סיכום מבוא לסטטיסטיקה והסתברות - קורס מבוא למדעי הנתונים

כפיר גולדפרב

שיעור ראשון - *מבוא לסטטיסטיקה והסתברות:*

בחבילת קלפים יש 52 קלפים רגילים ו-2 ג'וקרים (54 סה"כ), מה ההסתברות שאחרי ערבוב שרירותי של כל הקלפים יהיו 2 ג'וקרים בראש החבילה?

- בהנחה שהם דומים (לא ניתן להבדיל בניהם) $\frac{2}{54}\cdot\frac{1}{53}$.
- בהנחה שהם קלפים שונים: $\frac{1}{54}\cdot\frac{1}{53}+\frac{1}{54}\cdot\frac{1}{53}$.

אם אטיל קוביה איסוף פעמים הסיכוי שייצא מה שרציתי חחלך ויגדל לפי הנוסחה: $P(A)=\lim_{k\to\infty}\frac{kA}{k}$.

כאשר A הוא מאורע ו-P הוא הסתברות כלומר $P(A)$ פירושו הוא מה ההסתברות שמאורע A יקרה?

אקסיומות ההסתברות:

1. $P(A)\geq 0$, אנו תתופסים כי הסתברות של מאורע הוא מספר ממשי בין 0 ל-1.
2. $P(\Omega)=1$, אנו מצפים שלפחות אחד מהאירועים הבסיסיים שבמרחב המדם יתקיים תמיד – בכל ניסוי שנערוך, Ω – מרחב המדגם, קבוצת כל התוצאות האפשרויות בניסוי, למשל בהטלת מטבע יש שני אפשרויות עץ / פלי.
3. $P(A)^c=1-P(A)$, הסתברות שמאורע לא יתקיים שווה ל-1 פחות המשלים – ההסתברות שכן יקרה.
4. $P(A\cup B)=P(A)+P(B)-P(A\cap B)$, כאשר המאורעות זרים מתקיים $A\cap B=\phi$.

דוגמה:

בקוביה , A - הסיכוי לקבל 2 או 3, B - הסיכוי לקבל מספר זוגי.

כלומר $P(A)=\frac{1}{3}$, $P(A)=\frac{1}{2}$, הסיכוי לקבל מספר שיצא בשני הקבוצות $P(A\cup B)=\frac{1}{3}+\frac{1}{2}-\frac{1}{6}=\frac{4}{6}$.

5. חיסור / הפרש מאוררעות: המאורע A פחות המאורע B – הוא המאורע שכולל את כל התוצאות האפשרויות של מאורע A ולא של מאורע B, סימון: $A\backslash B$.

דוגמה:

נתון: $A=\{2,3,4,5\}, B=\{3,4,6\}, A\backslash B=\{2,5\}$.

שאלה:

בתהליך ייצור של מפעל קיימים 2 סוגים של פגמים: A ו- B, הסיכוי ש-A יתרחש בתהליך הייצור הוא 0.1, הסיכוי ש-B יתרחש בתהליך הייצור הוא 0.2, הסיכוי שגם A וגם B יתרחשו הוא 0.05,

לסיכום נתון: $P(A) = 0.1, P(B) = 0.2, P(A \cap B) = 0.05$.

חשבו את ההסתברויות:

א. לפחות פגם אחד.
ב. יש פגם A ואין פגם B.
ג. אין פגם בכלל.
ד. יש בדיוק פגם אחד.

פתרונות:

א. נחשב את $P(A \cup B) = 0.1 + 0.2 - 0.05 = 0.25$.
ב. נחשב את $P(A) - P(A \cap B) = 0.1 - 0.05 = 0.05$,
ניתן גם לחשב את $P(A \cup B) - P(B) = 0.25 - 0.2 = 0.05$.
ג. נחשב את $\overline{P(A \cup B)}$ כלומר את $1 - P(A \cap B)$, נקבל $1 - 0.25 = 0.75$.
ד. נחשב את $P(A \cup B) - P(A \cap B) = 0.25 - 0.05 = 0.2$.

תרגיל:

מושכים מחבילה 3 קלפים, חשבו את ההסתברות:

א. כל הקלפים הם לב:
בלי חזרות $\frac{11}{50} \cdot \frac{12}{51} \cdot \frac{13}{52}$, עם חזרות: $\left(\frac{13}{52}\right)^3$.
ב. אף קלף אינו לב: $\frac{2}{52} \cdot \frac{38}{52} \cdot \frac{39}{52}$.
ג. כל קלפים הם אסים: $\frac{2}{52} \cdot \frac{3}{52} \cdot \frac{4}{52}$.

שיעור שני – *המשך מבוא לסטטיסטיקה והסתברות:*

הסתברות מותנת:

התסברות מונתנת הוא הסתברות של מאורע כלשהו A, ונשאלת השאלה האם התנון של מאורע B משנה את את האינפורמציה ואז ההסתברות של A משתנה בהתאם.

חוק בייס - $\boldsymbol{Baye's\ Theorem}$:

חוק בייס או נוסחת בייס הוא תוצאה בתורת ההסתברות המאפשרת לחשב הסתברות מותנת של מאורע, הנוסחה:

$$P(A) = P(A\backslash B) = \frac{P(A \cap B)}{P(B)}$$

הסתברות של A בההתן B אינה בהכרח שווה להסתברות של B בהנתן A ולכן אינה סימטרית.

דוגמה:

הסתברות לקבל בקוביה 6: $P(A) = \frac{1}{6}$, הסתברות לקבל בקוביה מספר זוג: $P(B) = \frac{3}{6} = \frac{1}{2}$,

אם אומרים לי שהתוצאה יצאה זוגית אז ההסתברות שקבל 6 הוא $\frac{1}{3}$, כי מרחב המדגם שלי השתנה רק למספרים זוגיים.

$$P(A\backslash B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3}$$

נשים לב כי $P(A \cap B) = P(B \cap A)$,

ניקח את ההסתברות המותנת של B בהנתן A לפי נוסחת בייס: $P(B\backslash A) = \frac{P(B\cap A)}{P(A)}$

ניקח את ההסתברות המותנת של A בהנתן B לפי נוסחת בייס: $P(A\backslash B) = \frac{P(A\cap B)}{P(B)}$

מכיוון שמתקיים $P(A \cap B) = P(B \cap A)$ נקבל:

$$P(B) \cdot P(A\backslash B) = P(A) \cdot P(B\backslash A)$$

ולכן חוק בייס אומר גם:

$$P(A\backslash B) = \frac{P(B\backslash A) \cdot P(A)}{P(B)}$$

נשנה בחוק בייס את: $\mathrm{B} = \mathrm{E}, \mathrm{A} = \mathrm{H}$,

H – השארה/היפוטזה – מה שאני מאמין בו.

E – $evidence$, עדויות – מה שקורה בפועל.

למשל: אם אני מאמין שאצליח בתואר (היפוטזה), אך אם נכשלתי במבחנים $(evidence)$ זה יכול לגרום לי לשנות את ההיפוטזה שלי.

$$P(H\backslash E) = \frac{P(E\backslash H) \cdot P(H)}{P(E)}$$

$$P(A) = \sum P(A\backslash b_i)$$

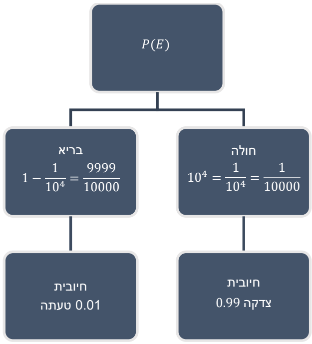
- הערה: $P(E\backslash H) \neq P(H\backslash E)$.

חוק בייס הופך את ההסתברות למשהו סובייקטיבי ופחות אובייקטיבי, כלומר ההסתברות הופכת להיות ממה הסיכוי שזה יקרה לכמה אני מאמין שהדבר הזה יקרה והאם יקרה מאורע שישנה את האמונה שלי.

תרגיל:

יש מחלה באוכליסיה, על כל 10^4 אזרחים יש 10 חולים, הבדיקה למחלה זו מדייקת בכ-99%,

נגדיר: H – האם אני חולה, E – האם הבדיקה חיובית.



הסיכוי שהדיקה חיובית כאשר אני לא יודע אם אני חולה או בריא:

$$P(E) = \frac{9999}{10000} \cdot 0.01 + \frac{1}{10000} \cdot 0.99 = 0.010098$$

הסיכוי שהבדיקה חיובית כאשר אני חולה $P(E \backslash H) = 0.99$,

הסיכוי שאני חולה $P(H) = 10^{-4}$,

לפפי נוסחת בייס:

$$P(H \backslash E) = P\left(\text{בדיקה חיובית}\backslash\text{חולה}\right) = \frac{0.99 \cdot 10^{-4}}{0.010098} \approx 0.9\%$$

- מונחים חשובים בהסתברות:
  - אפריורי – מידע שאני יודע לפני הנסיון.
  - פוסט-אפריורי – מידע שאני יודע אחרי הניסיון.

שיעור שלישי – *המשך מבוא לסטטיסטיקה והסתברות:*

$$P(H\backslash E) \propto P(E\backslash H) \cdot P(H)$$

כאשר $\propto$ אומר באופ פרופרוציונאלי – אם אגף ימין גדל אזי גם אגף שמאל גדל ולהפך בהתאם.

- הערה: $P(H\backslash E) + P(E\backslash H) = 100\%$, כיוון שהם משלימים אחד את השני.

חוק ההסתברות השלמה:

$$P(A) = P(A\backslash b_1) \cdot P(b_1) + P(A\backslash b_2) \cdot P(b_2) + \cdots + P(A\backslash b_n) \cdot P(b_n)$$

$$= \sum_{b\in B} P(A\backslash b) \cdot P(b)$$

משתנה אקראי:

לוקח תוצאה ומצמיד לה מספר אקראי $outcome \to nubmer$.

לדוגמה:

עבור הטלת מטבע: $\mathrm{x} = \begin{cases} 0, \text{עץ} \\ 1, \text{פלי} \end{cases}$ – תוצאה של x אינה קבועה.

כלומר משתנה רגיל מקבל ערך יחיד, משתנה אקראי יכול לקבל מספר משתנים.

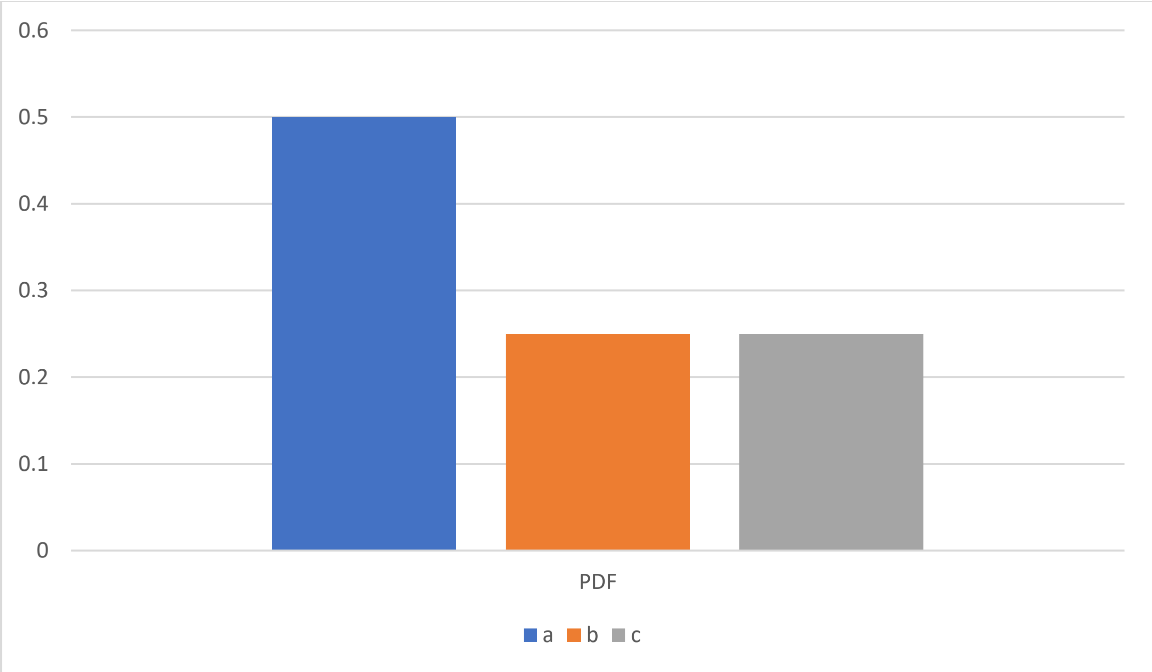
1. בדיד: מקבל מספר ערכים מוגבל של מספרים שלמים $\mathbb{N}$.
2. רציף: מספר בין 0 ל-1 כלומר אינסוף ערכים.

כאשר יש מרחב רציף אז הסיכוי לקבל ערך מסויים הוא 0 כי יש איסוף אפשרויות.

$Probability\ Density\ Function - PDF$ (פונקצית צפיפות הסתברות) – מתאר את ההסתברות לקבל ערך מסויים.

האינטגל של $PDF = 1$,

לדוגמה:



$$\int_{-\infty}^{\infty} PDF = 0.5 + 0.25 + 0.25 = 1$$

כאשר מטילים מטבע 3 פעמים, מה הסיכויים שלא ייצא פלי?

$$P(x=0)=\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}=\frac{1}{8}\,,P(x=3)=\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}=\frac{1}{8}$$

מה הסיכוי שייצא פעם אחת?

$$P(x=1)=\left(\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}\right)\cdot 3=\frac{3}{8},P(x=2)=\left(\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}\right)\cdot 3=\frac{3}{8}$$

בציור:

האפשרויות:

THH	HHH
TTH	HHT
THT	HTT
HTH	TTT

x – יצא פלי (H).

1. $x=0\to\frac{1}{8}$.
2. $x=1\to\frac{3}{8}$.

תוחלת של משתנה אקראי $P(x)\cdot x$ - $Expected\ Value$ (הערך שמציף מלקבל אחרי זמן).

הסתברות של מה שייצא ב-x הכי הרבה פעמים אחרי מספר פעמים (סימון: $E(x)$).

לדוגמה:

ההסתברות שייצא סכום 7 בהטלת 2 קוביות:

$$1\cdot\frac{1}{6}+2\cdot\frac{1}{6}+3\cdot\frac{1}{6}+4\cdot\frac{1}{6}+5\cdot\frac{1}{6}+6\cdot\frac{1}{6}=3.5$$

דוגמה:

משחקים משחק, מוציאים קרף באקראי מהחפיסה, אם הקלף הוא צורה (מלך, מלכה, נסיך), אתה מקבל 5 ש"ח, אם הקלף הוא מספר אתה משלם 2 ש"ח.

האם שווה לשחק או שלא?

$$x=\begin{cases}5, & \frac{12}{52}\\ 2, & -\frac{40}{52}\end{cases}=E(x)=\sum P(x)\cdot x$$

התוחלת של x - $5\cdot\frac{12}{52}-2\cdot\frac{40}{52}=-\frac{5}{13}$, מכיוון שהתוצאה שלילית לא שווה לשחק.

התפלגות אחידה בדידה:

התפלגות בה לכל האיברים בקבוצה סופית (בדידה) יש הסתברות שווה, כלומר לכל אחד מ-n האיברים שיכולים להתקבל יש הסתברות שווה שיתקבלו שהיא $\frac{1}{n}$.

התפלגות רציפה:

על טווח איסופי, רציף של ערכים כמו גובה, משקל, אורך חיי אדם וכו'.

שיעור רביעי – *המשך מבוא לסטטיסטיקה והסתברות:*

סטיית תקן:

מדד סטטיסטי לתיאור הפיזור של הנתונים המספריים סביב הממוצע שלהם, התלוי במרחק שלהם מן הממוצע, סטייית התקן נמדדת באותן היחידות כמו הנתונים עצמם, והוא שווה לשורש הריבועי של השונות ולכן היא חיובית, ושווה לאפס רק כאשר כל הנתונים שווים זה לזה, סימון σ (סיגמה קטנה).

יש להבחין בין סטיית התקן המחושבת לכל הנתונים לבין סטיית התקן המדגמית המחושבת על המדגם (תת-קבוצה), ומשמשת רק למדידה של סטיית התקן של התת-קבוצה.

- x_i – הנתונים,
- $\overline{x}$ – הממוצע.

$$\sigma = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2}$$

לדוגמה:

נחשב את סטיית התקן בין המשכורות $10000,0,8000$:

$$\sigma = \sqrt{\frac{1}{3} \cdot ((10-6)^2 + (0-6)^2 + (8-6)^2)} = \sqrt{\frac{1}{3} \cdot (16 + 36 + 4)} = \sqrt{\frac{56}{3}} = 4.2$$

שונות:

תוחלת ריבועית – הסטיות מהתוחלת – נותן אינדוקציה על הפיזור והסיכון של פונקציית ההסתברות (PDF), באופן אינטואטיבי, השונות הוא הגודל החיובי התלוי במרחק (הריבועי), הממוצע של כל ערך ממוצע של כל הערכים, ערך שונות גבוה מעיד על פיזום רחב של משתנים, ערך נמוך מעיד על פיזור צר, שונות שווה לאפס אם כל הערכים שווים וזהים ומתרכזים בנקודה אחת.

$$var(x) = \frac{1}{2} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2$$

סטיית התקן המדגמית: $S = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \overline{x})}$

התפלגות נורמלית:

בהתפלגות נורמלית ניתן להמיר את ציוני התקן לאחוזונים, אם ידועים לנו הממוצע וסטיית התקן – נוסחת הקו הידועה ולכן אפשר לחשב בדיוק את השטח תחת העקומה עד לציון מסויים באמצעות חישוב האינטגרל, שטח זה הוא בעצם האחוזון הציון.

בהתפלגות נורמלית הממוצע של משתנים בלתי תלויים בעלי אותה התפלגות, מתכנס בהתפלגות להתפלגות המואמליץ, לכן משתמשים בה המון כאשר לוקחים ממוצע של משתנים רבים כגון הממוצע של אנשים באוכלוסיה, ממדים פרמטרים שונים ועוד.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{\frac{-x(x-\mu)^2}{2\sigma^2}}$$

כאשר μ = התוחלת, σ = סטיית התקן.

התפלגות ברנולי:

מתארת התפלגות בדידה של משתנה אקראי המקל ערך 0 או 1, כלומר זו התאמה למערכות בהן יש שני מצבים – הצלחה או כישלון. במקרה זה מקובל לסמן הצלחה באות p וכישלון בתור ההסתברות המשלימה $q = 1 - p$, תוחלת של משתחנה אקראי x המתפלג ברנולי הוא $E(x) = p$, שונות של משתנה אקראי הוא $var(x) = p - q = p(1 - q)$.