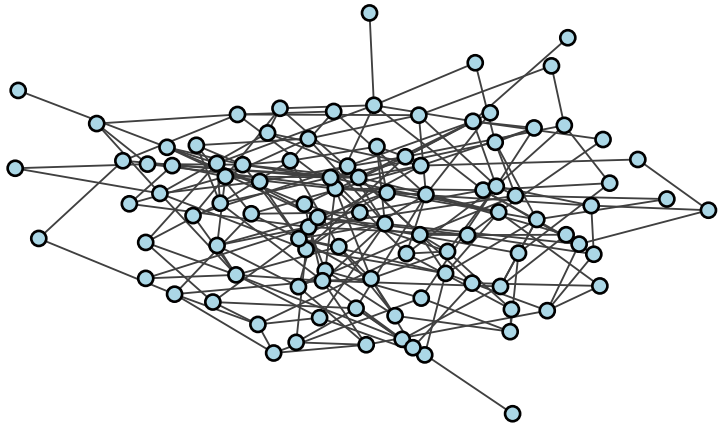


Figure 2.1: An example of an Erdős–Rényi random graph with 100 vertices and 236 edges. Despite looking rather intricate, the number of triangles amounts to 19 only, and $p = n^{-\alpha}$ with $\alpha = 2/3$ is approximately 0.05. The expected number of triangles with this configuration is about 16.

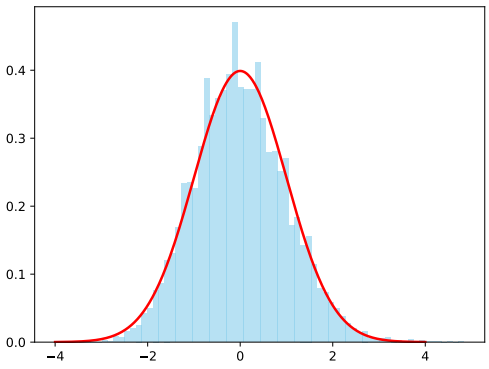
B_i is not influenced by the others, and so $\mathbb{E}[S] = np$ with $\mathrm{Var}(S) = np(1-p)$. Counting edges in a Erdős–Rényi graph is pretty straightforward, and instead in this section we will count triangles. A triangle is just a connected subgraph of G with three nodes and three edges or, in graph theory terminology, a K_3 subgraph of G.

If G is complete, or $G \simeq K_n$, the number of triangles is every possible unordered choice of three vertices, so $\binom{n}{3}$. For conciseness let $N := \binom{n}{3}$. Then, considering all the triangles in K_n in a fixed order, let Y_i be the indicator that the corresponding i-th triangle is present in G, or equivalently that those three edges are in $E(G)$, which is the set of all the edges in G. Then $Y_i = B_a B_b B_c$, for $i, a, b, c \in \{1, \dots, n\}$. Let $T := \sum_{i=1}^{N} Y_i$ the number of triangles in G.

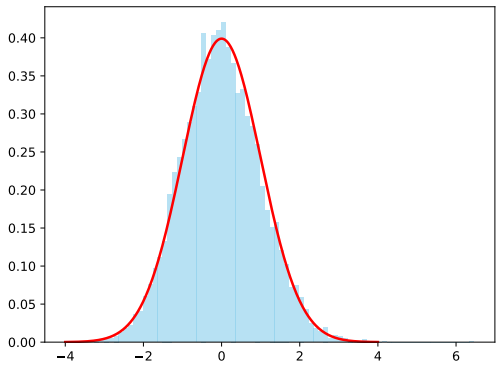
For $i \neq j$, Y_i, Y_j are independent if and only if they don't share any edge, or equivalently they don't share more than one vertex, therefore the size of the neighbourhood of a Y_i is the number of all triangles which can be constructed with two nodes (or one edge!) of the i-th triangle and one outside of it, plus the i-th triangle itself. Then $|N_i| = 3(n-3)+1$ and we can apply Theorem 5 with $X_i = Y_i - p^3$ and $D = 3n - 8$. Since

$$\mathbb{E}|X_i|^k = p^3(1-p^3)[(1-p^3)^{k-1} + p^{3(k-1)}], \quad k = 1, 2, \dots$$

we now only have to compute $\mathrm{Var}(T)$ to apply the theorem. $\mathrm{Var}(T) = \mathrm{Var}(\sum X_i) = \mathrm{Var}(\sum X_i)$ can be rewritten to be the sum of all N variances $p^3(1-p)^3$, plus two times



(a) $\alpha = 4/9,\ p \simeq 0.129$



(b) $\alpha = 2/9,\ p \simeq 0.359$

Figure 2.2: Histograms of Monte Carlo simulations for the standardized triangles count in Erdős–Rényi random graphs with $n = 100$ and $p = n^{-\alpha}$, 10000 iterations. The red curve is the standard normal density.

every covariance. If $j \in N_i$, then

$$\mathrm{Cov}(Y_i, Y_j) = \mathbb{E}[Y_i Y_j] - \mathbb{E}[Y_i]\mathbb{E}[Y_j] = \mathbb{E}[B_a B_b B_c B_d B_e] - p^6 = p^5 - p^6$$

for $Y_i = B_a B_b B_c,\ Y_j = B_c B_d B_e,\ a, b, c, d, e \in \{1, \dots, n\}$, otherwise the covariance is zero. We used the fact $B_c^2 = B_c$. Every Y_i is dependent on $D - 1$ variables, therefore

$$\sigma^2 := \mathrm{Var}(T) = Np^3(1-p^3) + N(D-1)(p^5 - p^6) = \binom{n}{3} p^3 \left[1 - p^3 + 3(n-3)p^2(1-p)\right],$$

and Theorem 5 implies that for $W = (T - \mathbb{E}[T])\sigma^{-1}$ and Z a standard normal random variable

$$\begin{aligned} d_W(W, Z) \leq{}& \frac{(3n-8)}{\sigma^3}\binom{n}{3} p^3(1-p^3)[(1-p^3)^2 + p^6] \\ &+ \frac{(3n-8)^{\frac{3}{2}}}{\sigma^2}\sqrt{\frac{28}{\pi}\binom{n}{3} p^3(1-p^3)[(1-p^3)^3 + p^9]}. \end{aligned} \tag{2.7}$$

Ross (2011) states this bound holds for all $n \geq 3$ and $0 \leq p \leq 1$, however analysing the asymptotic behaviour of it if, for example, p is of order $n^{-\alpha}$ for some $0 \leq \alpha < 1$ (so that $\mathrm{Var}(T) \to \infty$), shows that the number of triangles satisfies a CLT for $0 \leq \alpha < \frac{2}{9}$.

Simulations for the distribution of T with different choices for α can be seen in Figure 2.2. Moreover, in is found the Python code to produce an Erdős–Rényi graph like the one in Figure 2.1 and the simulations in Figure 2.2.

Appendix A

Simulations on Erdős–Rényi random graphs

In this appendix I wrote down the Python code I used to generate graphs and simulate distributions. The main external library imported is networkx, which was used to speed up the coding for the generation and visualisation of graphs, and besides that it also makes the code here easier to read. All of this could be made without that, the only tricky part is counting triangles, however even that is not too difficult. The other external library used are mainly for plotting.

A.1 Generating an Erdős–Rényi graph

```
import networkx as nx

n = int(input('Enter the number of vertices: '))
p = float(input('Enter the probability of edge creation: '))
G = nx.erdos_renyi_graph(n, p)
triangles = sum(nx.triangles(G).values()) // 3
print(G)
print('Number of triangles in the graph:', triangles)
```

```
Enter the number of vertices: 20
Enter the probability of edge creation: 0.5
Graph with 20 nodes and 94 edges
Number of triangles in the graph: 132
```

`nx.triangles(G)` returns a Python dictionary with the number of triangles intersecting a particular vertex for every vertex in G. Therefore, after extracting the values from the dictionary and summing over them, we divide the result by 3 to get the actual total number of triangles in the graph. The use of the double dash to divide by 3 is to integer divide instead of using the normal division, which would have returned a float number instead of an int (not necessary at all, normal division would have worked fine, with the only downside of having a number of triangles ending with '.0').

We can also print some more info, using the mathematical expressions found in subsection 2.3.1.

```
from math import sqrt, pi

N = (n * (n-1) * (n-2)) // 6
D = 3*n - 8
meanT = N * p**3
varT = N * p**3 * (1 - p**3 + 3*(n-3) * p**2 * (1 - p))
dW = (D**2 / (sqrt(varT)**3) * N * p**3 * (1-p**3)*((1-p**3)**2 + p**6)
           + sqrt(28)*D**1.5 / (sqrt(pi) * varT)
           * sqrt(N * p**3 * (1-p**3) * ((1 - p**3)**3 + p**9)))

print('Number of all possible triangles:', N)
print('Expected number of triangles:', ET)
print('Variance of T:', varT)
print('Wasserstein distance bound:', dW)
```

```
Number of all possible triangles: 1140
Expected number of triangles: 142.5
Variance of T: 1033.125
Wasserstein distance bound: 17.8498610647571
```

Finally, we use matplotlib to display the graph in a new window.

```
import matplotlib.pyplot as plt

plt.title(f"Erdős–Rényi Graph G({n}, {p}),"
    +f"number of triangles: {triangles}")
nx.draw(G)
plt.show()
```

It is possible to customise the displayed graph in a great variety of ways, but here I decided to write just a few lines of code sufficient to get what we wanted.

A.2 Generating simulated distributions through Monte Carlo

```
def random_T(n, p, sample_size=1000):
    N = (n * (n - 1) * (n - 2)) // 6
    meanT = N * p**3
    stdT = sqrt(N * p**3 * (1 - p**3 + 3*(n-3) * p**2 * (1 - p)))
    results = []
    for _ in range(sample_size):
        G = nx.erdos_renyi_graph(n, p)
        results.append((sum(nx.triangles(G).values())//3 - meanT) / stdT)
    return results
```

This function accepts a number of vertices, a probability and optionally the size of the sample as input parameters, and returns the standardised number of triangles in many Erdős–Rényi random graphs as the sample size (1000 by default), so basically it generates a sample from the distribution of T. These values can be plotted on a histogram to compare their empirical distribution to the standard normal density curve. Equivalently, it is possible to take the raw triangle count and then compare them to the density of a normal with $\binom{n}{3}p^3$ mean and variance as computed above.

To display the histogram with the standard normal density curve on top:

```
import numpy as np


plt.hist(random_T(n,p), density=True)
x = np.linspace(-4, 4, 500)
plt.plot(x, (1/sqrt(2*pi)) * np.exp(-(x**2) / 2))
plt.show()
```

where `np.linspace(-4, 4, 500)` returns an array of 500 numbers in the closed interval $(-4, 4)$ with a minimum distance from each other of $8/500$.

References

Chen, Louis HY, Larry Goldstein, and Qi-Man Shao (2010). *Normal approximation by Stein's method*. Springer Science & Business Media.

Klenke, Achim (2013). *Probability theory: a comprehensive course*. Springer Science & Business Media.

Ross, Nathan (2011). "Fundamentals of Stein's method". In: *Probability Surveys* 8, pp. 210–293. DOI: 10.1214/11-PS182. URL: https://doi.org/10.1214/11-PS182.

Stein, Charles (1972). "A bound for the error in the normal approximation to the distribution of a sum of dependent random variables". In: *Proceedings of the sixth Berkeley symposium on mathematical statistics and probability, volume 2: Probability theory*. Vol. 6. University of California Press, pp. 583–603.

— (1986). "Approximate computation of expectations". In: IMS.