הרצאה 13

1 משפט הגבול המרכזי – Central Limit Theorem

(זה משפט "הגבול המרכזי". כלומר זה משפט על הגבול המרכזי, ולא "משפט גבול" מרכזי. הגבול הוא מרכזי, לא המשפט.)
למרות שהמשפט לא "מרכזי" (מסתבר), הוא תוצאה משמעותית בתורת ההסתברות.

משפט 1.1 (משפט הגבול המרכזי)**:** תהי $\{X_i\}_{i=1}^{\infty}$ סדרה של משתנים מקריים שמקיימת:

א. בלתי תלויים
ב. כולם בעלי התפלגות זהה
ג. בעלי תוחלת סופית μ
ד. שונות סופית $\sigma^2 > 0$

לכל שלם חיובי n, יהי $Y_n = \frac{X_1+\cdots+X_n-n\mu}{\sigma\sqrt{n}}$ ותהי F_n ה-CDF של Y_n. כלומר $F_n(a) = \mathbb{P}(Y_n \le a)$ לכל $a \in \mathbb{R}$.
אזי מתקיים לכל $a \in \mathbb{R}$: $\lim_{n\to\infty} F_n(a) = \Phi(a)$.

למה הגדרנו ככה את Y_n? ניזכר קודם במ"מ נורמלי סטנדרטי. התוחלת שלו היא 0, והשונות היא 1.
אז כדי שנטען ש-Y_n מתפלג בצורה דומה למ"מ נורמלי סטנדרטי, אז נצטרך שהתוחלת תהיה 0, והשונות 1 (לפחות בקירוב).
אבל התחלנו עם מ"מ שיש לו תוחלת שונה מ-0, וסכמנו n כאלה. אז נוריד את התוחלת של הסכום – $n\mu$.
באופן דומה, בגלל שהם בת"ל אז השונות של הסכום זה סכום השונויות - $n \cdot \sigma^2$. אז כדי לקבל 1, נחלק באותו דבר.
אבל כשכופלים (או מחלקים) מ"מ בקבוע, הקבוע יוצא בריבוע. אז נחלק בשורש של זה.

דוגמה 1: נטיל מטבע הוגן 1000 פעמים, באופן בת"ל. יהי X המספר הכולל של הטלות שיצאו עץ.
בהרצאה 1 השתמשנו באי-שוויונות צ'בישב וצ'רנוף כדי לתת חסם תחתון על $\mathbb{P}(450 < X < 550)$.
על ידי צ'בישב קיבלנו 0.9, ועל ידי צ'רנוף קיבלנו $1 - 2e^{-5} \approx 0.98652$.
עכשיו נשתמש במשפט הגבול המרכזי.
לכל הטלה $1 \le i \le 1000$, יהי X_i האינדיקטור לעץ.
נשים לב שכל ה- X_i מקיימים את 4 התנאים, ובפרט $\mathbb{E}(X_i) = 1/2$, $Var(X_i) = \mathbb{E}(X_i^2) - \left(\mathbb{E}(X_i)\right)^2 = 1/4$.
יהי $Y = \frac{X-500}{\sqrt{250}}$, אז נקבל לפי 1.1: (1)

$$\begin{aligned}
\mathbb{P}(450 < X < 550) &= \mathbb{P}\left(\frac{450 - 1000\cdot\frac{1}{2}}{\sqrt{\frac{1}{4}}\cdot\sqrt{1000}} < \frac{X - 1000\cdot\frac{1}{2}}{\sqrt{\frac{1}{4}}\cdot\sqrt{1000}} < \frac{550 - 1000\cdot\frac{1}{2}}{\sqrt{\frac{1}{4}}\cdot\sqrt{1000}}\right)\\
&= \mathbb{P}\left(-\frac{50}{\sqrt{250}} < Y < \frac{50}{\sqrt{250}}\right) = \mathbb{P}\left(-\frac{2500}{250} < Y < \frac{50}{250}\right) = \mathbb{P}\left(-\sqrt{10} < Y < \sqrt{10}\right)\\
&= \mathbb{P}\left(Y < \sqrt{10}\right) - \mathbb{P}\left(Y \le -\sqrt{10}\right) \approx^{\text{א}} \Phi\left(\sqrt{10}\right) - \Phi\left(-\sqrt{10}\right) = \Phi\left(\sqrt{10}\right) - \left(1 - \Phi\left(\sqrt{10}\right)\right)\\
&= 2\Phi\left(\sqrt{10}\right) - 1 \approx^{\text{ב}} 0.9984
\end{aligned}$$

א – משפט 1.1. ב – לפי הטבלה.

קיבלנו חסם הדוק יותר. בנוסף, הוא חסם תחתון וגם עליון.
מצד שני, זו תוצאה איכותית ולא כמותית. זה אומר לנו מה קורה בגבול אבל לא ספציפית עבור ערך כלשהו.
הבעיה נמצאת ב"שוויון" א. אמרנו שזה *בערך* התוצאה, אבל התוצאה מדברת על אינסוף הטלות ולא ספציפית על 1000.
אנחנו יודעים שיש פער כלשהו ואנחנו לא יודעים בדיוק כמה.
נגיד אם הפער הוא 1/2, אז החסם גרוע. אם הפער הוא יותר מ1, אז בכלל חרגנו מהטווח שהגיוני להסתברות.
יכול להיות שהחסם האמיתי רחוק ממה שמצאנו.
תוצאה כמותית הייתה אולי נותנת חסם פחות טוב אבל היינו יודעים שהוא נכון.

לכאורה נשמע שאי אפשר להשתמש ב CLT לחישובים כאלה, ואכן יש גרסאות כמותיות של CLT. נראה אחת מהן.
(בפועל בקורס הזה, נעבוד עם ה CLT כאילו זה קירוב תקין).

משפט 1.2 (משפט ברי-אסן)**:** תהי $\{X_i\}_{i=1}^{\infty}$ סדרה של מ"מ בת"ל בעלי אותה התפלגות.
נניח שמתקיים: $\mathbb{E}(X_i) = 0$, $Var(X_i) = \sigma^2 > 0$, $\mathbb{E}(|X_i|^3) = \rho$ עבור כל $i \in \mathbb{N}$, כאשר $\sigma, \rho \in \mathbb{R}$.
לכל שלם חיובי n, יהי $Y_n = \frac{X_1+\cdots+X_n}{\sigma\sqrt{n}}$ ויהי F_n ה CDF של Y_n. כלומר, $F_n(a) = \mathbb{P}(Y_n \le a)$ לכל $a \in \mathbb{R}$.
אזי מתקיים לכל $a \in \mathbb{R}$ ולכל n שלם חיובי: $|F_n(a) - \Phi(a)| \le \frac{\rho}{\sigma^3\sqrt{n}}$.
נשים לב שזה מזכיר את הגדרת הגבול, ולא סתם: הטענה של CLT היא לפי גבול, קרי: לכל אפסילון, מתישהו ה CDF תהיה קרובה עד כדי אפסילון ל-Φ. אז הטענה פה קובעת את "אפסילון" לפי n. זו תוצאה כמותית. עבור n קטן החסם אולי לא יהיה הדוק, אבל הוא מדויק.

נשתמש במשפט 1.2 כדי לחשב שוב את החסם מדוגמה 1.
לכל $1 \le i \le 1000$, יהי $Z_i = X_i - 1/2$.
נשים לב ש $\mathbb{E}(Z_i) = 0$. השונות לא משתנה, אז $Var(Z_i) = Var(X_i) = 1/4$.
כלומר $\sigma := \sqrt{Var(Z_i)} = 1/2$. לבסוף, $\mathbb{P}(|Z_i| = 1/2) = 1$.
ולכן $\mathbb{E}(|Z_i|^3) = (1/2)^3 = 1/8$.
יהי $Y = \frac{Z_1+\cdots+Z_{1000}}{\sqrt{1/4}\cdot\sqrt{1000}} = \frac{X-500}{\sqrt{250}}$. ממשפט 1.2, נקבל: (2)

$$\mathbb{P}\left(Y \le \sqrt{10}\right) \ge \Phi\left(\sqrt{10}\right) - \frac{\frac{1}{8}}{\left(\frac{1}{2}\right)^3 \cdot \sqrt{1000}} \ge \Phi\left(\sqrt{10}\right) - 0.03163$$

וגם (3):

$$\mathbb{P}\left(Y \le -\sqrt{10}\right) \le \Phi\left(-\sqrt{10}\right) + \frac{\frac{1}{8}}{\left(\frac{1}{2}\right)^3 \cdot \sqrt{1000}} \le \Phi\left(-\sqrt{10}\right) + 0.03163$$

בדומה ל(1), נקבל עם (2),(3):

$$\begin{aligned}\mathbb{P}(450 < X < 550) &= \mathbb{P}\left(\frac{450 - 1000\cdot\frac{1}{2}}{\sqrt{\frac{1}{4}}\cdot\sqrt{1000}} < \frac{X - 1000\cdot\frac{1}{2}}{\sqrt{\frac{1}{4}}\cdot\sqrt{1000}} < \frac{550 - 1000\cdot\frac{1}{2}}{\sqrt{\frac{1}{4}}\cdot\sqrt{1000}}\right) \\ &= \mathbb{P}\left(-\frac{50}{\sqrt{250}} < Y < \frac{50}{\sqrt{250}}\right) = \mathbb{P}\left(-\frac{2500}{250} < Y < \frac{50}{250}\right) = \mathbb{P}\left(-\sqrt{10} < Y < \sqrt{10}\right) \\ &= \mathbb{P}\left(Y < \sqrt{10}\right) - \mathbb{P}\left(Y \le -\sqrt{10}\right) \ge \left[\Phi\left(\sqrt{10}\right) - 0.03163\right] - \left[\Phi\left(-\sqrt{10}\right) + 0.03163\right] \\ &= 2\Phi\left(\sqrt{10}\right) - 1 - 0.06326 \ge 0.935\end{aligned}$$

זה יותר טוב מהחסם שמצאנו עם צ'בישב, אבל פחות טוב מהחסם שמצאנו עם צ'רנוף.

דוגמה 2: נזרוק קובייה הוגנת 360,000 פעמים, בת"ל. יהי X מספר הזריקות שיצאו 6. נרצה למצוא $a, b \in \mathbb{R}$ שמקיימים:

$$\mathbb{P}(54000 \le X \le 63000) \approx \frac{1}{\sqrt{2\pi}}\int_a^b e^{-x^2/2}dx$$

(שאלה ממבחן, 2017. הטריק שאמור לעזור בשאלה הוא שהצד הימני זה בעצם $\Phi(b) - \Phi(a)$).
הרעיון הוא להשתמש ב CLT, ואפילו לא צריך להשתמש בטבלה.

לכל $1 \le i \le 360000$, יהי X_i האינדיקטור לכך שיצא 6. אזי $x = \sum_{i=1}^{360000} X_i$ הוא מספר הפעמים שיצא 6. ולכל i בתחום מתקיים: $\mathbb{E}(X_i) = 1/6$, $Var(X_i) = 5/36$. אז לפי CLT מתקיים:

$$\mathbb{P}(54000 \le X \le 63000) = \mathbb{P}\left(\frac{54000 - 360000 \cdot \frac{1}{6}}{\sqrt{\frac{5}{36}} \cdot \sqrt{360000}} \le \frac{X - 360000 \cdot \frac{1}{6}}{\sqrt{\frac{5}{36}} \cdot \sqrt{360000}} \le \frac{63000 - 360000 \cdot \frac{1}{6}}{\sqrt{\frac{5}{36}} \cdot \sqrt{360000}}\right)$$

$$= \mathbb{P}\left(-12\sqrt{5} \le \frac{X - 360000 \cdot \frac{1}{6}}{\sqrt{\frac{5}{36}} \cdot \sqrt{360000}} \le 6\sqrt{5}\right)$$

$$= \mathbb{P}\left(\frac{X - 360000 \cdot \frac{1}{6}}{\sqrt{\frac{5}{36}} \cdot \sqrt{360000}} \le 6\sqrt{5}\right) - \mathbb{P}\left(\frac{X - 360000 \cdot \frac{1}{6}}{\sqrt{\frac{5}{36}} \cdot \sqrt{360000}} < -12\sqrt{5}\right) \approx$$

$$\approx \Phi\left(6\sqrt{5}\right) - \Phi\left(-12\sqrt{5}\right) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{6\sqrt{5}} e^{-\frac{x^2}{2}}dx - \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{-12\sqrt{5}} e^{-x^2/2}dx = \frac{1}{\sqrt{2\pi}}\int_{-12\sqrt{5}}^{6\sqrt{5}} e^{-x^2/2}dx$$

נסיק שהתנאי הרצוי מתקיים עם: $a = -12\sqrt{5}, b = 6\sqrt{5}$.

דוגמה 3: נשתמש ב CLT כדי להוכיח ש:

$$\lim_{n\to\infty} e^{-n}\left(1 + n + \frac{n^2}{2!} + \cdots + \frac{n^n}{n!}\right) = \frac{1}{2}$$

תהי $\{X_i\}_{i=1}^{\infty}$ סדרה בת"ל של מ"מ כאשר $X_i \sim Poi(1)$ לכל $i \in \mathbb{N}$. לכל n שלם חיובי, יהי $Y_n = \sum_{i=1}^{n} X_i$. אזי, ראינו בהסתברות 1 ש: $Y_n \sim Poi(n)$ לכל $n \in \mathbb{N}$. (משתנה מקרי שהוא סכום של שני משתנים מקריים שמתפלגים פואסון, בעצמו מתפלג פואסון עם פרמטר שהוא סכום הפרמטרים. באינדוקציה אפשר להרחיב את זה לכל סכום של n מ"מ) בפרט, (4):

$$\mathbb{P}(Y_n \le n) = \sum_{j=0}^{n} \mathbb{P}(Y_n = j) =^{\aleph} \sum_{j=0}^{n} e^{-n} \cdot \frac{\lambda^j}{j!} = e^{-n}\left(1 + n + \frac{n^2}{2!} + \cdots + \frac{n^n}{n!}\right)$$

א – אם $X \sim Poi(\lambda)$, אז $\mathbb{P}(X = k) = e^{-\lambda} \cdot \frac{\lambda^k}{k!}$

וגם, לפי CLT מתקיים (5): ננרמל את Y_n לפי התנאים של CLT:

מתקיים: $\mathbb{E}(Y_n) = n, Var(Y_n) = n$. אז נחסיר n כדי שהתוחלת תהיה 0, ונחלק בשורש השונות כדי שנקבל שונות 1.

$$\lim_{n\to\infty} \mathbb{P}(Y_n \le n) = \lim_{n\to\infty} \mathbb{P}(Y_n - n \le 0) = \lim_{n\to\infty} \mathbb{P}\left(\frac{Y_n - n}{\sqrt{n}} \le 0\right) = \Phi(0) = \frac{1}{2}$$

אם נחבר את (4), (5) נקבל:

$$\lim_{n\to\infty} e^{-n}\left(1 + n + \frac{n^2}{2!} + \cdots + \frac{n^n}{n!}\right) = \lim_{n\to\infty} \mathbb{P}(Y_n \le n) = \frac{1}{2}$$

כנדרש. ■

עד כאן הסתברות 2.