הסקה סטטיסטית שיעור 1

הקדמה חומר קריאה 1א, חומר קריאה 1ב

1 הסתברות מול סטטיסטיקה

סטטיסטיקה – שימוש בכלים הסתברותיים כדי להסיק מסקנות מתוך דאטא.

דוגמה להסתברות: בהינתן מטבע, אם נזרוק אותו 100 פעמים, מה ההסתברות לקבל עץ 60 פעמים או יותר? יש תשובה אחת (בערך 0.028444).

דוגמה לסטטיסטיקה: נתון מטבע, שלא ידוע אם הוא הוגן. כדי לבדוק, נטיל אותו 100 פעמים. נגיד שיצא עץ 60 פעמים. אנחנו רוצים להסיק מסקנה מהדאטא. יש מספר דרכים שונות לבצע את זה, מבחינת הצורה שתהיה למסקנה והחישובים שנעשה. יש מצבים שנגיע בדרכים שונות לתוצאות שונות.

נשים לב שבדוגמה הראשונה, התהליך המקרי ידוע (ההסתברות לעץ = 0.5), ואנחנו רוצים למצוא את ההסתברות לתוצאה מסויימת. בדוגמה השנייה, התוצאה ידועה (60 פעמים עץ) ואנחנו רוצים לגלות את התהליך המקרי שהיה (ההסתברות לעץ).

2 הסקה שכיחותנית מול בייסיאנית (Frequentist vs. Bayesian Interpretations)

יש שני גישות שונות ולפעמים סותרות בהסקה סטטיסטית: **בייסיאנית** ו**שכיחותנית**. הגישות מגיעות מפירושים שונים למשמעות של הסתברות.

שכיחותנים אומרים שהסתברות מודדת את **השכיחות של תוצאות שונות של ניסוי**. לדוגמה, לומר שלמטבע יש הסתברות של 50% לעץ אומר שאם נזרוק אותו הרבה פעמים אז נצפה שבערך חצי מהפעמים יצאו עץ.

בייסיאנים אומרים שהסתברות היא רעיון מופשט שמודד את **מצב המידע שלנו או רמת אמונה** בטענה נתונה. בפועל, בייסיאנים לא נותנים ערך ספציפי להסתברות שמטבע ייצא עץ. אלא, הם מתייחסים לטווח של ערכים, כל אחד עם הסתברות מסויימת להיות אמיתי.

הגישה השכיחותנית נפוצה בתחומי ביולוגיה, רפואה, בריאות הציבור, ומדעי החברה. הגישה הבייסיאנית מתחזקת בעידן המחשבים החזקים ו-big data. היא שימושית במיוחד כשמכניסים דאטא חדש למודל סטטיסטי קיים. לדוגמה, בזמן אימון מודל זיהוי דיבור או פנים. כיום, משתמשים בשתי הגישות בצורה משלימה.

חזרה על מושגים בהסתברות: תורת הקבוצות, קומבינטוריקה.

נסמן nPk (פרמוטציות) את מספר הדרכים לבחור k איברים מתוך n, עם חשיבות לסדר. $nPk = \frac{n!}{(n-k)!}$.

נסמן nCk (תתי קבוצות) את מספר הדרכים לבחור k איברים מתוך n, בלי חשיבות לסדר. $nCk = \frac{nPk}{k!} = \frac{n!}{(n-k)!\cdot k!}$.

3 הסתברות מותנית – כלל הכפל

מההגדרה של הסתברות מותנית, נקבל את הנוסחה: $P(A \cap B) = P(A|B) \cdot P(B)$

דוגמה 1

נבחר שני קלפים מחפיסה. נגדיר מאורעות: S_1=הקלף הראשון יצא עלה, S_2=הקלף השני יצא עלה. נחשב את $P(S_2|S_1)$:

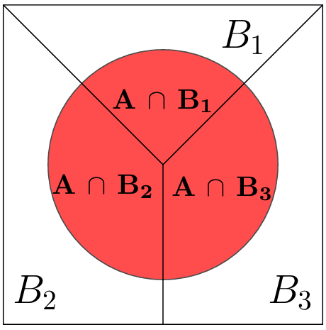
אפשר לספור ישירות. אם הקלף הראשון יצא עלה, נשארו 51 קלפים ו-12 מתוכם הם עלה, כלומר: $P(S_2|S_1) = 12/51$.

נחשב בעזרת הנוסחה: $P(S_1) = P(S_2) = 1/4\,, P(S_1 \cap S_2) = \frac{13\cdot 12}{52\cdot 51} = \frac{3}{51}$.

הערה: אולי מפתיע ש $P(S_2) = 1/4$, כי הקלף הראשון משפיע על השני. אבל לכל קלף יש את אותה הסתברות להיבחר בפעם הראשונה, אז זה משפיע על כל הקלפים באותה צורה. ולכן לכל הקלפים יש את אותה הסתברות להיבחר בפעם השנייה. בסה"כ:

$$P(S_2|S_1) = \frac{P(S_2 \cap S_1)}{P(S_1)} = \frac{3/51}{1/4} = \frac{12}{51}$$

4 נוסחת ההסתברות השלמה



עבור איברים A_i שמהווים חלוקה של העולם, מתקיים:

$$P(B) = \sum_i P(B|A_i)P(A_i) = \sum_i P(B \cap A)$$

אינטואיטיבית: אם העולם מחולק לשלושה חלקים:

אז ההסתברות של A היא סכום ההסתברויות של החיתוך של A עם כל אחד מהחלקים.

דוגמה 2

יש בכד 5 כדורים אדומים ו-2 ירוקים. נוציא 2 כדורים, ללא החזרה. מה ההסתברות שהשני אדום?

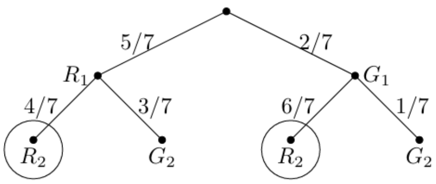
נחשב: $P(R_2|G_1) = 5/6\,, P(R_2|R_1) = 4/6$. מכיוון ש G_1, R_1 המהווים חלוקה של Ω, מתקיים:

$$P(R_2) = P(R_2|R_1)P(R_1) + P(R_2|G_1)P(G_1) = \frac{4}{6}\cdot\frac{5}{7} + \frac{5}{6}\cdot\frac{2}{7} = \frac{30}{42} = \frac{5}{7}$$

באותו כד, משחק אחר: נוציא כדור. אם הוא ירוק, מוסיפים כדור אדום. אם הוא אדום, מוסיפים ירוק (ללא החזרה). ואז מוציאים עוד כדור. מה ההסתברות שהכדור השני אדום? החישוב זהה, רק נשנה את הערכים:

$$P(R_2) = P(R_2|R_1)P(R_1) + P(R_2|G_1)P(G_1) = \frac{4}{7}\cdot\frac{5}{7} + \frac{6}{7}\cdot\frac{2}{7} = \frac{32}{49}$$

5 עצים



אפשר להשתמש בעצים כדי לתאר מהלך של ניסוי: קודקוד הוא מאורע, והצלע המובילה אליו היא ההסתברות. לדוגמה, הניסוי המתואר בדוגמה הקודמת.
לפי כלל הכפל, ההסתברות לקבל קודקוד מסויים הוא מכפלת ההסתברויות במסלול המוביל אליו.

הרצאה

מצגת שיעור א1, מצגת שיעור ב1

1 טרמינולוגיה ודוגמאות (מצגת א1)

דוגמה 1

בפוקר, לכמה "ידיים" של 5 קלפים יש בדיוק זוג אחד? מה ההסתברות לקבל יד עם זוג יחיד?

דרך א: נחשב לפי מספר דרכים:

1. נבחר את הדרגה של הזוג (כלומר 1,2,... או K, Q...). יש $\binom{13}{1}$ דרכים.
2. נבחר 2 קלפים מהדרגה. $\binom{4}{2}$.
3. נבחר עוד 3 דרגות **שונות**. $\binom{12}{3}$.
4. נבחר קלף אחד מכל דרגה. $\binom{4}{1}^3$.

סה"כ: $1098240 = \binom{13}{1}\binom{4}{2}\binom{12}{3}\binom{4}{1}^3$. נחשב את מספר הדרכים ליד כללית של 5 קלפים: $\binom{52}{5} = 2598960$. אז ההסתברות לזוג יחיד: $1098240/2598960 = 0.42257$.

דרך ב: נחשב לפי פרמוטציות:

1. נבחר את המקומות ביחד שבהן הזוג יהיה. $\binom{5}{2}$.
2. נשים קלף אחד מתוך 52 במיקום הראשון של הזוג.
3. במיקום השני של, נשים קלף אחד מתוך ה-3 שמתאימים.
4. במיקום הראשון **שלא** של הזוג, נשים קלף אחד מתוך ה-48 שנשארו (ששונים מהזוג).
5. במיקום השני שלא של הזוג, נשים קלף מתוך ה-44 שנשארו.
6. במיקום האחרון, קלף אחד מתוך 40.

בסה"כ: $\binom{5}{2} \cdot 52 \cdot 3 \cdot 48 \cdot 44 \cdot 40 = 131788800$ דרכים (כאשר יש חשיבות לסדר). ויש $52 \cdot 51 \cdot 50 \cdot 49 \cdot 48 = 311875200$ דרכים לסדר יד של 5 קלפים כאשר יש חשיבות לסדר. אז ההסתברות:

$$131788800/311875200 = 0.42257$$

דוגמה 2

בכיתה של 50 תלמידים, יש 20 בנים (M) ו-25 עם עיניים חומות (B). עבור תלמיד שנבחר באופן מקרי ואחיד, מה טווח הערכים עבור $p = \mathbb{P}(M \cup B)$?

נשים לב ש: $25 \le |M \cup B| \le 45$. כלומר, $\frac{25}{50} \le p \le \frac{45}{50} \Longrightarrow 0.5 \le p \le 0.9$.

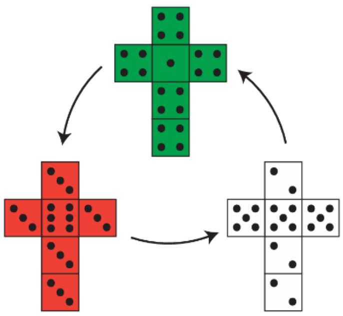
דוגמה 3

נזרוק קוביה עם 20 פאות, 9 פעמים. נבדוק אם כל הזריקות יצאו שונות. נחזור על זה 5 פעמים.

מרחב המדגם S הוא כל הרצפים של 9 מספרים בין 1 ל-20. כלומר $|S| = 20^9$. נגדיר את A להיות המאורע שבו יש שתי זריקות זהות. נרצה לחשב את A^c: $\frac{20!}{11!} = |A^c| = 20P9$. בסה"כ: $P(A) = 1 = P(A^c) = 1 - \frac{20!}{11!} = 0.881$.

דוגמה 4

נתונות הקוביות הבאות:



שני שחקנים יבחרו קוביה כל אחד. יזרקו פעם אחת והזריקה הגבוהה מנצחת. איזה קוביה כדאי לבחור?

נחשב את ההסתברות לכל מספר, בהינתן קוביה:

	Red die		White die		Green die	
Outcomes	3	6	2	5	1	4
Probability	5/6	1/6	3/6	3/6	1/6	5/6

אז ההסתברות לכל מאורע בכל סיטואציה (כלומר לדוגמה, אם נבחרו הקוביה האדומה והירוקה, ההסתברות שייצא 3 באדום ו-1 בירוק היא $\frac{5}{36}$), בכל מאורע הצבע מסמן את המנצח (שחור במקום לבן):

		White		Green	
		2	5	1	4
Red	3	15/36	15/36	5/36	25/36
	6	3/36	3/36	1/36	5/36
Green	1	3/36	3/36		
	4	15/36	15/36		

נשים לב שבכל זוג יש מנצח ברור, אבל זה לא טרנזיטיבי – אדום מנצח את לבן שמנצח את ירוק שמנצח את אדום. אין קוביה אחת הכי טובה.

דוגמה 5

בהינתן מטבע שלא הוגן: נזרוק אותה פעמיים, ונרצה לנחש מראש אם ייצא אותו דבר פעמיים או משהו שונה. נקרא להסתברות לעץ p. $q=1-p$. מתקיים:

		second flip H	second flip T
first flip	H	p^2	pq
	T	qp	q^2

כלומר, $P(same)=p^2+q^2, P(different)=pq+qp=2pq$.

באופן כללי מתקיים: עבור $a\neq b$, $(a-b)^2>0 \Longrightarrow a^2+b^2>2ab$.

עבור $0<p\neq q<1$, מתקיים: $p^2+q^2>2pq$ ולכן ההסתברות שהן זהות גדולה.

2 הסתברות מותנית, אי תלות, נוסחת בייס (מצגת 1ב)

תזכורת – הסתברות מותנית: עבור $P(B)\neq 0$, $P(A|B)=\frac{P(A\cap B)}{P(B)}$.

דוגמה 1

עבור 4 זריקות מטבע, נגדיר: A=לפחות 3 עץ, B=הזריקה הראשונה היא פלי. נחשב:

$$P(A|B)=\frac{P(A\cap B)}{P(B)}=\frac{\frac{1}{16}}{\frac{1}{2}}=\frac{1}{8},\qquad P(B|A)=\frac{P(A\cap B)}{P(A)}=\frac{\frac{1}{16}}{\frac{5}{16}}=\frac{16}{80}=\frac{1}{5}$$

דוגמה 2

בניסוי שנערך בארה"ב[1], משתתפים נשאלו: "סטיב הוא אדם ביישן ומופנם, אוהב לעזור אבל עם עניין מועט באנשים, או בעולם האמיתי. אדם מסודר וצנוע, יש לו צורך בסדר וארגון ותשומת לב לפרטים". מה ההסתברות שסטיב הוא ספרן, או חקלאי?

רוב האנשים ניחשו שסטיב הוא ספרן, למרות שיחס הספרנים לחקלאים בארה"ב הוא בערך 1/50. אחרי שגילו להם את זה, רוב האנשים החליפו את הניחוש לחקלאי.

הפער הוא זה: העובדה שלספרן נתון יש הסתברות גבוהה להיות בעל התכונות המתוארות, לא אומר שאדם עם התכונות הוא בהכרח ספרן. $P(A|B)\neq P(B|A)$.

תזכורת – חוק הכפל: $P(A\cap B)=P(A|B)\cdot P(B)$

נוסחת ההסתברות השלמה: $P(B)=P(B|A)P(A)+P(B|A^c)P(A^c)$. או בגרסה הכללית: עבור איברים A_i שמהווים חלוקה של העולם, מתקיים:

$$P(B)=\sum_i P(B|A_i)P(A_i)=\sum_i P(B\cap A)$$

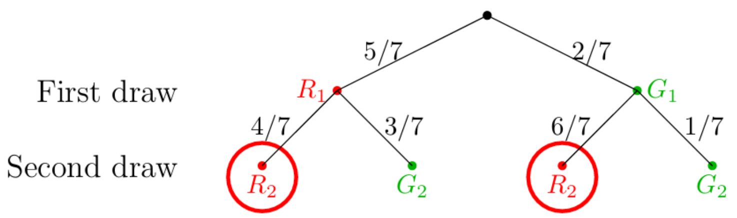
3 עצים

דוגמה 1

יש בכד 5 כדורים אדומים ו-2 ירוקים. נוציא כדור אקראי ונכניס במקומו כדור מהצבע השני. ואז נוציא כדור נוסף.

1. מה ההסתברות שהכדור השני אדום?
2. מה ההסתברות שהכדור הראשון היה אדום, בהינתן שהכדור השני היה אדום?

נשרטט את העץ המתאים:



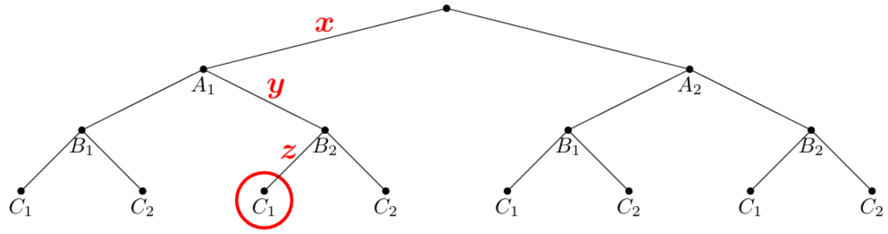
[1] Judgment under uncertainty: heuristics and biases by Tversky and Kahneman

לפי נוסחת ההסתברות השלמה, ונוסחת בייס:

$$P(R_2)=\frac{5}{7}\cdot\frac{4}{7}+\frac{2}{7}\cdot\frac{6}{7}=\frac{32}{49},\qquad P(R_2|R_1)=\frac{P(R_1\cap R_2)}{P(R_2)}=\frac{20/49}{32/49}=\frac{20}{32}$$

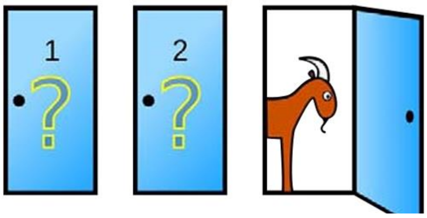
דוגמה 2

עבור העץ הנתון:



הצלע x מייצגת את ההסתברות $P(A_1)$. הצלע y מייצגת את ההסתברות $P(B_2|A_1)$. הצלע z מייצגת את ההסתברות $P(C_1|B_2\cap A_1)$. הקודקוד המסומן בעיגול מייצג את המאורע $C_1\cap B_2\cap A_1$.

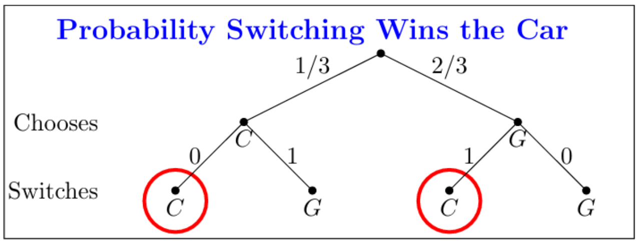
דוגמה 3 – חידת מונטי הול (מצגת 1ב)



נתונות 3 דלתות. מאוחרי אחת מהן יש רכב, מאחורי ה-2 האחרות יש עיזים. השחקן בוחר דלת אחת (בה"כ, דלת מספר 1). המארח פותח דלת אחרת (בה"כ, דלת 3) ומראה שיש מאחוריה עז. עכשיו הוא נותן לשחקן הזדמנות להחליף את הבחירה לדלת אחרת (כלומר לבחור בדלת 2). השאלה היא: האם כדאי לשחקן להחליף?

אם השחקן לא מחליף, ההסתברות למצוא את הרכב היא $1/3$.

נשרטט את העץ המתאים למצב שבו מחליפים:



אם השחקן מחליף, ההסתברות למצוא את הרכב היא $2/3$.

תזכורת – אי תלות:

מאורעות A,B ייקראו בלתי תלויים אם הידיעה שאחד מהם קרה לא משפיע על ההסתברות שהשני קרה. פורמלית:

$$P(A|B)=^1 P(A)\Leftrightarrow P(B|A)=^2 P(B)\Leftrightarrow P(A\cap B)=P(A)P(B)$$

1 – בהנחה ש $P(B)\neq 0$. 2 – בהנחה ש $P(A)\neq 0$.

דוגמה 4

נזרוק 2 קוביות ונתבונן במאורעות הבאים:

A = הקוביה הראשונה יצאה 3. B = הסכום הוא 6. C = הסכום הוא 7.

נחשב את ההסתברויות:

$$P(A)=\frac{1}{6},\qquad P(A|B)=\frac{1}{5},\qquad P(A|C)=\frac{1}{6}$$

כלומר, A ו-C בלתי תלויים. A ו-B תלויים. זה הגיוני כי אם אנחנו יודעים ש-B קרה, זה אומר שבזריקה הראשונה לא יכול לצאת 6. לעומת זאת, אם יודעים ש-C קרה, בזריקה הראשונה כל המספרים אפשריים.

לא תמיד האינטואיציה צודקת. כדי לדעת אי תלות, צריך לחשב לפי הגדרה.

נוסחת בייס

מאפשרת לנו "להחליף" הסתברות מותנית.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

תרגול

תרגיל 1

נתונים כד לבן וכד שחור. בכל כד יש כדורים אדומים וירוקים.
המשחק: נבחר את אחד הכדים, ונוציא ממנו כדור אחד באקראי. אם נבחר כדור אדום – ננצח. איזה כד כדאי לבחור?

	לבן	שחור
אדום	3	5
ירוק	4	6

בכד הלבן, ההסתברות לכדור אדום: 3/7. בכד השחור, 5/11.

מכיוון ש $\frac{3}{7} < \frac{5}{11}$, נבחר בכד השחור.

נשנה את המצב:

	לבן	שחור
אדום	9	6
ירוק	5	3

גם פה עדיף את הכד השחור.

אבל, אם נאחד את שני המקרים:

	לבן	שחור
אדום	12	11
ירוק	9	9

מתקיים ש $\frac{12}{21} > \frac{11}{20}$, כלומר עדיף את הכד הלבן.

זה נקרא **פרדוקס סימפסון**.

תרגיל 2 – פרדוקס יום ההולדת

בהינתן כיתה עם n אנשים (נניח $n \leq 365$), מה ההסתברות שיש שניים מהם עם אותו יום הולדת?

נחשב את המשלים: כדי שלא יהיו שניים עם אותו יום הולדת, לראשון יש 365 אפשרויות, לשני יש 364, וכו'. כלומר:

$$P(2 \text{ with same birthday}) = 1 - \frac{\frac{365!}{(365-n)!}}{365^n}$$

n	P
4	0.016
16	0.284
23	0.507
32	0.753
56	0.988

תרגיל 3 – הסתברות מותנית

נתון ש: 10% מעמד גבוה, 40% מעמד ביניים, 50% מעמד נמוך.
ונתונה הטבלה:

	$\boldsymbol{U_2}$	$\boldsymbol{M_2}$	$\boldsymbol{L_2}$
$\boldsymbol{U_1}$	0.45	0.48	0.07
$\boldsymbol{M_1}$	0.05	0.70	0.25
$\boldsymbol{L_1}$	0.1	0.50	0.49

כלומר, $P(U_2|U_1) = 0.45$

נחשב בנוסחת ההסתברות השלמה:

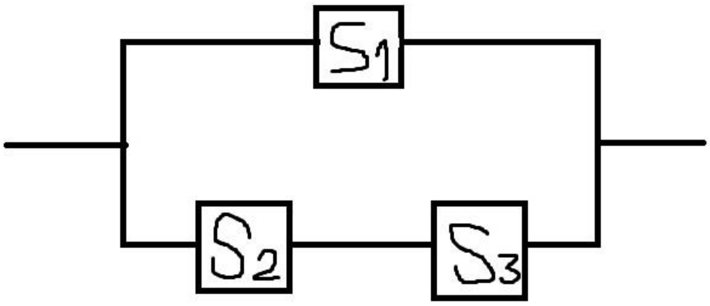
$$P(U_2) = P(U_2|U_1)P(U_1) + P(U_2|M_1)P(M_1) + P(U_2|L_1)P(L_1) = 0.45 \cdot 0.1 + 0.05 \cdot 0.4 + 0.01 \cdot 0.5 = 0.07$$

וגם:

$$P(U_1) = \frac{P(U_2|U_1)P(U_1)}{P(U_2)} = \frac{0.45 \cdot 0.50}{0.07} = 0.64$$

תרגיל 4

נתונה המערכת:



ההסתברות שכל חלק עובד: $P(S_1) = P(S_2) = P(S_3) = p$. המערכת עובדת כאשר: $W = S_1 \cup (S_2 \cap S_3)$

כלומר ההסתברות שהמערכת עובדת:

$$P(W) = P(S_1) + P(S_2 \cap S_3) - P(S_1 \cap S_2 \cap S_3) = p + p^2 - p^3$$