הסקה סטטיסטית שיעור 3

הקדמה חומר קריאה 3א, חומר קריאה 3ב, חומר קריאה 3ג

1 פונקציות של מ"מ רציפים (חומר קריאה 3א)

אם נתון $Y = aX + b$, אנחנו יודעים את התוחלת והשונות ($E(Y) = aE(X) + b, Var(Y) = a^2 Var(X)$). אבל מה ה-CDF, PDF? נוכל לחשב ע"י חישוב ישיר לפי הגדרה, או החלפת משתנה.

דוגמה 1

יהי $X \sim U(0,2)$. כלומר $F_X(x) = \frac{x}{2}, f_X(x) = \frac{1}{2}$ בתחום $[0,2]$. מה הטווח, PDF, CDF של $Y = X^2$?

קל לראות שהתחום הוא $[0,4]$. נחשב הCDF:

$$F_Y(y) = P(Y \leq y) = P(X^2 \leq y) = P\left(X \leq \sqrt{y}\right) = F_X\left(\sqrt{y}\right) = \sqrt{y}/2$$

כדי למצוא את הPDF, נגזור:

$$f_Y(y) = \frac{d}{dy} F_Y(y) = 1/4\sqrt{y}$$

אפשר לחשב גם בהחלפת משתנה:

$$y = x^2 \Longrightarrow \frac{dy}{dx} = 2x \Longrightarrow dy = 2xdx \Longrightarrow dx = \frac{dy}{2x} \Longrightarrow dx = \frac{dy}{2\sqrt{y}}$$

$$f_X(x)dx = f_Y(y)dy$$

$$f_X(x)dx = f_X(x)\frac{dy}{2x}$$

$$f_Y(y)dy = f_X(x)\frac{dy}{2x}$$

$$f_Y(y) = f_X(x)\frac{1}{2x} = \frac{f_X(x)}{2\sqrt{y}} = \frac{1/2}{2\sqrt{y}} = \frac{1}{4\sqrt{y}}$$

דוגמה 2

יהי $X \sim Exp(\lambda)$, אז $f_X(x) = \lambda e^{-\lambda x}$ עבור $x \in [0, \infty)$. מה הCDF של $Y = X^2$? נחשב בעזרת החלפת משתנה:

$$y = x^2 \quad \Longrightarrow \quad dy = 2xdx \quad \Longrightarrow \quad dx = \frac{dy}{2\sqrt{y}}$$

$$f_X(x)dx = \lambda e^{-\lambda x}dx = \lambda e^{-\lambda\sqrt{y}}\frac{dy}{2\sqrt{y}} = f_Y(y)dy$$

$$f_Y(y) = \lambda e^{-\lambda\sqrt{y}}\frac{1}{2\sqrt{y}} = \frac{\lambda}{2\sqrt{y}}e^{-\lambda\sqrt{y}}$$

דוגמה 3

יהי $X \sim N(5,3^2)$. צ"ל ש- $Z = \frac{X-5}{3}$ היא נורמלית סטנדרטית (כלומר ש $Z \sim N(0,1)$).

נבצע החלפת משתנים בנוסחה של $f_X(x)$:

$$z = \frac{x-5}{3} \quad \Longrightarrow \quad dz = \frac{dx}{3} \quad \Longrightarrow \quad dx = 3dz$$

$$f_X(x)dx = \frac{1}{3\sqrt{2\pi}} e^{-(x-5)^2/(2\cdot 3^2)} dx = \frac{1}{3\sqrt{2\pi}} e^{-z^2/2} 3dz = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = f_Z(z)dz$$

שזה בדיוק הנוסחה של $N(0,1)$.

דוגמה 4

יהי $X \sim N(\mu, \sigma^2)$. צ"ל: $Z = \frac{X-\mu}{\sigma}$ הוא נורמלי סטנדרטי, כלומר $Z \sim N(0,1)$. כמו בדוגמה הקודמת:

$$z = \frac{x-\mu}{\sigma} \quad \Longrightarrow \quad dz = \frac{dx}{\sigma} \quad \Longrightarrow \quad dx = \sigma dz$$

$$f_X(x)dx = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\cdot\sigma^2)} dx = \frac{1}{\sigma\sqrt{2\pi}} e^{-z^2/2} \sigma dz = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = f_Z(z)dz$$

2 **תוחלת של מ"מ רציף** (חומר קריאה 3ב)

עבור מ"מ רציף עם תחום $[a,b]$ ו-PDF $f(x)$, התוחלת מוגדרת:

$$E(X) = \int_a^b x\, f(x)\, dx$$

דוגמה 1

עבור $X \sim U(0,1)$, נמצא את $E(X)$: לפי הגדרה,

$$E(X) = \int_0^1 x\, dx = \left[\frac{x^2}{2}\right]_0^1 = \frac{1}{2}$$

דוגמה 2

יהי X בתחום $[0,2]$ עם PDF $f(x) = \frac{3}{8}x^2$. נמצא את $E(X)$:

$$E(X) = \int_0^2 \frac{3}{8} x^3\, dx = \left[\frac{3x^4}{32}\right]_0^2 = \frac{3}{2}$$

דוגמה 3

יהי $X \sim Exp(\lambda)$, נמצא את $E(X)$:

$$E(X) = \int_0^\infty \lambda e^{-\lambda x}\, dx = \left[-\lambda e^{-\lambda x} - \frac{e^{-\lambda x}}{\lambda}\right]_0^\infty = \frac{1}{\lambda}$$

תוחלת של פונקציה של מ"מ

$$E\big(h(X)\big) = \int_{-\infty}^{\infty} h(x) f_X(x) dx$$

דוגמה 6

עבור $X \sim Exp(\lambda)$, נעשה אינטגרציה בחלקים:

$$E(X^2) = \int_0^\infty x^2 \cdot \lambda e^{-\lambda x}\, dx = \left[-x^2 e^{-\lambda x} - \frac{2x}{\lambda} e^{-\lambda x} - \frac{2}{\lambda^2} e^{-\lambda x}\right]_0^\infty = \frac{2}{\lambda^2}$$

3 שונות

דוגמה 1

עבור $X \sim U(0,1)$,

$$Var(X) = E((X-\mu)^2) = \int_0^1 \left(x - \frac{1}{2}\right)^2 dx = \frac{1}{12}$$

דוגמה 2

עבור $X \sim Exp(\lambda)$, (ניעזר בדוגמה 6)

$$Var(X) = E(X^2) - \mu^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}, \qquad \sigma_X = \frac{1}{\lambda}$$

דוגמה 3

יהי $Z \sim N(0,1)$. נראה ש $Var(Z) = 1$:

$$Var(Z) = E(Z^2) - \underbrace{E(Z)}_{0} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-z^2/2} dz =$$

נעשה אינטגרציה בחלקים עם $u = z,\ v' = ze^{-z^2/2} \implies u' = 1,\ v = -e^{-z^2/2}$.

$$= \frac{1}{\sqrt{2\pi}} \left(\left[-ze^{-z^2/2} \right]_{-\infty}^{\infty} \right) + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2/2} dz$$

הצד השמאלי שווה לאפס, כי ה $-z^2/2$ שואף לאפס (גם בפלוס אינסוף וגם במינוס, כי זה $-(z^2)$). הצד הימני שווה 1 כי זה בדיוק האינטגרל של ה-PDF של $N(0,1)$.

דוגמה 4

יהי $X \sim N(\mu, \sigma^2)$. נראה ש $Var(X) = \sigma^2$. בעזרת החלפת משתנים $z = (x-\mu)/\sigma$

$$Var(X) = E((X-\mu)^2) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x-\mu)^2 e^{-(x-\mu)^2/2\sigma^2} dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (\sigma z)^2 e^{-(\sigma z)^2/2\sigma^2} dz$$
$$= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-z^2/2} dz = \sigma^2$$

כי האינטגרל האחרון זה מה שחישבנו ב $Var(Z)$ בדוגמה הקודמת.

4 שברון (*quantile*)

מונח בסטטיסטיקה, שמתייחס לנקודת חתך שמתחתיה נמצאה החלק ה-q ($0 \le q \le 1$) מהאוכלוסייה.

חציון של X הוא הערך של x שעבורו $P(X \le x) = 0.5$ כלומר, $P(X \le x) = P(X \ge x)$. בחציון, $F(x) = 0.5$.

דוגמה 1

נמצא את החציון של $X \sim Exp(\lambda)$: הCDF מוגדרת: $F(x) = 1 - e^{-\lambda x}$

נרצה: $1 - e^{-\lambda x} = 1/2$. נמצא את x:

$$1 - e^{-\lambda x} = 1/2 \implies 1 - 1/2 = e^{-\lambda x} \implies \ln(1/2) = \ln\left(e^{-\lambda x}\right) \implies -\ln(2) = -\lambda x \implies x = \frac{\ln(2)}{\lambda}$$

נשים לב שהחציון לא תמיד שווה לממוצע (או התוחלת).

הגדרה: השברון ה-p הוא הערך q_p כך ש $P(X = q_p) = p$. (לדוגמה, החציון הוא השברון החצי. $q_{0.5}$)

בדרך כלל נרשום את זה לפי הCDF: $F(q_p) = p$

השברון לפעמים מתואר גם בתור שבר ספציפי. לדוגמה, **האחוזון** ה-60 הוא בעצם השברון ה-0.6. אדם שנמצא באחוזון ה-60 בגובה, גבוה יותר מ-60% מהאוכלוסייה. כלומר ההסתברות שהוא גבוה יותר מאדם אקראי שנדגם היא 0.6.

5 חוק המספרים הגדולים ומשפט הגבול המרכזי (חומר קריאה 3ג)

חוק המספרים הגדולים אומר ש: בהסתברות גבוהה,

א) הממוצע של הרבה דגימות בת"ל יהיה קרוב לממוצע של ההתפלגות.
ב) היסטוגרמת הצפיפות של הרבה דגימות בת"ל יהיה קרוב לגרף הצפיפות של ההתפלגות.

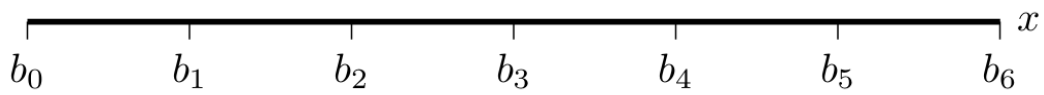
משפט הגבול המרכזי אומר שהסכום או הממוצע של הרבה עותקים בת"ל של מ"מ, הוא בערך התפלגות נורמלית. המשפט גם מגדיר את הממוצע וסטיית התקן של ההתפלגות הנורמלית הזו.

הטענות האלה מאוד שימושיות בסטטיסטיקה, ובדרך כלל לא צריך n ענק – ערכים של $n = 30$ בדרך כלל יספיקו.

היסטוגרמות

נרצה לבנות את ההיסטוגרמה כך שהן יזכירו את השטח מתחת לגרף הPDF. ונראה איך חוק המספרים הגדולים מתאים להיסטוגרמות. שלבי בניית היסטוגרמה:

1) נבחר אינטרוול של הממשיים ונחלק אותו ל-m אינטרוולים, עם קצוות $b_0, b_1, \dots, b_m$. בדרך כלל הם יהיו באותו גודל:



כל אחד נקרא **דלי** (bin).

2) נכניס כל x_i לדלי המתאים (אם הוא על הגבול, הוא ייכנס לשמאלי).

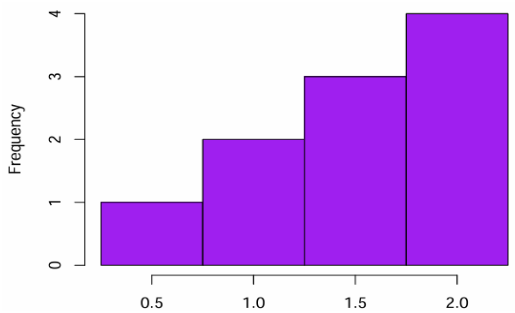
3) כדי לבנות היסטוגרמת **תדירות:** נשים עמודה מעל כל דלי, גובה העמודה הוא מספר הנקודות x_i ששמנו בדלי.

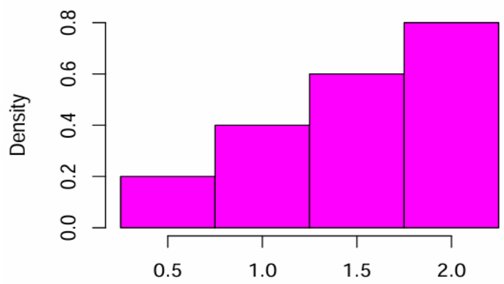
4) כדי לבנות היסטוגרמת **צפיפות:** נשים עמודה מעל כל דלי, כך ש**השטח** של העמודה הוא האחוז של הנקודות שיש בדלי.

כשכל הדליים באותו הרוחב, לעמודות של היסטוגרמת התדירות יש שטח יחסי לספירה. אז היסטוגרמת הצפיפות מתקבלת מחלוקה של כל הגובה של כל עמודה בשטח הכולל של היסטוגרמת התדירות. אם נתעלם מהסרגל האנכי, שתי ההיסטוגרמות זהות.

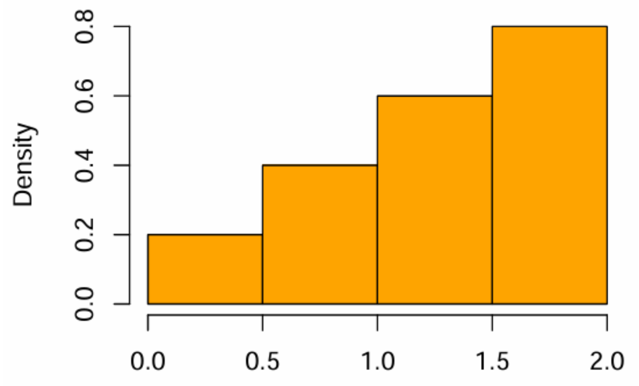
אם הדליים ברוחב שונה, שני ההיסטוגרמות נראות מאד שונה.

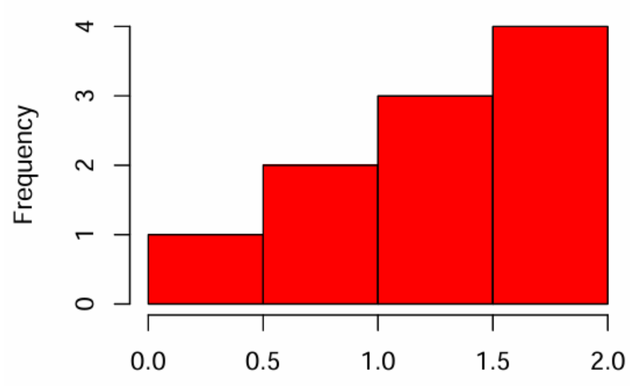
דוגמאות: עבור הדאטא [0.5, 1, 1, 1.5, 1.5, 1.5, 2, 2, 2, 2]. רוחב 0.5, קצוות 0.25, 0.75, 1.25, 1.75, 2.25.



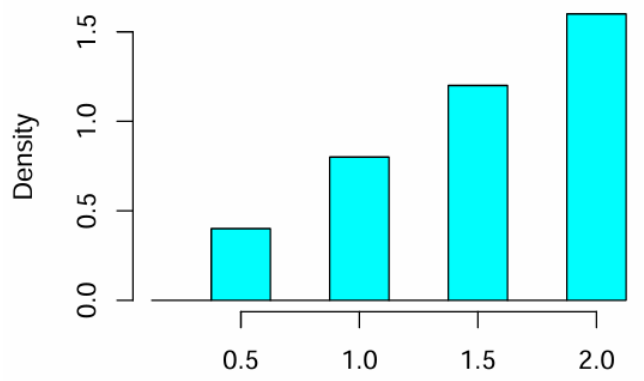


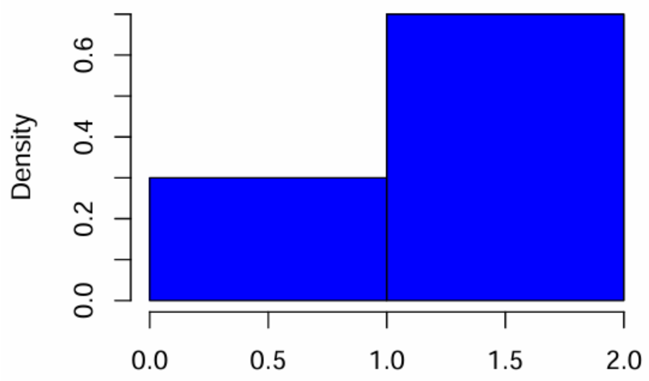
קצוות $0,\ 0.5,\ 1,\ 1.5,\ 2$





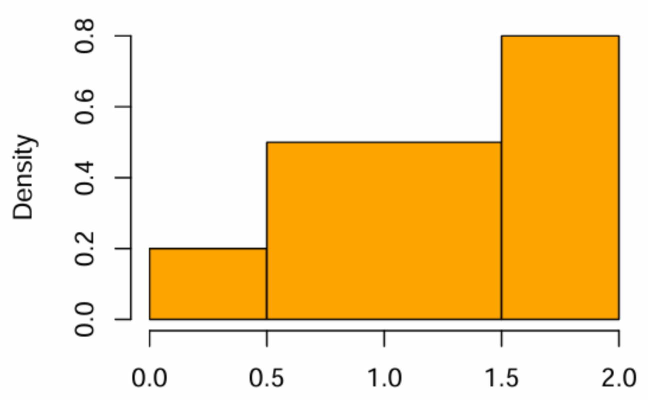
היסטוגרמות צפיפות של אותה הדאטא, עם רוחב דליים אחר (אחיד בתוך כל היסטוגרמה):

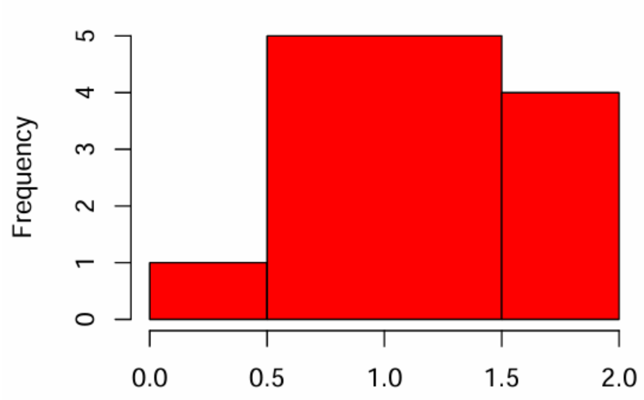




נשים לב שהסרגל שומר על כך שהשטח הכולל הוא 1. הרווחים הם עמודות בגובה 0.

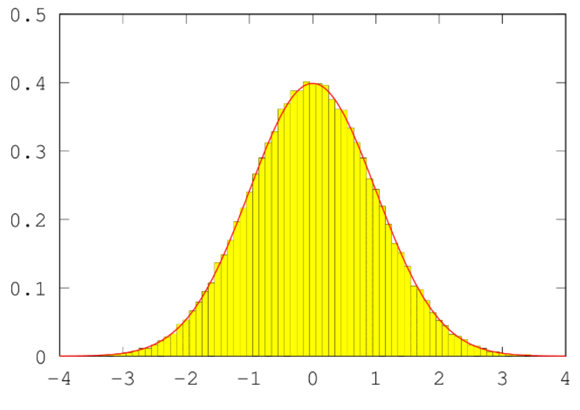
היסטוגרמות עם דליים ברוחב שונה: זה גורם להיסטוגרמת הצפיפות והתדירות להיות שונות.





חוק המספרים הגדולים והיסטוגרמות

בהסתברות גבוהה, היסטוגרמת הצפיפות של מספר גדול של דגימות הוא קירוב טוב לגרף הPDF. דוגמה: 100,000 דגימות מתוך התפלגות נורמלית. רוחב דלי 0.1:



סטנדרטיזציה: עבור מ"מ X עם תוחלת μ וסטיית תקן σ: $Y=\frac{X-\mu}{\sigma}$. ל-Y יש תוחלת (ממוצע) 0, ושונות (וסטיית תקן) 1.

משפט הגבול המרכזי

עבור ... X_1, X_2 מ"מ בת"ל בעלי אותה התפלגות עם תוחלת μ וסטיית תקן σ, נגדיר:

$$S_n = X_1 + X_2 + \cdots + X_n = \sum_{i=1}^{n} X_i\,, \qquad \overline{X_n} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{S_n}{n}$$

מתכונות תוחלת ושונות נקבל:

$$E(S_n) = n\mu, \qquad Var(S_n) = n\sigma^2, \qquad \sigma_{S_n} = \sigma\sqrt{n}$$

$$E(\overline{X_n}) = \mu, \qquad Var(\overline{X_n}) = \frac{\sigma^2}{n}, \qquad \sigma_{\bar{X}_n} = \frac{\sigma}{\sqrt{n}}$$

סטנדרטיזציה:

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{\overline{X_n} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

ומתקיים (עבור n גדול):

$$\overline{X_n} \approx N\left(\mu, \frac{\sigma^2}{n}\right), \qquad S_n \approx N(n\mu, n\sigma^2), \qquad Z_n \approx N(0,1)$$

במילים: $\overline{X_n}$ הוא בערך התפלגות נורמלית, עם אותו ממוצע כמו X ושונות קטנה יותר. S_n הוא בערך נורמלי. אז אחרי סטנדרטיזציה, שניהם בערך נורמלי סטנדרטי.

$$\lim_{n\to\infty} F_{Z_n}(z) = \Phi(z)$$

הסתברות נורמלית סטנדרטית

$$P(|Z| < 1) = 0.68, \qquad P(|Z| < 2) = 0.95, \qquad P(|Z| < 3) = 0.997$$

מכאן נובע גם:

$$P(Z < 1) = 0.84, \qquad P(Z < 2) = 0.977, \qquad P(Z < 3) = 0.999$$

נוכיח עבור הראשון:

אנחנו יודעים ש $P(|Z| < 1) = 0.68$, כלומר $P(Z < -1) + P(Z > 1) = 0.32$. אז:

$$P(Z < 1) = P(|Z| < 1) + P(Z < -1) = 0.84$$

התחום $P(Z < -1)$ נקרא **הזנב השמאלי**. $P(Z > 1)$ נקרא הזנב הימני.

שימוש ב*CLT*

דוגמה

נטיל מטבע הוגן 100 פעמים. נחשב את ההסתברות שיצא יותר מ55 עץ. נגדיר X_i את האינדיקטור להטלה ה-i. S הוא סכום האינדיקטורים.

$$E(X_i) = 0.5, \qquad Var(X_i) = 0.25, \qquad E(S) = 50, \qquad Var(S) = 25, \qquad \sigma_S = 5$$

הCLT אומר שהסטנדרטיזציה של S היא בערך $N(0,1)$. נחשב את $P(S > 55)$:

$$P(S>55)=P\left(\frac{S-50}{5}>1\right)\approx P(Z>1)=0.16$$

אם נבדוק מה ההסתברות ליותר מ-220 עץ ב-400 הטלות, נקבל 0.025. נשים לב: $\frac{55}{100}=\frac{220}{400}$, אבל ההסתברות לראשון גדולה יותר. זה בגלל חוק המספרים הגדולים – ככל שיש יותר דגימות, ההסתברות להתרחק מהתוחלת קטנה.

דוגמה – סקרים

באופן כללי, עבור n אנשים, מרווח הטעות הוא $\pm\frac{1}{\sqrt{n}}$. נסביר את המשמעות ואת הקשר לCLT.

אם נשאל n אנשים לגבי העדפה בין A ל-B, הסקר הוא סדרה של מ"מ ברנולי. אחוז האנשים שמעדיפים את A הוא הממוצע $\bar{X}$.

לכל X_i:

$$E(X_i)=p,\sigma_{X_i}=\sqrt{p(1-p)}$$

אז הCLT אומר ש:

$$\bar{X}\approx N\left(p,\frac{\sigma}{\sqrt{n}}\right)$$

בהתפלגות נורמלית, 95% מההסתברות נמצאת בטווח 2 סטיות תקן מהממוצע. כלומר, ב95% מהסקרים של n אנשים, הממוצע $\bar{X}$ יהיה בטווח $p\pm\frac{2\sigma}{\sqrt{n}}$. מכיוון ש $\sigma^2=p(1-p)\leq 0.25$, מתקיים שלכל ערך של p, $\sigma\leq 0.5$. כלומר $\frac{2\sigma}{\sqrt{n}}\leq\frac{1}{\sqrt{n}}$. סטטיסטיקאי שכיחותני יגיד שהאינטרוול $\bar{X}\pm\frac{1}{\sqrt{n}}$ הוא **הרווח סמך** (*confidence interval*) של 95% עבור p.

נשים לב: זה לא אומר שיש 95% שהערך האמיתי של p נמצא ברווח סמך. זה כמו לחשוב שאם לבדיקה יש 95% דיוק, והיא יצאה חיובית, אז יש הסתברות של 95% אחוז שאני חולה. זה נכון ש95% מהבדיקות יוצאות נכונות, אבל לא בהכרח 95% מהבדיקות החיוביות הן נכונות. (לדוגמה קיצונית, אם השכיחות של המחלה היא 95%, וכל הבדיקות יוצאות חיוביות, אז לבדיקות יש 95% דיוק אבל אם יצא לי חיובי זה לא אומר שיש לי את המחלה).

הרצאה

מצגת שיעור 3א, מצגת שיעור 3ב

1 תוחלת ושונות רציפה, חוק המספרים הגדולים, משפט הגבול המרכזי (מצגת 3א)

חוק המספרים הגדולים

"הממוצע של הרבה מדידות יהיה מדויק יותר ממדידה אחת". באופן פורמלי:

עבור ... X_1,X_2 מ"מ בת"ל בעלי אותה התפלגות עם תוחלת μ וסטיית תקן σ, נגדיר:

$$\overline{X_n}=\frac{X_1+X_2+\cdots+X_n}{n}=\frac{1}{n}\sum_{i=1}^{n}X_i$$

ואז לכל $\epsilon>0$, מתקיים:

$$\lim_{n\to\infty}P(|\overline{X_n}-\epsilon|<\epsilon)=1$$

תרגיל כיתה 1

יש לנו 100$, ואנחנו צריכים 1000$. יש לנו אפשרות להמר. אם אנחנו מהמרים k, בהסתברות p נרוויח k (כלומר סה"כ $2k$) ובהסתברות $1-p$ נפסיד k. נבחן שתי שיטות:

א) שיטה מקסימלית – כל פעם נהמר את כל מה שיש לנו (או מה שצריך כדי להגיע ל-1000)
ב) שיטה מינימלית – כל פעם נהמר משהו קטן (נגיד 5)

עבור p כלשהו, איזו שיטה עדיפה?

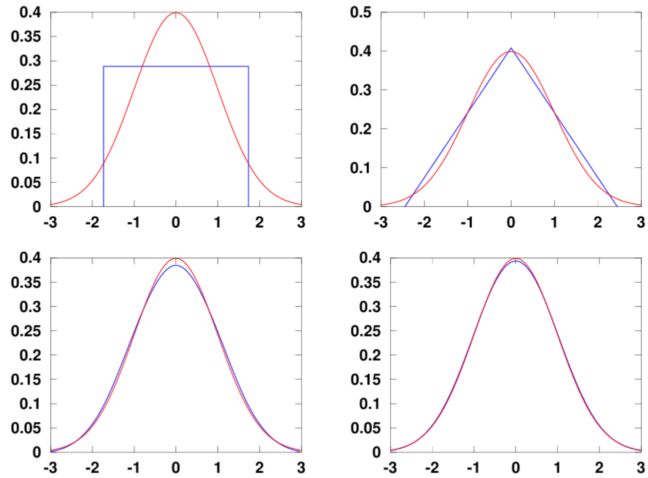
נחשב את התוחלת עבור הימור:

$$E(k) = pk + (1-p)(-k) = pk + (-k + pk) = pk - k + pk = 2pk - k$$

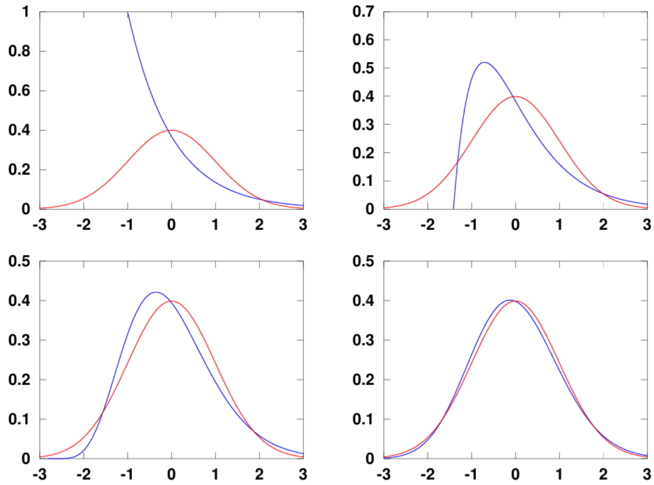
לדוגמה עבור $p = 0.45$, $E(k) = 0.9k - k = -0.1k$. אנחנו מצפים להפסיד 10% בכל הימור. כלומר עדיף להמר כמה שפחות פעמים. הגישה המקסימלית עדיפה. עבור $p = 0.8$, $E(k) = 0.6k$ אז נעדיף להמר כמה שיותר. הגישה המינימלית.

דוגמה 2

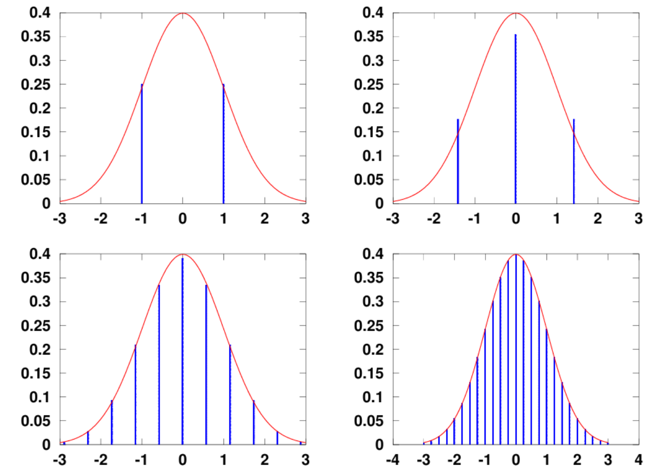
ממוצע של n מ"מ אחידים אחרי סטנדרטיזציה,
עבור $n = 1,2,4,12$



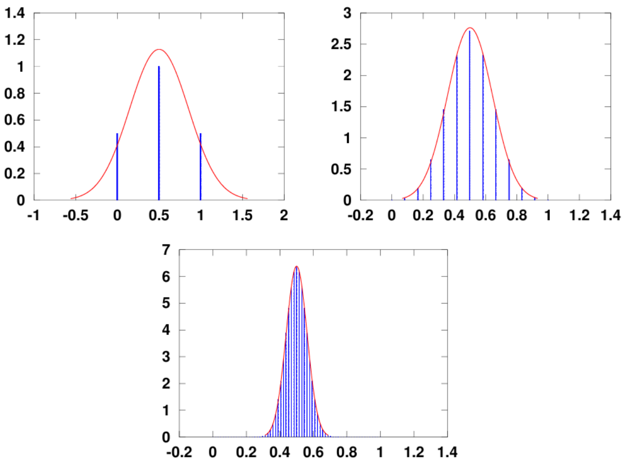
ממוצע של n מ"מ מעריכיים אחרי סטנדרטיזציה,
עבור $n = 1,2,8,64$



ממוצע של n מ"מ ברנולי אחרי סטנדרטיזציה,
עבור $n = 1,2,8,64$



ממוצע של n מ"מ ברנולי בלי סטנדרטיזציה,
עבור $n = 4,12,64$



הסטנדרטיזציה גורמת לפיזור להיות קרוב לנורמלי. אם לא עושים סטנדרטיזציה, הפיזור קטן.

תרגיל כיתה 2

בעזרת זריקות של קוביית 10 פאות בלבד, נרצה לייצר דגימה יחידה מהתפלגות נורמלית סטנדרטית.

נזרוק את הקוביה n פעמים, וזה נותן לנו n מ"מ מקריים בת"ל עם תוחלת $\mu = 5.5$ ושונות $\sigma^2 = 8.25$.

אם נעשה סטנדרטיזציה לממוצע $\bar{X}_n$, נקבל מ"מ מהתפלגות ששואפת לנורמלי סטנדרטי.

תרגיל כיתה 3

בהצבעה, נניח ש 50% תומכים ב-A, 25% ב-B, ו-25% מפוזרים בין השאר. סקר שאל 400 אנשים אקראיים במי הם תומכים. מה ההסתברות שלפחות 55% מהם תומכים ב-A?

נחשב: יהי a אחוז האנשים שתומכים ב-A. כלומר, ממוצע של 400 מ"מ שמתפלגים $Ber(0.5)$. לכל אחד מהם יש $\mu = 0.5, \sigma^2 = 0.25$, אז $\sigma_a^2 = \frac{0.25}{400}$, $\sigma_a = 0.025$. משפט הגבול המרכזי אומר שהוא שואף להתפלגות נורמלית. נעשה סטנדרטיזציה: $Z \approx \frac{a-0.5}{0.025}$.

כלומר:

$$P(a \geq 0.55) \approx P\left(\frac{a-0.5}{0.025} \geq \frac{0.55-0.5}{0.025}\right) = P\left(Z \geq \frac{0.55-0.5}{0.025}\right) = P(Z \geq 2) \approx 0.025$$

2 מבוא לסטטיסטיקה (מצגת 3ב)

מתי אפשר להסיק?

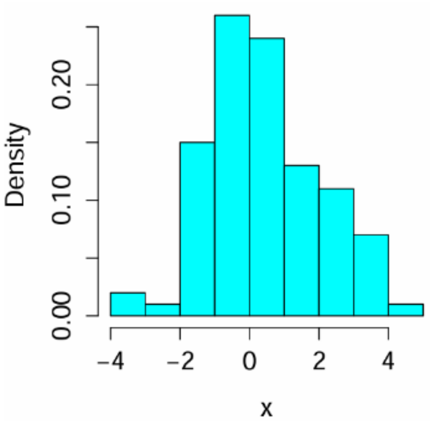
דוגמה 1

T	T	T	H	H	T	H	H	H	T
H	T	H	T	H	T	H	T	H	T
H	T	T	T	H	T	T	T	T	H
H	T	T	H	H	T	H	H	T	H
T	T	H	H	H	H	T	H	T	H
T	T	T	H	T	H	H	H	H	T
T	T	T	H	H	H	T	T	T	H
H	H	H	H	H	H	H	T	T	T
H	T	H	H	T	T	T	H	H	T
H	T	H	H	H	T	T	T	H	H

בהינתן סדרת ההטלות, האם נוכל להסיק שהמטבע הוגן?

אם יצא 50-50, נרצה כנראה לומר שהוא הוגן.
אם יצא 30-70, נרצה להגיד שלא.
מה אם יצא 55-65? 49-51? תמיד יש "רעש".
נרצה ללמוד מתי מחליטים לקבל היפותזה ומתי לא.

דוגמה 2



האם המשתנה המקרי הזה נחשב מתפלג נורמלי סטנדרטי?

הממוצע הוא 0.38, סטיית תקן 1.59

האם זה מספיק קרוב ל-0, 1?

שלבי הסקה סטטיסטית

1) איסוף דאטא. (בדיקות, ניסויים, סקרים). צריך לקחת בחשבון אי תלות, סיבתיות... לדוגמה: בניסוי שבודק תרופה ניסיונית, יש לאדם חולה הסתברות יותר גבוהה להסכים להשתתף.
2) תיאור הסטטיסטיקה. לנתח את הדאטא לחלקים רלוונטיים.
3) הסקה סטטיסטית. הסקת מסקנות לפי הסטטיסטיקה.

סטטיסטי

הגדרה: **סטטיסטי** (*statistic*) הוא כל דבר שאפשר לחשב מהדאטא בלבד.

- ערך שמחושב מהדאטא – ממוצע, סטיית תקן.
- אינטרוול או טווח של דאטא (לדוגמה $\bar{X} \pm s$).

סטטיסטי הוא בעצמו מ"מ.

דוגמה 3

נניח שזמן חיים של נורה מתפלג לפי $Exp(\lambda)$. כדי לבדוק את זה, מדדנו את זמן החיים של 5 נורות $X_1, \dots, X_5$.

הממוצע $\frac{X_1+\cdots+X_5}{5}$ הוא סטטיסטי, כי הוא מחושב ישירות מהדאטא.

התוחלת $(1/\lambda)$ היא לא סטטיסטי, כי אנחנו לא יכולים לחשב את למדא – רק להעריך אותו.

תרגול

תרגיל 1

נתון מטבע עם הסתברות p לעץ. נרצה לשער את p.

נטיל אותו n פעמים. $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ (סכום האינדיקטורים). זה הניחוש שלנו. לפי חוק המספרים הגדולים, ככל ש-n גדל זה שואף ל-p.

נרצה לחשב את הטעות: עבור $p = 0.5, n = 256, \epsilon = 0.1$, מה ההסתברות: $P(|\bar{X} - p| \geq \epsilon)$.

נחשב את השונות:

$$\frac{\sigma^2}{n} = Var(\bar{X}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) =^{\text{א}} \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n} = \frac{1}{1024}$$

א – רק אם הם בת"ל

נעשה חסם צ'בישב:

$$P(|\bar{X} - p| \geq \epsilon) \leq \frac{Var(\bar{X})}{0.1^2} = \frac{\frac{1}{1024}}{0.01} = 0.097$$

ננסה את משפט הגבול המרכזי. כי לפי המשפט, $\bar{X}$ מתפלג בערך נורמלי, עם $\mu = p = 0.5$:

נעשה סטנדרטיזציה על המשלים:

$$P(|\bar{X} - p| \leq 0.1) = P(0.1 \leq \bar{X} - p \leq 0.1) = P\left(\frac{0.1}{\frac{1}{32}} \leq \frac{\bar{X} - p}{\frac{1}{32}} \leq \frac{0.1}{\frac{1}{32}}\right) = P(3.2 \leq Z \leq 3.2)$$
$$= P(Z \leq 3.2) - P(Z \geq 3.2) = \Phi(3.2) - \Phi(-3.2) = 2\Phi(3.2) = 0.9986$$

תרגיל 2

במפעל, משקל האריזות מתפלג $Exp\left(\frac{3}{5}\right)$. נרצה לדעת מה ההסתברות שהמשקל של 900 אריזות יהיה בין 1425 ל-1600.

$$X_i = (\text{weight of package i}) \sim Exp\left(\frac{3}{5}\right), \qquad E(X) = \frac{5}{3}, \qquad Var(X_i) = \frac{25}{9}, \qquad \bar{X} = \sum_{i=1}^{900} X_i$$

$$E(\bar{X}) = E\left(\sum_{i=1}^{900} X_i\right) = \sum_{i=1}^{900} E(X_i) = 900 \cdot \frac{5}{3}$$

$$Var(\bar{X}) = Var\left(\sum_{i=1}^{900} X_i\right) = \sum_{i=1}^{900} Var(X_i) = \frac{25}{9} \cdot 900$$

$$P(1425 \le \bar{X} \le 1600) = p\left(\frac{1425 - 900 \cdot \frac{5}{3}}{\sqrt{\frac{25}{9} \cdot 900}} \le \frac{\bar{X} - 900 \cdot \frac{5}{3}}{\sqrt{\frac{25}{9} \cdot 900}} \le \frac{1600 - 900 \cdot \frac{5}{3}}{\sqrt{\frac{25}{9} \cdot 900}}\right) = P(-1.5 \le Z \le 2)$$
$$= \Phi(2) - \Phi(-1.5) \approx 0.91$$

תרגיל 3

שיכור הולך. כל דקה צעד אחד קדימה או אחורה בהסתברות חצי כל אחד, כל צעד 50 ס"מ. אחרי שעה, מה ההסתברות שהוא התרחק 10 מ' (לאיזשהו כיוון)?

$$X_i \sim \begin{cases} 50, & \frac{1}{2} \\ -50, & \frac{1}{2} \end{cases}, \qquad E(X) = 0, \qquad Var(X_i) = E(X^2) - \underbrace{E(X)^2}_{0} = 50^2 \cdot \frac{1}{2} + (-50)^2 \cdot \frac{1}{2} = 50^2$$

$$\bar{X} = \sum_{i=1}^{60} X_i\,, \qquad E(\bar{X}) = 0, \qquad Var(\bar{X}) = Var\left(\sum_{i=1}^{60} X_i\right) = 60 \cdot 50^2$$

$$P(|\bar{X}| \ge 1000) = 1 - P(|\bar{X}| < 1000)$$

$$P(|\bar{X}| < 1000) = P(-1000 < \bar{X} < 1000) = P\left(-\frac{1000}{\sqrt{60 \cdot 50^2}} < Z < \frac{1000}{\sqrt{60 \cdot 50^2}}\right) = 2\Phi(2.581) - 1$$
$$\approx 0.9901$$

תרגיל 4

דומה לתרגיל 3, אבל: $X_i \sim \begin{cases} 50, & \frac{2}{3} \\ -50, & \frac{1}{3} \end{cases}$, כלומר:

$$E(X) = 50 \cdot \frac{2}{3} - 50 \cdot \frac{1}{3} = \frac{50}{3}, \qquad Var(X_i) = E(X^2) - E(X)^2 = \frac{5000}{9}$$

$$\bar{X} \sim N\left(\frac{50 \cdot 60}{3}, 60 \cdot \frac{5000}{9}\right)$$

תרגיל 5

נניח שיש מ"מ X עם תוחלת μ, $\sigma^2 = Var(X) = 25$. $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. מה n צריך להיות כדי שיתקיים:

$$P(|\bar{X} - \mu| < 1) = 0.95$$

$$Var(\bar{X}) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{n \cdot \sigma^2}{n^2} = \frac{\sigma^2}{n} = \frac{25}{n}$$

$$P(-1 < \bar{X} - \mu < 1) = P\left(-\frac{1}{\frac{5}{\sqrt{n}}} < \frac{\bar{X} - \mu}{\frac{5}{\sqrt{n}}} < \frac{1}{\frac{5}{\sqrt{n}}}\right) = P\left(-\frac{\sqrt{n}}{5} < Z < \frac{\sqrt{n}}{5}\right) = 0.95$$

עבור משתנה $Z \sim N(0,1)$, מתקיים $P(-1.96 < Z < 1.96) \approx 0.95$ (נקבל כנתון). כלומר:

$$\frac{\sqrt{n}}{5} = 1.96 \quad \Longrightarrow \quad n \approx 96$$