

סטטיסטיקה למדעי המחשב – תרגיל בית שבוע 4

שאלה 1 (20 נקודות, 4 נקודות לכל סעיף)

הזמן שלוקח להגיע לאוניברסיטה מתפלג מעריכית עם תוחלת של 20 דקות.

- א. מהו פרמטר הקצב של ההתפלגות, λ ? (במילים אחרות, כמה אנשים צפויים להגיע בדקה אחת? התשובה יכולה גם להיות מספר לא שלם)
- ב. מהי שונות זמן ההגעה לאוניברסיטה? מהו חציון זמן ההגעה לאוניברסיטה?
- ג. כמה זמן לפני השיעור עלינו לצאת בכדי להגיע בזמן ב- 80% מהפעמים?
- ד. השיעור מתחיל עוד 15 דקות ויצאנו רק עכשיו, מה הסיכוי שלא נאחר? (רמז: יש לפתור בעזרת פונ' ההתפלגות המצטברת).
- ה. יצאנו הרגע לכיוון האוניברסיטה, מה הסיכוי שיקח לנו לפחות 10 דקות להגיע?
- ו. (4 נקודות בonus) הראו את תכונת חוסר הזיכרון על ידי דוגמא (הקשורה להתפלגות זמן ההגעה לאוניברסיטה בשאלה זו). על בסיס הדוגמא שלכם, האם סביר שזמן ההגעה לאוניברסיטה מתפלג מעריכית?

שאלה 2 (25 נקודות, 5 נקודות לכל סעיף)

שאלה זאת משתמשת בקובץ הנתונים Age_And_Time.csv בסעיפים א'-ג'. סעיפים ד'-ה' משתמשים בקובץ הנתונים countries.csv.

- א. הציגו גרף פיזור לגיל (X) וזמן (Y), מצאו את הקו הלינארי $Y = bX$ לפי שיטת הריבועים הפחותים והציגו אותו בתוך תרשים הפיזור. מהו אחוז השונות המוסברת?
- ב. בידקו האם יש צורך בטרנספורמציה והאם יש תצפיות חריגות. הסירו תצפיות חריגות ועשו טרנספורמציה במידת הצורך, נמקו את בחירתכם.
- ג. חיזרו על סעיף א' לאחר ביצוע ההתאמות בסעיף ב'. איזה מודל מתאים יותר לנתונים?
- ד. עבור נתוני countries.csv, מצאו את קו הריבועים הפחותים המתאים לכל אחד מהמודלים הבאים:

$$Y = \beta_0 + \beta_1 X_1 \quad \diamond$$

$$Y = \beta_0 + \beta_1 X_2 \quad \diamond$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \quad \diamond$$

כאשר $X_2 = education$, $X_1 = income$, $Y = life_expectancy$. דווחו את השונות המוסברת בכל מודל. מי המודל בעל השונות המוסברת הגדולה ביותר? האם זה מפתיע?

ה. האם דרושה טרנספורמציה של אחד המשתנים? בצעו את הטרנספורמציה המתאימה וחשבו מודלים לינאריים חדשים. האם יש שיפור בשונות המוסברת?

שאלה 3 (30 נקודות, 5 נקודות לכל סעיף)

ידוע כי ההתפלגות של משקל של עציץ היא נורמלית בעלת תוחלת של 81.5 ק"ג וסטיית תקן של 13 ק"ג. חשבו: (ניתן להשתמש בפייתון)

- מה ההסתברות שעציץ בודד ישקול מעל 82 ק"ג?
- מה ההסתברות ש-50 עציצים ישקלו כולם מעל 82 ק"ג?
- מה ההסתברות שהממוצע של 50 עציצים יהיה גבוה מ-82 ק"ג?
- מה ההסתברות שמשקלם הממוצע של 25 עציצים יהיה בין 80 ל-82 ק"ג?
- מה המשקל ש-10% עציצים גדולים ממנו?
- בבדיקה חוזרת התבררה טעות במדידת הקודמות של העציצים שעליהם התבססה ההתפלגות עד כה. המידע היחיד שנותר הוא שמשקלם מתפלג נורמאלי, האחוזון ה-70 הוא 90 ק"ג והאחוזון ה-20 הוא 70 ק"ג. מהי התוחלת וסטיית התקן של התפלגות משקלי העציצים כעת?

שאלה 4 (25 נקודות, 5 נקודות לכל סעיף)

הדגימו את משפט הגבול המרכזי בפייתון:

- הראו בטבלה את פונקציית ההתפלגות של תוצאת הטלת קובייה (X). מה הממוצע וסטיית התקן של X ? נסמן ב- X_2 ממוצע מקרי של 2 הטלות קובייה. מה הממוצע וסטיית התקן של X_2 ?
- הגרילו את X_2 עשרות אלפי פעמים וציירו את התפלגות שלו בעזרת הפונקציה `kdeplot` של `seaborn`.
- מה הממוצע וסטיית התקן האמפיריות של ההתפלגות שיצרתם? האם זה מתאים לחישוב מסעיף א'?
- הוסיפו לשרטוט את הצפיפות של התפלגות נורמלית עם אותם פרמטרים (תוחלת ושונות) כמו X_2 . האם ניתן להגיד ש X_2 מתפלג בקירוב נורמלית?
- הגדילו בהדרגה את גודל המדגם עבורו מחושב הממוצע X_n (זאת אומרת, את מספר הטלות הקובייה המרכיבות את הממוצע) וחזרו על סעיפים ב' ו-ג'. הראו לפחות 4 גדלי מדגם שונים.

- ו. מאיזה גודל מדגם (בערך) הייתם אומרים שהממוצעים מתפלגים נורמאלית? עבור המקרה הזה, השתמשו ב CLT על מנת למצוא את השברון ה 0.95 באופן מקורב.
- ז. (5 נקודות בונוס) בחרו משתנה מיקרי רציף המתפלג לא נורמלית וחיזרו על סעיפים א' עד ה' (לא כולל הצגת טבלה של פונקציית ההתפלגות).