Κεντρικό Οριακό Θεώρημα (Central Limit Theorem – Lindeberg Levy)

Εάν $X_i \overset{iid}{\sim} f(\cdot)$, $i \geq 1$ με $\mathbb{E}[X_i] = \mu$, $Var[X_i] = \sigma^2 < \infty$ και $S_n = \sum_{i=1}^{n} X_i$ τότε η τ.μ. $Z_n = \frac{S_n - \mathbb{E}(S_n)}{\sqrt{Var(S_n)}} = \frac{S_n - n\mu}{\sigma\sqrt{n}}$ συγκλίνει ως προς κατανομή (converges in distribution) στη $N(0,1)$, δηλαδή $Z_n \xrightarrow[n\to\infty]{d} N(0,1)$ ή ισοδύναμα $\lim_{n\to\infty} F_{Z_n}(x) = \lim_{n\to\infty} P\{Z_n \leq x\} = \Phi(x)$ με $\Phi(x) = \int_{-\infty}^{x} N(u \mid 0,1)\,du$.

Επειδή το Z_n αναπαρίσταται σαν $Z_n = \frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu)$ πρώτα θα υπολογίσουμε την ροπογεννήτρια της τ.μ. $Y = X_i - \mu$

$$M_Y(t) = \mathbb{E}\left[e^{t(X_i-\mu)}\right] = \mathbb{E}\left[1 + (X_i-\mu)t + \frac{1}{2}(X_i-\mu)^2 t^2 + O\left(t^3\right)\right]$$
$$= 1 + \mathbb{E}(X_i-\mu)t + \frac{1}{2}\mathbb{E}\left[(X_i-\mu)^2\right]t^2 + O\left(t^3\right) = 1 + \frac{1}{2}\sigma^2 t^2 + O\left(t^3\right)$$

Οπότε για την ροπογεννήτρια της Z_n θα έχουμε:

$$M_{Z_n}(t) = \mathbb{E}\left[\exp(tZ_n)\right] = \mathbb{E}\left[\exp\left(\frac{t}{\sigma\sqrt{n}}\sum_{i=1}^{n}(X_i-\mu)\right)\right] = \left[M_Y\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n$$
$$= \left\{1 + \frac{1}{2}\sigma^2\left(\frac{t}{\sigma\sqrt{n}}\right)^2 + O\left(\left(\frac{t}{\sigma\sqrt{n}}\right)^3\right)\right\}^n = \left\{1 + \frac{t^2/2}{n} + O\left(\frac{t^3}{n^{3/2}}\right)\right\}^n$$

$$\lim_{n\to\infty} M_{Z_n}(t) = \exp\left\{\lim_{x\to\infty} \frac{1}{x^{-1}} \log\left[1 + \frac{t^2/2}{x} + O\left(\frac{t^3}{x^{3/2}}\right)\right]\right\}$$
$$= \exp\left\{\lim_{x\to\infty} \frac{-\frac{t^2/2}{x^2} + O\left(\frac{t^3}{x^{5/2}}\right)}{-\frac{1}{x^2}\left[1 + \frac{t^2/2}{x} + O\left(\frac{t^3}{x^{3/2}}\right)\right]}\right\}$$
$$= \exp\left\{\lim_{x\to\infty} \frac{t^2/2 + O\left(\frac{t^3}{x^{1/2}}\right)}{1 + \frac{t^2/2}{x} + O\left(\frac{t^3}{x^{3/2}}\right)}\right\} = \exp\left(\frac{t^2}{2}\right).$$

Δηλαδή δείξαμε ότι το όριο της ροπογεννήτριας της Z_n είναι η ροπογεννήτρια της τυπικής κανονικής. Αυτό σημάνει ότι η τ.μ. $\lim_{n\to\infty} Z_n$ κατανέμεται σαν $N(0,1)$.

<u>Παρατήρηση 1</u>: Η τ.μ. Z_n είναι η τυποποίηση του δειγματικού μέσου $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

$$Z_n = \frac{S_n - \mathbb{E}[S_n]}{\sqrt{Var[S_n]}} = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\overline{X}_n - \mathbb{E}\left[\overline{X}_n\right]}{\sqrt{Var\left[\overline{X}_n\right]}}$$

<u>Παρατήρηση 2</u>: Πρακτικά για δείγμα μεγέθους περίπου $n \geq 30$ μπορούμε να θεωρήσουμε ότι η τ.μ. Z_n βρίσκεται «πολύ κοντά» στη τυπική κανονική, δηλαδή προσεγγιστικά δεχόμαστε ότι

$$Z_n = \frac{\overline{X}_n - \mathbb{E}\left[\overline{X}_n\right]}{\sqrt{Var\left[\overline{X}_n\right]}} = \frac{\overline{X}_n - \mu}{\sqrt{\sigma^2/n}} \sim N(0,1) \Leftrightarrow \overline{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Δηλαδή η κατανομή του δειγματικού μέσου $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ είναι προσεγγιστικά κανονική με μέσο το μ και διασπορά $\frac{\sigma^2}{n}$.

<u>Εφαρμογή</u>

Έχουμε δείξει ότι, εάν $X_i \overset{iid}{\sim} Bernoulli(p),\ i \geq 1$ τότε $S_n \sim Bin(n,p),\ n \geq 1$ και από το Κ.Ο.Θ. θα έχουμε

$$Z_n = \frac{S_n - \mathbb{E}[S_n]}{\sqrt{Var[S_n]}} = \frac{S_n - np}{\sqrt{np(1-p)}} \xrightarrow[n\to\infty]{d} N(0,1).$$

Εάν $n \geq 30$ προσεγγιστικά $Z_n = \frac{S_n - np}{\sqrt{np(1-p)}} \overset{d}{\approx} N(0,1)$ ή ισοδύναμα $S_n \overset{d}{\approx} N(np, np(1-p))$.

Έτσι για διωνυμικές πιθανότητες με αριθμό δοκιμών $n \geq 30$ θα έχουμε για την πιθανότητα του ενδεχομένου $\{x_1 < S_n \leq x_2\}$ για $0 \leq x_1 < x_2 \leq n$, την προσέγγιση:

$$P\{x_1 < S_n \leq x_2\} = P\left\{\frac{x_1 - np}{\sqrt{np(1-p)}} < \frac{S_n - np}{\sqrt{np(1-p)}} \leq \frac{x_2 - np}{\sqrt{np(1-p)}}\right\}$$

$$= P\left\{\frac{x_1 - np}{\sqrt{np(1-p)}} < Z_n \leq \frac{x_2 - np}{\sqrt{np(1-p)}}\right\}$$

$$= P\left\{ Z_n \le \frac{x_2 - np}{\sqrt{np(1-p)}} \right\} - P\left\{ Z_n \le \frac{x_1 - np}{\sqrt{np(1-p)}} \right\}$$

$$\approx \Phi\left(\frac{x_2 - np}{\sqrt{np(1-p)}} \right) - \Phi\left(\frac{x_1 - np}{\sqrt{np(1-p)}} \right).$$

Παρατηρήστε ότι εφόσον δεχόμαστε ότι $S_n \overset{d}{\approx} N\left(np, np(1-p)\right)$, τότε ισοδύναμα θα έχουμε:

$$P\{x_1 < S_n \le x_2\} \approx \int_{x=x_1}^{x_2} N\left(x \mid np, np(1-p)\right) dx \text{ , εάν } y = \frac{x - np}{\sqrt{np(1-p)}}$$

$$P\{x_1 < S_n \le x_2\} \approx \int_{y=\frac{x_1 - np}{\sqrt{np(1-p)}}}^{\frac{x_2 - np}{\sqrt{np(1-p)}}} N\left(y \mid 1, 0\right) dy = \Phi\left(\frac{x_2 - np}{\sqrt{np(1-p)}} \right) - \Phi\left(\frac{x_1 - np}{\sqrt{np(1-p)}} \right).$$

<u>Αριθμητική εφαρμογή</u>

Παίκτης κερδίζει παίγνιο, με πιθανότητα $p = 0.6$. Εάν παίξει $n = 100$ ανεξάρτητα παίγνια να βρεθεί η πιθανότητα ο αριθμός των κερδισμένων παιγνίων να είναι τουλάχιστον 55 και το πολύ 70.

Θέλουμε να υπολογίσουμε την πιθανότητα του ενδεχομένου $\{54 < S_{100} \le 70\}$.

$S_{100} \sim Bin(100, 0.6)$, με μέση τιμή $\mu = np = 60$ και τυπική απόκλιση $\sigma = \sqrt{np(1-p)} = \sqrt{24}$. Από τα προηγούμενα έχουμε ότι:

$$P\{54 < S_{100} \le 70\} = P\left\{ \frac{54 - 60}{\sqrt{24}} < \frac{S_{100} - 60}{\sqrt{24}} \le \frac{70 - 60}{\sqrt{24}} \right\}$$

$$= P\{-1.12 < Z_{100} \le 2.04\}.$$

Κάνοντας την υπόθεση $Z_{100} \overset{d}{\approx} N(0,1)$ παίρνουμε:

$$P\{54 < S_{100} \le 70\} \approx \Phi(2.04) - \Phi(-1.12) = \Phi(2.04) - \left[1 - \Phi(1.12)\right] \approx 0.85.$$

<u>Εφαρμογή</u>

Εάν $X_i \overset{iid}{\sim} Po(\lambda)$, $i \ge 1$, δείξτε ότι για $n \ge 30$ έχουμε τις προσεγγίσεις:

$$\sum_{x=x_1+1}^{x_2} e^{-n\lambda} \frac{(n\lambda)^x}{x!} \approx \Phi\left(\frac{x_2 - n\lambda}{\sqrt{n\lambda}} \right) - \Phi\left(\frac{x_1 - n\lambda}{\sqrt{n\lambda}} \right), \text{ και } \sum_{x=0}^{x_2} e^{-n\lambda} \frac{(n\lambda)^x}{x!} \approx \Phi\left(\frac{x_2 - n\lambda}{\sqrt{n\lambda}} \right)$$

Γνωρίζουμε ότι $S_n \sim Po(n\lambda)$. Από το Κ.Ο.Θ. έχουμε ότι $Z_n = \dfrac{S_n - n\lambda}{\sqrt{n\lambda}} \xrightarrow[n\to\infty]{d} N(0,1)$,

έτσι για $n \geq 30$ θα έχουμε $Z_n = \dfrac{S_n - n\lambda}{\sqrt{n\lambda}} \overset{d}{\approx} N(0,1)$ ή ότι $S_n \overset{d}{\approx} N(n\lambda, n\lambda)$ που δίνει

$$\sum_{x=x_1+1}^{x_2} e^{-n\lambda} \frac{(n\lambda)^x}{x!} = P\{x_1 < S_n \leq x_2\} = P\left\{\frac{x_1 - n\lambda}{\sqrt{n\lambda}} < \frac{S_n - n\lambda}{\sqrt{n\lambda}} \leq \frac{x_2 - n\lambda}{\sqrt{n\lambda}}\right\}$$
$$= P\left\{\frac{x_1 - n\lambda}{\sqrt{n\lambda}} < Z_n \leq \frac{x_2 - n\lambda}{\sqrt{n\lambda}}\right\} \approx \Phi\left(\frac{x_2 - n\lambda}{\sqrt{n\lambda}}\right) - \Phi\left(\frac{x_1 - n\lambda}{\sqrt{n\lambda}}\right),$$

ενώ

$$\sum_{x=0}^{x_2} e^{-n\lambda} \frac{(n\lambda)^x}{x!} = P\{S_n \leq x_2\} = P\left\{\frac{S_n - n\lambda}{\sqrt{n\lambda}} \leq \frac{x_2 - n\lambda}{\sqrt{n\lambda}}\right\}$$
$$= P\left\{Z_n \leq \frac{x_2 - n\lambda}{\sqrt{n\lambda}}\right\} \approx \Phi\left(\frac{x_2 - n\lambda}{\sqrt{n\lambda}}\right).$$

<u>Παρατήρηση</u>

Χρησιμοποιήσαμε ότι εάν $X_i \overset{iid}{\sim} Bernoulli(\vartheta)$ τότε $S_n \sim Bin(n, \vartheta)$.

Η ροπογεννήτρια της Bernoulli είναι:

$$M_{X_i}(t) = M(t) = e^{t\cdot 0} P\{X_i = 0\} + e^{t\cdot 1} P\{X_i = 1\} = 1 - \vartheta + e^t \vartheta,$$

ενώ η ροπογεννήτρια της S_n

$$M_{S_n}(t) = \mathbb{E}\left[e^{tS_n}\right] = \mathbb{E}\left[e^{tX_1} \cdots e^{tX_n}\right] = \mathbb{E}\left[e^{tX_1}\right] \cdots \mathbb{E}\left[e^{tX_n}\right]$$
$$= M(t)^n = \left\{1 - \vartheta + e^t \vartheta\right\}^n$$

Για την ροπογεννήτρια της διωνυμικής $X \sim Bin(n, \vartheta)$ έχουμε

$$M_X(t) = \mathbb{E}\left[e^{tX}\right] = \sum_{x=0}^{n} e^{tx} \binom{n}{x} \vartheta^x (1-\vartheta)^{n-x}$$
$$= \sum_{x=0}^{n} \binom{n}{x} \left(\vartheta e^t\right)^x (1-\vartheta)^{n-x} = \left\{1 - \vartheta + e^t \vartheta\right\}^n$$

Επειδή $M_{S_n}(t) = M_X(t)$ παίρνουμε ότι $S_n \overset{d}{=} X \sim Bin(n, \vartheta)$.

<u>Ορισμός</u>: Λέμε ότι η ακολουθία τυχαίων μεταβλητών $\{X_n\}_{n\ge 1}$ συγκλίνει ως προς πιθανότητα στην τ.μ. X (ασθενής σύγκλιση) όταν

$$\lim_{n\to\infty} P\{|X_n - X| \ge \varepsilon\} = 0,\ \forall \varepsilon > 0 \Leftrightarrow X_n \xrightarrow[n\to\infty]{P} X .$$

<u>Ο ασθενής νόμος των μεγάλων αριθμών</u> Δίνεται ακολουθία τ.μ. $\{X_i\}_{i\ge 1}$ που είναι ανεξάρτητες και ταυτοτικά κατανεμημένες (iid) με μέση τιμή $\mu < \infty$ τότε για τον δειγματικό μέσο $\bar{X}_n$ ισχύει ότι $\bar{X}_n \xrightarrow[n\to\infty]{P} \mu$.

Απόδειξη με την επιπρόσθετη συνθήκη $Var[X_i] = \sigma^2 < \infty$.

Από την IID ακολουθία τ.μ. $\{X_i\}_{i\ge 1}$, φτιάχνουμε την ακολουθία $\{\bar{X}_n\}_{n\ge 1}$ όπου $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Η μέση τιμή και η διασπορά της $\bar{X}_n$ είναι

$$\mathbb{E}\left[\bar{X}_n\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i] = \mu ,$$

$$Var\left[\bar{X}_n\right] = \mathbb{E}\left[\left(\bar{X}_n - \mu\right)^2\right] = \frac{1}{n^2}\mathbb{E}\left[\left(\sum_{i=1}^n X_i - n\mu\right)^2\right] = \frac{1}{n^2}\mathbb{E}\left[\left(\sum_{i=1}^n (X_i - \mu)\right)^2\right]$$

$$= \frac{1}{n^2}\left\{\sum_{i=1}^n \mathbb{E}\left[\left(X_i - \mu\right)^2\right] + 2\sum_{1\le i<j\le n} Cov\left(X_i, X_j\right)\right\} = \frac{1}{n^2}\sum_{i=1}^n Var[X_i] = \frac{\sigma^2}{n} .$$

Από την ανισότητα Chebyshev για την $\bar{X}_n$ έχουμε ότι:

$$P\left\{\left|\bar{X}_n - \mathbb{E}\left[\bar{X}_n\right]\right| \ge \varepsilon\right\} \le \frac{Var\left[\bar{X}_n\right]}{\varepsilon^2} \text{ ή ότι } P\left\{\left|\bar{X}_n - \mu\right| \ge \varepsilon\right\} \le \frac{\sigma^2}{n\varepsilon^2} \xrightarrow[n\to\infty]{} 0 .$$

Απόδειξη χωρίς την επιπρόσθετη συνθήκη $\sigma^2 < \infty$ (Khintchin)

Εδώ χρειάζεται μόνο η ύπαρξη της μέσης τιμής $\mathbb{E}[X_i] = \mu$. Η απόδειξη μοιάζει με την απόδειξη του Κεντρικού Οριακού Θεωρήματος.

Εάν $\varphi_{X_j}(t) = \varphi(t)$, τότε είναι εύκολο να δείξουμε ότι $\varphi_{\bar{X}_n}(t) = \varphi(t/n)^n$. Επειδή

$$\varphi(t) = \mathbb{E}\left[e^{itX_j}\right] = \mathbb{E}\left[1 + itX_j + \mathcal{O}\left(t^2\right)\right] = 1 + it\mu + \mathcal{O}\left(t^2\right),$$

έχουμε $\varphi_{\bar{X}_n}(t) = \varphi(t/n)^n = \left(1 + \frac{it\mu}{n} + \mathcal{O}\left(\frac{t^2}{n^2}\right)\right)^n$, και παίρνοντας το όριο για $n \to \infty$

$$\lim_{n\to\infty}\varphi_{\bar{X}_n}(t)=\lim_{n\to\infty}\left(1+\frac{it\mu}{n}+\mathcal{O}\left(\frac{t^2}{n^2}\right)\right)^n=\lim_{x\to\infty}\left(1+\frac{it\mu}{x}+\mathcal{O}\left(\frac{t^2}{x^2}\right)\right)^x$$
$$=\exp\left\{\lim_{x\to\infty}\frac{1}{x^{-1}}\log\left[1+\frac{it\mu}{x}+\mathcal{O}\left(\frac{t^2}{x^2}\right)\right]\right\}=\exp(it\mu).$$

Αλλά η μόνη τ.μ. με χαρακτηριστική συνάρτηση $\exp(it\mu)$ είναι η (τετριμμένη) τ.μ. $X=\mu$ με $P\{X=\mu\}=1$. Έτσι έχουμε ότι $\bar{X}_n \xrightarrow[n\to\infty]{d} \mu$, δηλαδή η τ.μ. $\bar{X}_n$ συγκλίνει κατά νόμο (κατά κατανομή) στην τ.μ. $X=\mu$. Για να τελειώσουμε την απόδειξη, θα πρέπει να δείξουμε και ότι $\bar{X}_n \xrightarrow[n\to\infty]{P} \mu$.

Γενικά για ακολουθία τ.μ. $Y_n,\ n\geq 1$, ισχύει ότι $Y_n \xrightarrow[n\to\infty]{P} Y \Rightarrow Y_n \xrightarrow[n\to\infty]{d} Y$, αλλά όχι και το αντίστροφο. Στην ειδική περίπτωση όμως που $Y=\sigma\tau\alpha\theta.$ ισχύει και το αντίστροφο, και έχουμε $Y_n \xrightarrow[n\to\infty]{P} \sigma\tau\alpha\theta. \Leftrightarrow Y_n \xrightarrow[n\to\infty]{d} \sigma\tau\alpha\theta.$.

Αποδεικνύουμε το τελευταίο για την περίπτωση $\Leftarrow$

$$Y_n \xrightarrow[n\to\infty]{d} Y=c=\sigma\tau\alpha\theta. \Leftrightarrow \lim_{n\to\infty}P\{Y_n\leq y\}=P\{c\leq y\}=1(c\leq y)$$

$$\lim_{n\to\infty}P\{|Y_n-c|\geq\varepsilon\}=\lim_{n\to\infty}P\{Y_n-c\geq\varepsilon\}+\lim_{n\to\infty}P\{Y_n-c\leq-\varepsilon\}$$
$$=\lim_{n\to\infty}P\{Y_n\geq c+\varepsilon\}+\lim_{n\to\infty}P\{Y_n\leq c-\varepsilon\}$$
$$=1-\lim_{n\to\infty}P\{Y_n< c+\varepsilon\}+\lim_{n\to\infty}P\{Y_n\leq c-\varepsilon\}$$
$$=1-1(c<c+\varepsilon)+1(c\leq c-\varepsilon)=1-1-0=0,\ \forall\varepsilon>0.$$