A simple CLT: de Moivre-Laplace theorem. Application to statistics

The *Central Limit Theorem* (CLT) is one of the most important asymptotic results in probability theory. Although one often refers to "the" CLT, there are in fact several versions of the theorem, each stating – under various sets of assumptions – that the properly normalized average of a sample converges in distribution to a Gaussian random variable. Classical formulations, such as the *Lindeberg-Lévy*, *Lyapunov*, and *Lindeberg* versions of the CLT, are typically proved using characteristic functions.

In the special case of i.i.d. Bernoulli random variables, the CLT – known in this context as the *de Moivre-Laplace theorem* – can be proved using Stirling's formula and classical analysis. This is the main objective of the present problem.

It is worth noting that the CLT for i.i.d. Bernoulli random variables is often interpreted as a Gaussian approximation to the binomial distribution – a technique frequently used in statistical applications. In this problem, we also explore a practical application of this approximation in the context of airline overbooking.

Towards de Moivre-Laplace theorem: a local estimate

We consider in this problem a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and a sequence of i.i.d. Bernoulli $\mathcal{B}(p)$ random variables $(X_n)_{n\in\mathbb{N}^*}$ (with $p \in (0,1)$) defined on that probability space. We define, for all $n \in \mathbb{N}^*$, the random variable $S_n = X_1 + \ldots + X_n$.

1. Prove that $\forall k \in \{0, \ldots, n\}, \mathbb{P}(S_n = k) = C_n^k p^k (1-p)^{n-k}$.

We recall (see the problem on Stirling's formula) that there exists a positive constant A such that

$$\forall n \in \mathbb{N}^*, \left| \frac{n!}{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n} - 1 \right| \leq \frac{A}{n}.$$

For two real numbers $a < b$ we define the sets

$$I_n^{a,b} = \{k \in \{0, \ldots, n\} | np + a\sqrt{n} \leq k \leq np + b\sqrt{n}\}, \qquad n \in \mathbb{N}^*.$$

2. Prove that there exists $n^{a,b} \in \mathbb{N}^*$, such that $\forall n \geq n^{a,b}$:

$$n > A, \quad pn + a\sqrt{n} > A, \quad (1-p)n - b\sqrt{n} > A$$

$$-a\sqrt{n} < \frac{1}{2}np, \quad -a\sqrt{n} < \frac{1}{2}n(1-p), \quad b\sqrt{n} < \frac{1}{2}np, \quad b\sqrt{n} < \frac{1}{2}n(1-p).$$

3. Prove that $\forall n \geq n^{a,b}, \forall k \in I_n^{a,b}$,

$$\mathbb{P}(S_n = k) \geq \frac{1}{\sqrt{2\pi}}\sqrt{\frac{n}{k(n-k)}}\left(\frac{np}{k}\right)^k \left(\frac{n(1-p)}{n-k}\right)^{n-k} \frac{1 - \frac{A}{n}}{\left(1 + \frac{A}{np+a\sqrt{n}}\right)\left(1 + \frac{A}{n(1-p)-b\sqrt{n}}\right)}$$

and

$$\mathbb{P}(S_n = k) \leq \frac{1}{\sqrt{2\pi}}\sqrt{\frac{n}{k(n-k)}}\left(\frac{np}{k}\right)^k \left(\frac{n(1-p)}{n-k}\right)^{n-k} \frac{1 + \frac{A}{n}}{\left(1 - \frac{A}{np+a\sqrt{n}}\right)\left(1 - \frac{A}{n(1-p)-b\sqrt{n}}\right)}.$$

4. Prove that $\forall n \geq n^{a,b}, \forall k \in I_n^{a,b}$,

$$\sqrt{\frac{n}{k(n-k)}} \geq \frac{1}{\sqrt{np(1-p)}} \frac{1}{\sqrt{\left(1 + \frac{b}{p\sqrt{n}}\right)\left(1 - \frac{a}{(1-p)\sqrt{n}}\right)}}$$

and

$$\sqrt{\frac{n}{k(n-k)}} \leq \frac{1}{\sqrt{np(1-p)}} \frac{1}{\sqrt{\left(1+\frac{a}{p\sqrt{n}}\right)\left(1-\frac{b}{(1-p)\sqrt{n}}\right)}}.$$

5. Prove that there exists $B > 0$ such that

$$\forall x \in \left(-\frac{1}{2}, \frac{1}{2}\right), \left|\log(1+x) - x + \frac{x^2}{2}\right| \leq B|x|^3.$$

6. Deduce that there exists $C^{a,b} > 0$ such that $\forall n \geq n^{a,b}, \forall k \in I_n^{a,b}$,

$$\left|\log\left(\frac{k}{np}\right) - \frac{k-np}{np} + \frac{(k-np)^2}{2n^2p^2}\right| \leq \frac{C^{a,b}}{n^{\frac{3}{2}}}$$

and

$$\left|\log\left(\frac{n-k}{n(1-p)}\right) - \frac{np-k}{n(1-p)} + \frac{(k-np)^2}{2n^2(1-p)^2}\right| \leq \frac{C^{a,b}}{n^{\frac{3}{2}}}$$

7. Prove that $\forall n \geq n^{a,b}, \forall k \in I_n^{a,b}$,

$$k\frac{(k-np)^2}{2n^2p^2} + (n-k)\frac{(k-np)^2}{2n^2(1-p)^2} = \frac{(k-np)^2}{2np(1-p)} + \frac{(k-np)^3}{2n^2}\left(\frac{1}{p^2} - \frac{1}{(1-p)^2}\right).$$

8. Deduce that there exists $D^{a,b} > 0$ such that $\forall n \geq n^{a,b}, \forall k \in I_n^{a,b}$,

$$\left|k\log\left(\frac{np}{k}\right) + (n-k)\log\left(\frac{n(1-p)}{n-k}\right) + \frac{(k-np)^2}{2np(1-p)}\right| \leq \frac{D^{a,b}}{\sqrt{n}}.$$

9. Conclude that there exists a sequence $\left(\epsilon_n^{a,b}\right)_{n\in\mathbb{N}^*}$ converging towards 0 such that

$$\frac{1}{\sqrt{2\pi np(1-p)}} e^{-\frac{(k-np)^2}{2np(1-p)}}(1-\epsilon_n^{a,b}) \leq \mathbb{P}(S_n = k) \leq \frac{1}{\sqrt{2\pi np(1-p)}} e^{-\frac{(k-np)^2}{2np(1-p)}}(1+\epsilon_n^{a,b}).$$

de Moivre-Laplace theorem

Let $\alpha < \beta$ be two real numbers and let us choose $a = \alpha\sqrt{p(1-p)}$ and $b = \beta\sqrt{p(1-p)}$.

10. Prove that $\forall n \geq n^{a,b}$,

$$I_n^{a,b} = \{k \in \mathbb{Z} | np + a\sqrt{n} \leq k \leq np + b\sqrt{n}\}.$$

11. Deduce that $\forall n \geq n^{a,b}$

$$\mathbb{P}\left(\alpha \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq \beta\right) \geq \sum_{\substack{k\in\mathbb{Z} \\ np+a\sqrt{n}\leq k\leq np+b\sqrt{n}}} \frac{1}{\sqrt{2\pi np(1-p)}} e^{-\frac{(k-np)^2}{2np(1-p)}}(1-\epsilon_n^{a,b}).$$

and

$$\mathbb{P}\left(\alpha \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq \beta\right) \leq \sum_{\substack{k\in\mathbb{Z} \\ np+a\sqrt{n}\leq k\leq np+b\sqrt{n}}} \frac{1}{\sqrt{2\pi np(1-p)}} e^{-\frac{(k-np)^2}{2np(1-p)}}(1+\epsilon_n^{a,b}).$$

12. Prove that

$$\lim_{n\to+\infty} \sum_{\substack{k\in\mathbb{Z} \\ np+a\sqrt{n}\leq k\leq np+b\sqrt{n}}} \frac{1}{\sqrt{2\pi np(1-p)}} e^{-\frac{(k-np)^2}{2np(1-p)}} = \frac{1}{\sqrt{2\pi}} \int_\alpha^\beta e^{-\frac{t^2}{2}}\, dt.$$

Hint: Use the same idea as for proving the convergence of a Riemann sum.

13. Conclude that[1]

$$\lim_{n\to+\infty} \mathbb{P}\left(\alpha \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq \beta\right) = \frac{1}{\sqrt{2\pi}} \int_\alpha^\beta e^{-\frac{t^2}{2}}\, dt.$$

Application to overbooking

14. Explain why de Moivre-Laplace theorem justifies the approximation of a $\mathcal{B}(n,p)$ binomial random variable by a $\mathcal{N}(np, np(1-p))$ Gaussian random variable when n is large.[2]

A Paris-London flight is operated with a 150-seat aircraft. It is known that a person buying a ticket will show up at the airport with probability $p = 75\%$.

The company wants to sell more than 150 tickets. Let Z_n be the number of people showing up for the flight if n tickets are sold.

15. What is the distribution of Z_n?
16. Use the Gaussian approximation to the binomial to determine (approximately) the maximum number of tickets the company can sell to be sure (with 95% confidence level) that there will be enough seats in the plane for all the clients who show up (i.e. $\mathbb{P}(Z_n \leq 150) \geq 95\%$)?

[1] This is de Moivre-Laplace theorem.
[2] This approximation is usually referred to as the Gaussian approximation to the binomial.