README: מאגר לידות חי לשנת 2014

מחברים

שלומי הוד, אוניברסיטת בוסטון
ד"ר מיטל אבגיל צדוק, תמנ"ע
פרופ' רן קנטי, אוניברסיטת בוסטון

תאריך עדכון אחרון

14 במרץ 2024

קישור למאגר ולמסמך זה
(בעברית ובאנגלית)

https://data.gov.il/dataset/birth-data

רפרנס למאגר

Shlomi Hod, Meytal Avgil Tsadok, and Ran Canetti. "Israel's National Registry of Live Births." *Ministry of Health, Government of Israel*. https://data.gov.il/dataset/birth-data

SHA-256 של קובץ הנתונים

`c493b8f09ca24f90621124b26332e63b29689814f575b9c7a1d4447b5753f751`

זהו **פרויקט פיילוט**: על מנת לשפר את תהליך פרסום הנתונים בעתיד, חשוב לנו לקבל הערות ושאלות מהציבור. מייל ליצירת קשר: birth_data@moh.gov.il

תוכן עניינים

פרק א' – מבוא

מאגר לידות חי הוא המקור הרשמי המכיל תיעוד של כל לידות החי שאירעו בישראל, באשר לפחות אחד מההורים הינו בעל תעודת זהות ישראלית (ללא לידות של ישראלים שאירעו בחו"ל)[21]. לנתוני המאגר חשיבות רבה, בין השאר, לבחינת סוגיות דמוגרפיות, לקביעת מדדי איכות, קביעת מדיניות ולמחקר רפואי. לכן, הנגשת נתונים אלו לציבור היא בעלת ערך רב ומאפשרת לגורמים מגוונים, שעד כה לא היתה להם גישה לנתונים אלו, לחקור את הנתונים ולהפיק תובנות בנושא.

עם זאת, באופן טבעי, נתונים רפואיים בכלל ונתוני לידות בפרט, מכילים מידע פרטי ורגיש אודות אמהות וילודים, ולכן אי אפשר לפרסם אותם במלואם כמות שהם. כדי לאפשר פרסום של נתונים אלו בצורה מועילה ככל האפשר מבלי לפגוע בפרטיות האמהות והילודים, אנו מציגים לראשונה פרסום של נתונים מותממים מקובץ הלידות הלאומי של ישראל. **הנתונים הופקו בעזרת שיטות התממה מודרניות שמאפשרות ליצור נתונים סינטטיים באיכות גבוהה. שיטות ההתממה אלו מספקות הגנה חזקה ואפקטיבית על הפרטיות של האמהות והילודים, ובד בבד מבטיחות שמגוון רחב של ניתוחים סטטיסטיים יניבו תוצאות דומות דיין לניתוחים המבוצעים על הנתונים הגולמיים.**

על מנת להבטיח את האיכות של תוצאות מבחנים סטטיסטיים המופעלים על הנתונים המותממים, הוגדרו מספר **קריטריוני קבלה**. קריטריונים אלו מבטיחים את השגיאה המירבית של מספר שאילתות סטטיסטיות בין ההרצה על הנתונים הגולמיים וההרצה על הנתונים המותממים. קריטריוני הקבלה נקבעו בעצת מומחי תוכן, שלפי דעתם המקצועית הגדירו את הקריטריונים כמרכזיים וחשובים, והם מכסים מגוון רחב של שאילתות סטטיסטיות, בדגש על שכיחויות חד/רב-כיווניות ומדדי מרכז. השאילתות וספי הקבלה פותחו בעזרת בחינה של פרסומים המבוססים על נתונים דומים והיוועצות עם מומחי תוכן בתחום. קריטריון קבלה נוסף בודק שהנתונים המותממים נאמנים למקור גם ברמת השורה הבודדת. **הנתונים המותממים המפורסמים לציבור עברו בהצלחה את כל קריטריוני הקבלה.**

הגנת הפרטיות של שיטות ההתממה בהן נעשה שימוש עבור נתוני לידות החי מבוססות על מדד פרטיות המכונה פרטיות **דיפרנציאלית**[3]. פרטיות דיפרנציאלית היא מדד שפותח על מנת לתת מענה לאתגרי הפרטיות בעולם נתוני העתק. בפרט, מידת הפרטיות המובטחת[4] נשמרת גם כאשר מאגר הנתונים המותמם מוצלב עם מידע נוסף כלשהו, בין אם מידע זמין בהווה או מידע עתידי.

---

[1] משרד הבריאות – אגף המידע – לידות בישראל, ראו פרק "שיטות והגדרות" לתיאור מפורט על אופן איסוף הנתונים והגדרות המשתנים במאגר

[2] הלשכה המרכזית לסטטיסטיקה – לידות חי

[3] https://en.wikipedia.org/wiki/Differential_privacy

[4] פרפרזה מתוך Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends® in Theoretical Computer Science, 9(3–4), 211–407.

חשוב לציין שזהו **פרויקט פיילוט**. על מנת לשפר את תהליך פרסום הנתונים בעתיד, חשוב לנו לקבל הערות ושאלות מהציבור, במיוחד אודות **אילו שאילתות ואילו ניתוחים סטטיסטיים אתם מעוניינים להריץ על הנתונים**.
מוזמנים לשלוח לנו את שאלות או הערות במייל: birth_data@moh.gov.il

מסמך README זה מספק הסבר על הנתונים ואופן הפקתם.

במסמך זה ישנם שני חלקים:

1. **החלק ראשון** מספק את המידע הנדרש על מנת לעבוד עם הנתונים המפורסמים לציבור: מאפייני הנתונים המותממים (פרק ב'), סוגי השאילתות עליהן הנתונים יכולים להשיב ועל אילו הם אינם מתוכננים לענות (פרק ג') ופנייה לציבור בנוגע למשוב המבוקש (פרק ד').

2. **החלק השני** מעמיק בפרטים הטכניים: הגדרה של קריטריוני הקבלה והערכת טיב הנתונים המותממים (פרק ה'), הסבר מעמיק יותר על פרטיות דיפרנציאלית (פרק ו') ולבסוף תיאור שלם של תהליך הפקת הנתונים המותממים, כולל תיעוד שלבי העיבוד המקדים של הנתונים הגולמיים (פרק ז').

חלק ראשון: עבודה עם הנתונים

פרק ב' – תיאור הנתונים המותממים המפורסמים לציבור

הנתונים מפורסמים לציבור בטבלה המכילה 165,915 שורות של לידות יחיד (Singleton)[5] של ילוד חי[6] בישראל בשנת 2014. בטבלה ישנן שש עמודות כמפורט בהמשך. טרם הפקת הנתונים המפורסמים לציבור, הנתונים הגולמיים נוקו וטיובו כפי שמתואר בסעיף "הכנת הנתונים הגולמיים".

שם העמודה בקובץ הנתונים	תיאור העמודה בעברית	ערכים אפשריים
birth_month	חודש הלידה בשנת 2014	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12
mother_age	מספר השנים שמלאו לאם בתאריך הלידה	<18, 18–19, 20–24, 25–29, 30–34, 35–36, 37–39, 40–42, 43–44, 44<
parity	מספר לידת חי[7] של האם	1, 2–3, 4–6, 7–10, 10<
gestation_week	שבוע ההריון שבמהלכו התרחשה הלידה	<29, 29–31, 32–33, 34–36, 37–41, 41<
birth_sex	מין הילוד בלידה	M (זכר), F (נקבה)
birth_weight	משקל הילוד בגרמים במעמד הלידה	ברזולוציה של 100 גרם בין 1500 ל-4500 גרם: <1500, 1500–1599, ..., 4400–4499, 4499<

[5] לידת יחיד – לידה הכוללת עובר אחד בלבד.

[6] לידת חי – לידה של עובר, ללא קשר למשך ההריון, שאחרי היפרדו מהאם בין אם חבל הטבור נחתך ובין אם לא, בין אם השליה עדיין קשורה ובין אם לא, מגלה לפחות אחד מסימני החיים הבאים: נשימה, דפיקות לב, דופק חבל הטבור או תנועות ברורות של שרירים רצוניים.

[7] מספר לידת החי מתייחס למספר העוברים החיים שנולדו, ולא למספר ההריונות. הלידה המדווחת בשורה נכללת במספר לידת חי.

פרק ג' – שימוש

הנתונים המותממים מיועדים בראש ובראש לחישובי **שכיחויות יחסיות חד-כיווניות ורב-כיווניות** (k-way marginal frequencies).

לפי קריטריון קבלה 1 (פרק ה'), הטעות **המוחלטת המקסימלית** בין השכיחויות היחסיות בנתונים הגולמיים לשכיחויות היחסיות בנתונים המותממים עבור שאילתות אלו אינה עולה על 0.44% (מתוך כלל השורות).

קריטריון קבלה 2 מוודא שהטעות **היחסית בשכיחות החד-כיוונית** של הערכים הנפוצים פחות היא קטנה יחסית, ואינה עולה על 28.4% (מתוך מספר השורות עם ערך זה).

קריטריוני קבלה 3-5 מראים שהנתונים מאפשרים גם חישובי **מדדי מרכז** כמו חציונים וממוצעים, של מספר עמודות ב**פילוחים** מגוונים.

קריטריוני קבלה 6-7 מראים כי הנתונים מאפשרים גם הרצת רגרסיה לינארית, עם הבדל קטן בממוצע שגיאות הניבוי ובדיוק ערכי מקדמי הרגרסיה.

מסימולציות שערכנו, הנתונים מתאימים גם לחישובי **קורלציות**: ההפרש בין מקדמי קורלציות על הנתונים המותממים והנתונים הגולמיים הוא לכל היותר 0.05 (לדוגמא, אם מקדם הקורלציה בנתונים המותממים הוא 0.3, אזי בנתונים המקוריים המקדם נע בין 0.25 לבין 0.35). נדגיש שחישוב קורלציות אינו חלק מקריטריוני הקבלה.

יש לציין שהנתונים המותממים המפורסמים לציבור אינם מיועדים לניתוח חריגים ואנומליות, רגרסיות מורכבות, בדיקת השערות ולמידת מכונה. בפרט, ניתוחים אלו עלולים להפיק תוצאות לא מדויקות ביחס לנתונים הגולמיים.

ניתוחים סטטיסטיים הכוללים ערכים נדירים בקצוות ההתפלגות של העמודות (למשל, גיל האם <18, שבוע לידה <29, מספר לידת חי >10) ו**בחיתוכים של מספר עמודות הכוללים מספר קטן של רשומות**, עלולים להיות **לא מדוייקים** ו**יש להימנע** מהסקת מסקנות על פיהם. קריטריון קבלה 1 מבטיח שהשגיאה במספר הרשומות **לכל חיתוך** היא לכל היותר של 730 רשומות.

בנוסף, הנתונים לא מאפשרים <u>הצלבה ברמת השורה</u> עם מאגרי נתונים אחרים הודות לתהליך ההתממה. כל ניסיון להתאמה בהצלבה הוא חסר משמעות.

פרק ד' – פנייה לציבור למשוב על הפיילוט

כאמור, מדובר בפרויקט **פיילוט**. על מנת לשפר את תהליך פרסום הנתונים בעתיד, חשוב לנו לקבל הערות ושאלות מנומקות מהציבור, במיוחד לגבי הנקודות הבאות:

1. כיצד הנתונים הועילו לכם ולשם מה עשיתם בהם שימוש?

2. אילו שאילתות אתם מריצים על הנתונים בהתאם לפרק ג' – שימוש? אילו שאילתות נוספות תרצו להריץ על הנתונים?

3. האם קריטריוני הקבלה הולמים את סוגי השאילתות שתרצו להריץ? אלו קריטריוני קבלה מיותרים, ואילו חסרים? האם הרף שנקבע בכל קריטריון קבלה מתאים?

4. אלו עמודות או נתונים נוספים תרצו לקבל בהמשך?

5. האם הרזולוציה של הנתונים (חלוקה לקטגוריות) טובה דיה למטרותיכם? אם לא, מדוע? מהי הרזולוציה המתאימה לכך?

6. האם מסמך ה-Readme מקיף מספיק? האם הוא ברור? מה חסר או דורש הרחבה?

נשמח לשמוע את הערותיכם. מייל ליצירת קשר: birth_data@moh.gov.il

חלק שני: העמקה טכנית

פרק ה' – הערכת טיב הנתונים לפרסום וקריטריוני הקבלה

על מנת לוודא שהנתונים המותממים מאפשרים ניתוח סטטיסטי טוב דיו, הנתונים עברו סדרה של בדיקות על סמך קריטריוני קבלה שהוגדרו מראש בשיתוף מומחי תוכן מתחום הפריון. הקריטריונים משקפים את סוגי השאילתות שהנתונים המותממים מאפשרים לענות עליהם (ראו פרק ג').

קריטריוני הקבלה מתחלקים לארבעה סוגים: (1) טעות מקסימלית בטבלת שכיחויות, (2) טעות מקסימלית בממוצע לפי פילוחים, (3) טעות ברגרסיה לינארית ו-(4) נאמנות (faithfulness) לנתונים הגולמיים.
הנתונים המותממים המפורסמים לציבור עברו בהצלחה את כל קריטריוני הקבלה.

קריטריון קבלה מורכב מ**מטריקה** ו**רף עליון**. המטריקה בוחנת את הביצועים של הנתונים המותממים על שאילתה סטטיסטית אחת או על אוסף שלהן **בהשוואה** לנתונים הגולמיים. אם תוצאת המטריקה היא אפס (עבור קריטריונים 1, 3-7) או 1 (עבור קריטריון 2), אז השאילתות הסטטיסטיות מחזירות תוצאה זהה עבור הנתונים המותממים והנתונים הגולמיים – כלומר אין שגיאה. לכל אחד מקריטריוני הקבלה הוגדר מראש רף עליון שאסור לתוצאת המטריקה להיות גדולה ממנו על מנת שהנתונים המותממים יעמדו בקריטריון הקבלה. המטריקות וספי הקבלה נקבעו מראש טרם הפקת הנתונים המותממים והם מתוארים בהמשך. גם תוצאות המטריקות מפורסמות לציבור בעמודים הבאים.

בכל הקריטריונים אנו משווים בין הנתונים המותממים לבין הנתונים הגולמיים לאחר ביצוע binning (חלוקה לקטגוריות המפורטות בעמודה "ערכים אפשריים") המתואר בפרק ב'.

קריטריוני קבלה מסוג טעות מקסימלית בטבלת שכיחויות

מס'	שם	תיאור	רף עליון (שהתוצאה צריכה להיות קטנה ממנו)	תוצאה עבור הנתונים בקובץ המפורסם לציבור (ס"ת[8])
1	מקסימום טעות מוחלטת בשכיחות חד/רב-כיוונית	**קריטריון זה מוודא שחישובי שכיחויות של הנתונים המותממים יהיו קרובים מאוד לחישוב על הנתונים הגולמיים.** 1. מחשבים את בל טבלאות השכיחות חד/רב-כיווניות האפשריות על הנתונים המותממים ועל הנתונים הגולמיים. 2. מחשבים את ההפרש המוחלט בין כל זוג תאים מתאימים (בלומר, אותו החיתוך); הראשון מטבלת השכיחות של הנתונים המותממים והשני מטבלת השכיחות של הנתונים הגולמיים. 3. קריטריון הקבלה מתייחס למקסימום בין כל ההפרשים המוחלטים, בלומר טעות מקסימלית מוחלטת.	1%	0.44% בלומר 730 רשומות (ס"ת <0.001 נקודות אחוז)

[8] סטיית התקן המדווחת עבור התוצאות של קריטריוני הקבלה נובעת ממנגנון ההתממה של תוצאות קריטריוני הקבלה. ראו הערת שוליים מספר 23.

2	מקסימום טעות יחסית בשביחות חד-כיוונית	**קריטריון זה מודא שחישובי שביחויות חד-כיווניות של הנתונים המותממים יהיו קרובים מאוד לחישוב על הנתונים הגולמיים באופן יחסי. מטרת הקריטריון היא לבחון את הדיוק בעיקר עבור ערכים נדירים (לרוב בקצוות ההתפלגות).** 1. מחשבים את טבלאות השביחויות החד-כיווניות האפשריות (בלומר, של משתנה אחד) על הנתונים המותממים ועל הנתונים הגולמיים . 2. מחשבים את היחס בין כל זוג תאים מתאימים; הראשון מטבלת השביחות של הנתונים המותממים והשני מטבלת השביחות של הנתונים הגולמיים. אם היחס קטן מאחד, בוחרים בהופכי כדי לקבל מדד גדול מאחד. 3. קריטריון הקבלה מתייחס למקסימום בין כל היחסים הגדולים מאחד, כלומר טעות מקסימלית יחסית.	פי 2	1.284 (ס"ת 0.135)

קריטריוני קבלה מסוג טעות מקסימלית בממוצע לפי פילוחים

קריטריונים אלו בחנו ממוצעים ולא חציונים כי הממוצע רגיש יותר לערכי קיצון מאשר חציון, ולרוב, בטבלאות עם מספר רב של שורות, טעות בממוצע תהיה גדולה יותר מאשר טעות בחציון. שימוש בממוצעים מפשט את תהליך בקרת האיכות המתבצע בעזרת קריטריוני קבלה, אך בשלב ניתוח הנתונים, **חציונים** מתאימים יותר בגלל האופי של העמודות בקטגוריות.

על מנת לחשב ממוצעים על נתונים קטגוריאליים, יש לתרגם את הקטגוריות לערכים מספריים. הסבר על התהליך מפורט לאחר הטבלה הבאה.

מס'	שם	תיאור	רף עליון (שהתוצאה צריכה להיות קטנה ממנו)	תוצאה[9] עבור הנתונים בקובץ המפורסם לציבור (ס"ת[8])
3	מקסימום טעות מוחלטת בממוצע מספר הלידה מפולח לפי גיל האם	**קריטריון זה מוודא שחישובי ממוצעי מספר הלידה על הנתונים המותממים יהיו קרובים מאוד לחישוב על הנתונים הגולמיים.** 1. מפלחים את הנתונים המותממים והנתונים הגולמיים לפי גיל האם בהתאם לפילוח המפורט. 2. מחשבים את הממוצע[10] של מספר הלידות עבור כל קבוצה בפילוח. 3. מחשבים את ההפרש המוחלט בין הממוצעים של כל זוג הקבוצות המתאימות בפילוח; הראשון מהנתונים המותממים והשני מהנתונים הגולמיים. 4. קריטריון הקבלה מתייחס למקסימום הטעות בין כל ההפרשים המוחלטים. כלומר טעות מקסימלית מוחלטת.	0.3 לידות חי	-0.014 (ס"ת 0.044)

[9] בגלל מנגנון ההתממה של תוצאות קריטריוני הקבלה, ייתכן ויתקבלו תוצאות שליליות ולכן יש להיעזר בסטיית התקן על מנת לפרש את התוצאות. ראו הערת שוליים מספר 23.

[10] המנגנון לחישוב ממוצע מפולח של הנתונים הגולמיים בעזרת פרטיות דיפרנציאלית משתמש בדגימה אקראית (ללא החזרה) של 85%-98% מהרשומות בכל פילוח.

4	מקסימום טעות בממוצע משקל הילוד במעמד הלידה מפולח לפי מין הילוד בלידה, מספר לידה, שבוע הלידה וגיל האם (כל משתנה בנפרד)	**קריטריון זה מוודא שחישובי ממוצעי משקל הילוד על הנתונים המותממים יהיו קרובים מאוד לחישוב על הנתונים הגולמיים.** 1. עבור כל אחד ממשתני הפילוח (מין, מספר לידת חי, שבוע לידה וגיל האם) א. מפלחים את הנתונים המותממים והנתונים הגולמיים לפי גיל האם בהתאם לפילוח המפורט. ג. מחשבים את הממוצע[10] של מספר הלידות עבור כל קבוצה בפילוח. ד. מחשבים את ההפרש המוחלט בין הממוצעים של כל זוג הקבוצות המתאימות בפילוח; הראשון מהנתונים המותממים והשני מהנתונים הגולמיים. 2. קריטריון הקבלה מתייחס למקסימום הטעות בין כל ההפרשים המוחלטים בין כל משתני הפילוח. כלומר טעות מקסימלית מוחלטת.	100 גרם	28.634 (ס"ת 3.821)
5	מקסימום טעות מוחלטת בממוצע שבוע הלידה מפולח לפי מספר לידת חי וגיל האם (כל משתנה בנפרד)	**קריטריון זה מוודא שחישובי ממוצעי שבוע הלידה על הנתונים המותממים יהיו קרובים מאוד לחישוב על הנתונים הגולמיים.** 1. עבור כל אחד ממשתני הפילוח (מספר לידת חי וגיל האם): א. מפלחים את הנתונים המותממים והנתונים הגולמיים לפי שבוע הלידה בהתאם לפילוח המפורט לאחר הטבלה. ג. מחשבים את הממוצע[10] של מספר הלידות עבור כל קבוצה בפילוח. ד. מחשבים את ההפרש המוחלט בין הממוצעים של כל זוג בקבוצות המתאימות בפילוח; הראשון מהנתונים המותממים והשני מהנתונים הגולמיים . 2. קריטריון הקבלה מתייחס למקסימום הטעות בין כל ההפרשים המוחלטים בין כל משתני הפילוח. כלומר טעות מקסימלית מוחלטת.	1 שבוע	0.062 (ס"ת 0.033)

1. כדי לחשב ממוצע של משתנה המחולק לקטגוריות (binning) יש להמיר כל קטגוריה לערך מספרי אחד. ההמרה בוצעה לפי הכללים הבאים:
   a. אם יש רק ערך אחד בקטגוריה, נעשה שימוש בערך זה (למשל הערך 1 במספר לידת חי)

b. אם הקטגוריה מייצגת טווח סגור (מערך ... עד ערך ...), נעשה שימוש בערך הממוצע של קצוות הקטגוריה (למשל 2.5 עבור הטווח 2-3 במספר לידת חי)

c. אם הקטגוריה מייצגת טווח פתוח (גדול מ-..., קטן מ-...), נעשה שימוש בקצה הקטגוריה (למשל 10 עבור הטווח 10< במספר לידת חי)

2. כדי לבצע את הפילוחים, נעשה שימוש בקטגוריות הבאות שהן זהות לקטגוריות מקובץ הנתונים המותממים או איחוד שלהן (רזולוציה זהה או נמוכה יותר):

שם העמודה בקובץ הנתונים	תיאור העמודה בעברית	פילוח
mother_age	גיל האם בשנים בעת הלידה	≤24, 25-29, 30-34, 35≤
parity	מספר הלידה	1, 2-3, 4≤
gestation_week	שבוע ההריון בעת הלידה	<37, 37≤
sex	מין הילוד	M (זכר), F (נקבה)
birth_weight	משקל הילוד בגרמים	<2500, 2500-3999 , 4000≤

קריטריוני קבלה מסוג טעות ברגרסיה לינארית

שני קריטריוני הקבלה עוסקים ברגרסיה לינארית ל**ניבוי משקל הלידה בעזרת שאר המשתנים** בקובץ הנתונים.

מס'	שם	תיאור	רף עליון (שהתוצאה צריכה להיות קטנה ממנו)	**תוצאה עבור הנתונים בקובץ המפורסם לציבור** (ס"ת[8])
6	סכום הטעויות מוחלטת במקדמי רגרסיה לינארית	**קריטריון זה מוודא שהמקדמים של רגרסיה לינארית על הנתונים המותממים יהיו קרובים מאוד למקדמים על הנתונים הגולמיים.** 1. מתקננים (ציון תקן) את כל אחת מהעמודות בנתונים המותממים והנתונים הגולמיים. בשני המקרים משתמשים בממוצע ובס"ת של כל העמודה המחושבים על הנתונים המותממים. 2. מחשבים רגרסיה לינארית[11] על הנתונים המותממים המתוקננים. מחשבים רגרסיה לינארית עם פרטיות דיפרנציאלית על הנתונים הגולמיים המתוקננים. 3. מחשבים את ההפרש המוחלט בין כל אחד מהמקדמים של שתי הרגרסיות. 4. קריטריון הקבלה מתייחס לסכום בין כל ההפרשים המוחלטים (טעויות).	30	27.185

[11] רגרסיה לינארית שחושבה הנתונים המותממים המתוקננים ומוערכת על הנתונים המותממים מציגה את הביצועים הבאים:

R^2 = 0.210
MAE = 336.499

7	טעות מוחלטת בשגיאת הניבוי בין רגרסיות לינאריות שאומנו על הנתונים המותממים והגולמיים	**קריטריון זה מוודא שמודל רגרסיה לינארית שאומן על הנתונים המותממים הוא בעל שגיאת ניבוי נמוכה כשהוא מופעל על הנתונים הגולמיים ביחס למודל שאומן על הנתונים הגולמיים.** 1. עובדים עם הרגרסיות שהתקבלו בקריטריון קבלה 6. 2. מחשבים את ה- Mean Average Error (ממוצע השגיאה המוחלטת - MAE) בניבוי משקל הלידה של כל אחת מהרגרסיות על הנתונים הגולמיים המתוקננים. 3. קריטריון הקבלה מתייחס להפרש המוחלט בין ה-MAE של שתי הרגרסיות. שימו לב: אימון המודל וחישוב שגיאת הניבוי נעשית ללא held-out dataset.	5 גרם	0.351 (ס"ת 1.364)

קריטריון קבלה מסוג נאמנות (faithfulness) לנתונים הגולמיים

קריטריוני הקבלה הקודמים מוודאים שניתן לחשב שאילתות סטטיסטיות אגרגטיביות באיכות טובה דיה על הנתונים המותממים. לעומת זאת, קריטריון הקבלה האחרון, נאמנות, בוחן את נאמנות הנתונים המותממים לנתונים הגולמיים ברמת השורה הבודדת. כלומר, כמעט כל השורות בטבלה המותממת דומות מאוד לשורות מהטבלה המקורית.

מס'	שם	תיאור	רף עליון (שהתוצאה צריכה להיות קטנה ממנו)	תוצאה עבור הנתונים בקובץ המפורסם לציבור (ס"ת^2)
8	נאמנות (faithfulness)	קיימת התאמה של אחד לאחד בין טבלת הנתונים המותממים לטבלת הנתונים הגולמיים, כך שכל שורה מותממת דומה דיה לשורה גולמית (ראו בהמשך את ההגדרה לדמיון בין זוג שורות), פרט ל-Q% מהשורות. הקריטריון מודד את ערך ה-Q%.	5%	3.876% בלומר 6,431 רשומות (ס"ת <0.001 נקודות אחוז)

קריטריון הנאמנות מבוסס על השוואת שורה מותממת לשורה גולמית, ובחינת הדימיון בניהן.

נאמר שזוג שורות דומות דיו אם העמודות הבאות זהות בדיוק ברמת הקטגוריה:

1. birth_month (חודש לידה בשנת 2014)
2. parity (מספר לידת חי)
3. birth_sex (מין הילוד בלידה)

ובעמודות הבאות יכולה להיות תזוזה של קטגוריה אחת בלבד למטה או למעלה. תנודה זו יכולה להתרחש אך ורק בעמודה אחת לכל היותר:

1. mother_age (מספר השנים שמלאו לאם בתאריך הלידה)
2. gestation_week (שבוע ההריון שבמהלכו התרחשה הלידה)
3. birth_weight (משקל הילוד בגרמים במעמד הלידה)

פרק ו' – פרטיות

על מנת להנגיש את נתוני הלידות לציבור ולאפשר ניתוחים סטטיסטיים למטרות מגוונות, יש להגן על פרטיות האמהות והילודים הנמצאים במאגר. ישנן שיטות התממה רבות, אולם רבות מהן לא נותנות הגנה מספקת – במיוחד כאשר המידע שעבר התממה מוצלב עם מידע נוסף, פומבי או פרטי, שנמצא בידי צד שלישי. דוגמא לאוסף שיטות הנמצא בשימוש (אך לא נותן הגנה מספקת) היא k-anonymity[12].

כדי לענות על אתגר זה, חוקרי הגנת פרטיות פיתחו אוסף שיטות התממה מודרניות המבוססות על הכנסת רעש סטטיסטי מדוד ומבוקר לתהליך הפקת הנתונים המותממים החישוב בצורה שתבטיח פרטיות וגם תשמור על איכות הנתונים. בנוסף, חוקרים פיתחו מדד כמותי לרמת הפרטיות המובטחת על ידי מנגנוני התממת מידע. למדד זה יש הרבה תכונות רצויות, בין השאר, המדד מבטיח שרמת הפרטיות תישמר גם כאשר מאגר הנתונים המותמם מוצלב עם מידע נוסף כלשהו, בין אם בהווה או בכל נקודת זמן בעתיד.

מדד **הפרטיות הדיפרנציאלית**[13] מתאים ערך מספרי ε (אפסילון) לכל שיטת התממה באופן הבא. המדד משווה את המידע המותמם המתקבל מהפעלת שיטת ההתממה על כל שתי טבלאות ששונות במידע על אדם אחד (או, במקרה שלנו, לידה אחת) בלבד, ומודד את השינוי היחסי. ככל שהשינוי גדול יותר ערך ε (אפסילון) גדול יותר.

המשמעות של פרטיות דיפרנציאלית היא שכל מסקנה על אדם אחד (או לידה אחת) על בסיס הטבלה המותממת תלויה במידע המקורי על אותו אדם רק באופן מוגבל – שכן הטבלה המותממת יכולה הייתה להתקבל בהסתברות דומה גם אילו מידע זה היה אחר לחלוטין[14]. כזכור, ערך ה-ε (אפסילון) מודד את מידת ההשפעה של שורה בודדת על הנתונים המותממים. לכן, ככל ש-ε (אפסילון) קטן השינוי היחסי בין ההסתברויות גם הוא קטן ולכן הבטחת הפרטיות מתחזקת.

---

[12] Cohen, A. (2022). Attacks on Deidentification's Defenses. (2022). 31st USENIX Security Symposium.

[13] Wood, A., Altman, M., Bembenek, A., Bun, M., Gaboardi, M., Honaker, J., Nissim, K., O'Brien, D.R., Steinke, T. and Vadhan, S. (2018). Differential privacy: A primer for a non-technical audience. Vand. J. Ent. & Tech. L., 21, 209

[14] בניסוח טכני: נניח את הנתונים הגולמיים בטבלה T, ונחליף בה שורה מסויימת בנתונים אחרים (לא חייבים להיות מתוך הטבלה T) כך שנקבל טבלה T'. כלומר הטבלאות T ו-T' נבדלות רק בשורה אחת. עכשיו נריץ את שיטת ההתממה המבוססת על פרטיות דיפרנציאלית על הטבלאות T ו-T', ונניח שקיבלנו את הטבלאות המותממות X ו- X' בהתאמה. שיטת ההתממה כוללת מרכיב אקראי מדוד ומכוייל (כמו הוספת רעש), לכן הפלטים X ו-X' הם משתנים מקריים המתקבלים באיזושהי הסתברות. ההגנה של פרטיות דיפרנציאלית מודדת את השינוי המקסימלי בהסתברות לקבל את הפלטים f(X) ו-f(X') לכל פונקציה f. פרטיות דיפרנצאלית מבטיחה שהתפלגות של X היא קרובה יחסית להתפלגות של X'. ההבטחה הזו, שהשינוי המקסימלי הוא קטן, נותנת הגנת פרטיות חזקה מאוד ברמת השורה הבודדת (כלומר אדם, או במקרה שלנו לידה), ובפרט היא מונעת מגוון של סיכוני פרטיות כמו זיהוי מחדש.

בדי להפיק את הנתונים המותממים של מאגר הלידות, נעשה שימוש בערך ε (אפסילון) של **9.98**, שהוא בסדר גודל דומה של יישומי פרטיות דיפרנציאלית אחרים בעולם[15]. לצורך השוואה, מפקד האוכלוסין האמריקאי של שנת 2020 התפרסם לציבור תוך שימוש בפרטיות דיפרנציאלית עם ערך אפסילון של 19.6.

מאז הצעת מדד הפרטיות הדיפרנציאלית ב-2006, פותחו עשרות שיטות התממה שעונות על הדרישה, כשכל שיטה מותאמת לסיטואציה וסוג נתונים אחרים. כדי לשחרר את מאגר הלידות לציבור נעשה שימוש בשיטת PrivBayes כפי שמופרט בפרק ז'.

לכל השיטות המבוססות על פרטיות דיפרנציאלית יש רעיון משותף: הוספה של כמות מדודה ומבויילת של רעש אקראי לתוך הנתונים. הרעש הוא אקראי ולא ניתן לחיזוי, כך שאי-אפשר לגלות את ערכם המדויק של הנתונים המקוריים. כאמור, קיים מגוון של שיטות להכנסת הרעש בשלבים שונים של תהליך עיבוד הנתונים, כאשר שיטות שונות מאפשרות שמירת הדיוק של ניתוחים סטטיסטים שונים על הנתונים (באופן יותר כללי, שמירת אספקטים שונים של אוסף הנתונים המקורי), גם עבור ערכים קטנים של ε (אפסילון). באופן כללי, ככל שבמאגר יש יותר נתונים על יותר אנשים, מידת הרעש היא קטנה יותר. ואכן, מאגר הלידות לשנת 2014 מכיל מספר רב של שורות, כך שמידת הרעש היא נמוכה, ולכן איכות הנתונים ונאמנותם גבוהות.

דוגמאות לשימושים בפרטיות דיפרנציאלית במגזר הציבורי והפרטי

- בעשור האחרון ישנן יותר ויותר דוגמאות מהעולם לשימוש מוצלח של שיטות המבוססות על מדד פרטיות דיפרנציאלית במגזר הציבורי והעסקי. לשכת מפקד האוכלוסין של ארה"ב[16] עברה לשימוש בפרטיות דיפרנציאלית לשחרור נתונים לציבור החל מהמפקד של שנת 2020. נתוני המפקד הם בעלי חשיבות מהותית לדמוקרטיה האמריקאית, ובין השאר משמשים לקביעת אזורי בחירה והקצאת מימון ממשלתי. ההתקדמות במחקרי הפרטיות הראו כיצד שיטות ההתממה הקודמות, המבוססות בין השאר על הכללה, אינן נותנות הגנה מספקת לפי דרישות החוק האמריקאי, ופרטיות דיפרנציאלית היא המענה ההולם. כאמור, ה-ε (אפסילון) במפקד האוכלוסין של ארה"ב נקבע ל-19.6.

[15] https://desfontain.es/privacy/real-world-differential-privacy.html

[16] https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance.html

- ה-IRS (רשות המיסים האמריקאית) עשו שימוש בקונפיגורציה הכוללת נתונים סינטטיים ופרטיות דיפרנציאלית כדי להנגיש לציבור ולחוקרים מידע אודות תשלומי מיסים[17]. מחלקת החינוך של ארה"ב פיתחה אתר אינטרקטיבי המאפשר לציבור להשוות בין מוסדות אקדמיים, למשל לפי סיכויי קבלה, אחוז הבוגרים והכנסה של בוגרים. מקור חלק מהמידע מגיע מדיווחי המס של בוגרים, ופרטיות דיפרנציאלית שימשה להתממת הנתונים[18].

- בין אפריל 2020 לאוקטובר 2022 גוגל[19] פירסמה באופן תדיר 'דו"חות מוביליות בקהילה' (Community Mobility Reports) המאפשרים לחוקרים ולציבור הרחב לנתח מידע שינוי בהתנהגות הניידות של אנשים בעקבות COVID-19 בתגובה למדיניות אפידמיולוגיות. המידע מבוסס על נתוני-המיקום שגוגל אוספת, והוא מותמם בעזרת הוספת רעש אקראי מדוד באופן שמבטיח פרטיות דיפרנציאלית. חברות נוספות כמו אפל[20] ולינקדאין[21] גם הן משתמשות בפרטיות דיפרנציאלית כדי להבטיח הגנה על פרטיות המשתמשים.

[17] https://www.bea.gov/system/files/2021-02/Burman-Presentation-ACDEB-021921.pdf
[18] https://www.usenix.org/system/files/pepr22_slides_miklau.pdf
[19] /https://blog.google/technology/health/covid-19-community-mobility-reports
[20] https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf
[21] https://engineering.linkedin.com/blog/2019/04/privacy-preserving-analytics-and-reporting-at-linkedin

פרק ז' – מנתונים גולמיים לנתונים מותממים

הכנת הנתונים הגולמיים

לאחר שליפת טבלה ממאגר הלידות של לידות יחיד (singleton) של ילוד חי לשנת 2014, השורות הבאות הוסרו על מנת לטייב את הנתונים:

1. שורות עם נתונים חסרים באחד המשתנים או יותר (לשם קבלת קובץ עם מידע מלא)
2. שורות עם משקל לידה הקטן מ-500 **או** גדול מ-5500 (בשל מקרים קיצוניים שלא יופקו באופן מדויק דיו לאחר התממה וכן בשל חשש לטעות בנתון)
3. שורות עם שבוע הריון קטן מ-22 **או** גדול מ-44 (הטווח הנהוג בלידות חי)
4. שורות עם גיל אם במהלך הלידה קטן מ-23 שנים **וגם** מספר לידת חי גדול מ-6 לידות (בשל מקרים קיצוניים שלא יופקו באופן מדויק דיו לאחר התממה)
5. שורות עם גיל האם במהלך הלידה קטן מ-20 **וגם** מספר לידת חי גדול מ-3 (בשל מקרים קיצוניים שלא יופקו באופן מדויק דיו לאחר התממה)
6. שורות עם שבוע הריון במהלך הלידה קטן מ-26 **וגם** משקל לידה גדול מ-1499 (בשל מקרים קיצוניים שלא יופקו באופן מדויק דיו לאחר התממה)
7. שורות עם שבוע הריון במהלך הלידה קטן מ-29 **וגם** משקל לידה גדול מ-2999 (בשל מקרים קיצוניים שלא יופקו באופן מדויק דיו לאחר התממה)
8. שורות עם שבוע הריון במהלך הלידה קטן מ-34 **וגם** משקל לידה גדול מ-3999 (בשל מקרים קיצוניים שלא יופקו באופן מדויק דיו לאחר התממה)
9. שורות עם משקל לידה קטן מ-600 **וגם** שבוע הריון במהלך הלידה גדול מ-29 (בשל מקרים קיצוניים שלא יופקו באופן מדויק דיו לאחר התממה)
10. שורות עם משקל לידה קטן מ-700 **וגם** שבוע הריון במהלך הלידה גדול מ-32 (בשל מקרים קיצוניים שלא יופקו באופן מדויק דיו לאחר התממה)

סך הכל, הוסרו פחות מ-1.5% מהשורות בשל קריטריוני הסרה אלו. הטבלה הגולמית הסופית לפני הפקת הנתונים המותממים מכילה N=165,915 שורות.

הפקת הנתונים המותממים

לאחר הכנת הנתונים הגולמיים, הנתונים המותממים הופקו בהתאם לשלבים הבאים:

1. יצירת הקטגוריות בנתונים הגולמיים.
2. אימון מודל של רשת בייסיאנית בעזרת פרטיות דיפרנציאלית בעזרת הנתונים הגולמיים שעברו קטגוריזציה עם $\varepsilon = 4$ בעזרת אלגוריתם PrivBayes[22].
3. ביצוע השלבים הבאים עד שנאספו N שורות סינטטיות:
   a. דגימת שורה סינטטית מהרשת הבייסיאנית
   b. הוסף את השורה לנתונים הסינטטיים רק אם אינה מקיימת את האילוצים הבאים (חופפים לשלב הכנת הנתונים הגולמיים):
      i. שורות עם גיל האם קטן מ-23 וגם מספר לידת חי גדול מ-6
      ii. שורות עם גיל האם קטן מ-20 וגם מספר לידת חי גדול מ-3
      iii. שורות עם שבוע הריון במהלך הלידה קטן מ-29 וגם משקל לידה גדול מ-2999
      iv. שורות עם שבוע הריון במהלך הלידה קטן מ-34 וגם משקל לידה גדול מ-3999
4. שכפול או הסרה אקראיים של שורות סינטטיות המופיעות רק פעם אחת או פעמיים, כך שכל שורה סינטטית מופיעה לכל הפחות שלוש פעמים תוך שמירה על מספר השורות הכולל N - **זהו סט הנתונים המותמם שפורסם**.
5. חישוב ערכי קריטריוני הקבלה בעזרת פרטיות דיפרנציאלית[23] עם $\varepsilon = 0.99$ ובדיקתם מול הרף העליון - **אלו תוצאות קריטריוני הקבלה**.

---

[22] Zhang, J., Cormode, G., Procopiuc, C. M., Srivastava, D., & Xiao, X. (2017). PrivBayes: Private data release via bayesian networks. ACM Transactions on Database Systems (TODS), 42(4), 1-41.

[23] חישוב המטריקות של קריטריוני הקבלה דורש, בין השאר, גישה ישירה לנתונים הגולמיים. כל גישה כזו צריכה להיעשת באמצעות שאילתות המבוססות על פרטיות דינפרנציאלית כדי להבטיח את הגנת הפרטיות עבור האמהות והילודים. לכן, לתוצאות המטריקות נוסף רעש אקראי מדוד ומכוייל בהתאם למדד הפרטיות דיפרנציאלית (אפסילון).

בכל קריטריוני הקבלה פרט למספר 6 (שגיאה במקדמי רגרסיה לינארית), הרעש המתווסף לתוצאת המטריקה מגיע מהתפלגות לפלס (Laplace Distribution) עם תוחלת אפס וסטיית תקן תלויית קריטריון, כפי שמופיע בטבלאות בפרק ה'. התפלגות לפלס היא סימטרית סביב התוחלת ובאופן כללי דומה להתפלגות נורמלית. לכן, אם תוצאת המטריקה קטנה יחסית (כלומר, איכות טובה יותר), הוספה של רעש סימטרי עם תוחלת אפס יכולה להפיק מספר שלילי בתוצאת מטריקה. הודות למספר השורות הרב בנתונים הגולמיים, כמות הרעש היא קטנה ביותר. למשל, קריטריון קבלה 1 מעריך את השגיאה המוחלטת בשכיחויות יחסיות חד/רב-ביוונית (k-way marginal frequencies), וסטיית התקן של הרעש היא קטנה מ-0.001 נקודות אחוז. למשל, קריטריון הקבלה 4 מעריך את השגיאה בממוצע משקל הילוד במעמד הלידה בגרמים, וסטיית התקן של הרעש היא כ-3.821 גרם.

גם מקדמי הרגרסיה הלינארית על הנתונים הגולמיים מחושבים בעזרת מנגנון המקיים פרטיות דיפרנציאלית מהמאמר Zhang, J., Zhang, Z., Xiao, X., Yang, Y., & Winslett, M. (2012). Functional Mechanism: Regression Analysis under Differential Privacy. Proceedings of the VLDB Endowment, 5(11).

שימו לב, כאן מדובר ברעש שהתווסף לתוצאות המטריקות אחרי שהנתונים המותממים הופקו, ולא לתהליך יצירת הנתונים המותממים.