דף נוסחאות במבוא להסתברות ולסטטיסטיקה

<u>מרחב מדגם</u>: אוסף כל התוצאות האפשריות לניסוי. Ω
1. בדיד סופי (טבעיים); 2. בדיד אינסופי (טבעיים) 3. רציף (ממשיים) (צפיפות) אם אינסופי אז נקרא שדה סיגמא ע"פ אמות מידה (שטח)
<u>מרחב מדגם סימטרי</u>: כשלכל מאורע פשוט יש הסתברות כמו לאחר.

<u>תכונות ההסתברות</u> P למאורע A:
1. $P(\Omega)=1$
2. $P(A)\ge 0$, $P(A)\le 1$, לכל A
3. קלמוגורוב: אם A ו B **<u>זרים</u>** אז $P(A\cup B)=P(A)+P(B)$
4. החוק המשלים: $P(\bar{A})=1-P(A)$
5. מונוטוניות: אם $A\subseteq B$ אז $P(A)\le P(B)$

<u>משפט האיחוד</u>: $P(A\cup B)=P(A)+P(B)-P(A\cap B)$
אם A,B **<u>זרים</u>** אז מתקיים: $P(A\cup B)=P(A)+P(B)$
הסתברות האיחוד לשלושה מאורעות:
$$P(A\cup B\cup C)=P(A)+P(B)+P(C)$$
$$-P(A\cap B)-P(A\cap C)-P(B\cap C)+P(A\cap B\cap C)$$
אם A,B **<u>בלתי תלויים</u>** אז מתקיים: $P(A\cap B)=P(A)\cdot P(B)$
אז: $P(B)=P(B|A)\wedge P(A)=P(A|B)$

<u>הסתברות מותנית</u> (בדר"כ ניתנת באחוזים)(A בתנאי B) :
$$P(A|B)=\frac{P(A\cap B)}{P(B)} \quad , \quad P(A|B)\ne P(B|A)$$
אם אין קשר בין המאורעות, ההסתברות המותנית והרגילה שוות.

<u>משפט המכפלה</u>: $P(A\cap B)=P(A)\bullet P(B|A)$
A ו B **בלתי תלויים** אם ורק אם
$P(B|A)=P(B)$ ואז $P(A\cap B)=P(A)\bullet P(B)$
בלי החזרה – המאורעות תלויים, עם החזרה – בלתי תלויים.

<u>משפט ההסתברות השלמה:</u>
שלב I : $A_1,\ldots,A_n$ חלוקה של Ω - מאורעות זרים
$$P(A_1)+P(A_2)+\ldots+P(A_n)=1$$
שלב II :
$$P(B)=\sum_{i=1}^{n}P(B|A_i)\cdot P(A_i)$$

<u>חוק בייס</u>:
$$P(A_i|B)=\frac{P(A_i)\cdot P(B|A_i)}{\sum_{j=1}^{n}P(A_j)\cdot P(B|A_j)}$$

<u>קומבינטוריקה</u> (צירופים סדורים):
- עם החזרה, עם חשיבות לסדר: $a\cdot a\cdot a\ldots\cdot a=a^n$
- עם החזרה, ללא חשיבות לסדר
- ללא החזרה, עם חשיבות לסדר:

$$N(N-1)(N-2)\cdot\ldots\cdot(N-n+1)=\frac{N!}{(N-n)!}$$
ללא החזרה, ללא חשיבות לסדר: $C_k^n=(\frac{n!}{(n-k)!k!})$
סידור של n עצמים שונים בשורה נקרא תמורה של העצמים.

<u>פונקציית התפלגות (מצטברת):</u> $F_x(t)=P(X\le t)$
הסתברות מצטברת $F_x(t)=P(a\le x\le b)=\sum_{a}^{b}P(X=x)$

מציאת התפלגות מ"מ בדיד המקבל ערכים שלמים באמצעות $F_x(t)$:
$$P(X=k)=F_x(k)-F_x(k-1)$$

<u>משתנה מקרי דו-מימדי בדיד</u>
על מרחב המדגם מוגדרים **שני** משתנים מקריים.
התפלגות משותפת והתפלגות שולית:

$x\backslash^{y}$	...	j	...	שולית X
⋮				⋮
i		$P(X=i\wedge Y=j)$		$P(X=i)$
⋮				⋮
שולית Y	...	$P(Y=j)$	...	1

i ערך עבור X , j ערך עבור Y
סך ההסתברויות המשותפות בתוך הטבלה שווה ל1.
סכום ההסתברויות השוליות שווה ל1.
הסתברות שולית: מקבעים ערך לאחד מהמשתנים והשני "רץ" עליו.
- הערך במשבצת i,j שווה ל- $P(X=i\wedge Y=j)$
- הערך במשבצת i בשולית X שווה ל- $P(X=i)$, ולסכום שורה i
* ניתן למצוא את ההתפלגות של $P(X|Y=j)$ ע"י נרמול העמודה ה- j בהכפלה בקבוע, כך שסכום ההסתברויות שווה 1 (כנ"ל עבור $P(Y|X=i)$).
אם יש אפס בטבלה כאשר ההתפלגויות השוליות שונות מאפס, אזי יש תלות . אם x ו y הם בלתי תלויים אז
$$P(X=x,Y=y)=P(X=x)\cdot P(Y=y)$$
נוסחא להתפלגות שולית: $P_{X,Y}(x,y)$
התפלגות משותפת: $P_X(x)=\sum_{y\in Y}P_{X,Y}(x,y)$

חישוב ההסתברות של **<u>מקסימום</u>** z=max(x,y):
$$P(Z=z)=P(X=z)\cdot P(Y\le z)+P(Y=z)\cdot P(X<z)$$
חישוב ההסתברות של **<u>מינימום</u>** m=min(x,y):
$$P(M=m)=P(X=m)\cdot P(Y\ge m)+P(Y=m)\cdot P(X>m)$$

<u>נוסחאות כלליות</u>
חוקי דה-מורגן: $\overline{A\cap B}=\bar{A}\cup\bar{B},\ \overline{A\cup B}=\bar{A}\cap\bar{B}$
סכום סדרה הנדסית סופית:
$$a+aq+aq^2+..+aq^n=a\cdot\frac{1-q^n}{1-q}$$
סכום סדרה הנדסית אינסופית מתכנסת: כלומר, $|q|<1$
$$a+aq+aq^2+\ldots=\frac{a}{1-q}$$
סכום סדרה חשבונית:
$$a+a\cdot d+a\cdot 2d+..+a\cdot(n-1)d=\frac{(a+a(n-1)d)\cdot n}{2}$$

אחידה בדידה: $X \sim U(N)$

משמעות: קיימות N נקודות ולכולן אותן ההסתברות.

$$P(X=k)=\frac{1}{N}$$

$$E(X)=\frac{N+1}{2},V(X)=\frac{N^2-1}{12}$$

N - מספר הערכים האפשריים

k - ערך אפשרי לְ- x, ($k=1,..,N$)

היפר-גאומטרית: $X \sim H(N,D,n)$

משמעות: רוצים למצוא הסתברות למשיכת k מתוך D מיוחדים, בְּ- n דגימות, מתוך אוכלוסייה בגודל N.

$$P(X=k)=\frac{\binom{D}{k}\binom{N-D}{n-k}}{\binom{N}{n}}$$

$$E(X)=n\cdot\frac{D}{N}$$

$$V(X)=n\cdot\frac{D}{N}\left(1-\frac{D}{N}\right)\left(1-\frac{n-1}{N-1}\right)$$

N – מספר הפריטים באוכלוסיה כולה

D – מספר המיוחדים מתוך האוכלוסייה

n - מספר הדגימות (זה נחשב *ללא החזרה* וללא חשיבות לסדר)

k – מספר המיוחדים שנקבל בדגימה($k=0,..,n$)

בינומית: $X \sim Bin(n,p)$

משמעות: מבצעים n ניסויים עם הסתברות להצלחה p (קבועה לכל ניסוי), ורוצים למצוא הסתברויות לְ- k הצלחות.

$$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$$

$$E(X)=np \ , \ V(X)=np(1-p)$$

$$M_X(t)=\left[pe^t+(1-p)\right]^n$$

n - מספר ניסויים (זה נחשב *עם החזרה*)

p - הסתברות להצלחה, (q=1-p הסתברות לכשלון)

p - מספר הצלחות,($k=0,..,n$)

חיבור מ"מ בינומיים:

$X \sim Bin(n,p), Y \sim Bin(m,p)$ אותו P, אז $X+Y \sim Bin(n+m,p)$

גאומטרית: $X \sim G(p)$

משמעות: מבצעים ניסוי שוב ושוב עד ההצלחה. רוצים למצוא הסתברות לְ- k ניסויים.

$$P(X=k)=(1-p)^{k-1}p$$

$$E(x)=\frac{1}{p}$$

$$V(x)=\frac{1-p}{p^2}$$

p - הסתברות להצלחה

k – מספר ניסויים עד הצלחה ראשונה (כולל), ($k=1,2,...$)

תכונת חוסר זיכרון: $P(X=a+k \mid X>a)=P(X=k)$

מעברים שימושיים:

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

$$P(X\le k)=1-(1-p)^k$$

בינומית שלילית: $X \sim NB(m,p)$

משמעות: מבצעים ניסוי שוב ושוב עד הגעה לְ- m הצלחות. רוצים למצוא הסתברות לְ- k ניסויים.

$$P(X=k)=\binom{k-1}{m-1}p^m(1-p)^{k-m}$$

$$E(X)=\frac{m}{p},V(X)=\frac{m(1-p)}{p^2}$$

m - כמות הצלחות נדרשת

p - הסתברות להצלחה

k - מס' ניסויים, ($k=m,m+1,...$)

חשוב: סכום של m מ"מ ב"ת שווי התפלגות המתפלגים גיאומטרית, מתפלג בינומי שלילי עם פרמטרים m,p

פואסונית: $X \sim Poi(\lambda)$

משמעות: מספר אירועים ביחידת זמן. ידוע שבתוחלת מגיעים λ מצטרפים, ורוצים לדעת הסתברות שבפועל הגיעו k.

$$P(X=k)=e^{-\lambda}\frac{\lambda^k}{k!}$$

$$E(X)=\lambda \ , \ V(X)=\lambda$$

$$M_X(t)=e^{-\lambda}\cdot e^{\lambda e^t}=e^{\lambda(e^t-1)}$$

λ - ממוצע של אירועים ליחידה (בדר"כ זמן)

k - מספר האירועים ליחידה, ($k=0,1,2,...$)

חיבור מ"מ פואסונים:

$$X \sim Poi(\lambda_1), Y \sim Poi(\lambda_2)$$

$$\Rightarrow X+Y \sim Poi(\lambda_1+\lambda_2)$$

קירוב בינומי לפואסוני:

$X \sim Bin(np)$ ומתקיים שֶ- n מספיק גדול, אזי $X \sim Poi(\lambda=np)$

קירוב בינומי להיפרגאו:

אם $n<N+M$ אז $X \sim HG(n,N,M) \Rightarrow X \sim Bin\left(n,\frac{M}{N+M}\right)$

הוצאה של n פריטים ללא החזרה בזה אחר זה שקולה להוצאת n פריטים יחד, שקולה לסידור n פריטים בשורה.

$$P(A_i)=\frac{a}{N}$$

מספר האפשרויות לסידור כאשר הפריט ה i מיוחד, חלקי מספר האפשרויות לסידור N פריטים, שמתוכם a מיוחדים, בשורה.

<u>תוחלת:</u>

אם נבצע ניסוי פעמים רבות, התוצאות יתכנסו למספר מסוים (תוחלת)

x מ"מ המקבל ערכים $a_1, a_2, \ldots$ $\quad E(X)=\sum_{i=1}^{\infty} a_i \cdot P(X=a_i)$

תכונות התוחלת:

$$E(c)=c$$
$$E(aX+b)=aE(X)+b$$
$$E(X+Y)=E(X)+E(Y)$$
$$E(g(X))=\sum_{x\in X} g(X)P(X)$$

תשומת לב: $E(x^2)=\sum_x x^2P(x)\neq (E(x))^2$

אם X וY בלתי תלויים אז: $E(X\cdot Y)=E(X)\cdot E(Y)$ (לא נכון הפוך)

תוחלת מותנה: $E(Y)=E[E(Y\mid X)]$

$$E(X,Y)=\sum_x\sum_y x\cdot y\cdot P(X=x,Y=y)$$

אם הם ב"ת: $E(X,Y)=\sum_x x\cdot P(X=x)\cdot \sum_y y\cdot P(Y=y)$

<u>שונות:</u>

מדד לפיזור סביב התוחלת (=כמה אנחנו מרוחקים מהתוחלת?)

$$V(X)=E(X^2)-[E(X)]^2$$
$$V(X)=E[(X-E(X))^2]$$

תכונות השונות:

$$V(X)\geq 0$$
$$V(c)=0$$
$$V(c+x)=V(x)$$
$$V(cx)=V(-cx)=c^2V(x)$$

<u>סטיית תקן:</u> מדד לפיזור $\quad \sigma=\sqrt{Var(x)}$

אם X וY בלתי תלויים אז: $V(X+Y)=V(X)+V(Y)$

שונות מותנה: $V(Y)=E[V(Y\mid X)]+V[E(Y\mid X)]$

שונות של חיבור שני מ"מ:

$$V(X+Y)=V(X)+V(Y)-2COV(X,Y)$$

<u>אי שיוון מרקוב</u>:

חסם עליון להסתברות. התוחלת נתונה, ההתפלגות אינה ידועה.

לכל c קבוע: $\quad P(X\geq c)\leq \frac{E(X)}{c}$

<u>אי שיוון צ'ביצ'ב</u>:

חישוב מרחקים מהתוחלת. התוחלת והשונות (סטיית התקן) נתונות.

$$P(|x-\mu|<a\cdot\sigma)\geq 1-\frac{1}{a^2}$$
$$p(|x-\mu|<\varepsilon)\geq 1-\frac{\sigma^2}{\varepsilon^2}$$

$\varepsilon=a\cdot\sigma$ מרחק מהתוחלת, a מספר סטיות תקן מהתוחלת.

(תוחלת=אמצע ההפרש, ומחלצים את a לפי סטיית התקן הנתונה.)

<u>שונות משותפת:</u>

עד כמה שני משתנים מתואמים / לא מתואמים

$$COV(X,Y)=E[(X-E(X))(Y-E(Y))]$$
$$COV(X,Y)=E(X\cdot Y)-E(X)\cdot E(Y)$$

x,y בלתי מתואמים:

אם $COV(X,Y)=0$, או אם $E(XY)=E(X)E(Y)$

או אם $V(X+Y)+V(X)+V(Y)$

אם x,y בלתי תלויים אז הם בהכרח בלתי מתואמים (להפך לא בהכרח נכון).

שונות משותפת עוזרת לנו לחשב שונות בביטוי:

$$V(X+Y)=V(X)+V(Y)+2COV(X,Y)$$

תכונות השונות המשותפת:

$$COV(X,X)=V(X)$$
$$COV(X,Y)=COV(Y,X)$$

אם a קבוע, $COV(aX,Y)=a\cdot COV(X,Y)$

X,Y,Z מ"מ, $COV(X+Z,Y)=COV(X,Y)+COV(Z,Y)$

אם a,b קבועים, $COV(aX,bY)=a\cdot b\cdot COV(X,Y)$

<u>סכום וממוצע של משתנה מקרי:</u>

סכום: $S_n=X_1+..+X_n$ $\quad$ ממוצע: $\overline{X}_n=\frac{X_1+..+X_n}{n}$

שונויות של סכום וממוצע: $V(S_n)=\sum_{i=1}^{n}V(X_i)+\sum_{i\neq j}COV(X_i,X_j)$

$$V(\overline{X}_n)=V\left(\frac{S_n}{n}\right)=\frac{1}{n^2}\cdot V(S_n)$$

הערה: אם X,Y <u>בלתי מתואמים</u> או <u>בלתי תלויים</u>:

אזי $COV(X,Y)=0$ לכל $i\neq j$ ולכן: $V(S_n)=\sum_{i=1}^{n}V(X_i)$

<u>מקדם המתאם:</u>

מדד למידת התאמה בין משתנים

$$\rho=\frac{COV(x,y)}{\sigma(x)\cdot\sigma(y)}=\frac{COV(x,y)}{\sqrt{V(x)}\cdot\sqrt{V(y)}} \qquad -1\leq\rho\leq 1$$

אם שווה 0: המשתנים בלתי מתואמים (לינארית);

אם שווה 1: קשר בין X לY חיובי חזק (X עולה אז Y עולה);

אם שווה 1-: קשר בין X לY שלישי חזק (X עולה אז Y יורד);

סיכום: הערך המוחלט של מקדם המתאם מתאר את חוזק הקשר, והסימן מתאר את הכיוון.

אם $X_1,..,X_n$ בלתי תלויים ושווי התפלגות אזי:

$$E(\overline{X}_n)=E(X_1)\,,\;V(\overline{X}_n)=\frac{V(X_1)}{n}$$

<u>משתנה מקרי רציף:</u>
אין משמעות להסתברות בנקודה מסוימת אלא בתחום מסוים.
פונקצית צפיפות: (השטח בין a לb מתחת לf):

$$P(a \le x \le b) = \int_a^b f(x)dx$$

תכונות: א. f חיובית בכל תחום ההגדרה;

ב. $\int_{-\infty}^{\infty} f(x)dx = 1$ (אחרת זו לא פונקציה חוקית)

פונקצית התפלגות (הצטברות):
כדי להגיע לפונקצית התפלגות נחשב אינטגרל לפי התחומים ונקבל פונקציה של השטח:

$$F(t) = P(x \le t) = \int_{-\infty}^{t} f(x)dx, \quad , \quad f(x) = F'(x)$$

(תשומת לב לרציפות האינטגרל! לפעמים צריך לחבר שטחים שונים)

<u>משתנה מקרי רציף דו מימדי:</u>
פונקצית צפיפות: $f(x,y)$

$$\int_y \int_x f(x,y)dxdy$$

צפיפות שולית:

$$f(x) = \int_y f(x,y)dy$$

$$f(y) = \int_x f(x,y)dx$$

אי תלות: $f(x,y) = f(x)\cdot f(y)$

התפלגות מותנית: $f(y|x) = \frac{f(x,y)}{f(x)}$

תזכורת: חישוב תוחלת:

$$E(x,y) = \int_y \int_x x\cdot y\cdot f(x,y)dxdy$$

התפלגויות ידועות

<u>התפלגות אחידה רציפה:</u> $X \sim U(a,b)$

$$f(x) = \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & else \end{cases}$$

$$F(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \le x \le b \\ 1 & x > b \end{cases}$$

תוחלת ושונות:

$$E(x) = \frac{(a+b)}{2}, V(x) = \frac{(b-a)^2}{12}$$

<u>התפלגות אקפוננציאלית (מעריכית):</u>
מודדת זמן בין אירועים פואסונים
X – זמן בין אירועים פואסונים, λ - מס' אירועים ממוצע ביחידת זמן.
דוגמא: אם λ ממוצע תאונות דרכים בחודש אז $X \sim P(\lambda)$ מספר התאונות בחודש, ו $X \sim \exp(\lambda)$ - זמן בין 2 תאונות.

$$f(x) = \begin{cases} \lambda \cdot e^{-\lambda x} & 0 \le x \le \infty \\ 0 & else \end{cases}$$

$$F(x) = \begin{cases} 0 & x < 0 \\ 1-e^{-\lambda x} & x \ge 0 \end{cases}$$

תוחלת ושונות:

$$E(x) = \frac{1}{\lambda}, V(x) = \frac{1}{\lambda^2}$$

<u>התפלגות נורמלית:</u> $X \sim N(\mu,\sigma^2)$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} \cdot e^{-0.5(\frac{x-\mu}{\sigma})^2} \quad -\infty < x < \infty$$

משתנה נורמלי סטנדרטי: $Z \sim N(0,1)$ $\quad Z = \frac{x-\mu}{\sigma}$

פונקצית ההסתברות המצטברת של z:

$$\Phi(z) = P(z \le t)$$

$$\Phi(-z) = 1-\Phi(z)$$