הסקה סטטיסטית שיעור 2

הקדמה חומר קריאה 2א, חומר קריאה 2ב, חומר קריאה 2ג, חומר קריאה 2ד, חומר קריאה 2ה

המשך חזרה על הסתברות: ניסוי, מרחב מדגם, מאורע, פונקציית ההסתברות, מרחב מדגם בדיד או רציף, משלים, הכלה והדחה.

1 כשל הסתברות קודמת (כשל שיעור הבסיס) *base rate fallacy*

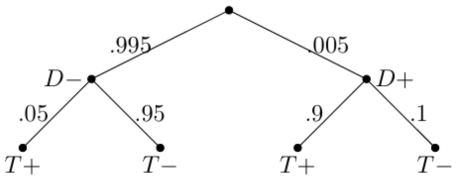
דוגמה לכך שקל להתבלבל בין $P(A|B), P(B|A)$, במיוחד כאשר המצב מתואר במילים. דוגמה:

נניח שיש בדיקה למחלה כלשהי. המחלה נפוצה ב-0.5% מהאוכלוסייה (זה שיעור הבסיס). דיוק הבדיקה נותן רק 5% *false positive*, ו-10% *false negative*. אם נבדקנו ונמצאנו חיוביים, מה ההסתברות שאנחנו באמת חולים?

נסמן: D^+=חולים, D^-=בריאים, T^+=בדיקה חיובית, T^-=בדיקה שלילית.

מתקיים: $P(D^+) = 0.005, P(D^-) = 0.995, P(T^+|D^-) = 0.05, P(T^-|D^+) = 0.1$.

נחשב את המשלימים:



$$P(T^-|D^-) = 0.95, P(T^+|D^+) = 0.9$$

אנחנו רוצים לחשב את $P(D^+|T^+)$. בעזרת נוסחת בייס נקבל:

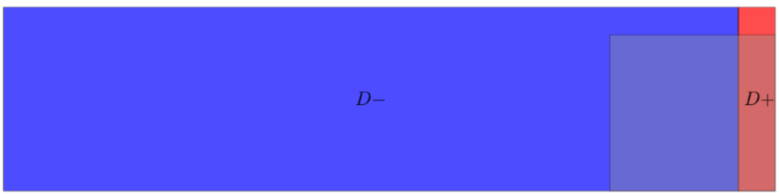
$$P(D^+|T^+) = \frac{P(T^+|D^+)P(D^+)}{P(T^+)}$$

נחשב את $P(T^+)$ בעזרת נוסחת ההסתברות השלמה:

$$P(T^+) = P(T^+|D^-)P(D^-) + P(T^+|D^+)P(D^+) = 0.05 \cdot 0.995 + 0.9 \cdot 0.005 = 0.5425$$

ובסה"כ נקבל:

$$P(D^+|T^+) = \frac{0.9 \cdot 0.005}{0.05425} = 0.082949 \approx 8.3\%$$



השטח המסומן זה האנשים שיצאו חיוביים. הוא מכסה את רוב השטח האדום ומעט מהשטח הכחול, אבל עדיין רובו כחול.

זה נקרא כשל שיעור הבסיס כי שיעור הבסיס של המחלה באוכלוסייה כל כך נמוך, כך שרוב האנשים שנבדקים הם בריאים. ולכן גם עם בדיקה מדוייקת, רוב האנשים שיוצאים חיוביים הם בריאים.
בפשטות: בדיקה שצודקת ב 95% אחוז לא אומר ש 95% מהבדיקות החיוביות הן נכונות.

2 משתנים מקריים

התפלגויות נפוצות:

ברנולי - $Ber(P)$: $X = 1$ (הצלחה) בהסתברות p. הסתברות להצלחה בניסיון יחיד.

בינומית - $Bin(n,p)$: X הוא מספר ההצלחות מתוך n ניסיונות, כל אחד בהסתברות p להצלחה. n פעמים ברנולי.

$P(X = k) = \binom{n}{k}p^k(1-p)^{n-k}$. מתוך n ניסויים, נבחר k מקומות להצלחה.

גיאומטרית - $Geom(p)$: X הוא מספר הכישלונות לפני ההצלחה הראשונה, כל ניסוי בהסתברות p (כל ניסוי הוא ברנולי).

$P(X = k) = (1-p)^k p$.

תוחלת

$$E(X) = \sum_{i=1}^{n} p(x_i)x_i$$

לינאריות התוחלת:

$$E(X+Y) = E(X) + E(Y), \qquad E(aX+b) = aE(X) + b, \qquad E\big(h(X)\big) = \sum_i h(x_i)p(x_i)$$

שונות וסטיית תקן

שונות היא בעצם הממוצע המשוקלל של ריבוע המרחק מהתוחלת: $Var(X) = E((X-\mu)^2) = \sum_{i=1}^{n} p(x_i)(x_i-\mu)^2$.

סטיית תקן: $\sigma = \sqrt{Var(X)}$.

אם X, Y בת"ל אז $Var(X+Y) = Var(X) + Var(Y)$.

$$Var(aX+b) = a^2 Var(X), \qquad Var(X) = E(X^2) - \big(E(X)\big)^2$$

Distribution	range X	pmf $p(x)$	mean $E(X)$	variance $\mathrm{Var}(X)$
Bernoulli(p)	0, 1	$p(0) = 1-p, \quad p(1) = p$	p	$p(1-p)$
Binomial(n, p)	0, 1,..., n	$p(k) = \binom{n}{k} p^k (1-p)^{n-k}$	np	$np(1-p)$
Uniform(n)	1, 2, ..., n	$p(k) = \frac{1}{n}$	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$
Geometric(p)	0, 1, 2,...	$p(k) = p(1-p)^k$	$\frac{1-p}{p}$	$\frac{1-p}{p^2}$

3 משתנים מקריים רציפים

מ"מ ייקרא רציף אם קיימת פונקציה $f(x)$ כך שלכל $c \leq d$ מתקיים:

$$P(c \leq x \leq d) = \int_c^d f(x)dx$$

הפונקציה $f(x)$ נקראת ה- *PDF - probability density function*, והיא מקיימת:

$$f(x) \geq 0, \qquad P(-\infty \leq x \leq \infty) = \int_{-\infty}^{\infty} f(x)dx = 1$$

ה- *CDF - cumulative distribution function* של מ"מ רציף מוגדרת כמו מ"מ בדיד:

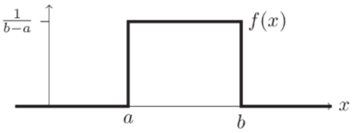
$$F(b) = P(X \leq b) = \int_{-\infty}^{b} f(x)dx$$

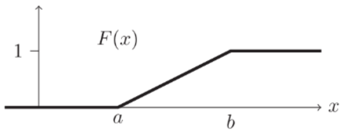
במעבר מ-PDF ל-CDF חשוב לציין את התחומים של a.

הPDF $f(x)$ היא המקבילה הרציפה ל-PMF $p(x)$ הבדידה. אבל היא לא בעצמה הסתברות – צריך לחשב את ה-CDF $F(x)$.

התפלגות אחידה

$$X\sim U(a,b),\qquad f(x)=1/(b-a)\,,\qquad F(x)=(x-a)/(b-a)\,,\qquad (for\ a\le x\le b)$$





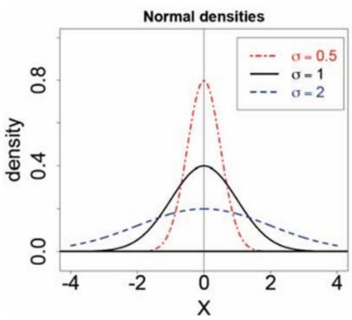
התפלגות מעריכית

$$X\sim Exp(\lambda),\qquad f(x)=\lambda e^{-\lambda x},\qquad F(x)=1-e^{-\lambda x},\qquad (for\ x\ge 0)$$

המקבילה הרציפה להתפלגות גיאומטרית. מייצגת את הזמן שנחכה לתהליך רציף לשנות מצב. כלומר, אם מספר המוניות שעובר ברחוב מסויים בדקה הוא n, אז $\lambda=\frac{1}{n}$, ו-X הוא מספר הדקות שנחכה למונית.

התפלגות מעריכית היא חסרת זיכרון. אם ההסתברות שמונית תגיע תוך 5 דקות היא p, וחיכיתי 5 דקות ולא עברה מונית, ההסתברות שתעבור מונית ב-5 דקות הבאות היא עדיין p. מנגד, הזמן שנחכה לרכבת הוא לא חסר זיכרון, כי הן מגיעות לפי לו"ז מסויים. אם חיכיתי 5 דקות בלי שהגיעה רכבת, יש סיכוי גדול יותר שתגיע רכבת ב-5 דקות הבאות. זה יותר מתאים להסתברות אחידה.

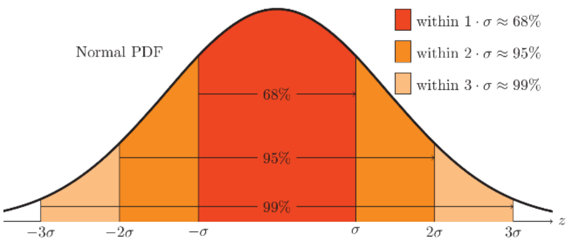
התפלגות נורמלית (1809, קרל פרדריך גאוס)



$$N(\mu,\sigma^2),\qquad f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/2\sigma^2},\qquad (for-\infty\le x\le\infty)$$

אין נוסחה ל-$F(x)$ – משתמשים בטבלה.

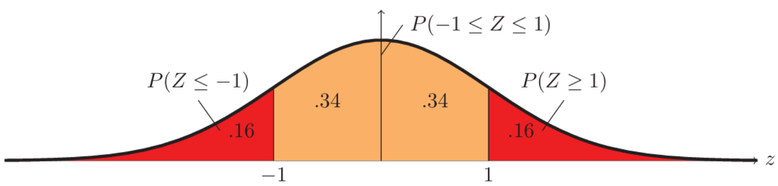
התפלגות נורמלית סטנדרטית – $N(0,1)$. ממוצע (תוחלת) 0, סטיית תקן (שונות) 1.



נסמן ע"י Z. את ה-CDF נסמן $\Phi(z)$.

$$\phi(z)=\frac{1}{\sigma\sqrt{2\pi}}e^{-z^2/2}$$

היא סימטרית ביחס לציר הY, ואפשר להשתמש בזה בשביל לחשב:



התפלגות פארטו $(pareto)$

$$Pareto(m,\alpha),\qquad m,\alpha>0,\qquad [m,\infty),\qquad f(x)=\frac{\alpha m^\alpha}{x^{\alpha+1}},\qquad F(x)=1-\frac{m^\alpha}{x^\alpha}\ for\ x\ge m$$

הרצאה

מצגת שיעור 2א (שקופית 20 והלאה), מצגת שיעור 2ב, מצגת שיעור 2ג

דוגמה 1 – סנאים

בקמפוס יש 1,000,000 סנאים. רובם טובים ו-100 מתוכם רעים. סטודנטית פיתחה אזעקת שמגלה סנאים רעים. האוניברסיטה בדקה אותה ומצאה: בהינתן סנאי טוב, האזעקה מופעלת 1% מהזמן. בהינתן סנאי רע, האזעקה מופעלת 99% מהזמן. נשאל:

א) בהינתן שסנאי הפעיל את האזעקה, מה ההסתברות שהוא רע?
ב) האם המערכת טובה?

לפני שניגש לחישוב, נשים לב שרק 100/1000000 הסנאים הם רעים, כך שאם בחרנו סנאי, קרוב לוודאי שהוא טוב.

נחשב: $N \coloneqq nice, E \coloneqq evil, A \coloneqq alarm$

$$P(N) = 0.9999, \qquad P(E) = 0.0001$$

אנחנו רוצים לדעת: $P(E|A) = ?$. לפי נוסחת בייס:

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

נציב את נוסחת ההסתברות השלמה ונציב את הערכים הידועים:

$$P(E|A) = \frac{P(A|E)P(E)}{P(A|E)P(E) + P(A|N)P(N)} = \frac{0.99 \cdot 0.0001}{0.99 \cdot 0.0001 + 0.01 \cdot 0.9999} \approx 0.01$$

וקיבלנו שהדיוק מאד נמוך. זה נקרא *base rate fallacy*.

דוגמה 2 – קוביות

יש 2 קוביות – אחת בעלת 6 פאות ואחת 8. בוחרים באופן מקרי ואחיד אחת מהקוביות (ולא מגלים לנו). זורקים את הקובייה ואומרים לנו איזה מספר יצא. עבור כל מספר, מה ההסתברות שנבחרה הקובייה של 6 פאות?

נחשב, לדוגמה אם יצא 4: לפי נוסחת בייס, ונציב במכנה את נוסחת ההסתברות השלמה:

$$P(D_6|R_4) = \frac{P(R_4|D_6)P(D_6)}{P(R_4)} = \frac{\frac{1}{6}\cdot\frac{1}{2}}{\frac{1}{6}\cdot\frac{1}{2} + \frac{1}{8}\cdot\frac{1}{2}} = \frac{4}{7}$$

כנ"ל עבור כל מספר 1-6. אם יצא 7 או 8, ההסתברות בוודאי 0. וזה גם מופיע בנוסחה, כי: $P(R_7|D_6) = 0$.

2 משתנים מקריים בדידים, תוחלת (מצגת 2ב)

משתנה מקרי X נותן מספר לכל תוצאה: $X : \Omega \to \mathbb{R}$. $X = a$ מייצג את המאורע $\{\omega : X(\omega) = a\}$.

ה- *PMF – Probability Mass Function* של X נתונה ע"י: $p(a) = P(X = a)$.

ה- *CDF - Cumulative distribution function* של X נתונה ע"י: $F(a) = P(X \le a)$.

לדוגמה:

values of X:	-2	-1	0	4
pmf $p(a)$:	1/4	1/4	1/4	1/4
cdf $F(a)$:	1/4	2/4	3/4	4/4

סכום משתנים מקריים

עבור $X\sim Bin(n,p), Y\sim Bin(m,p), Z\sim Bin(n,q)$, מתקיים:

$(X+Y)\sim Bin(n+m,p)$. כי כל אחד הוא סכום של מ"מ ברנולי שכולם בת"ל וכולם באותה הסתברות להצלחה. אם יש שני מטבעות עם אותה הסתברות לעץ, זה לא באמת משנה איזה מהם מטילים. אז הסכום שלהם הוא פשוט מספר ההצלחות הכולל.

לעומת זאת, $(X+Z)$ לא מתפלג באותה צורה.

תרגיל כיתה 1

יהי X מ"מ שסופר את מספר ההצלחות לפני הכישלון השני בסדרת ניסויים שכל אחד מתפלג $Ber(p)$. ה-PMF היא:

$$p(n)=(n+1)p^n(1-p)^2$$

הסבר: המיקום האחרון בסדרה יהיה כישלון (הכישלון השני). כלומר רק צריך לבחור עוד מקום אחד לכישלון הראשון, מתוך $n+1$ מקומות. צריך שיצא n פעמים הצלחה (בהסתברות p) ועוד פעמיים כישלון (בהסתברות $1-p$).

חוסר זיכרון

ברולטה, גם אם שחור יצא 26 פעמים ברצף (מונטה קרלו, 1913), זה לא אומר שיש לו יותר או פחות סיכוי להיות בפעם הבאה.

נוכיח שמ"מ גיאומטרי הוא חסר זיכרון, כלומר: $P(X=n+k|X\geq n)=P(X=k)$.

$$P(X=n+k|X\geq n)=\frac{P(X=n+k\cap X\geq n)}{P(X\geq n)}=\frac{p^{n+k}(1-p)}{p^n}=p^k(1-p)=P(X=k)$$

תרגיל כיתה 2

נתונים שני הימורים: מה עדיף? בניסוי שנערך, מתוך האנשים שהייתה להם עדיפות לאחד על גבי השני, הרוב העדיפו את ב.

א. סיכוי של 10% להרוויח 95$ ו-90% להפסיד 5$,
ב. לשלם 5$ בשביל סיכוי של 10% להרוויח 100$ ו-90% לכלום.

נחשב את התוחלת של כל אחד, כשנכתוב את הנתונים בנוסחה, נשים לב שהם בעצם אותו דבר:

$$0.1\cdot 95+0.9\cdot(-5)=9.5-4.5=5$$

תרגיל כיתה 3

בשולחן עם n אנשים, אם כולם יקומו ויתערבבו ואז ישבו חזרה באקראי. מה התוחלת של מספר האנשים שיחזרו למקום שלהם?

מצב שבו אף אחד לא חוזר למקום שלו נקרא אי סדר מלא. מספר הדרכים לזה הוא המספר השלם הקרוב ל- $n!/e$. יש סה"כ $n!$ דרכים להסתדר במקומות, כלומר:

$$P(everyone\ in\ different\ seats)=\frac{n!/e}{n!}=\frac{1}{e}\approx 0.3679$$

נמספר את האנשים 1 עד n. יהי X_i האינדיקטור למאורע שהאדם ה-i חזר למקום שלו. ההסתברות לכל X_i היא $1/n$. כלומר, $X_i\sim Ber(1/n)$ ולכן $E(X_i)=1/n$. יהי X מספר האנשים שחזרו למקומות שלהם, כלומר $X=\sum_{i=1}^n x_i$. מלינאריות התוחלת, $E(X)=\sum_{i=1}^n E(X_i)=\sum_{i=1}^n 1/n=n$.

נשים לב שאם $n=2$, אז או ששניהם חוזרים למקום או ששניהם לא. כלומר $P(X=0)=0.5=P(X=2)$. X אף פעם לא יהיה שווה לתוחלת במקרה הזה.

3 שונות , משתנים מקריים רציפים (מצגת 2ג)

תרגיל כיתה 4

יהיו $X_1, X_2, \dots, X_n$ מ"מ בת"ל בעלי $\sigma_i = 2$ (כלומר $Var(X_i) = 4$). יהי $\bar{X} = \sum_{i=1}^{n} X_i/n$ הממוצע. מה סטיית התקן של $\bar{X}$?

מכיוון שהמ"מ בת"ל, מתקיים: $Var(X+Y) = Var(X) + Var(Y)$, כלומר $Var(\sum_{i=1}^{n} X_i) = n \cdot Var(X_i) = 4n$.

$$Var(\bar{X}) = Var\left(\sum_{i=1}^{n} X_i/n\right) = Var\left(\frac{1}{n} \cdot \sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} Var\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} 4n = \frac{4}{n}$$

כלומר, $\sigma_{\bar{X}} = \frac{2}{\sqrt{n}}$. הממוצע של מ"מ בת"ל משתנה פחות מאשר מ"מ יחיד.

תרגיל כיתה 5

נניח ש-X הוא מ"מ בתחום $[0,2]$, עם PDF $f(x) = cx^2$.

נחשב את c. ההסתברות הכוללת צריכה להיות 1, אז: $1 = \frac{c8}{3} = c\left[\frac{x^3}{3}\right]_0^2 = \int_0^2 cx^2 dx = \int_0^2 f(x)dx$, כלומר $c = \frac{3}{8}$.

נחשב את ה-CDF: ה-PDF היא 0 עבור כל x מחוץ לתחום $[0,2]$, אז נחשב: $\frac{x^3}{8} = \frac{3}{8} \cdot \left[\frac{u^3}{3}\right]_0^x = \int_0^x \frac{3}{8} u^2 du$. ובסה"כ:

$$F(x) = \begin{cases} 0, & x < 0 \\ \frac{x^3}{8}, & 0 \le x \le 2 \\ 1, & 2 < x \end{cases}$$

נחשב $P(1 \le X \le 2)$: לפי ה-CDF: $P(1 \le X \le 2) = F(2) - F(1) = 1 - \frac{1}{8} = \frac{7}{8}$.

תרגיל כיתה – מ"מ מעריכי

בממוצע, מונית עוברת ברחוב מסויים כל 10 דקות. נניח שהזמן שנחכה למונית ממודל ע"י מ"מ מעריכי. כלומר $X \sim Exp\left(\frac{1}{10}\right)$.

$$f(x) = \frac{1}{10} e^{-\frac{x}{10}}$$

$$P(3 \le X \le 7) = \int_3^7 \frac{1}{10} e^{-\frac{x}{10}} = \left[-e^{-\frac{x}{10}}\right]_3^7 = e^{-\frac{3}{10}} - e^{-\frac{7}{10}} \approx 0.244$$

$$F(x) = 1 - e^{-\frac{x}{10}}$$

תרגול

תרגיל 1

בעיר יש 99 מכוניות ירוקות, ואחת כחולה. הרכב הכחול עשה תאונה. אנחנו רוצים להוכיח (בבית משפט) שזה לא היה הוא.

בדקנו את העד היחיד לתאונה ומצאנו ש:

99% מהזמן, העד רואה מכונית כחולה ככחולה. 2% מהזמן, הוא רואה את המכונית הירוקה ככחולה.

נגדיר: T_b=היה רכב כחול, T_g=היה רכב ירוק, W_b=העד ראה רכב כחול, W_g=העד ראה רכב ירוק.

אנחנו רוצים למזער את $P(T_b|W_b)$. לפי נוסחת בייס:

$$P(T_b|W_b) = \frac{P(W_b|T_b)P(T_b)}{P(W_b)}$$

נמצא את $P(W_b)$ עם נוסחת ההסתברות השלמה:

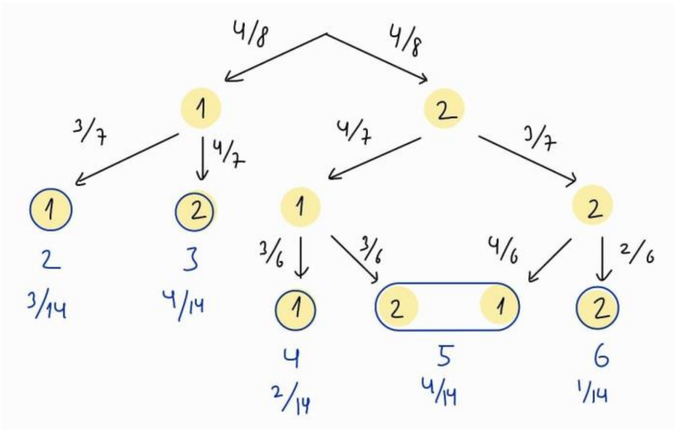
$$P(W_b) = P(W_b|T_b)P(T_b) + P\left(W_b|T_g\right)P\left(T_g\right) = 0.99 \cdot 0.01 + 0.02 \cdot 0.99 = 0.99 \cdot 0.03$$

נציב בנוסחת בייס:

$$\frac{0.99 \cdot 0.01}{0.99 \cdot 0.03} = \frac{1}{3}$$

תרגיל 2

יש חפיסה עם 8 קלפים: 1ע, 1י, 1ת, 1ל, 2ע, 2י, 2ת, 2ל. מהלך המשחק: אם שולפים 1, שולפים עוד קלף. אם שולפים 2, שולפים עוד 2 קלפים. פעם אחת, ללא החזרה. נגדיר: X הוא סכום המספרים על הקלפים. נשרטט את העץ:



התוחלת היא:

$$E(X) = \sum_{i \in S} i \cdot P(X = i) =$$
$$= \frac{6}{14} + \frac{12}{14} + \frac{8}{14} + \frac{20}{14} + \frac{6}{14} = \frac{52}{14} = \frac{26}{7}$$

תרגיל 3

יש מפעל שמייצר בלונים, לכל בלון יש הסתברות p שיהיה בו חור. ברגע שיש חור בבלון הפס ייצור נעצר. נגדיר: X=מספר הבלונים שנבדוק. נחשב את התוחלת:

$$E(X) = p \cdot E(X|X = 1) + (1 - p)E(X|X > 1) = p + (1 - p)\big(1 + E(X - 1|X > 1)\big)$$
$$= p + (1 - p)\big(1 + E(X)\big) = p + 1 + E(X) - p - pE(X) = 1 + E(X) - pE(X)$$

$$\Downarrow$$

$$-1 = -pE(X) \Longrightarrow E(X) = \frac{1}{p}$$