מבוא

למספרים אין דרך לדבר בעד עצמם. אנחנו מדברים בשמם. אנחנו יוצקים להם משמעות.

– נייט סילבר, *האות והרעש*[1]

למה אנחנו צריכים סטטיסטיקה

הרולד שיפמן היה הרוצח המורשע הפורה ביותר בבריטניה, אם כי הוא אינו מתאים לפרופיל ארכיטיפי של רוצח סדרתי. רופא משפחה מתון שעבד בפרבר של מנצ'סטר, בין 1975 ל-1998 הוא הזריק לפחות ל-215 ממטופליו, רובם קשישים, מנת יתר מסיבית של אופיאטים. לבסוף הוא טעה כשזייף את צוואתה של אחת מקורבנותיו כדי להשאיר לו קצת כסף: בתה הייתה עורכת דין, התעוררו חשדות, וניתוח פורנזי של המחשב שלו הראה שהוא שינה בדיעבד את רישומי המטופלים כדי לגרום לקורבנותיו להיראות חולים יותר ממה שהם באמת. הוא היה ידוע כמאמץ מוקדם נלהב של טכנולוגיה, אבל הוא לא היה מספיק מבין בטכנולוגיה כדי להבין שכל שינוי שהוא עשה היה חותמת זמן (אגב, דוגמה טובה לנתונים שחושפים משמעות נסתרת).

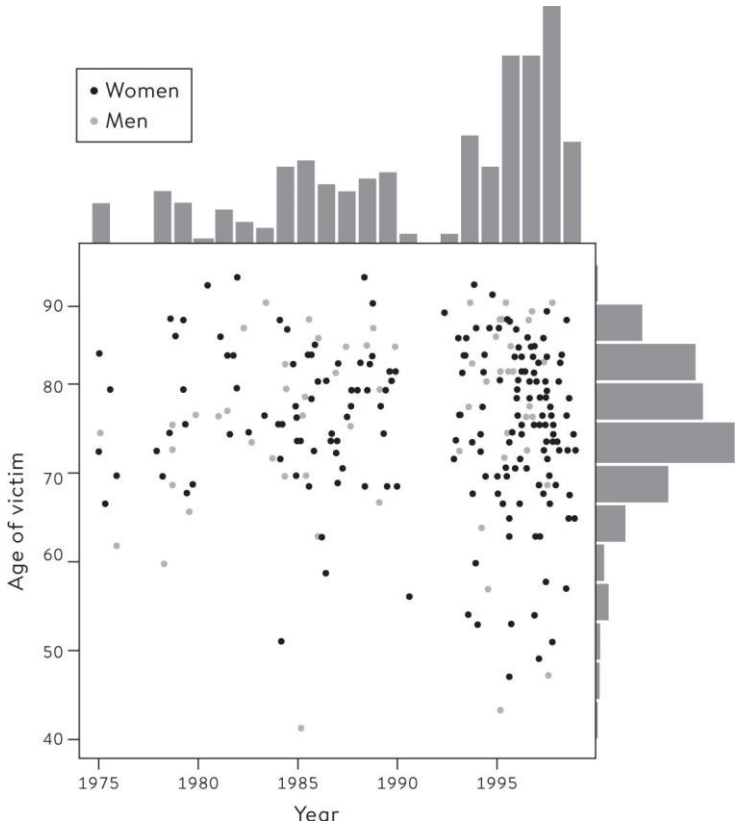
מבין מטופליו שלא נשרפו, חמישה עשר הוצאו ונמצאו בגופם רמות קטלניות של דיאמורפין, הצורה הרפואית של הרואין. שיפמן נשפט לאחר מכן על חמישה עשר מעשי רצח בשנת 1999, אך בחר שלא להציע כל הגנה ולא הוציא מילה במשפטו. הוא נמצא אשם ונכלא לכל החיים, והוקמה ועדת חקירה ציבורית כדי לקבוע אילו פשעים ייתכן שביצע מלבד אלה שבגינם נשפט, והאם ניתן היה לתפוס אותו מוקדם יותר. הייתי אחד מכמה סטטיסטיקאים שנקראו למסור עדות בוועדת החקירה הציבורית, שהגיעה למסקנה שהוא רצח בוודאות 215 ממטופליו, ואולי 45 נוספים.[2]

ספר זה יתמקד בשימוש **מדע סטטיסטי** * כדי לענות על סוג השאלות שעולות כאשר אנו רוצים להבין טוב יותר את העולם – חלק מהשאלות האלה יודגשו בקופסה. כדי לקבל קצת תובנה על התנהגותו של שיפמן, שאלה ראשונה טבעית היא:

איזה מין אנשים רצח הרולד שיפמן, ומתי הם מתו?

ועדת החקירה הציבורית סיפקה פרטים על גילו, מינו ותאריך פטירתו של כל אחד מהקורבנות. איור 0.1 הוא הדמיה מתוחכמת למדי של נתונים אלה, המציגה תרשים פיזור של גיל הקורבן לעומת תאריך מותו, עם הצללת הנקודות המציינת אם הקורבן היה זכר או נקבה. תרשימי עמודות הונחו על הצירים ומראים את תבנית הגילאים (בלהקות של 5 שנים) והשנים.

כמה מסקנות ניתן להסיק פשוט על ידי לקיחת קצת זמן להסתכל על הדמות. יש יותר נקודות שחורות מלבנות, ולכן הקורבנות של שיפמן היו בעיקר נשים. תרשים העמודות מימין לתמונה מראה שרוב קורבנותיו היו בשנות ה-70 וה-80 לחייהם, אך התבוננות בפיזור הנקודות מגלה כי למרות שבתחילה כולם היו קשישים, חלק מהמקרים הצעירים התגנבו פנימה עם השנים. תרשים העמודות בראש מראה בבירור פער סביב 1992 כאשר לא היו רציחות. התברר שלפני כן עבד שיפמן במרפאה משותפת עם רופאים אחרים, אבל אז, אולי משום שחש חשד, עזב כדי להקים מרפאה כללית ביד אחת. לאחר מכן הואצה פעילותו, כפי שמוכיח תרשים העמודות העליון.

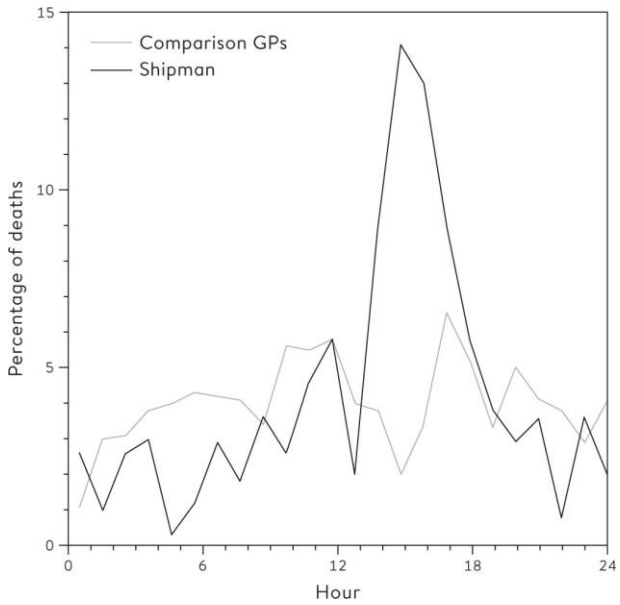


תרשים 0.1

עלילת פיזור המציגה את גילם ושנת מותם של 215 קורבנותיו המאושרים של הרולד שיפמן. על הצירים נוספו טבלאות עמודות כדי לחשוף את דפוס הגילאים ואת דפוס השנים שבהן ביצע מעשי רצח.

ניתוח זה של הקורבנות שזוהו בחקירה מעלה שאלות נוספות לגבי האופן שבו ביצע את רציחותיו. ראיות סטטיסטיות מסוימות מסופקות על ידי נתונים על השעה ביום המוות של קורבנותיו לכאורה, כפי שנרשם בתעודת הפטירה. איור 0.2 הוא גרף קווי המשווה את השעות ביום שבהן מתו מטופליו של שיפמן לזמנים שבהם מדגם של חולים של רופאי משפחה מקומיים אחרים מתו. הדפוס אינו דורש ניתוח עדין: המסקנה מכונה לעתים 'בין-עינית', שכן היא פוגעת בך בין העיניים. מטופליו של שיפמן נטו ברובם המכריע למות בשעות אחר הצהריים המוקדמות.

הנתונים אינם יכולים לומר לנו *מדוע* הם נטו למות באותו זמן, אך חקירה נוספת העלתה כי הוא ביצע את ביקורי הבית שלו לאחר ארוחת הצהריים, כאשר בדרך כלל היה לבד עם מטופליו הקשישים. הוא היה מציע להם זריקה שלדבריו נועדה להקל עליהם, אבל למעשה הייתה מנה קטלנית של דיאמורפין: אחרי שמטופל מת בשלווה מולו, הוא היה משנה את התיק הרפואי שלהם כך שייראה כאילו מדובר במוות טבעי צפוי. דיים ג'נט סמית', שעמדה בראש ועדת החקירה הציבורית, אמרה מאוחר יותר: "אני עדיין מרגישה שזה היה נורא באופן שלא יתואר, פשוט בלתי נתפס ובלתי נתפס שהוא צריך להסתובב יום אחרי יום ולהעמיד פנים שהוא הרופא האכפתי להפליא הזה ולהחזיק איתו בתיק את הנשק הקטלני שלו... שהוא פשוט היה מוציא בצורה הכי ענייינית".



תרשים 0.2

הזמן שבו מתו מטופליו של הרולד שיפמן, בהשוואה לזמנים שבהם מתו חולים של רופאים כלליים מקומיים אחרים. הדפוס אינו דורש ניתוח סטטיסטי מתוחכם.

הוא לקח סיכון מסוים, שכן ניתוח אחד לאחר המוות היה חושף אותו, אך בהתחשב בגיל מטופליו ובסיבות המוות הטבעיות לכאורה, אף אחת מהן לא בוצעה. והסיבות שלו לביצוע הרציחות האלה מעולם לא הוסברו: הוא לא העיד במשפטו, מעולם לא דיבר על מעלליו עם איש, כולל משפחתו, והתאבד בכלא, בדיוק בזמן שאשתו תקבל את הפנסיה שלו.

אנו יכולים לחשוב על סוג זה של עבודה איטרטיבית, גישושית כסטטיסטיקה "משפטית", ובמקרה זה זה היה נכון פשוטו כמשמעו. אין מתמטיקה, אין תיאוריה, רק חיפוש אחר תבניות שעשויות להוביל לשאלות מעניינות יותר. פרטי מעלליו של שיפמן נקבעו באמצעות ראיות ספציפיות לכל מקרה ומקרה, אך ניתוח נתונים זה תמך בהבנה כללית של האופן שבו הוא ביצע את פשעיו.

בהמשך הספר, בפרק 10, נראה אם ניתוח סטטיסטי פורמלי יכול היה לעזור לתפוס את שיפמן מוקדם יותר.[בינתיים] , הסיפור של שיפמן מדגים היטב את הפוטנציאל הגדול של שימוש בנתונים כדי לעזור לנו להבין את העולם ולקבל החלטות טובות יותר. זוהי מהותו של המדע הסטטיסטי.

להפוך את העולם לנתונים

גישה סטטיסטית לפשעיו של הרולד שיפמן חייבה אותנו להתרחק מהרשימה הארוכה של טרגדיות אישיות שהוא היה אחראי להן. כל אותם פרטים אישיים וייחודיים על חייהם ומותם של אנשים היו צריכים להצטמצם למערכת של עובדות ומספרים שניתן היה לספור ולצייר על גרפים. זה אולי נראה בהתחלה קר ודה-הומניזציה, אבל אם אנחנו רוצים להשתמש במדע סטטיסטי כדי להאיר את העולם, אז החוויות היומיומיות שלנו צריכות להפוך לנתונים, וזה אומר לסווג ולתייג אירועים, להקליט מדידות, לנתח את התוצאות ולהעביר את המסקנות.

עם זאת, פשוט סיווג ותיוג יכולים להוות אתגר רציני. קח את השאלה הבסיסית הבאה, אשר צריך לעניין את כל מי שמודאג עם הסביבה שלנו:

כמה עצים יש על פני כדור הארץ?

לפני שאפילו נתחיל לחשוב איך נוכל לענות על שאלה זו, ראשית עלינו ליישב סוגיה בסיסית למדי. מהו 'עץ'? אתם עשויים להרגיש שאתם מכירים עץ כשאתם רואים אותו, אבל השיפוט שלכם עשוי להיות שונה במידה ניכרת מאחרים שעשויים לראות בו שיח או שיח. אז כדי להפוך ניסיון לנתונים, אנחנו צריכים להתחיל עם הגדרות קפדניות.

מתברר כי ההגדרה הרשמית של 'עץ' היא צמח בעל גזע עצי בעל קוטר גדול מספיק בגובה החזה, המכונה DBH. שירות היערות האמריקני דורש שלצמח יהיה DBH של יותר מ-5 אינץ' (12.7 ס"מ) לפני הכרזתו הרשמית כעץ, אך רוב הרשויות משתמשות ב-DBH של 10 ס"מ (4 אינץ').

אבל אנחנו לא יכולים לשוטט סביב כדור הארץ כולו, למדוד בנפרד כל צמח בעל גזע עצי ולספור את אלה שעונים על הקריטריון הזה. לכן, החוקרים שחקרו את השאלה הזו נקטו בגישה פרגמטית יותר: הם לקחו תחילה סדרה של אזורים עם סוג נפוץ של נוף, המכונה ביום, וספרו את מספר העצים הממוצע שנמצא לקילומטר רבוע. לאחר מכן הם השתמשו בצילומי לוויין כדי להעריך את השטח הכולל של כוכב הלכת המכוסה על ידי כל סוג של ביום, ביצעו כמה מודלים סטטיסטיים מורכבים, ובסופו של דבר הגיעו לסך מוערך של 3.04 טריליון (כלומר 3,040,000,000,000) עצים על פני כדור הארץ. זה נשמע הרבה, אלא שהם חשבו שפעם היה מספר כפול מזה.[3*]

אם הרשויות חלוקות לגבי מה שהן מכנות עץ, אין זה מפתיע שמושגים מעורפלים יותר מאתגרים עוד יותר להצמדה. אם ניקח דוגמה קיצונית, ההגדרה הרשמית של "אבטלה" בבריטניה שונתה לפחות שלושים ואחת פעמים בין 1979 ל -1996.[4] ההגדרה של תוצר מקומי גולמי (תמ"ג) מתעדכנת ללא הרף, כמו כאשר סחר בסמים לא חוקיים וזנות נוסף לתמ"ג הבריטי בשנת 2014; האומדנים השתמשו בכמה מקורות נתונים יוצאי דופן - למשל Punternet, אתר ביקורות המדרג שירותי זנות, סיפק מחירים לפעילויות שונות.[5]

אפילו הרגשות האישיים ביותר שלנו יכולים להיות מקודדים ונתונים לניתוח סטטיסטי. בשנה שהסתיימה בספטמבר 2017, 150,000 אנשים בבריטניה נשאלו במסגרת סקר: "בסך הכל, כמה מאושר הרגשת אתמול?"[6] תגובתם הממוצעת, בסולם שבין אפס לעשר, הייתה 7.5, שיפור לעומת 2012, אז עמדה על 7.3, שעשויה להיות קשורה להתאוששות הכלכלית מאז המשבר הפיננסי של 2008. הציונים הנמוכים ביותר

דווחו עבור בני 50 עד 54, והגבוהים ביותר בין 70 ל -74, דפוס אופייני לבריטניה.[*]

קשה למדוד אושר, ולהחליט אם מישהו חי

או מת צריך להיות פשוט יותר: כפי שיוכיחו הדוגמאות בספר זה, הישרדות ותמותה הן דאגה נפוצה של המדע הסטטיסטי. אבל בארה"ב לכל מדינה יכולה להיות הגדרה משפטית משלה למוות, ולמרות שחוק הצהרת המוות האחידה הוצג בשנת 1981 כדי לנסות לבסס מודל משותף, נותרו כמה הבדלים קטנים. מי שהוכרז כמת באלבמה יכול, לפחות באופן עקרוני, להפסיק להיות מת מבחינה חוקית אם היה חוצה את גבול המדינה בפלורידה, שם הרישום חייב להיעשות על ידי שני רופאים מוסמכים.[7]

דוגמאות אלה מראות כי סטטיסטיקה תמיד נבנית במידה מסוימת על בסיס שיפוטים, וזו תהיה אשליה ברורה לחשוב שניתן לקודד באופן חד משמעי את מלוא המורכבות של החוויה האישית ולהכניס אותה לגיליון אלקטרוני או לתוכנה אחרת. למרות שזה מאתגר להגדיר, לספור ולמדוד מאפיינים של עצמנו ושל העולם הסובב אותנו, זה עדיין רק מידע, ורק נקודת המוצא להבנה אמיתית של העולם.

לנתונים יש שתי מגבלות עיקריות כמקור לידע כזה. ראשית, זה כמעט תמיד מדד לא מושלם למה שאנחנו באמת מתעניינים בו: לשאול כמה אנשים היו מאושרים בשבוע שעבר בסולם מאפס עד עשר בקושי מתמצת את הרווחה הרגשית של האומה. שנית, כל דבר שנבחר למדוד יהיה שונה ממקום למקום, מאדם לאדם, מעת לעת, והבעיה היא לחלץ תובנות משמעותיות מכל השונות האקראית לכאורה הזו.

במשך מאות שנים, המדע הסטטיסטי התמודד עם אתגרים תאומים אלה, ומילא תפקיד מוביל בניסיונות מדעיים להבין את העולם. היא סיפקה את הבסיס לפרש נתונים, שהם תמיד לא מושלמים, כדי להבדיל בין יחסים חשובים לבין השתנות הרקע שהופכת את כולנו לייחודיים. אבל העולם משתנה כל הזמן, כאשר שאלות חדשות נשאלות ומקורות נתונים חדשים הופכים זמינים, וגם מדע הסטטיסטיקה נאלץ להשתנות.

אנשים תמיד ספרו ומדדו, אבל הסטטיסטיקה המודרנית כדיסציפלינה החלה למעשה בשנת 1650, כאשר, כפי שנראה בפרק 8,

ההסתברות הובנה כהלכה בפעם הראשונה על ידי בלייז פסקל ופייר דה פרמה. בהינתן הבסיס המתמטי המוצק הזה להתמודדות עם שונות, ההתקדמות הייתה אז מהירה להפליא. בשילוב עם נתונים על הגילאים שבהם אנשים מתים, תיאוריית ההסתברות סיפקה בסיס איתן לחישוב הפנסיה והקצבאות. האסטרונומיה עברה מהפכה כאשר מדענים הבינו כיצד תורת ההסתברות יכולה להתמודד עם שונות במדידות. חובבי הוויקטוריאנים הפכו אובסיביים לאיסוף נתונים על גוף האדם (וכל השאר), ויצרו קשר חזק בין ניתוח סטטיסטי לגנטיקה, ביולוגיה ורפואה. במאה העשרים הפכה הסטטיסטיקה למתמטית יותר, ולרוע מזלם של סטודנטים ומתרגלים רבים, הנושא הפך לשם נרדף ליישום מכני של שק כלים סטטיסטיים, שרבים מהם קרויים על שם סטטיסטיקאים אקסצנטריים ווכחניים שנפגוש בהמשך הספר.

תפיסה נפוצה זו של סטטיסטיקה כ"שק כלים" בסיסי ניצבת כעת בפני אתגרים גדולים. ראשית, אנו נמצאים בעידן של **מדעי הנתונים**, שבו מערכי נתונים גדולים ומורכבים נאספים ממקורות שגרתיים כמו מוניטרי תנועה, פוסטים ברשתות החברתיות ורכישות באינטרנט, ומשמשים בסיס לחידושים טכנולוגיים כמו אופטימיזציה של מסלולי נסיעה, פרסום ממוקד או מערכות המלצה על רכישות – נבחן **אלגוריתמים** המבוססים על '**ביג דאטה**' בפרק 6. הכשרה סטטיסטית נתפסת יותר ויותר כמרכיב הכרחי אחד של להיות מדען נתונים, יחד עם מיומנויות בניהול נתונים, תכנות ופיתוח אלגוריתמים, כמו גם ידע נכון של הנושא.

אתגר נוסף לתפיסה המסורתית של סטטיסטיקה נובע מהעלייה העצומה בכמות המחקר המדעי המתבצע, במיוחד במדעי הביו-רפואה והחברה, בשילוב עם לחץ לפרסם בכתבי עת בכירים. זה הוביל לספקות לגבי אמינותם של חלקים מהספרות המדעית, עם טענות כי "תגליות" רבות אינן ניתנות לשחזור על ידי חוקרים אחרים – כגון המחלוקת המתמשכת האם אימוץ עמדה אסרטיבית הידועה בכינויה העממי "תנוחת כוח" יכול לגרום לשינויים הורמונליים ואחרים.[8] השימוש הבלתי הולם בשיטות סטטיסטיות סטנדרטיות קיבל חלק לא מבוטל מהאשמה למה שנודע כמשבר השכפול או השכפול במדע.

עם הזמינות הגוברת של ערכות נתונים עצומות וידידותיות למשתמש

תוכנת ניתוח, ניתן לחשוב כי יש פחות צורך בהכשרה בשיטות סטטיסטיות. זה יהיה נאיבי בקיצוניות. רחוק מלשחרר אותנו מהצורך בכישורים סטטיסטיים, נתונים גדולים יותר והעלייה במספר ובמורכבות של מחקרים מדעיים מקשים עוד יותר על הסקת מסקנות מתאימות. יותר נתונים פירושם שאנחנו צריכים להיות אפילו יותר מודעים למה שהראיות באמת שוות.

לדוגמה, ניתוח אינטנסיבי של מערכי נתונים הנגזרים מנתונים שגרתיים יכול להגדיל את האפשרות לתגליות שווא, הן בשל הטיה שיטתית הטבועה במקורות הנתונים והן מביצוע ניתוחים רבים ודיווח רק על מה שנראה הכי מעניין, פרקטיקה המכונה לעתים "גרימת נתונים". כדי שנוכל לבקר עבודות מדעיות שפורסמו, ועוד יותר את הדיווחים התקשורתיים שכולנו נתקלים בהם על בסיס יומיומי, צריכה להיות לנו מודעות חריפה לסכנות הטמונות בדיווח סלקטיבי, לצורך בשכפול טענות מדעיות על ידי חוקרים עצמאיים, ולסכנה של פירוש יתר של מחקר בודד מחוץ להקשרו.

ניתן לרכז את כל התובנות הללו תחת המונח **אוריינות נתונים**, המתאר את היכולת לא רק לבצע ניתוח סטטיסטי של בעיות בעולם האמיתי, אלא גם להבין ולבקר כל מסקנות שהסיקו אחרים על בסיס סטטיסטיקה. אבל שיפור אוריינות נתונים פירושו לשנות את הדרך שבה מלמדים סטטיסטיקה.

סטטיסטיקה של הוראה

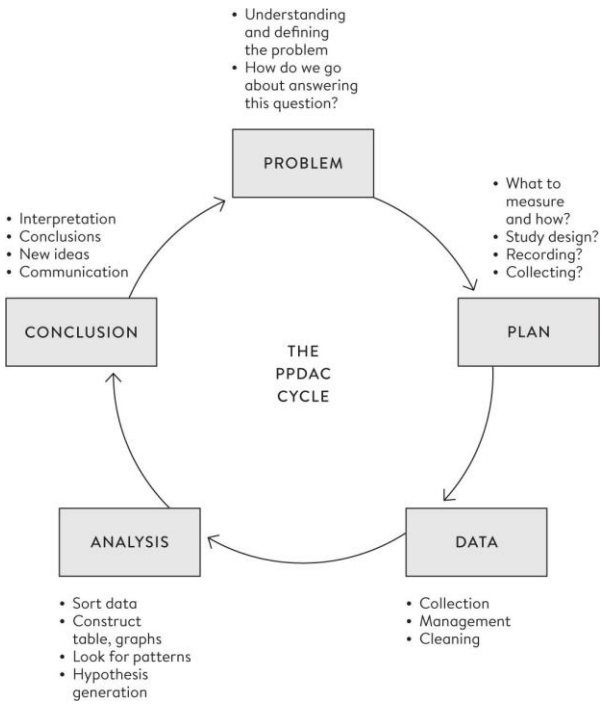
דורות של תלמידים סבלו באמצעות קורסים סטטיסטיים יבשים המבוססים על לימוד קבוצה של טכניקות שיש ליישם במצבים שונים, עם התייחסות רבה יותר לתיאוריה מתמטית מאשר להבין הן מדוע הנוסחאות משמשות, ואת האתגרים המתעוררים כאשר מנסים להשתמש בנתונים כדי לענות על שאלות.

למרבה המזל זה משתנה. הצרכים של מדעי הנתונים ואוריינות נתונים דורשים גישה מונחית בעיות יותר, שבה היישום של כלים סטטיסטיים ספציפיים נתפס רק מרכיב אחד של מחזור שלם של חקירה. מבנה **PPDAC** הוצע כדרך לייצג מחזור פתרון בעיות, אותו נאמץ לאורך ספר זה.[9] תרשים 0.3 מבוסס על דוגמה מניו-זילנד, שהייתה מובילה עולמית בחינוך סטטיסטי בבתי ספר.

השלב הראשון של המחזור הוא הגדרת בעיה; חקירה סטטיסטית

תמיד מתחיל בשאלה, כמו השאלה שלנו על דפוס הרציחות של הרולד שיפמן או על מספר העצים בעולם. בהמשך הספר נתמקד בבעיות החל מהתועלת הצפויה של טיפולים שונים מיד לאחר ניתוח סרטן השד, ועד לשאלה מדוע לגברים זקנים יש אוזניים גדולות.

מפתה לדלג על הצורך בתוכנית זהירה. שאלת שיפמן פשוט דרשה איסוף של כמה שיותר נתונים על קורבנותיו. אבל האנשים שספרו עצים הקדישו תשומת לב קפדנית להגדרות מדויקות ולאופן ביצוע המדידות, שכן מסקנות בטוחות ניתן להסיק רק ממחקר שתוכנן כראוי. למרבה הצער, במרוץ לקבל נתונים ולהתחיל ניתוח, תשומת הלב לעיצוב היא לעתים קרובות מבריק.



תרשים 0.3

מחזור פתרון הבעיות של PPDAC, עובר מבעיה, תוכנית, נתונים, ניתוח למסקנה ותקשורת, ומתחיל שוב במחזור אחר.

איסוף נתונים טובים דורש סוג של מיומנויות ארגון וקידוד שנתפסות כחשובות יותר ויותר במדעי הנתונים, במיוחד כאשר נתונים ממקורות שגרתיים עשויים להזדקק לניקוי רב על מנת להכין אותם לניתוח. ייתכן שמערכות איסוף הנתונים השתנו עם הזמן, ייתכנו שגיאות ברורות וכן הלאה – הביטוי 'נתונים שנמצאו' משדר בצורה מסודרת שהם עשויים להיות די מבולגנים, כמו משהו שנאסף ברחוב.

שלב האנליזה היה באופן מסורתי הדגש העיקרי של קורסי סטטיסטיקה, ואנו נסקור מגוון של טכניקות אנליטיות בספר זה; אבל לפעמים כל מה שנדרש הוא הדמיה שימושית, כמו באיור 0.1. לבסוף, המפתח למדע סטטיסטי טוב הוא הסקת מסקנות מתאימות שמכירות באופן מלא במגבלות הראיות, והעברתן בצורה ברורה, כמו באיורים הגרפיים של נתוני שיפמן. כל מסקנה בדרך כלל מעלה יותר שאלות, וכך המעגל מתחיל מחדש, כמו כשהתחלנו להסתכל על השעה ביום שבה מתו החולים של שיפמן.

למרות שבפועל לא ניתן לעקוב במדויק אחר מחזור PPDAC המתואר באיור 0.3 , הוא מדגיש כי טכניקות פורמליות לניתוח סטטיסטי משחקות רק תפקיד אחד בעבודתו של סטטיסטיקאי או מדען נתונים. מדע סטטיסטי הוא הרבה יותר מאשר ענף במתמטיקה המערב נוסחאות אזוטריות שעמן נאבקו דורות של תלמידים (לעתים קרובות בחוסר רצון).

ספר זה

כשהייתי סטודנט בבריטניה בשנות השבעים, היו רק שלושה ערוצי טלוויזיה, מחשבים היו בגודל של ארון בגדים כפול, והדבר הכי קרוב שהיה לנו לוויקיפדיה היה במכשיר כף היד הדמיוני במדריך *הטרמפיסט לגלקסיה* של דאגלס אדמס.לשיפור עצמי פנינו אפוא לספרי שקנאי, והקוצים הכחולים האיקוניים שלהם היו מאפיין סטנדרטי של כל מדף ספרים של תלמידים.

מכיוון שלמדתי סטטיסטיקה, אוסף השקנאים שלי כלל *את Facts from Figures* מאת M. J. Moroney (1951) ו-*How to Lie with Statistics* מאת Darrell Huff (1954). פרסומים מכובדים אלה נמכרו במאות אלפים, ושיקפו הן את רמת העניין בסטטיסטיקה והן את חוסר הברירה העגום באותה תקופה. קלאסיקות אלה עמדו בצורה יוצאת דופן ל

שישים וחמש שנים, אך העידן הנוכחי דורש גישה שונה להוראת סטטיסטיקה המבוססת על העקרונות שהותוו לעיל.

ספר זה משתמש אפוא בפתרון בעיות בעולם האמיתי כנקודת מוצא להצגת רעיונות סטטיסטיים. חלק מהרעיונות האלה אולי נראים מובנים מאליהם, אבל חלקם עדינים יותר ועשויים לדרוש קצת מאמץ מנטלי, אם כי מיומנויות מתמטיות לא יהיה צורך. בהשוואה לטקסטים מסורתיים, ספר זה מתמקד בנושאים רעיוניים ולא טכניים, וכולל רק כמה משוואות תמימות למדי הנתמכות על ידי מילון מונחים. תוכנה היא חלק חיוני בכל עבודה במדעי הנתונים והסטטיסטיקה, אך היא אינה מוקד של ספר זה – ערכות לימוד זמינות בקלות עבור סביבות זמינות באופן חופשי כגון R ו- Python.

על השאלות המוצגות בקופסאות ניתן לענות במידה מסוימת באמצעות ניתוח סטטיסטי, אם כי הן שונות מאוד בהיקפן. חלקן השערות מדעיות חשובות, כגון האם בוזון היגס קיים, או אם באמת יש ראיות משכנעות לתפיסה על-חושית (ESP). אחרות הן שאלות הנוגעות לטיפול רפואי, כגון האם בבתי חולים עמוסים יותר יש שיעורי הישרדות גבוהים יותר, ואם בדיקות סקר לסרטן השחלות מועילות. לפעמים אנחנו רק רוצים להעריך כמויות, כגון הסיכון לסרטן מכריכי בייקון, מספר השותפים המיניים שיש לאנשים בבריטניה במהלך חייהם, ואת היתרון של נטילת סטטין יומי.

וכמה שאלות הן פשוט מעניינות, כמו לזהות את הניצול בר המזל ביותר *מהטיטניק*; האם הרולד שיפמן יכול היה להיתפס מוקדם יותר; ולהעריך את ההסתברות ששלד שנמצא בחניון בלסטר הוא באמת של ריצ'רד השלישי.

ספר זה מיועד הן לסטודנטים לסטטיסטיקה המחפשים מבוא לא טכני לסוגיות הבסיסיות, והן לקוראים כלליים שרוצים להיות מעודכנים יותר בסטטיסטיקה שהם נתקלים בה הן בעבודתם והן בחיי היומיום. הדגש שלי הוא על טיפול בסטטיסטיקה במיומנות ובזהירות: מספרים אולי נראים כעובדות קרות וקשות, אבל הניסיונות למדוד עצים, אושר ומוות כבר הוכיחו שצריך להתייחס אליהם בעדינות.

סטטיסטיקה יכולה להביא בהירות ותובנה לבעיות שאנו מתמודדים איתן, אך כולנו מכירים את האופן שבו ניתן לנצל אותן לרעה, לעתים קרובות כדי לקדם דעה או פשוט כדי למשוך תשומת לב. היכולת להעריך את אמינותן של טענות סטטיסטיות נראית מיומנות מפתח בעולם המודרני, ואני מקווה שספר זה יעזור להעצים אנשים להטיל ספק במספרים שהם

מפגש בחיי היומיום שלהם.

תקציר

- הפיכת חוויות לנתונים אינה פשוטה, ונתונים מוגבלים באופן בלתי נמנע ביכולתם לתאר את העולם.
- למדע הסטטיסטי יש היסטוריה ארוכה ומוצלחת, אך כיום הוא משתנה לאור הזמינות המוגברת של נתונים.
- מיומנות בשיטות סטטיסטיות משחקת חלק חשוב בלהיות מדען נתונים.
- הוראת סטטיסטיקה משתנה מהתמקדות בשיטות מתמטיות לכזו המבוססת על מחזור שלם של פתרון בעיות.
- מחזור PPDAC מספק מסגרת נוחה: בעיה—תוכנית—נתונים—ניתוח
  – סיכום ותקשורת.
- אוריינות נתונים היא מיומנות מפתח בעולם המודרני.

פרק 1

קבלת דברים בפרופורציה: נתונים קטגוריאליים ואחוזים

מה קרה לילדים שעברו ניתוח לב בבריסטול בין 1984 ל-1995?

יהושע ל' היה בן 16 חודשים וסבל מטרנספוזיציה של העורקים הגדולים, צורה חמורה של מחלת לב מולדת שבה כלי הדם העיקריים המגיעים מהלב מחוברים לחדר הלא נכון. הוא נזקק לניתוח כדי "להחליף" את העורקים, וקצת אחרי השעה 7 בבוקר ב-12 בינואר 1995 הוריו נפרדו ממנו וצפו בו נלקח לניתוח במרפאה המלכותית בבריסטול. אבל הוריו של ג'ושוע לא היו מודעים לכך שסיפורים על שיעורי ההישרדות הכירורגיים הנמוכים בבריסטול הופצו מאז תחילת שנות התשעים. אף אחד לא סיפר להם שהאחיות עזבו את היחידה ולא המשיכו לספר להורים שילדם נפטר, או שערב קודם לכן התקיימה ישיבה בשעת לילה מאוחרת שבה התלבט אם לבטל את הניתוח של יהושע.[1]

יהושע מת על שולחן הניתוחים. בשנה שלאחר מכן פתחה המועצה הרפואית הכללית (הרגולטור הרפואי) בחקירה בעקבות תלונות של יהושע והורים שכולים אחרים, ובשנת 1998 נמצאו שני מנתחים והמנכ"ל לשעבר אשמים בהתנהגות רפואית בלתי הולמת חמורה. הדאגה הציבורית לא שככה, והוזמנה חקירה רשמית: זה הביא צוות של סטטיסטיקאים שקיבלו את המשימה הקודרת להשוות את שיעורי ההישרדות בבריסטול עם מקומות אחרים בבריטניה בין 1984 ל - 1995. הובלתי את הקבוצה הזו.

תחילה היינו צריכים לקבוע כמה ילדים עברו ניתוח לב וכמה מתו. זה נשמע כאילו זה צריך להיות פשוט, אבל, כמו

כפי שהוצג בפרק הקודם, פשוט לספור אירועים יכול להיות מאתגר. מהו 'ילד'? מה נחשב 'ניתוח לב'? מתי ניתן לייחס מוות לניתוח? וגם כאשר הגדרות אלה הוכרעו, האם נוכל לקבוע כמה מכל אחת מהן היו?

לקחנו 'ילד' כמו כל מי שמתחת לגיל 16, והתמקדנו בניתוח 'פתוח' שבו הלב הופסק ותפקודו הוחלף במעקף לב-ריאה. יכולות להיות פעולות מרובות לכל הודאה, אך אלה נחשבו כאירוע אחד. מקרי מוות נספרו אם התרחשו בתוך 30 יום מהניתוח, בין אם בבית החולים ובין אם עקב הניתוח. ידענו שמוות הוא מדד לא מושלם לאיכות התוצאה, מכיוון שהוא התעלם מילדים שסבלו מנזק מוחי או נכים אחרים כתוצאה מהניתוח, אבל לא היו לנו נתונים על תוצאות ארוכות טווח.

מקור הנתונים העיקרי היה סטטיסטיקה לאומית של אפיזודות בתי חולים (HES), שנגזרו מנתונים מנהליים שהוזנו על ידי מתכנתים בשכר נמוך. ל-HES היה מוניטין גרוע בקרב רופאים, אך למקור זה היה יתרון גדול בכך שניתן היה לקשר אותו לרישומי מוות לאומיים. הייתה גם מערכת מקבילה של נתונים שהוגשו ישירות לרישום ניתוחי לב (CSR) שהוקם על ידי האגודה המקצועית של המנתחים.

שני מקורות נתונים אלה, אף שהיו אמורים להיות על אותה פרקטיקה בדיוק, הראו חילוקי דעות ניכרים: בשנים 1991-1995, HES אמר שהיו 62 מקרי מוות מתוך 505 ניתוחים פתוחים (14%), ואילו CSR אמר שהיו 71 מקרי מוות מתוך 563 ניתוחים (13%). לא פחות מחמישה מקורות מידע מקומיים נוספים היו זמינים, החל מרישומי הרדמה וכלה ביומנים האישיים של המנתחים עצמם. בריסטול הייתה מוצפת בנתונים, אבל אף אחד ממקורות הנתונים לא יכול להיחשב ל"אמת", ואף אחד לא לקח אחריות על ניתוח תוצאות הניתוח ועל פיהן.

חישבנו שאם לחולים בבריסטול היה את הסיכון הממוצע השורר במקומות אחרים בבריטניה, בריסטול הייתה מצפה ל-32 מקרי מוות במהלך תקופה זו, במקום 62 שנרשמו ב-HES, עליהם דיווחנו כ"30 מקרי מוות עודפים" בין השנים 1991 ל-1995.[*] המספרים המדויקים השתנו בהתאם למקורות הנתונים, וזה אולי נראה יוצא דופן שלא יכולנו אפילו לקבוע את העובדות הבסיסיות על מספר המבצעים ותוצאותיהם, למרות שמערכות התיעוד הנוכחיות אמורות להיות טובות יותר.

ממצאים אלה זכו לסיקור תקשורתי נרחב, והחקירה בבריסטול הובילה לשינוי משמעותי בגישה לניטור ביצועים קליניים: כבר לא היה

מקצוע הרפואה אמון במשטרה עצמה. נקבעו מנגנונים לדיווח פומבי על נתוני הישרדות בבתי חולים, אם כי, כפי שנראה כעת, האופן שבו נתונים אלה מוצגים יכול כשלעצמו להשפיע על תפיסת הקהל.

תקשורת ספירות ופרופורציות

נתונים המתעדים אם אירועים בודדים התרחשו או לא ידועים כנתונים **בינאריים,** מכיוון שהם יכולים לקבל רק שני ערכים, המסומנים בדרך כלל ככן ולא. קבוצות של נתונים בינאריים ניתן לסכם על ידי מספר פעמים ואחוז המקרים שבהם התרחש אירוע.

הנושא של פרק זה הוא שההצגה הבסיסית של סטטיסטיקה חשובה. במובן מסוים אנו קופצים לשלב האחרון של מחזור PPDAC שבו מועברות מסקנות, ובעוד שצורת תקשורת זו לא נחשבה באופן מסורתי לנושא חשוב בסטטיסטיקה, העניין הגובר בהדמיה של נתונים משקף שינוי בגישה זו. לכן, הן בפרק זה והן בפרק הבא נתרכז בדרכים להצגת נתונים, כך שנוכל לקבל במהירות את תמצית המתרחש ללא ניתוח מפורט, החל ממבט על דרכים חלופיות להצגת נתונים, שבמידה רבה בגלל חקירת בריסטול, זמינים כעת לציבור.

טבלה 1.1 מציגה את התוצאות של כמעט 13,000 ילדים שעברו ניתוח לב בבריטניה ובאירלנד בין השנים 2012 ו-2015.[2] מאתיים שישים ושלושה תינוקות מתו תוך 30 יום מניתוחם, וכל אחד ממקרי המוות הללו הוא טרגדיה למשפחה המעורבת. זו תהיה נחמה קטנה עבורם ששיעורי ההישרדות השתפרו מאוד מאז החקירה בבריסטול, וכעת הם עומדים על ממוצע של 98%, ולכן יש סיכוי מלא תקווה למשפחות של ילדים העומדים בפני ניתוח לב.

טבלה יכולה להיחשב כסוג של גרפיקה, והיא דורשת בחירות עיצוב קפדניות של צבע, גופן ושפה כדי להבטיח מעורבות וקריאות. התגובה הרגשית של הקהל לטבלה עשויה להיות מושפעת גם מהבחירה אילו עמודות להציג. טבלה 1.1 מציגה את התוצאות במונחים של מחלימים ומקרי מוות, אך בארה"ב מדווחים שיעורי *התמותה* מניתוחי לב של ילדים, בעוד שבבריטניה שיעורי *ההישרדות*. זה ידוע כמסגור שלילי או חיובי, והשפעתו הכוללת על האופן שבו אנו מרגישים היא אינטואיטיבית ומתועדת היטב: "5% תמותה" נשמע גרוע יותר מאשר "95% הישרדות". דיווח על מספר מקרי המוות בפועל וכן על

אחוז יכול גם להגדיל את הרושם של סיכון, כמו סך זה ניתן לדמיין אז כמו קהל של אנשים אמיתיים.

Hospital	Number of babies having surgery	Number surviving for at least 30 days after surgery	Number dying within 30 days of surgery	Percentage surviving	Percentage dying
London, Harley Street	418	413	5	98.8	1.2
Leicester	607	593	14	97.7	2.3
Newcastle	668	653	15	97.8	2.2
Glasgow	760	733	27	96.3	3.7
Southampton	829	815	14	98.3	1.7
Bristol	835	821	14	98.3	1.7
Dublin	983	960	23	97.7	2.3
Leeds	1,038	1,016	22	97.9	2.1
London, Brompton	1,094	1,075	19	98.3	1.7
Liverpool	1,132	1,112	20	98.2	1.8
London, Evelina	1,220	1,185	35	97.1	2.9
Birmingham	1,457	1,421	36	97.5	2.5
London, Great Ormond Street	1,892	1,873	19	99.0	1.0
Total	12,933	12,670	263	98.0	2.0

*

טבלה 1.1

תוצאות ניתוחי לב לילדים בבתי חולים בבריטניה ובאירלנד בין השנים 2012 ל-2015, במונחים של הישרדות או לא, 30 יום לאחר הניתוח.

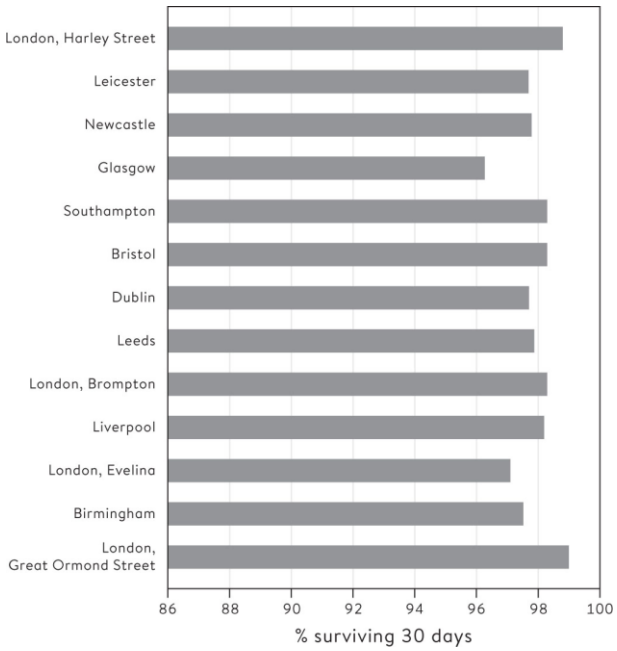
דוגמה קלאסית לאופן שבו מסגור אלטרנטיבי יכול לשנות את ההשפעה הרגשית של מספר היא פרסומת שהופיעה ברכבת התחתית של לונדון בשנת 2011, והכריזה כי "99% מהצעירים הלונדונים אינם מבצעים אלימות נוער חמורה". פרסומות אלה נועדו לכאורה להרגיע את הנוסעים לגבי עירם, אבל יכולנו להפוך את הרגש שלה

השפעה באמצעות שני שינויים פשוטים. ראשית, משמעות ההצהרה היא ש-1% מהצעירים הלונדונים *אכן* מבצעים אלימות חמורה. שנית, מכיוון שאוכלוסיית לונדון מונה כ-9 מיליון, ישנם כמיליון אנשים בגילאי 15 עד 25, ואם ניקח בחשבון אותם כ"צעירים", המשמעות היא שיש 1% ממיליון או בסך הכל 10,000 צעירים אלימים מאוד בעיר. זה לא נשמע מרגיע בכלל. שימו לב לשני הטריקים המשמשים למניפולציה של ההשפעה של נתון זה: להמיר ממסגרת חיובית למסגרת שלילית, ולאחר מכן להפוך אחוז למספר ממשי של אנשים.

באופן אידיאלי, יש להציג מסגרות חיוביות ושליליות אם ברצוננו לספק מידע חסר פניות, אם כי סדר העמודות עשוי עדיין להשפיע על אופן פירוש הטבלה. סדר השורות של טבלה גם צריך להיחשב בזהירות. לוח 1.1 מציג את בתי החולים לפי סדר מספר הניתוחים בכל אחד מהם, אך אילו היו מוצגים, למשל, לפי סדר שיעורי התמותה הגבוהים ביותר בראש הטבלה, הדבר היה יוצר את הרושם שמדובר בדרך תקפה וחשובה להשוואה בין בתי החולים. טבלאות ליגה כאלה מועדפות על התקשורת ואפילו על ידי פוליטיקאים מסוימים, אך עלולות להטעות באופן גס: לא רק משום שההבדלים יכולים לנבוע משונות מקרית, אלא משום שבתי החולים עשויים לקלוט סוגים שונים מאוד של מקרים. בטבלה 1.1, למשל, אנו עשויים לחשוד שבירמינגהם, אחד מבתי החולים לילדים הגדולים והידועים ביותר, מתמודד עם המקרים החמורים ביותר, ולכן יהיה זה לא הוגן, בלשון המעטה, להדגיש את שיעורי ההישרדות הכוללים הלא מרשימים לכאורה שלהם.[*]

ניתן להציג את שיעורי ההישרדות בתרשים עמודות אופקי כגון אחד מהם מוצג באיור 1.1. בחירה מכרעת היא היכן להתחיל את הציר האופקי: אם הערכים מתחילים מ -0%, כל העמודות יהיו כמעט לכל אורך הגרפיקה, אשר יציג בבירור את שיעורי ההישרדות הגבוהים במיוחד, אך לא ניתן יהיה להבחין בין הקווים. אבל הטריק הוותיק ביותר של גרפיקה מטעה הוא להתחיל את הציר ב-95%, מה שיגרום לבתי החולים להיראות אחרת לגמרי, גם אם השונות היא למעשה רק מה שניתן לייחס למקריות בלבד.

בחירת תחילת הציר מציבה אפוא דילמה. אלברטו קהיר, מחברם של ספרים רבי השפעה על ויזואליזציה של נתונים,3 מציע שתמיד כדאי להתחיל עם "קו בסיס הגיוני ומשמעותי", שבמצב זה נראה קשה לזיהוי – הבחירה השרירותית למדי שלי של 86% מייצגת בערך את ההישרדות הנמוכה באופן בלתי מתקבל על הדעת בבריסטול עשרים שנה קודם לכן.



תרשים 1.1

תרשים עמודות אופקי של 30 – שיעורי הישרדות ביום עבור 13 בתי חולים. הבחירה של תחילת הציר האופקי, כאן 86%, יכולה להיות השפעה מכרעת על הרושם שניתן על ידי גרפיקה. אם הציר יתחיל ב-0%, כל בתי החולים ייראו בלתי ניתנים להבחנה, ואילו אם נתחיל ב-95% ההבדלים ייראו דרמטיים באופן מטעה.

פתחתי את הספר הזה בציטוט של נייט סילבר, מייסד פלטפורמת הנתונים *FiveThirtyEight* והתפרסם לראשונה בזכות חיזוי מדויק של הבחירות לנשיאות ארה"ב ב-2008, שביטא ברהיטות את הרעיון שמספרים אינם מדברים בעד עצמם – אנחנו אחראים לתת להם משמעות. משמעות הדבר היא שתקשורת היא חלק מרכזי במחזור פתרון הבעיות, והראיתי בחלק זה כיצד המסר מקבוצה של פרופורציות פשוטות יכול להיות מושפע מהבחירות שלנו בהצגה.

כעת עלינו להציג קונספט חשוב ונוח שיעזור לנו להתקדם מעבר לשאלות פשוטות של כן/לא.

משתנים קטגוריאליים

משתנה מוגדר ככל מדידה שיכולה לקבל ערכים שונים בנסיבות שונות; זהו מונח מקוצר שימושי מאוד לכל סוגי התצפיות המרכיבות נתונים. משתנים בינאריים הם שאלות של כן/לא כגון האם מישהו חי או מת ואם הוא נקבה או לא: שניהם משתנים בין אנשים, ויכולים, אפילו עבור מגדר, להשתנות בתוך אנשים בזמנים שונים. **משתנים קטגוריאליים** הם מדדים שיכולים לקחת על שתי קטגוריות או יותר, אשר עשויות להיות

- קטגוריות לא מסודרות: כגון ארץ המוצא של אדם, צבע המכונית או בית החולים שבו מתבצע ניתוח.
- קטגוריות מסודרות: כגון דרגת אנשי צבא.
- מספרים מקובצים: כגון רמות של השמנת יתר, אשר מוגדר לעתים קרובות במונחים של סף עבור מדד מסת הגוף (BMI).[*]

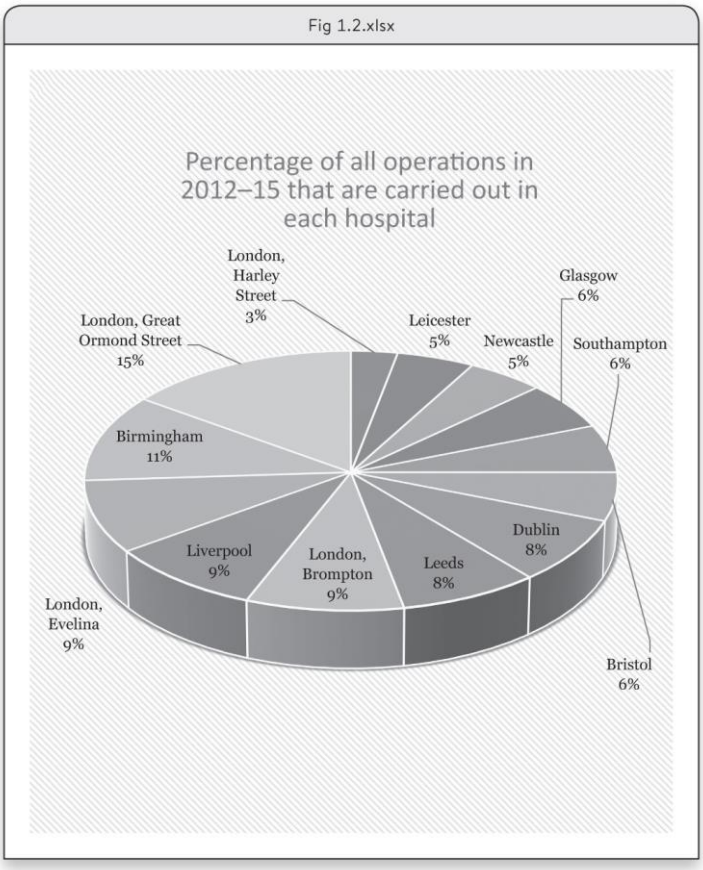
כשמדובר בהצגת נתונים קטגוריאליים, תרשימי עוגה מאפשרים להתרשם מהגודל של כל קטגוריה ביחס לכל העוגה, אך לעתים קרובות הם מבלבלים מבחינה חזותית, במיוחד אם הם מנסים להציג קטגוריות רבות מדי באותו תרשים, או משתמשים בייצוג תלת ממדי שמעוות אזורים. איור 1.2 מציג דוגמה מחרידה למדי מהסוג המוצע על-ידי Microsoft Excel, המראה את הפרופורציות של 12,933 ילדים חולי לב מהטבלה 1.1 המטופלים בכל בית חולים.

תרשימי עוגה מרובים הם בדרך כלל לא רעיון טוב, מכיוון שהשוואות נפגעות בגלל הקושי להעריך את הגדלים היחסיים של אזורים בצורות שונות. השוואות טובות יותר בהתבסס על גובה או אורך בלבד בתרשים עמודות.

איור 1.3 מציג דוגמה פשוטה וברורה יותר של תרשים עמודות אופקי של הפרופורציות המטופלות בכל בית חולים.

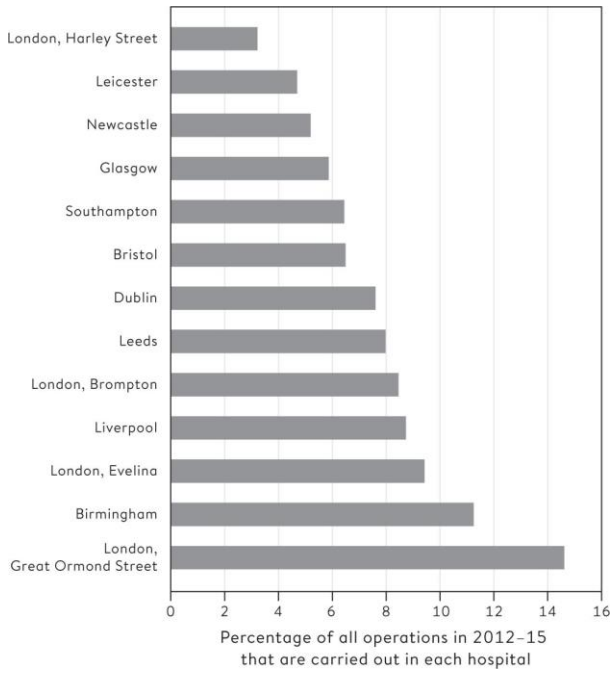
השוואת זוג פרופורציות

ראינו כיצד ניתן להשוות באלגנטיות קבוצה של פרופורציות באמצעות תרשים עמודות, ולכן יהיה זה הגיוני לחשוב שהשוואה בין שתי פרופורציות תהיה עניין של מה בכך. אבל כאשר פרופורציות אלה מייצגות הערכות של הסיכונים לחוות נזק כלשהו, אז האופן שבו סיכונים אלה מושווים הופך לנושא רציני ושנוי במחלוקת. הנה שאלה טיפוסית:



תרשים 1.2

החלק היחסי של כלל ניתוחי הלב של ילדים המבוצעים בכל בית חולים, מוצג בתרשים עוגה תלת-ממדי מ- Excel. התרשים המאוד לא נעים הזה גורם לקטגוריות הסמוכות לחזית להיראות גדולות יותר, ולכן לא מאפשר לערוך השוואות ויזואליות בין בתי חולים.



תרשים 1.3

אחוז מכלל ניתוחי הלב בילדים המבוצעים בכל בית חולים: ייצוג ברור יותר באמצעות תרשים עמודות אופקי.

כולנו מכירים את הכותרות ההיפרבוליות בתקשורת שמזהירות אותנו שמשהו שגרתי מגביר את הסיכון להתרחשות מפחידה כלשהי: אני אוהב לקרוא לסיפורים האלה 'חתולים גורמים לסרטן'. לדוגמה, בנובמבר 2015 הודיעה הסוכנות הבינלאומית לחקר הסרטן של ארגון הבריאות העולמי (IARC) כי בשר מעובד הוא "מסרטן מקבוצה I", והציבה אותו באותה קטגוריה כמו סיגריות ואסבסט. זה הוביל באופן בלתי נמנע לכותרות מבוהלות כמו הטענה של *הדיילי רקורד* כי "בייקון, בשר חזיר ונקניקיות יש את אותו סיכון סרטן כמו סיגריות מזהירים מומחים".[4]

IARC ניסה להרגיע את המהומה על ידי הדגשה כי סיווג קבוצה 1 נועד להיות בטוח כי סיכון מוגבר לסרטן קיים בכלל, ולא אמר דבר על גודל הסיכון בפועל. בהודעה לעיתונות דיווחה IARC כי 50 גרם בשר מעובד ביום קשורים לסיכון מוגבר לסרטן המעי של 18%. זה נשמע מדאיג, אבל האם זה צריך להיות?

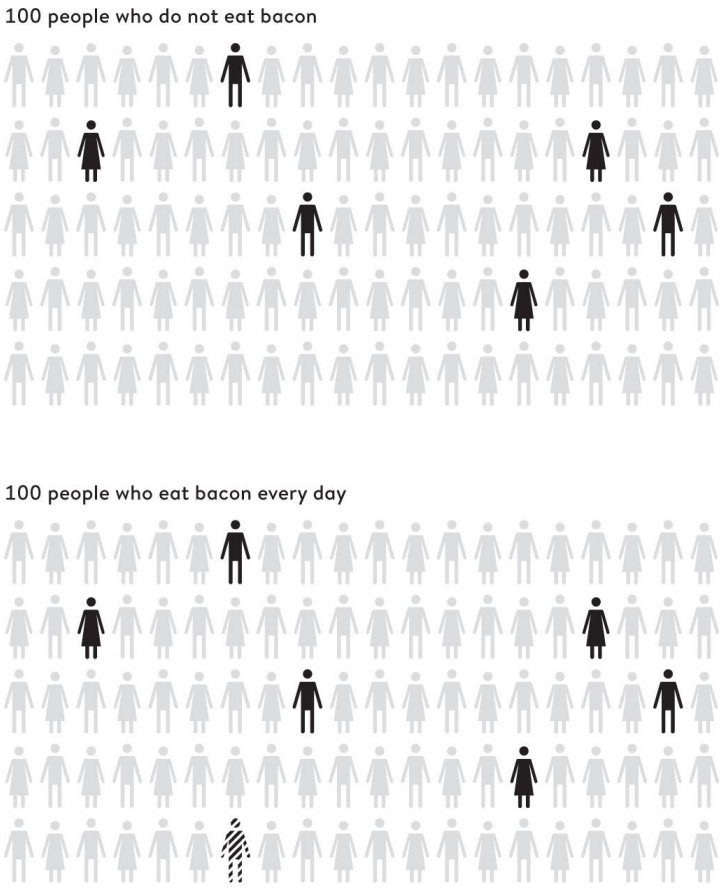
הנתון של 18% ידוע כסיכון **יחסי** מכיוון שהוא מייצג את העלייה בסיכון לחלות בסרטן המעי בין קבוצת אנשים שאוכלים 50 גרם בשר מעובד ביום, שיכולים, למשל, לייצג כריך בייקון יומי של שתי פריחות, לבין קבוצה שלא. פרשנים סטטיסטיים לקחו את הסיכון היחסי הזה וניסחו אותו מחדש לשינוי **בסיכון האבסולוטי**, כלומר השינוי בשיעור בפועל בכל קבוצה שהייתה צפויה לסבול מהאירוע השלילי.

הם הגיעו למסקנה כי בטווח הרגיל של הדברים, כ -6 מכל 100 אנשים שאינם אוכלים בייקון מדי יום צפויים לחלות בסרטן המעי במהלך חייהם. אם 100 אנשים דומים אכלו כריך בייקון בכל יום ויום בחייהם, אז על פי דו"ח IARC היינו מצפים כי 18% יותר יחלו בסרטן המעי, כלומר עלייה מ 6 ל 7 מקרים מתוך 100.[*] זהו מקרה אחד נוסף של סרטן המעי בכל אותם 100 אוכלי בייקון לכל החיים, וזה לא נשמע מרשים כמו הסיכון היחסי (עלייה של 18%), ועשוי לשמש לשים את הסיכון הזה בפרספקטיבה. עלינו להבחין בין מה שבאמת מסוכן לבין מה שנשמע מפחיד.[5]

דוגמה זו של כריך בייקון ממחישה את היתרון של תקשורת

במקום לדון באחוזים או בהסתברויות, אנחנו פשוט שואלים 'מה זה אומר לגבי 100 (או 1,000) אנשים?'. מחקרים פסיכולוגיים הראו כי טכניקה זו משפרת את ההבנה: למעשה רק לתקשר כי אכילת בשר נוספת זו הובילה ל "18% סיכון מוגבר" יכול להיחשב מניפולטיבי, שכן אנו יודעים ניסוח זה נותן רושם מוגזם של חשיבות הסיכון.[6] איור 1.4 משתמש **במערכי סמלים** כדי לייצג ישירות את התדירויות הצפויות של סרטן המעי אצל 100 אנשים.

באיור 1.4 סמלי ה"סרטן" מפוזרים באופן אקראי בין 100. בעוד פיזור כזה הוכח כמגביר את הרושם של בלתי צפוי, יש להשתמש בו רק כאשר יש סמל מודגש אחד נוסף. לא צריך להיות צורך לספור סמלים כדי לבצע השוואה חזותית מהירה.



תרשים 1.4

דוגמה לכריך בייקון באמצעות זוג מערכי אייקונים, עם סמלים מפוזרים באופן אקראי המראים את הסיכון המצטבר של אכילת בייקון מדי יום. מתוך 100 אנשים שאינם אוכלים בייקון, 6 (סמלים מוצקים) מפתחים סרטן המעי במהלך האירועים הרגיל. מתוך 100 אנשים שאוכלים בייקון כל יום בחייהם, יש מקרה אחד נוסף (מפוספס).*

עם זאת, דרכים נוספות להשוות בין שתי פרופורציות מוצגות בטבלה 1.2, המודגמת על ידי הסיכונים עבור אנשים שאוכלים בייקון ואינם אוכלים.

"1 ב-X" היא דרך נפוצה להביע סיכון, כמו לומר "1 מתוך 16 אנשים" כדי לייצג סיכון של 6%. אבל באמצעות ריבוי '1 ב...' הצהרות אינן מומלצות, מכיוון שאנשים רבים מתקשים להשוות. לדוגמה, כשנשאלו את השאלה "מהו הסיכון הגדול יותר, 1 ל-100, 1 ל-10 או 1 ל-1,000?", כרבע מהאנשים ענו תשובה שגויה: הבעיה היא שהמספר הגדול יותר קשור לסיכון הקטן יותר, ולכן נדרשת מיומנות מנטלית מסוימת כדי לשמור על דברים ברורים.

מבחינה טכנית, הסיכויים לאירוע הם היחס בין הסיכוי שהאירוע יקרה לבין הסיכוי שהוא לא יקרה. לדוגמה, מכיוון שמתוך 100 אנשים שאינם אוכלי בייקון, 6 יחלו בסרטן המעי ו-94 לא, הסיכוי לחלות בסרטן המעי בקבוצה זו הוא 94/6, המכונה לעתים '6 עד 94'. יחסי זכייה נפוצים בהימורים בבריטניה, אך הם משמשים באופן נרחב גם במודלים סטטיסטיים של פרופורציות, ומשמעות הדבר היא שמחקר רפואי מבטא בדרך כלל את ההשפעות הקשורות לטיפולים או להתנהגות במונחים של **יחסי סיכויים**.

למרות שהדבר נפוץ מאוד בספרות המחקרית, יחסי סיכויים הם דרך לא אינטואיטיבית לסכם הבדלים בסיכון. אם האירועים נדירים למדי אז יחסי הסיכויים יהיו קרובים מספרית לסיכונים היחסיים, כמו במקרה של כריכי בייקון, אבל עבור אירועים נפוצים יחס הסיכויים יכול להיות שונה מאוד מהסיכון היחסי, והדוגמה הבאה מראה שזה יכול להיות מאוד מבלבל עבור עיתונאים (ואחרים).

Method	Non-bacon eaters	Daily bacon eaters
Event rate	6%	7%
Expected frequency	6 out of 100	7 out of 100
	1 in 16	1 in 14
Odds	6/94	7/93

Comparative measures	
Absolute risk difference	1%, or 1 out of 100
Relative risk	1.18, or an 18% increase
'Number Needed to Treat'	100
Odds ratio	(7/93) / (6/94) = 1.18

*

טבלה 1.2

דוגמאות לשיטות להעברת הסיכון לכל החיים לסרטן המעי עם ובלי כריך בייקון יומי. "המספר הדרוש לטיפול" הוא מספר האנשים שצריכים לאכול כריך בייקון בכל יום בחייהם, כדי לצפות למקרה אחד נוסף של סרטן המעי (ולכן אולי מוטב להגדיר אותו כמספר הדרוש לאכילה).

איך אפשר לקרוא לעלייה מ-85% ל-87% עלייה של 20%?

סטטינים נלקחים באופן נרחב כדי להפחית את הכולסטרול ואת הסיכון להתקפי לב ושבץ, אבל כמה רופאים הביעו דאגה לגבי תופעות לוואי. מחקר שפורסם בשנת 2013 מצא כי 87% מהאנשים שנטלו סטטינים דיווחו על כאבי שרירים, לעומת 85% מאלה שלא נטלו סטטינים. בהסתכלות על האפשרויות להשוואת סיכונים המוצגות בטבלה 1.2, אנו עשויים לדווח על עלייה של 2% בסיכון האבסולוטי, או על סיכון יחסי של 0.85/0.87 = 1.02, כלומר עלייה יחסית של 2% בסיכון. הסיכויים בשתי הקבוצות ניתנים על ידי 0.13/0.87 = 6.7 ו-0.15/0.85 = 5.7, ולכן יחס הסיכויים הוא
5.7/6.7 = 1.18: בדיוק כמו כריכי בייקון, אך מבוסס על סיכונים מוחלטים שונים מאוד.

הדיילי *מייל* פירש באופן שגוי את יחס הסיכויים הזה של 1.18 כסיכון יחסי, והפיק כותרת שטענה כי סטטינים "מעלים את הסיכון בעד 20 אחוזים", וזה מצג שווא חמור של מה שהמחקר מצא בפועל. אבל לא את כל האשמה אפשר להטיל על העיתונאים: תקציר העיתון הזכיר רק את יחס הסיכויים מבלי להזכיר שזה תואם להבדל בין סיכונים מוחלטים של 85% לעומת 87%.[7]

זה מדגיש את הסכנה בשימוש ביחסי סיכויים בכל דבר מלבד בהקשר מדעי, ואת היתרון של דיווח תמיד על סיכונים מוחלטים ככמות הרלוונטית לקהל, בין אם מדובר בבייקון, סטטינים או כל דבר אחר.

הדוגמאות בפרק זה הדגימו כיצד המשימה הפשוטה לכאורה של חישוב והעברת פרופורציות יכולה להפוך לעניין מורכב. זה צריך להתבצע בזהירות ובמודעות, ואת ההשפעה של סיכומי נתונים מספריים או גרפיים ניתן לחקור על ידי עבודה עם פסיכולוגים מיומנים בהערכת התפיסה של חלופה

תבניות. תקשורת היא חלק חשוב ממחזור פתרון הבעיות, ולא צריכה להיות רק עניין של העדפה אישית.

תקציר

- משתנים בינאריים הם שאלות כן/לא, שקבוצות מהן ניתנות לסיכום כפרופורציות.
- מסגור חיובי או שלילי של פרופורציות יכול לשנות את ההשפעה הרגשית שלהם.
- סיכונים יחסיים נוטים לשדר חשיבות מוגזמת, ויש לספק סיכונים מוחלטים לבהירות.
- תדרים צפויים מקדמים הבנה ותחושת חשיבות מתאימה.
- יחסי הסיכויים נובעים ממחקרים מדעיים, אך אין להשתמש בהם לתקשורת כללית.
- גרפיקה צריכה להיבחר בזהירות ומודעות להשפעתם.

פרק 2

סיכום והעברת מספרים. הרבה מספרים

האם אנחנו יכולים לסמוך על חוכמת ההמונים?

בשנת 1907 פרנסיס גלטון, בן דודו של צ'ארלס דרווין ואיש אשכולות שפיתח זיהוי באמצעות טביעות אצבע, תחזיות מזג אוויר ואאוגניקה, כתב מכתב לכתב העת המדעי היוקרתי *Nature* על ביקורו בתערוכת מלאי השומן והעופות בעיר הנמל פלימות'. שם הוא ראה שור גדול מוצג והמתמודדים משלמים שישה פני כדי לנחש את משקלו ה"לבוש" של הבשר שנוצר לאחר שהבהמה המסכנה נשחטה. הוא השיג 787 מהכרטיסים שמולאו ובחר בערך האמצעי של 1,207 ליברות (547 ק"ג) כבחירה הדמוקרטית, "כל הערכה אחרת נידונה כגבוהה מדי או נמוכה מדי על ידי רוב המצביעים". משקל הלבוש התברר כ-543 ק"ג, שהיה קרוב להפליא לבחירתו על סמך 787 הקולות.[1] גלטון כינה את מכתבו "*Vox Populi*" (קול העם), אך תהליך זה של קבלת החלטות ידוע כיום יותר **כחוכמת ההמונים.**

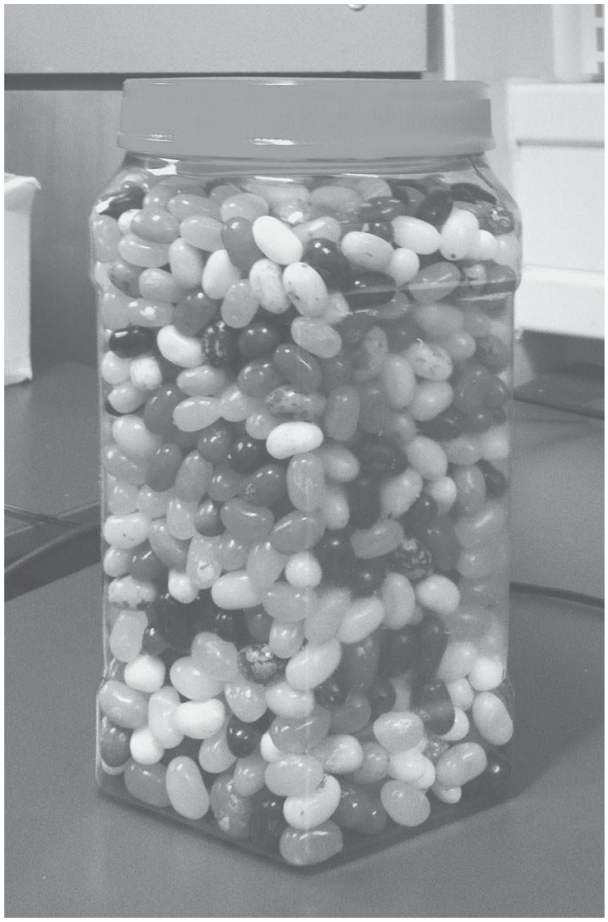
גלטון ביצע את מה שניתן לכנות כיום סיכום נתונים: הוא לקח מסה של מספרים שנכתבו על כרטיסים והפחית אותם למשקל משוער אחד של 1,207 ליברות. בפרק זה נבחן את הטכניקות שפותחו במאה שלאחר מכן לסיכום ותקשורת ערימות הנתונים שהפכו לזמינות. אנו נראה כי סיכומים מספריים של מיקום, פריסה, מגמה ומתאם קשורים קשר הדוק לאופן שבו ניתן לשרטט את הנתונים על נייר או מסך. ונתבונן במעבר העדין בין תיאור הנתונים בפשטות, לבין ניסיון לספר

סיפור באמצעות אינפוגרפיקה.

ראשית נתחיל בניסיון שלי לניסוי חוכמת ההמונים, המדגים רבות מהבעיות שצצות כאשר העולם האמיתי, הבלתי ממושמע, עם כל יכולתו למוזרות וטעות, משמש כמקור נתונים.

הסטטיסטיקה אינה עוסקת רק באירועים חמורים כגון סרטן וניתוחים. בניסוי די טריוויאלי, מתקשר המתמטיקה ג'יימס גריים ואני פרסמנו סרטון ביוטיוב וביקשנו מכל מי שצופה לנחש את מספר פולי הג'לי בצנצנת. ייתכן שתרצו לנסות את התרגיל הזה בעצמכם כשתראו את התמונה באיור 2.1 (המספר האמיתי יתגלה מאוחר יותר). תשע מאות וחמישה עשר אנשים סיפקו את הניחושים שלהם, שנעו בין 219 ל -31,337, ובפרק זה נבחן כיצד ניתן לתאר משתנים כאלה באופן גרפי ולסכם מספרית.

ראשית, איור 2.2 מציג שלוש דרכים להצגת תבנית הערכים שסיפקו 915 המשיבים: דפוסים אלה יכולים להיקרא באופן שונה התפלגות הנתונים, **התפלגות המדגם** או התפלגות אמפירית.*

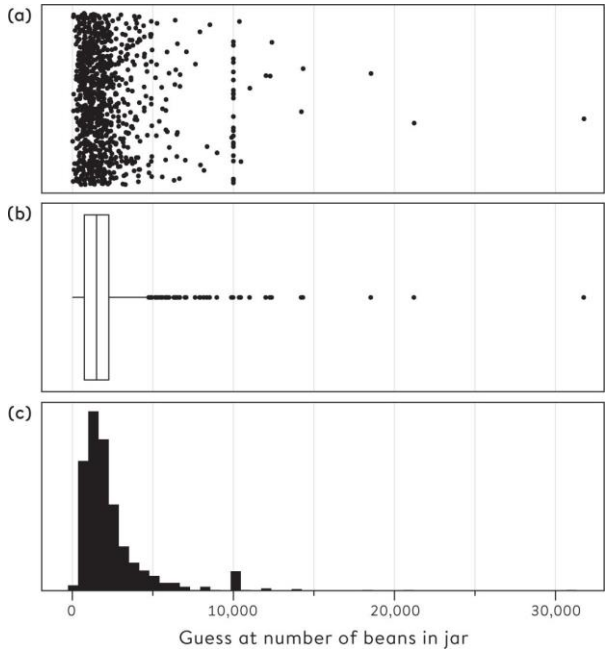


תרשים 2.1

כמה שעועית ג'לי יש בצנצנת הזאת?

שאלנו את זה בסרטון יוטיוב וקיבלנו 915

תגובות. התשובה תינתן בהמשך.



תרשים 2.2

דרכים שונות להראות את הדפוס של 915 ניחושים של מספר פולי הג'לי בצנצנת. (א) תרשים רצועה או דיאגרמת נקודות, עם ריצוד למניעת נקודות המונחות זו על גבי זו; (ב) חלקת קופסה ושפם; (ג) היסטוגרמה

(א) תרשים הרצועה, או דיאגרמת הנקודות, פשוט מציג כל נקודת נתונים כנקודה, אך כל אחת מהן מקבלת ריצוד אקראי כדי למנוע ניחושים מרובים של אותו מספר השוכבים זה על גבי זה ומטשטשים את התבנית הכוללת. זה מראה בבירור מספר רב של ניחושים בטווח של עד סביב 3,000, ולאחר מכן 'זנב' ארוך של ערכים ממש עד מעל 30,000, עם אשכול בדיוק 10,000.
(ב) עלילת הקופסה והשפם מסכמת כמה תכונות חיוניות של התפלגות הנתונים.*
(ג) היסטוגרמה זו פשוט סופרת כמה נקודות נתונים נמצאות בכל קבוצה של מרווחים – זה נותן מושג גס מאוד על צורת ההתפלגות.

תמונות אלה מיד להעביר כמה תכונות ייחודיות. התפלגות הנתונים **מוטה** מאוד, כלומר היא אפילו לא סימטרית בערך סביב ערך מרכזי כלשהו, ויש לה 'זנב ימני' ארוך בגלל התרחשותם של כמה ערכים גבוהים מאוד. סידרה אנכית של נקודות בתרשים הרצועה מציגה גם העדפה מסוימת למספרים עגולים.

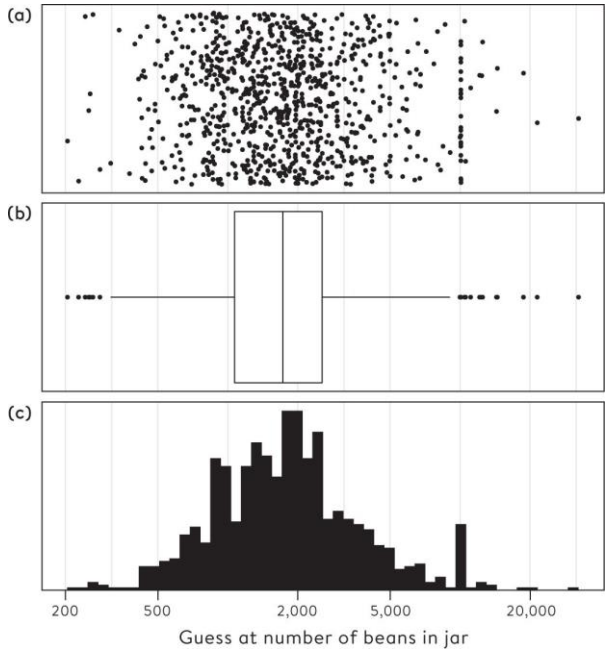
אבל יש בעיה עם כל התרשימים האלה. תבנית הנקודות פירושה שכל תשומת הלב ממוקדת בניחושים הגבוהים ביותר, כאשר רוב המספרים נדחסים לקצה השמאלי. האם נוכל להציג את הנתונים בצורה אינפורמטיבית יותר? יכולנו לזרוק את הערכים הגבוהים מאוד כמגוחכים (וכשניתחנו את הנתונים האלה במקור, שללתי באופן שרירותי למדי את כל מה שמעל 9,000). לחלופין, אנו יכולים לשנות את הנתונים באופן שיפחית את ההשפעה של הקצוות האלה, למשל על ידי התווייתם על מה שנקרא **סולם לוגריתמי**, שבו המרחב בין 100 ל-1,000 זהה למרחב שבין 1,000 ל-10,000.*

איור 2.3 מראה תבנית מעט ברורה יותר, עם התפלגות סימטרית למדי וללא חריגות קיצוניות. זה חוסך מאיתנו להוציא נקודות, וזה בדרך כלל לא רעיון טוב אלא אם כן מדובר בטעויות ברורות.

אין דרך "נכונה" להציג קבוצות של מספרים: לכל אחת מהחלקות שהשתמשנו בהן יש כמה יתרונות: תרשימי חשפנות מציגים נקודות בודדות, תרשימי קופסה ושפם נוחים לסיכומים חזותיים מהירים, והיסטוגרמות נותנות תחושה טובה לצורה הבסיסית של התפלגות הנתונים.

משתנים אשר נרשמים כמספרים מגיעים בזנים שונים:

- **ספירת משתנים**: כאשר המדידות מוגבלות למספרים השלמים 0, 1, 2... לדוגמה, מספר מקרי הרצח בכל שנה, או ניחושים לגבי מספר פולי הג'לי בצנצנת.
- **משתנים רציפים**: מדידות שניתן לעשות, לפחות באופן עקרוני, לדיוק שרירותי. לדוגמה, גובה ומשקל, שכל אחד מהם עשוי להשתנות הן בין אנשים והן מעת לעת. אלה עשויים, כמובן, להיות מעוגלים למספר שלם של סנטימטרים או קילוגרמים.



תרשים 2.3

תצוגות גרפיות של ניחושי שעועית הג'לי המשורטטות בקנה מידה לוגריתמי. (א) תרשים רצועות; (ב) חלקת קופסה ושפם; (ג) כל ההיסטוגרמה מציגה תבנית סימטרית למדי.

כאשר קבוצה של ספירות או תצפיות רציפות מצטמצמות לסטטיסטיקה מסכמת אחת, זה מה שאנו מכנים בדרך כלל הממוצע שלהן. כולנו מכירים את הרעיון של, למשל, שכר ממוצע, ציונים ממוצעים בבחינות וטמפרטורות ממוצעות, אבל לעתים קרובות לא ברור איך לפרש את הנתונים האלה (במיוחד אם האדם שמצטט את הממוצעים האלה לא מבין אותם).

ישנם שלושה פירושים בסיסיים למונח 'ממוצע', המכונה לעתים בצחוק על ידי המונח היחיד 'מצב ממוצע-חציון':

- **ממוצע**: סכום המספרים חלקי מספר המקרים.
- **חציון**: הערך האמצעי כאשר המספרים מסודרים. כך סיכם גלטון את קולות הקהל שלו.[*]
- **מצב**: הערך הנפוץ ביותר.

אלה ידועים גם כמדדים של מיקום התפלגות הנתונים.

פירוש המונח "ממוצע" כממוצע מוליד את הבדיחות הישנות על כך שכמעט לכולם יש יותר ממספר הרגליים הממוצע (שהוא כנראה בסביבות 1.99999), ולאנשים יש בממוצע אשך אחד. אבל זה לא רק עבור הרגליים והאשכים כי ממוצע ממוצע יכול להיות לא הולם. המספר הממוצע של פרטנרים מיניים מדווחים, וההכנסה הממוצעת במדינה, עשויים להיות דומים מעט לחוויה של רוב האנשים. הסיבה לכך היא שהאמצעים מושפעים יתר על המידה מכמה ערכים גבוהים במיוחד שגוררים את הסכום הכולל:[*] חשבו וורן ביטי או ביל גייטס (עבור שותפים מיניים והכנסה בהתאמה, עלי להוסיף).

ממוצעים ממוצעים יכולים להיות מטעים מאוד כאשר הנתונים הגולמיים אינם יוצרים תבנית סימטרית סביב ערך מרכזי אלא מוטים לצד אחד כמו ניחושי שעועית הג'לי, בדרך כלל עם קבוצה גדולה של מקרים סטנדרטיים אך עם זנב של כמה ערכים גבוהים מאוד (לדוגמה, הכנסה) או נמוכים (לדוגמה, רגליים). אני כמעט יכול להבטיח שבהשוואה לאנשים בגילך ובמינך, יש לך הרבה פחות מהסיכון הממוצע (הממוצע) למות בשנה הבאה. לדוגמה, טבלאות החיים בבריטניה מדווחות כי 1% מהגברים בני 63 מתים מדי שנה לפני יום הולדתם ה - 64, אך רבים מאלה שימותו כבר חולים קשה, ולכן הרוב המכריע שהם בריאים למדי יהיו בסיכון נמוך יותר מהממוצע הזה.

למרבה הצער, כאשר "ממוצע" מדווח בתקשורת, זה לעתים קרובות

לא ברור אם יש לפרש זאת כממוצע או כחציון. לדוגמה, המשרד הבריטי לסטטיסטיקה לאומית מחשב את הרווח השבועי הממוצע, שהוא *ממוצע*, תוך דיווח גם על הרווח השבועי החציוני על ידי הרשות המקומית. במקרה זה עשוי לעזור להבחין בין "הכנסה ממוצעת" (ממוצעת) לבין "הכנסתו של האדם הממוצע" (חציון). למחירי הבתים יש התפלגות מוטה מאוד, עם זנב ימני ארוך של נכסים יוקרתיים, ולכן מדדי מחירי הבתים הרשמיים מדווחים כחציונים. אבל אלה מדווחים בדרך כלל כ"מחיר הבית הממוצע", שהוא מונח מעורפל מאוד. האם זה מחיר הבית הממוצע (כלומר, החציון)? או מחיר הבית הממוצע (כלומר, הממוצע)? מקף יכול לעשות הבדל גדול.

הגיע הזמן לחשוף את תוצאות ניסויי חוכמת ההמונים שלנו עם פולי הג'לי: לא מרגשים כמו משקלו של שור, אבל עם מעט יותר קולות ממה שהיה לגאלטון.

מכיוון שהתפלגות הנתונים היא בעלת זנב ימני ארוך, הממוצע של 2,408 יהיה סיכום גרוע, ונראה שהמצב של 10,000 משקף בחירה קיצונית כלשהי של מספרים עגולים. ולכן כנראה עדיף לעקוב אחרי גלטון ולהשתמש בחציון כניחוש קבוצתי. מתברר שמדובר ב-1,775 שעועית. הערך האמיתי היה. . . 1,616.2 רק אדם אחד ניחש זאת במדויק, 45% מהאנשים ניחשו מתחת ל-1,616, ו-55% ניחשו למעלה, כך שלא הייתה נטייה שיטתית שהניחושים יהיו בצד הגבוה או הנמוך – אנו אומרים שהערך האמיתי נמצא באחוזון ה-45 של התפלגות הנתונים האמפירית. החציון, שהוא האחוזון ה-50, העריך יתר על המידה את הערך האמיתי ב-1,775-1,616=159, כך שיחסית לתשובה האמיתית החציון היה הערכת יתר בכ-10%, ורק כ-1 מכל 10 אנשים התקרב לכך. אז חוכמת ההמונים הייתה טובה למדי, והתקרבה לאמת יותר מ-90% מהאנשים הבודדים.

תיאור ההתפשטות של התפלגות נתונים

זה לא מספיק לתת סיכום יחיד להתפלגות – אנחנו צריכים לקבל מושג על ההתפשטות, המכונה לפעמים שונות. לדוגמה, ידיעת מידת הנעליים הממוצעת של גבר בוגר לא תעזור לחברת נעליים להחליט על הכמויות של כל מידה לייצר. מידה אחת אינה מתאימה לכולם, עובדה המומחשת היטב על ידי מושבי הנוסעים במטוסים.

טבלה 2.1 מציגה מגוון נתונים סטטיסטיים מסכמים עבור ניחושי פולי הג'לי, כולל שלוש דרכים לסכם את המרווח. המגוון הוא בחירה טבעית, אך ברור שהוא רגיש מאוד לערכים קיצוניים כמו הניחוש המוזר לכאורה של 31,337 שעועית.* לעומת זאת, הטווח **הבין-רבעוני** (IQR) אינו מושפע מקיצוניות. זהו המרחק בין האחוזון ה-25 לאחוזון ה-75 של הנתונים, ולכן מכיל את 'המחצית המרכזית' של המספרים, במקרה זה בין 1,109 ל-2,599 שעועית: ה'קופסה' המרכזית של חלקות הקופסה והשפם המוצגות לעיל מכסה את הטווח הבין-רבעוני. לבסוף, **סטיית התקן** היא מדד נפוץ של התפשטות. זהו המדד המורכב ביותר מבחינה טכנית, אך הוא מתאים רק לנתונים סימטריים המתנהגים היטב* , שכן הוא מושפע יתר על המידה גם מערכי הפריפריה. לדוגמה, הסרת הערך היחיד (השגוי כמעט בוודאות) של 31,337 מהנתונים מפחיתה את סטיית התקן מ-2,422 ל-1,398.*

Summary statistics for judgements of the number of jelly beans in a jar	Full data
Mean	2,408
Median	1,775
Mode	10,000
Range	219 to 31,337
Inter-quartile range	1,109 to 2,599
Standard deviation	2,422

*

טבלה 2.1

סטטיסטיקה מסכמת של 915 פסקי דין של שעועית ג'לי. המספר האמיתי היה 1,616.

הקהל בניסוי הקטן שלנו הראה שיש לו חוכמה לא מבוטלת, למרות כמה תגובות מוזרות. זה מראה כי נתונים לעתים קרובות יש כמה שגיאות, חריגים וערכים מוזרים אחרים, אבל אלה לא בהכרח צריך להיות מזוהה בנפרד ולא נכלל. הוא גם מצביע על היתרונות של שימוש במדדי סיכום שאינם מושפעים יתר על המידה מתצפיות מוזרות כגון 31,337 – מדדים אלה ידועים כמדדים חזקים, וכוללים את החציון ואת הטווח הבין-רבעוני. לבסוף, זה מראה את הערך הגדול של פשוט להסתכל על הנתונים, לקח שיחוזק על ידי הדוגמה הבאה.

תיאור הבדלים בין קבוצות מספרים

כמה פרטנרים מיניים מדווחים אנשים בבריטניה שהיו להם במהלך חייהם?

מטרתה של שאלה זו אינה רק להיות חטטנית לגבי חייהם הפרטיים של אנשים. כאשר איידס הפך לראשונה לדאגה רצינית בשנות השמונים, פקידי בריאות הציבור הבינו כי אין ראיות אמינות על התנהגות מינית בבריטניה, במיוחד במונחים של התדירות שבה אנשים החליפו בני זוג, כמה היו שותפים מרובים בו זמנית, ואילו פרקטיקות מיניות אנשים עסקו. הידע הזה היה חיוני כדי לחזות את התפשטות מחלות המין בחברה ולתכנן שירותי בריאות, ובכל זאת אנשים עדיין ציטטו מהנתונים הלא אמינים שאסף אלפרד קינסי בארה"ב בשנות הארבעים – שלא עשה שום ניסיון להשיג מדגם מייצג.

לכן, החל מסוף שנות ה-80 הוקמו בבריטניה ובארה"ב סקרים גדולים, זהירים ויקרים של התנהגות מינית, למרות התנגדות עזה מצד חוגים מסוימים. בבריטניה, מרגרט תאצ'ר משכה את תמיכתה מסקר גדול של אורח חיים מיני ברגע האחרון, אך אלה שערכו את המחקר הצליחו למרבה המזל למצוא מימון צדקה במקום, וכתוצאה מכך סקר עמדות מיניות וסגנון חיים לאומי (Natsal) שנערך בבריטניה כל עשר שנים מאז 1990.

הסקר השלישי, המכונה Natsal-3, בוצע בסביבות 2010 ועלה 7 מיליון ליש"ט.[3] לוח 2.2 מציג את הנתונים המסכמים לגבי מספר השותפים המיניים (המנוגדים) שדווחו על ידי בני 35-44 בנת"צ-3. זה תרגיל טוב להשתמש בסיכומים האלה לבד כדי לנסות לשחזר איך דפוס הנתונים עשוי להיראות. נציין כי הערך היחיד הנפוץ ביותר (מצב) הוא 1, המייצג את אותם אנשים שיש להם רק שותף אחד בחייהם, ובכל זאת יש גם טווח עצום. הדבר בא לידי ביטוי גם בהבדל המהותי בין האמצעים לבין החציונים, שהוא סימן מובהק להתפלגות נתונים עם זנב ימני ארוך. סטיות התקן גדולות, אך זהו מדד לא מתאים להתפשטות להפצת נתונים כזו, שכן היא תושפע יתר על המידה מכמה ערכים גבוהים במיוחד.

ניתן להשוות את תגובותיהם של גברים ונשים על ידי ציון כי גברים

דיווח בממוצע על 6 פרטנרים מיניים יותר מנשים, או לחילופין על כך שהגבר הממוצע (החציון) דיווח על 3 יותר פרטנרים מיניים מהאישה הממוצעת. או שבמונחים יחסיים, גברים מדווחים על כ-60% יותר פרטנרים מאשר נשים הן בממוצע והן בחציון.

Reported number of sexual partners in lifetime	Men aged 35–44	Women aged 35–44
Mean	14.3	8.5
Median	8	5
Mode	1	1
Range	0 to 500	0 to 550
Inter-quartile range	4 to 18	3 to 10
Standard deviation	24.2	19.7

*

טבלה 2.2

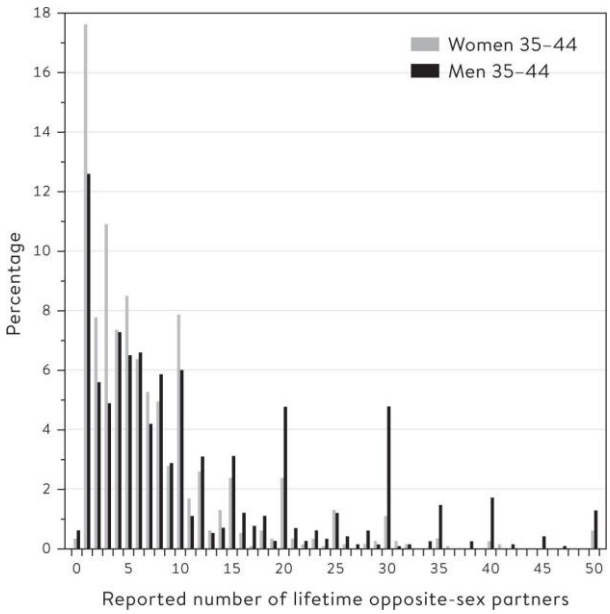
נתונים סטטיסטיים מסכמים על מספר הפרטנרים המיניים (המנוגדים) במהלך חייהם, כפי שדווחו על ידי 806 גברים ו-1,215 נשים בגילאי 35-44, בהתבסס על ראיונות שנערכו בנטסל-3 בין השנים 2010-2012. סטיות תקן נכללות לשלמות, אם כי הן סיכומים לא הולמים של התפשטות נתונים כאלה.

הבדל זה עשוי לעורר את חשדנו לגבי הנתונים. באוכלוסייה סגורה עם אותו מספר של גברים ונשים עם פרופיל גיל דומה, עובדה מתמטית היא שהמספר הממוצע של בני זוג מהמין השני צריך להיות זהה במהותו עבור גברים ונשים![*] אז למה גברים מדווחים על כל כך הרבה יותר בנות זוג מאשר נשים בקבוצת הגיל הזו של 35-44? זה יכול להיות חלקית בגלל שלגברים יש פרטנרים צעירים יותר, אבל גם בגלל שנראה שיש הבדלים שיטתיים באופן שבו גברים ונשים סופרים ומדווחים על ההיסטוריה המינית שלהם. אנו עשויים לחשוד שגברים נוטים יותר להגזים במספר בנות הזוג שלהם, או שנשים ממעיטות בחשיבותן, או בשתיהן.

תרשים 2.4 חושף את התפלגות הנתונים בפועל, התומכת ברושם המתקבל מהסטטיסטיקה המסכמת של זנב ימני קיצוני. אבל רק על ידי התבוננות בנתונים הגולמיים האלה מתגלים פרטים חשובים נוספים, כמו הנטייה החזקה של גברים ונשים כאחד לספק מספרים מעוגלים כאשר היו עשרה בני זוג או יותר (למעט הגבר הפדנטי למדי, אולי סטטיסטיקאי, שאמר במדויק, 'ארבעים -

שבע"). אתם יכולים, כמובן, לתהות לגבי המהימנות של דיווחים עצמיים אלה, והטיות פוטנציאליות בנתונים אלה יידונו בפרק הבא.

אוספים גדולים של נתונים מספריים מסוכמים ומועברים באופן שגרתי באמצעות כמה סטטיסטיקות של מיקום והתפשטות, והדוגמה של בן הזוג המיני הראתה כי אלה יכולים לקחת אותנו דרך ארוכה בתפיסת דפוס כולל. עם זאת, אין תחליף פשוט להסתכל על נתונים כראוי, והדוגמה הבאה מראה כי ויזואליזציה טובה היא בעלת ערך במיוחד כאשר אנו רוצים לתפוס את הדפוס בקבוצה גדולה ומורכבת של מספרים.



תרשים 2.4

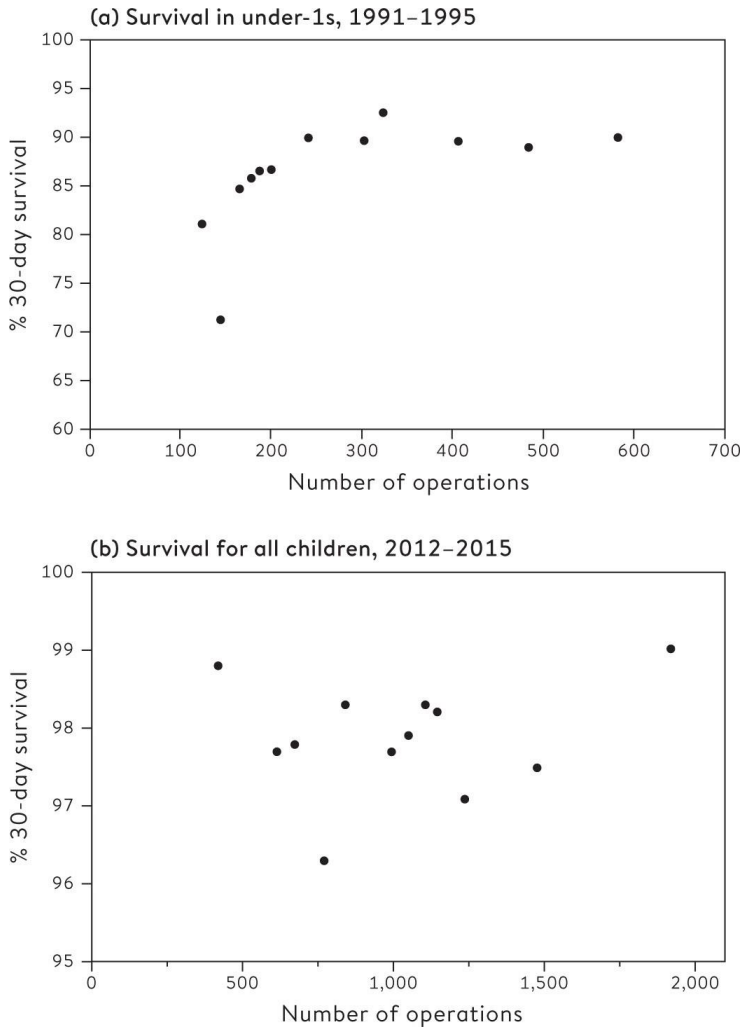
הנתונים שנמסרו על ידי נת"ל-3 מבוססים על ראיונות שנערכו בין השנים -2012 2010. הסדרה קוצצה ל-50 מטעמי מקום - המספרים מגיעים ל-500 עבור גברים ונשים כאחד. שימו לב לשימוש הברור במספרים עגולים עבור עשרה בני זוג או יותר, ולנטייה של גברים לדווח על יותר פרטנרים מאשר נשים.

תיאור קשרים בין משתנים

האם בבתי חולים עמוסים יותר יש שיעורי הישרדות גבוהים יותר?

יש עניין רב במה שמכונה "אפקט הנפח" בניתוחים - הטענה שבתי חולים עמוסים יותר מקבלים שיעורי הישרדות טובים יותר, אולי משום שהם משיגים יעילות רבה יותר ויש להם ניסיון רב יותר. תרשים 2.5 מראה שיעורי הישרדות של 30 יום בבתי חולים בבריטניה המבצעים ניתוחי לב בילדים לעומת מספר הילדים המטופלים. תרשים 2.5(א) מציג את הנתונים על ילדים מתחת לגיל שנה בתקופה 1991-1995 שהוצגו בתחילת הפרק האחרון, שכן קבוצת גיל זו נמצאת בסיכון גבוה יותר ועמדה במוקד חקירת בריסטול. תרשים 2.5(ב) מציג את הנתונים עבור כל הילדים מתחת לגיל 16 בשנים 2012-2015 שהוצגו קודם לכן בטבלה 1.1 – נתונים ספציפיים עבור ילדים מתחת לגיל שנה אינם זמינים לתקופה זו. הנפח משורטט על ציר *ה-x* האופקי, ושיעור ההישרדות על *ציר ה-y* האנכי.*

בנתוני 1991-1995 בתרשים 2.5(א) יש חריגה ברורה, בית חולים קטן יותר עם 71% הישרדות בלבד. זו הייתה בריסטול, ששיעורי ההישרדות הנמוכים שלה והחקירה הציבורית שבאה בעקבותיה סוקרו בפרק 1. אבל גם אם מסירים את בריסטול (נסו לשים את האגודל על נקודת הפריפריה), דפוס הנתונים לשנים 1991-1995 מצביע על כך שיש שיעורי הישרדות גבוהים יותר בבתי חולים המבצעים יותר ניתוחים.



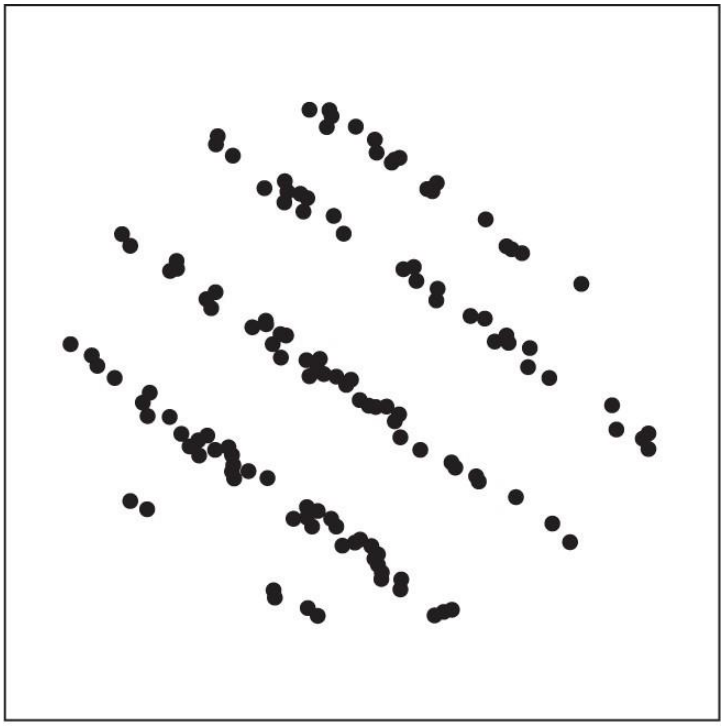
תרשים 2.5

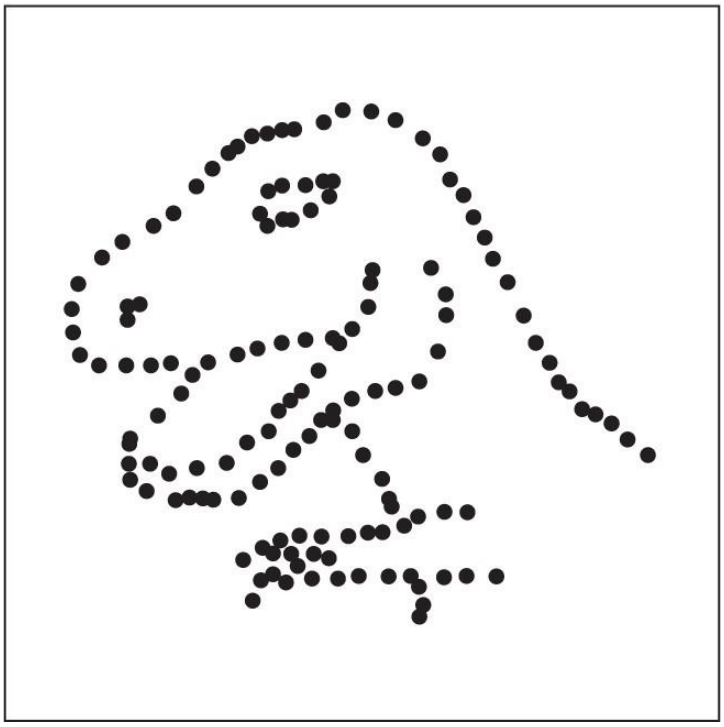
פיזור שיעורי הישרדות מול מספר ניתוחים בניתוחי לב בילדים. עבור (א) 1991–1995, מתאם פירסון הוא 0.59 ומתאם הדירוג הוא 0.85, עבור (ב) 2012–2015, מתאם פירסון הוא 0.17 ומתאם הדירוג הוא 0.03-.

נוח להשתמש במספר בודד כדי לסכם קשר עולה או יורד בהתמדה בין זוגות המספרים המוצגים בתרשים פיזור. זה נבחר בדרך כלל להיות **מקדם מתאם פירסון**, רעיון שהוצע במקור על ידי פרנסיס גלטון אך פורסם רשמית בשנת 1895 על ידי קארל פירסון, אחד ממייסדי הסטטיסטיקה המודרנית.[*]

מתאם פירסון נע בין 1- ל-1, ומבטא עד כמה קרוב לקו ישר נופלות הנקודות או נקודות הנתונים. מתאם של 1 מתרחש אם כל הנקודות נמצאות על קו ישר העולה כלפי מעלה, ואילו מתאם של 1- מתרחש אם כל הנקודות נמצאות על קו ישר היורד כלפי מטה. מתאם קרוב ל-0 יכול להגיע מפיזור אקראי של נקודות, או מכל דפוס אחר שבו אין מגמה שיטתית כלפי מעלה או כלפי מטה, שכמה דוגמאות לכך מוצגות באיור 2.6.

מתאם פירסון הוא 0.59 עבור הנתונים של 1991-1995 המוצגים באיור 2.5(a), דבר המצביע על קשר של הגדלת נפח עם הישרדות גוברת. אם מורידים את בריסטול מתאם פירסון עולה ל-0.67, מכיוון שהנקודות הנותרות נמצאות יותר על קו ישר. מדד חלופי נקרא **מתאם הדרגות של ספירמן** על שם הפסיכולוג האנגלי צ'ארלס ספירמן (שפיתח את הרעיון של אינטליגנציה כללית בסיסית), והוא תלוי רק בדרגות הנתונים ולא בערכים הספציפיים שלהם. משמעות הדבר היא שהוא יכול להיות קרוב ל-1 או -1 אם הנקודות קרובות לקו שעולה או יורד בהתמדה, גם אם קו זה אינו ישר; מתאם הדירוג של ספירמן לנתונים באיור 2.5(a) הוא 0.85, גבוה משמעותית ממתאם פירסון, שכן הנקודות קרובות יותר לעקומה עולה מאשר לקו ישר.





תרשים 2.6

שתי קבוצות של נקודות נתונים (פיקטיביות) שעבורן

מקדמי המתאם של פירסון הם שניהם

. זה בבירור לא אומר שאין קשר בין שני

המשתנים להיות

זמם. מתוך Datasaurus נפלא של אלברטו קהיר תריסר

.[4]

מתאם פירסון הוא 0.17 עבור נתוני 2012-2015 בתרשים 2.5(b), ומתאם הדירוג של ספירמן הוא -0.03, מה שמרמז על כך שאין עוד קשר ברור בין מספר המקרים לבין שיעורי ההישרדות. עם זאת, עם כל כך מעט בתי חולים מקדם המתאם יכול להיות רגיש מאוד לנתונים בודדים - אם נוציא את בית החולים הקטן ביותר, שיש לו שיעור הישרדות גבוה, מתאם פירסון קופץ ל -0.42.

מקדמי מתאם הם פשוט סיכומים של אסוציאציה, ולא ניתן להשתמש בהם כדי להסיק שבהחלט קיים קשר בסיסי בין נפח לשיעורי הישרדות, שלא לדבר על מדוע אחד מהם עשוי להתקיים.[*] ביישומים רבים ציר *ה-x* מייצג כמות המכונה **המשתנה הבלתי תלוי**, והעניין מתמקד בהשפעתו על המשתנה התלוי המשורטט על ציר *ה-y*. אבל, כפי שנחקור עוד בפרק 4 על סיבתיות, זה מניח מראש את הכיוון שבו עשויה להיות ההשפעה. אפילו בתרשים 2.5(א) איננו יכולים להסיק ששיעורי ההישרדות הגבוהים יותר נגרמו בשום מובן על ידי העלייה במספר המקרים – למעשה זה יכול להיות אפילו הפוך: בתי חולים טובים יותר פשוט משכו יותר חולים.

תיאור מגמות

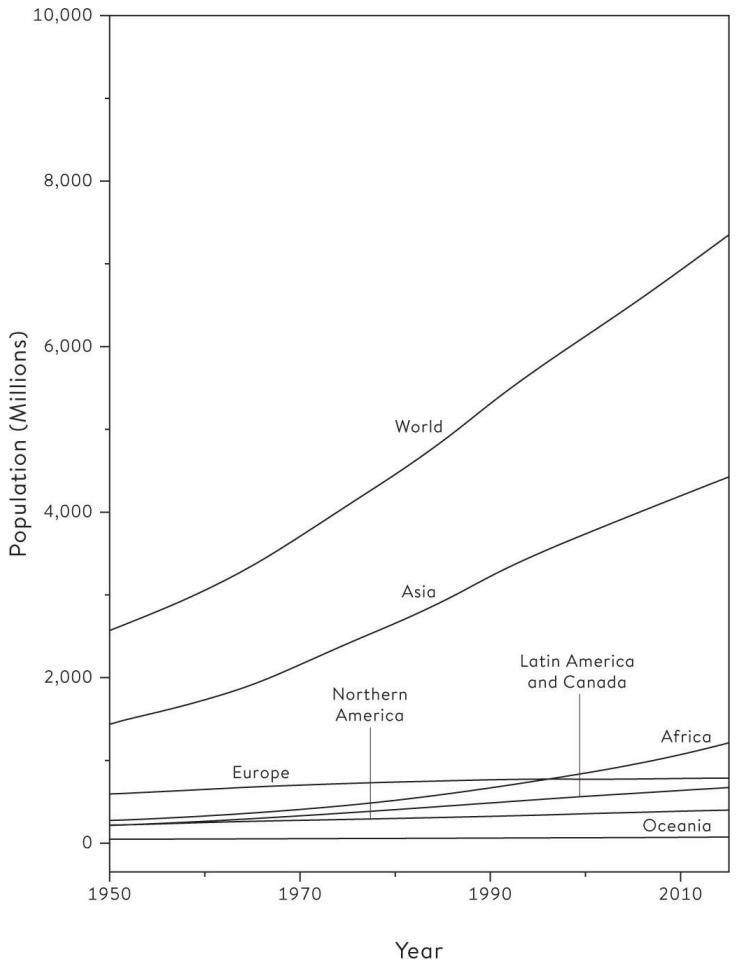
מהו דפוס גידול האוכלוסייה העולמי בחצי המאה האחרונה?

אוכלוסיית העולם גדלה, והבנת המניעים לשינוי האוכלוסייה היא בעלת חשיבות קריטית להיערכות לאתגרים העומדים בפני מדינות שונות בהווה ובעתיד. מחלקת האוכלוסין של האו"ם מפיקה אומדנים של ספירת האוכלוסייה בכל מדינות העולם משנת 1951 ועד היום, יחד עם תחזיות עד שנת 2100.[5] כאן אנו בוחנים את המגמות העולמיות מאז 1951.

איור 2.7(a) מציג גרפים קוויים פשוטים עבור אוכלוסיית העולם מאז 1951, ומראה עלייה של בערך פי שלושה לכמעט 7.5 מיליארד בתקופה זו. העלייה נבעה בעיקר ממדינות באסיה, אך קשה להבחין בדפוסים עבור יבשות אחרות באיור 2.7(a). עם זאת, סולם לוגריתמי באיור 2.7(b) מפריד בין היבשות, וחושף את השיפוע התלול יותר באפריקה, ואת המגמה השטוחה יותר ביבשות אחרות, במיוחד באירופה שבה האוכלוסייה הייתה לאחרונה

בירידה.

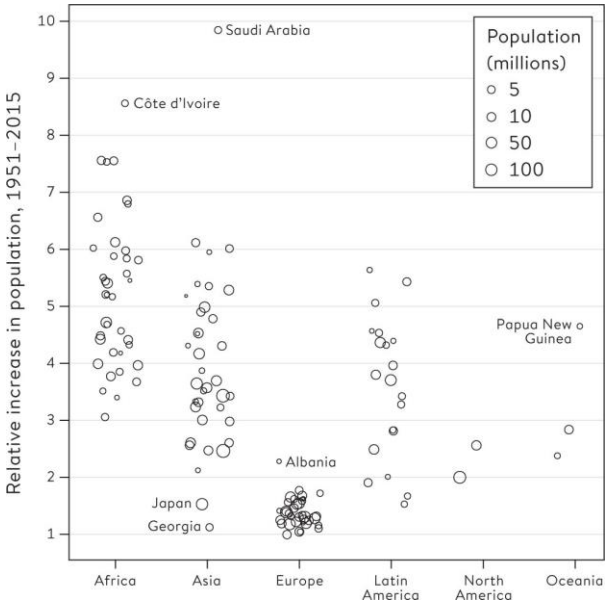
הקווים האפורים באיור 2.7(ב) מייצגים את השינויים במדינות בודדות, אך לא ניתן להצביע על סטיות ממגמת העלייה הכללית. תרשים 2.8 משתמש בסיכום פשוט של המגמה בכל מדינה – הגידול היחסי בין 1951 ל-2015 – כאשר עלייה יחסית של 4 פירושה שב-2015 יש פי ארבעה יותר אנשים מאשר ב-1951 (כפי שקרה למשל בליבריה, מדגסקר וקמרון). הפיכת הסמלים לפרופורציונליים לגודלה של מדינה מושכת את העין למדינות הגדולות יותר, וקיבוץ המדינות לפי יבשות מאפשר לנו לזהות באופן מיידי הן אשכולות כלליים והן מקרים מרוחקים. תמיד חשוב לפצל נתונים לפי גורם – כאן היבשות – שמסביר חלק מהשונות הכוללת.



תרשים 2.7

סך אוכלוסיית העולם, היבשות והמדינות בין השנים 1950 ל-2015, שני המינים יחד: (א) מראה מגמות בקנה מידה סטנדרטי, (ב) בקנה מידה לוגריתמי, יחד עם קווי המגמה של מדינות בודדות שאוכלוסייתן מנתה לפחות מיליון איש בשנת 1951.

10,000
World
Asia
Latin America
and Canada
Africa
Europe
Northern America
500
Oceania
100
50
10
5
2010
1990
1970
1950



תרשים 2.8

גידול יחסי באוכלוסייה בין 1951 ל -2015

עבור מדינות עם לפחות מיליון אנשים

בשנת 1951.

העליות הגדולות באפריקה בולטות, אך עם שונות רחבה וחוף השנהב הוא מקרה קיצון. גם באסיה ניכרת שונות עצומה, המשקפת את המגוון הרחב של המדינות ביבשת זו, כאשר יפן וגיאורגיה נמצאות בקצה אחד וערב הסעודית בקיצוניות השנייה, עם הגידול המדווח הגבוה ביותר בעולם. העליות באירופה היו נמוכות יחסית.

כמו כל גרפיקה טובה, זה מעלה יותר שאלות ומעודד חקירה נוספת, הן מבחינת זיהוי מדינות בודדות, וכמובן בחינת תחזיות של מגמות עתידיות.

ברור שיש מספר עצום של דרכים לבחון מערך נתונים כה מורכב כמו נתוני אוכלוסיית האו"ם, שאף אחת מהן אינה יכולה להיחשב "נכונה". עם זאת, אלברטו קהיר זיהה ארבע תכונות נפוצות של תצוגה חזותית טובה של נתונים:

א. הוא מכיל מידע אמין.
ב. העיצוב נבחר כך שדפוסים רלוונטיים יהיו בולטים.
ג. הוא מוצג בצורה מושכת, אבל המראה לא צריך להפריע לכנות, בהירות ועומק.
ד. כאשר הדבר מתאים, הוא מאורגן באופן המאפשר חקירה מסוימת.

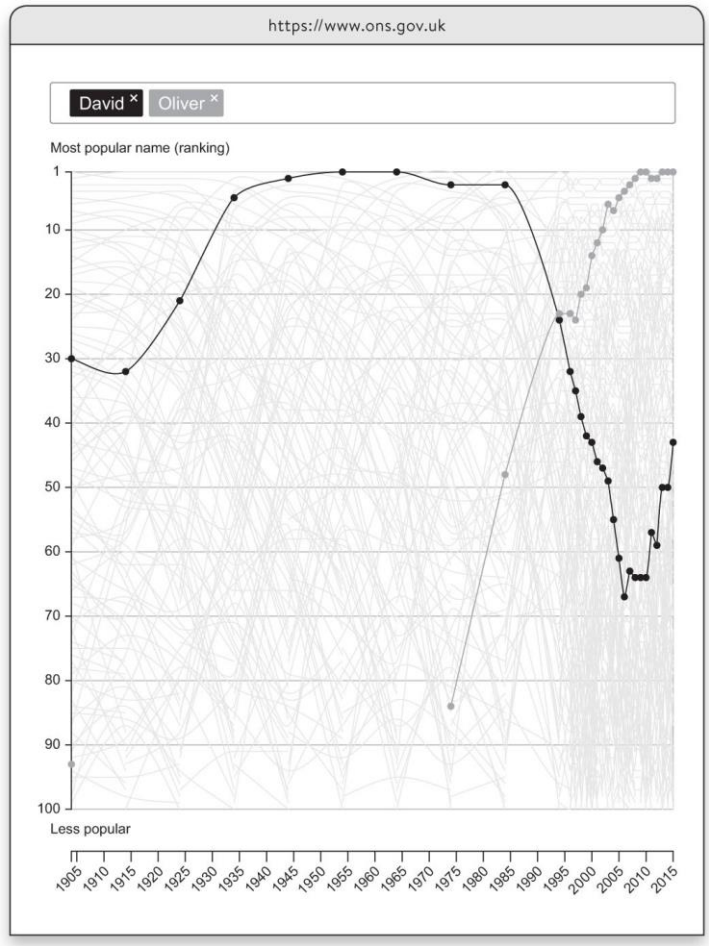
ניתן להקל על התכונה הרביעית על ידי מתן אפשרות לקהל לקיים אינטראקציה עם ההדמיה, ולמרות שקשה להמחיש זאת בספר, הדוגמה הבאה מראה את הכוח של התאמה אישית של תצוגה גרפית.

עד כמה השם שלי היה פופולרי לאורך זמן?

עלילות מסוימות הן כל כך מורכבות שקשה לזהות דפוסים מעניינים בעין בלתי. קחו לדוגמה את איור 2.9, שבו כל שורה מציגה את דרגת הפופולריות של שם פרטי מסוים לבנים שנולדו באנגליה ובווילס בין 1905 ל-2016.[6] זוהי היסטוריה חברתית יוצאת דופן, ועם זאת כשלעצמה רק מתקשרת את האופנות המשתנות במהירות במתן שמות, כאשר הקווים המאוחרים והצפופים יותר מרמזים על רוחב גדול יותר ו

מגוון שמות מאז אמצע שנות התשעים.

רק על ידי מתן אפשרות לאינטראקטיביות אנו יכולים לבחור קווים ספציפיים של עניין אישי. לדוגמה, מסקרן אותי לראות את המגמה של דיוויד, שם שהפך פופולרי במיוחד בשנות העשרים והשלושים, אולי בגלל שהנסיך מוויילס (לימים אדוארד השמיני קצר-המלוכה) נקרא דייוויד. אבל הפופולריות שלו ירדה מאוד – ב-1953 הייתי אחד מעשרות אלפי דודים, אבל ב-2016 רק 1,461 קיבלו את השם הזה, ויותר מ-40 שמות היו פופולריים יותר.



תרשים 2.9

צילום מסך של גרף אינטראקטיבי שסופק על ידי המשרד הבריטי לסטטיסטיקה לאומית, המציג את מגמת המיקום של שמו של כל ילד בטבלת הפופולריות של הליגה. הוריי חסרי הדמיון נתנו לי את שמו של הילד הפופולרי ביותר בשנת 1953, אבל מאז יצאתי מהאופנה, ב ניגוד ישיר לאוליבר. עם זאת, דיוויד הראה לאחרונה כמה סימני התאוששות, אולי בהשפעת דיוויד בקהאם.

תקשורת

פרק זה התמקד בסיכום והעברת נתונים באופן פתוח ולא מניפולטיבי; אנחנו לא רוצים להשפיע על הרגשות והעמדות של הקהל שלנו, או לשכנע אותו בפרספקטיבה מסוימת. אנחנו רק רוצים לספר את זה איך זה, או לפחות איך זה נראה, ובעוד שאנחנו אף פעם לא יכולים לטעון שאנחנו אומרים את האמת האבסולוטית, אנחנו יכולים לפחות לנסות להיות אמיתיים ככל האפשר.

כמובן שקל יותר לומר מאשר לעשות את הניסיון הזה לאובייקטיביות מדעית. כאשר החברה הסטטיסטית של לונדון (לימים החברה הסטטיסטית המלכותית) הוקמה בשנת 1834 על ידי צ'ארלס בבג', תומאס מלתוס ואחרים, הם הכריזו בנשגב כי "החברה הסטטיסטית תראה בכך את הכלל הראשון והחיוני ביותר בהתנהגותה להוציא בזהירות כל דעה מעסקאותיה ומפרסומיה – להגביל את תשומת לבה בקפדנות לעובדות
– וככל שיימצא, לעובדות שניתן לנסח אותן מספרית ולסדר אותן בטבלאות."[7] מלכתחילה הם לא שמו לב כלל להחמרה זו, ומיד החלו להכניס את דעתם על משמעות הנתונים שלהם על פשיעה, בריאות וכלכלה ומה יש לעשות בתגובה לכך. אולי הדבר הטוב ביותר שאנחנו יכולים לעשות עכשיו הוא להכיר בפיתוי הזה ולעשות כמיטב יכולתנו לשמור את דעותינו לעצמנו.

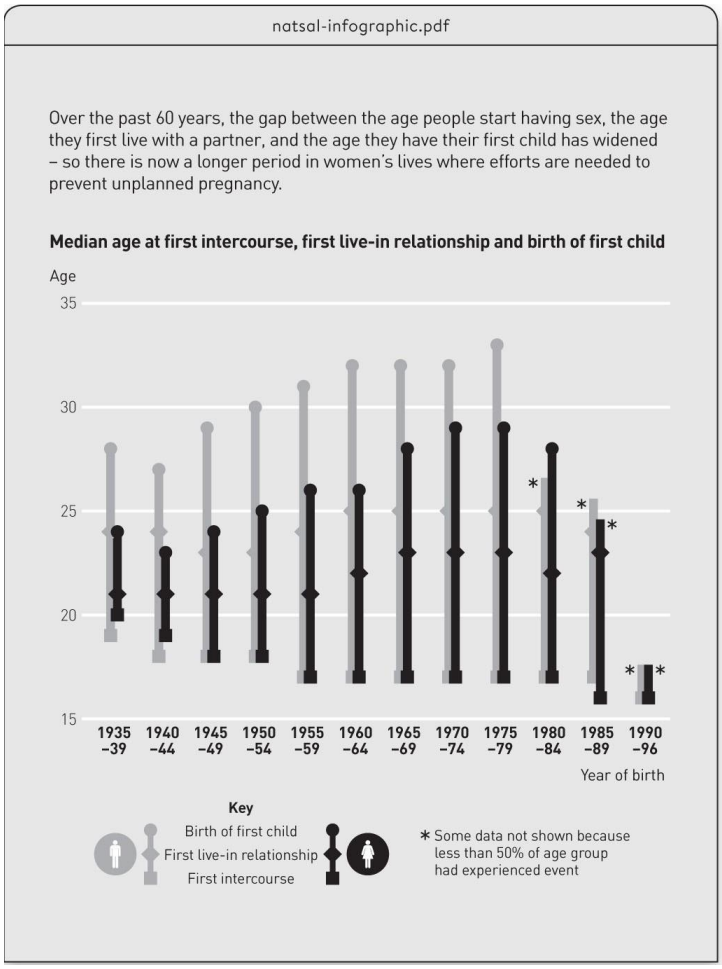
הכלל הראשון של תקשורת הוא לשתוק ולהקשיב, כך שתוכל להכיר את הקהל עבור התקשורת שלך, בין אם זה יכול להיות פוליטיקאים, אנשי מקצוע או הציבור הרחב. עלינו להבין את מגבלותיהם הבלתי נמנעות ואת כל אי ההבנות, ולהילחם בפיתוי להיות מתוחכמים וחכמים מדי, או לשים יותר מדי פרטים.

הכלל השני של תקשורת הוא לדעת מה אתה רוצה להשיג. יש לקוות שהמטרה היא לעודד דיון פתוח, וקבלת החלטות מושכלת. אבל נראה שאין שום נזק לחזור שוב על כך שהמספרים אינם מדברים בעד עצמם; ההקשר, השפה והעיצוב הגרפי תורמים כולם לאופן קבלת התקשורת. עלינו להכיר בכך שאנו מספרים סיפור, וזה בלתי נמנע שאנשים יעשו השוואות ושיפוטים, לא משנה כמה נרצה רק ליידע ולא לשכנע. כל מה שאנחנו יכולים לעשות הוא לנסות למנוע מראש תגובות בטן בלתי הולמות על ידי תכנון או אזהרה.

סיפור סיפורים עם סטטיסטיקה

פרק זה הציג את הרעיון של תצוגה חזותית של נתונים, המכונה לעתים dataviz. טכניקות אלה משמשות לעתים קרובות לחוקרים, או לקהלים מתוחכמים למדי, תוך שימוש בשריון סטנדרטי של חלקות שנבחרו על פי ערכן בהשגת הבנה וחקירת הנתונים, ולא על פי המשיכה החזותית גרידא שלהן. לאחר שעיבדנו את המסרים החשובים בנתונים שאנו רוצים להעביר, אנו עשויים להשתמש באינפוגרפיקה, או infoviz, כדי לתפוס את תשומת הלב של הקהל ולספר סיפור טוב.

אינפוגרפיקה מתוחכמת מופיעה באופן קבוע בתקשורת, אך איור 2.10 מציג דוגמה בסיסית למדי המספרת סיפור חזק של מגמות חברתיות על ידי ריכוז התשובות לשלוש שאלות בסקר הלאומי של בריטניה לעמדות מיניות וסגנון חיים (Natsal-3) בשנת 2010; באיזה גיל נשים וגברים קיימו לראשונה יחסי מין, התחילו לראשונה לחיות יחד, ויש להם ילד ראשון?[8] הגילאים החציוניים של כל אחד מאירועי החיים הללו משורטטים כנגד שנת לידתה של האישה, ושלוש הנקודות מחוברות בקו אנכי כבד. התארכותו המתמדת של קו זה בין נשים שנולדו בשנות ה-30 לבין נשים בשנות ה-70 מעידה על התארכות התקופה שבה יש צורך באמצעי מניעה יעילים.



תרשים 2.10

אינפוגרפיקה המבוססת על נתונים מהסקר הלאומי השלישי של בריטניה לעמדות מיניות וסגנונות חיים מיניים (Natsal-3) - הלקח מהנתונים מצביע הן חזותית והן מילולית.

מתקדמים עוד יותר הם הגרפיקה הדינמית, שבה ניתן להשתמש בתנועה כדי לחשוף דפוסים בשינויים לאורך זמן. המאסטר של טכניקה זו היה הנס רוסלינג, שהרצאותיו וסרטוני TED שלו קבעו סטנדרט חדש של סיפור סיפורים עם סטטיסטיקה, למשל על ידי הצגת הקשר בין עושר משתנה ובריאות באמצעות תנועה מונפשת של בועות המייצגות את ההתקדמות של כל מדינה משנת 1800 ועד היום. רוסלינג השתמש בגרפיקה שלו כדי לנסות לתקן תפיסות מוטעות לגבי ההבחנה בין מדינות "מפותחות" ו"לא מפותחות", כאשר העלילות הדינמיות חשפו כי לאורך זמן, כמעט כל המדינות נעו בהתמדה לאורך נתיב משותף לעבר בריאות ושגשוג גדולים יותר.[9*]

פרק זה הדגים רצף מתיאורים פשוטים ועלילות של נתונים גולמיים, ועד לדוגמאות מורכבות של סיפור סיפורים עם סטטיסטיקה. מחשוב מודרני פירושו שהדמיית נתונים הופכת קלה וגמישה יותר; ומכיוון שסטטיסטיקות סיכום יכולות להסתיר ולהאיר, תצוגות גרפיות מתאימות הן חיוניות. עם זאת, סיכום והעברת המספרים הגולמיים הוא רק השלב הראשון בתהליך הלמידה מהנתונים. כדי להתקדם בדרך זו, עלינו להתייחס לרעיון הבסיסי של מה שאנו מנסים להשיג מלכתחילה.

תקציר

- ניתן להשתמש במגוון נתונים סטטיסטיים כדי לסכם את ההתפלגות האמפירית של נקודות נתונים, כולל מדדים של מיקום והתפשטות.
- התפלגות נתונים מוטה היא נפוצה, וחלק מהסטטיסטיקות המסכמות רגישות מאוד לערכי פריפריה.
- סיכומי נתונים תמיד מסתירים פרטים מסוימים, ונדרשת זהירות כדי שמידע חשוב לא ילך לאיבוד.
- ניתן להציג באופן חזותי קבוצות בודדות של מספרים בתרשימי חשפנות, תרשימי קופסה ושפם והיסטוגרמות.
- שקול טרנספורמציות כדי לחשוף טוב יותר דפוסים, והשתמש בעין כדי לזהות דפוסים, חריגים, קווי דמיון ואשכולות.
- הסתכלו על זוגות מספרים כתרשימי פיזור, ועל סדרות זמן כתרשימי קווים.
- כאשר בוחנים נתונים, המטרה העיקרית היא למצוא גורמים המסבירים את השונות הכוללת.
- גרפיקה יכולה להיות אינטראקטיבית ומונפשת.
- אינפוגרפיקה מדגישה תכונות מעניינות ויכולה להנחות את הצופה בסיפור, אך יש להשתמש בה מתוך מודעות למטרתן ולהשפעתן.

פרק 3

למה אנחנו בכלל מסתכלים על נתונים? אוכלוסיות ומדידה

כמה פרטנרים מיניים באמת היו לאנשים בבריטניה ?

הפרק האחרון הראה כמה תוצאות מדהימות מסקר שנערך לאחרונה בבריטניה שבו אנשים דיווחו על מספר השותפים המיניים שהיו להם במהלך חייהם. שרטוט התגובות הללו חשף מאפיינים שונים, כולל זנב ארוך (מאוד), נטייה להשתמש במספרים עגולים כמו 10 ו-20, ויותר בני זוג שדווחו על ידי גברים מאשר נשים. אבל החוקרים שהוציאו מיליוני ליש"ט על איסוף הנתונים האלה לא ממש התעניינו במה שהמשיבים הספציפיים האלה אמרו – אחרי הכל, הובטחה להם אנונימיות מוחלטת. תגובותיהם היו אמצעי להשגת מטרה, מה שאומר משהו על הדפוס הכללי של שותפויות מיניות בבריטניה – אלה של מיליוני אנשים *שלא* נחקרו על התנהגותם המינית.

אין זה עניין של מה בכך לעבור מהתשובות בפועל שנאספו בסקר למסקנות לגבי בריטניה כולה. למעשה, זה לא נכון – קל מאוד פשוט לטעון שמה שהמשיבים האלה אומרים מייצג במדויק את מה שבאמת קורה במדינה. סקרי תקשורת על מין, שבהם אנשים מתנדבים למלא טפסים באתרי אינטרנט על מה שהם אומרים שהם עושים מאחורי דלתיים סגורות, עושים את זה כל הזמן.

את תהליך המעבר מהתשובות הגולמיות בסקר לטענות על התנהגות המדינה כולה ניתן לחלק לסדרה של שלבים:

א. הנתונים הגולמיים *המתועדים* על מספר השותפים המיניים שמשתתפי הסקר שלנו מדווחים עליהם אומרים לנו משהו על...
ב. המספר *האמיתי* של שותפים של אנשים במדגם שלנו, מה שאומר לנו משהו על...
ג. מספר השותפים של אנשים באוכלוסיית המחקר - אלה שהיו יכולים להיכלל בסקר שלנו - מה שאומר לנו משהו על ...
ד. מספר השותפים המיניים לאנשים בבריטניה, שהיא *אוכלוסיית היעד שלנו*.

היכן נמצאות נקודות התורפה בשרשרת החשיבה הזו? מעבר מהנתונים הגולמיים (שלב 1) לאמת על המדגם שלנו (שלב 2) פירושו להניח כמה הנחות חזקות לגבי מידת הדיוק של המשיבים כשהם אומרים כמה שותפים היו להם, ויש סיבות רבות לפקפק בהם. כבר ראינו נטייה ברורה של גברים להגזים, ונשים להמעיט בחשיבותן של בנות זוגן, אולי בגלל שנשים לא כוללות שותפויות שהן מעדיפות לשכוח, נטיות שונות לעגל כלפי מעלה או לעגל כלפי מטה, זיכרון לקוי ופשוט "הטיית מקובלות חברתית".[*]

המעבר מהמדגם שלנו (שלב 2) לאוכלוסיית המחקר (שלב 3) הוא אולי השלב המאתגר ביותר. ראשית עלינו להיות בטוחים שהאנשים שהתבקשו להשתתף בסקר הם מדגם אקראי של הזכאים: זה אמור להיות בסדר עבור מחקר מאורגן היטב כמו נצ"ל. אבל אנחנו גם צריכים להניח שהאנשים שבאמת מסכימים לקחת חלק הם מייצגים, וזה פחות פשוט. לסקרים יש כ -66% שיעור היענות, וזה טוב להפליא בהתחשב באופי השאלות. עם זאת, יש כמה ראיות לכך ששיעורי ההשתתפות מעט נמוכים יותר בקרב אלה שאינם פעילים מינית כל כך, אולי מאוזנים על ידי הקושי להגיע לראיונות עם חברים לא קונבנציונליים יותר בחברה.

לבסוף, המעבר מאוכלוסיית המחקר (שלב 3) לאוכלוסיית היעד (שלב 4) הוא פשוט יותר, בתנאי שאנו יכולים להניח שהאנשים שיכלו להתבקש להשתתף מייצגים את האוכלוסייה הבוגרת של בריטניה. במקרה של נצל יש להבטיח זאת על ידי תכנון ניסויי זהיר, המבוסס על מדגם אקראי של משקי בית, אם כי זה אומר שאנשים במוסדות כמו בתי סוהר, השירותים או

נזירות לא נכללו.

עד שנעבור על כל הדברים שיכולים להשתבש, זה עשוי להיות מספיק כדי לגרום למישהו להיות ספקן לגבי העלאת טענות כלליות כלשהן לגבי ההתנהגות המינית האמיתית של המדינה, בהתבסס על מה שנאמר לנו על ידי המשיבים לסקר. אבל כל המטרה של מדע סטטיסטי היא להחליק את ההתקדמות דרך השלבים האלה ולבסוף, עם הענווה הראויה, להיות מסוגלים לומר מה אנחנו יכולים ולא יכולים ללמוד מנתונים.

למידה מנתונים – תהליך של 'היסק אינדוקטיבי'

הפרקים הקודמים הניחו שיש לך בעיה, אתה מקבל נתונים, אתה מסתכל עליהם, ואז מסכם אותם בתמציתיות. לפעמים הספירה, המדידה והתיאור הם מטרה בפני עצמה. לדוגמה, אם אנחנו רק רוצים לדעת כמה אנשים עברו במחלקת תאונות וחירום בשנה שעברה, הנתונים יכולים לספר לנו את התשובה.

אבל לעתים קרובות השאלה חורגת מעבר לתיאור פשוט של נתונים: אנחנו רוצים ללמוד משהו גדול יותר מאשר רק התצפיות שלפנינו, בין אם זה לבצע תחזיות (כמה יגיעו בשנה הבאה?), או לומר משהו בסיסי יותר (מדוע המספרים גדלים?).

ברגע שאנחנו רוצים להתחיל להכליל מהנתונים – ללמוד משהו על העולם מחוץ לתצפיות המיידיות שלנו – אנחנו צריכים לשאול את עצמנו את השאלה, 'ללמוד על מה?' וזה מחייב אותנו להתמודד עם הרעיון המאתגר של **היסק אינדוקטיבי**.

לאנשים רבים יש מושג מעורפל של *ניכוי*, הודות לשרלוק הולמס המשתמש בהיגיון דדוקטיבי כאשר הוא מכריז בקור רוח כי חשוד חייב היה לבצע פשע. בחיים האמיתיים, ניכוי הוא תהליך של שימוש בכללי ההיגיון הקר כדי לעבוד מהנחות כלליות למסקנות מסוימות. אם החוק של המדינה הוא כי מכוניות צריך לנסוע בצד ימין, אז אנחנו יכולים להסיק כי בכל הזדמנות מסוימת עדיף לנהוג בצד ימין. אבל *אינדוקציה* עובדת בכיוון ההפוך, בלקיחת מקרים מסוימים ובניסיון להסיק מסקנות כלליות. לדוגמה, נניח שאיננו מכירים את המנהגים בקהילה לגבי נשיקת חברות על הלחי, ועלינו לנסות לפתור זאת על ידי התבוננות אם אנשים מתנשקים פעם, פעמיים, שלוש פעמים או בכלל לא. ההבחנה המכרעת היא שדדוקציה היא ודאית מבחינה לוגית, בעוד שהאינדוקציה היא בדרך כלל לא ודאית.

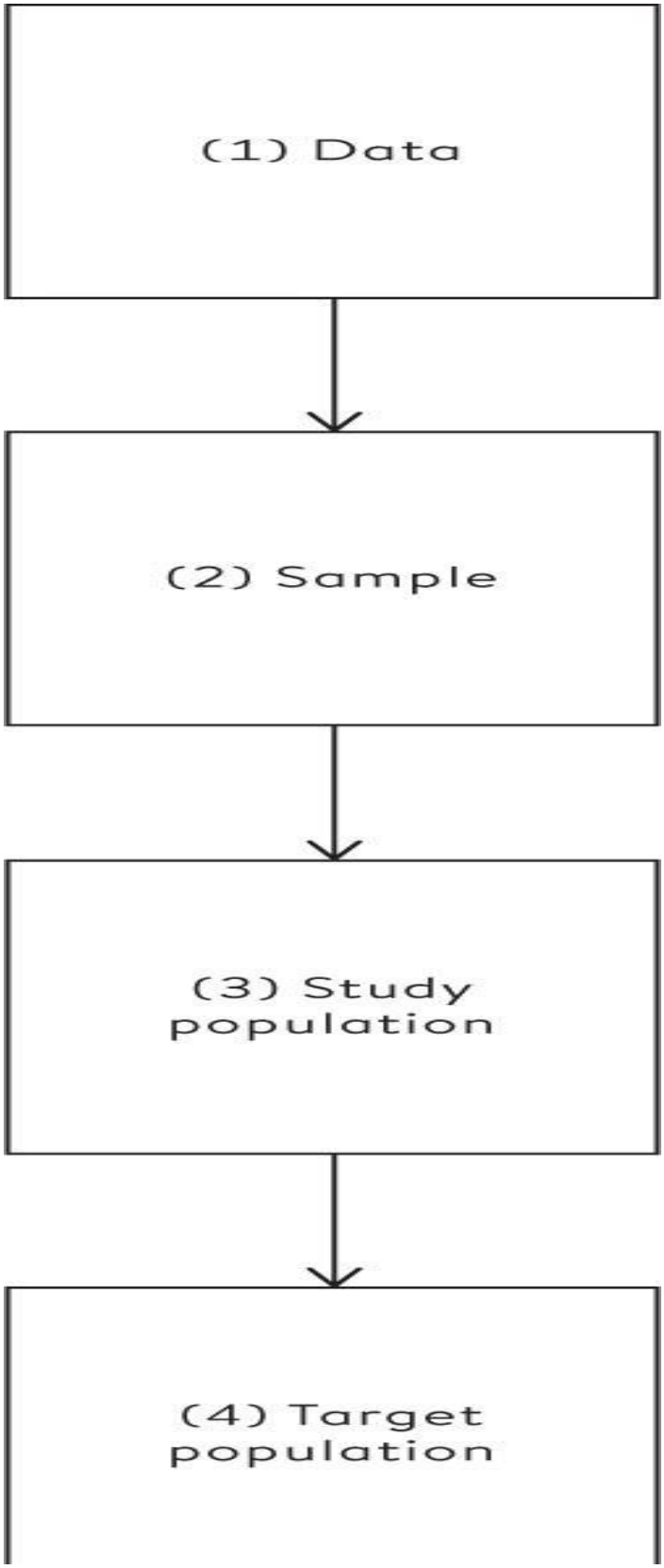
איור 3.1 מייצג היסק אינדוקטיבי כדיאגרמה כללית, ומציג

השלבים הכרוכים במעבר מנתונים ליעד הסופי של החקירה שלנו: כפי שראינו, הנתונים שנאספו בסקר המין מספרים לנו על התנהגות המדגם שלנו, שבו אנו משתמשים כדי ללמוד על האנשים שניתן היה לגייס לסקר, שממנו אנו מסיקים כמה מסקנות טנטטיביות על התנהגות מינית בכל הארץ.

כמובן, זה יהיה אידיאלי אם נוכל לעבור ישר מהסתכלות על הנתונים הגולמיים להעלאת טענות כלליות על אוכלוסיית היעד. בקורסים סטנדרטיים בסטטיסטיקה, מניחים שהתצפיות נמשכות באופן אקראי וישיר לחלוטין מהאוכלוסייה בעלת העניין הישיר. אבל זה רק לעתים רחוקות המקרה בחיים האמיתיים, ולכן אנחנו צריכים לשקול את כל התהליך של מעבר מנתונים גולמיים למטרה הסופית שלנו. וכפי שראינו בסקר המין, בעיות יכולות להתרחש בכל אחד מהשלבים השונים.

מעבר לנתונים (שלב 1) למדגם (שלב 2): אלה בעיות מדידה: האם מה שאנחנו רושמים בנתונים שלנו משקף במדויק את מה שאנחנו מעוניינים בו? אנחנו רוצים שהנתונים שלנו יהיו:

- אמין, במובן של שונות נמוכה מאירוע לאירוע, ולכן להיות מספר מדויק או חוזר.
- תקף, במובן של למדוד את מה שאתה באמת רוצה למדוד, ולא שיש הטיה שיטתית.


(1) Data
(2) Sample
(3) Study population
(4) Target population

תרשים 3.1

תהליך של היסק אינדוקטיבי: כל חץ יכול להתפרש

כ"אומר לנו משהו על"[1]

לדוגמה, הלימות סקר המין תלויה בכך שאנשים נותנים תשובות זהות או דומות מאוד לאותה שאלה בכל פעם שהם נשאלים, וזה לא צריך להיות תלוי בסגנון המראיין או בגחמות מצב הרוח או הזיכרון של המשיב. זה יכול להיבחן במידה מסוימת על ידי שאילת שאלות ספציפיות הן בתחילת הראיון והן בסופו. איכות הסקר גם מחייבת את המרואיינים להיות כנים כאשר הם מדווחים על פעילותם המינית, ולא להגזים או להמעיט באופן שיטתי בחוויותיהם. כל אלה הן דרישות חזקות למדי.

סקר לא יהיה תקף אם השאלות היו מוטות לטובת תגובה מסוימת. לדוגמה, בשנת 2017 הודיעה חברת התעופה המוזלת ריינאייר כי 92% מנוסעיה מרוצים מחוויית הטיסה שלהם. התברר שסקר שביעות הרצון שלהם איפשר רק את התשובות, "מצוין, טוב מאוד, טוב, הוגן, בסדר".[*]

ראינו כיצד מסגור חיובי או שלילי של מספרים יכול להשפיע על הרושם שניתן, ובאופן דומה מסגור של שאלה יכול להשפיע על התשובה. לדוגמה, סקר שנערך בבריטניה בשנת 2015 שאל אנשים אם הם תומכים או מתנגדים ל"מתן זכות הצבעה לבני 16 ו-17" במשאל העם על עזיבת האיחוד האירופי, ו-52% תמכו ברעיון בעוד 41% התנגדו לו. לכן הרוב תמך בהצעה זו כאשר היא ממוסגרת במונחים של הכרה בזכויות והעצמת צעירים.

אך כאשר נשאלו אותם משיבים את השאלה (הזהה לוגית) האם הם תומכים או מתנגדים ל"הורדת גיל ההצבעה מ-18 ל-16" במשאל העם, שיעור התומכים בהצעה ירד ל-37%, לעומת 56% מתנגדים. לכן, כאשר היא ממוסגרת במונחים של ליברליזציה מסוכנת יותר, ההצעה נתקלה בהתנגדות הרוב, היפוך בדעה שנגרם על ידי ניסוח מחדש פשוט של השאלה.[2]

התשובות לשאלות יכולות להיות מושפעות גם ממה שנשאל מראש, תהליך המכונה פריימינג. סקרי רווחה רשמיים מעריכים כי כ-10% מהצעירים בבריטניה רואים עצמם בודדים, אך שאלון מקוון של ה-BBC מצא שיעור גבוה בהרבה של 42% בקרב אלה שבחרו לענות. ייתכן שנתון זה נופח על ידי שני גורמים: אופי הדיווח העצמי של ה"סקר" מרצון, והעובדה שלשאלת הבדידות קדמה סדרה ארוכה של בירורים האם המשיב באופן כללי חש חוסר חברות, מבודד, עזוב וכו', וכל אלה עשויים היו להיות ראשוניים

אותם כדי לתת תשובה חיובית לשאלה המכרעת של תחושת בדידות.[3]

מעבר למדגם (שלב 2) לחקר האוכלוסייה (שלב 3): הדבר תלוי באיכות הבסיסית של המחקר, הידועה גם בשם **תוקפו הפנימי**: האם המדגם שאנו צופים בו משקף במדויק את המתרחש בקבוצה שאנו חוקרים בפועל? כאן אנו מגיעים לדרך המכרעת להימנע מהטיות: דגימה אקראית. אפילו ילדים מבינים מה זה אומר לבחור משהו באקראי: לעצום עיניים ולהושיט יד לתוך שקית ממתקים מעורבבת ולראות איזה צבע יוצא, או לשלוף מספר מהכובע כדי לראות מי מקבל פרס או פינוק (או לא). הוא שימש במשך אלפי שנים כדרך להבטיח הגינות וצדק, כאשר הוא ידוע כמיון,[*] ושימש כדרך להקצאת פרסים,[*] הפעלת הגרלות, ומינוי אנשים בעלי כוח כגון פקידים ומושבעים. היא גם הייתה מעורבת במטלות מפוכחות יותר, כמו לבחור אילו צעירים ייצאו למלחמה, או את מי לאכול בסירת הצלה שאבדה בים.

ג'ורג' גאלופ, שבעצם המציא את הרעיון של סקר דעת הקהל ב שנות השלושים, הגיעו עם אנלוגיה יפה לערך של דגימה אקראית. הוא אמר שאם בישלתם מחבת מרק גדולה, אתם לא צריכים לאכול את כולה כדי לברר אם היא זקוקה לתיבול נוסף. אתה יכול פשוט לטעום כפית, *בתנאי שנתת לו ערבוב טוב*. הוכחה מילולית לרעיון זה סופקה על ידי הגרלת הגיוס של מלחמת וייטנאם בשנת 1969, אשר היה צריך לספק רשימה מסודרת של ימי הולדת, ולאחר מכן גברים שיום הולדתם היה בראש הרשימה היו מגויסים תחילה לנסוע לווייטנאם, וכן הלאה בהמשך הרשימה. בניסיון פומבי להפוך את התהליך להוגן, הוכנו 366 קפסולות, שכל אחת מהן הכילה יום הולדת ייחודי, וקפסולות נועדו להיבחר מתוך קופסה באופן אקראי. אבל הקפסולות הוכנסו לקופסה לפי סדר חודש יום ההולדת, ולא עורבבו כראוי. זה אולי לא היה גורם לבעיה אם הגברים ששלפו את הקפסולות היו צוללים לתוך הקופסה, אבל כפי שמראה סרטון יוצא דופן, הם נטו לקחת אותן מלמעלה.[4] התוצאה הייתה שהיה זה מזל רע להיוולד בהמשך השנה: 26 מתוך 31 ימי הולדת בדצמבר גויסו בסופו של דבר, לעומת 14 בלבד בינואר.

הרעיון של "ערבוב" הולם הוא קריטי: אם אתה רוצה להיות מסוגל הכללה מהמדגם לאוכלוסייה, עליך לוודא שהמדגם שלך מייצג. עצם העובדה שיש מאסות של נתונים לא בהכרח עוזרת להבטיח מדגם טוב ואפילו יכולה לתת ביטחון שווא. לדוגמה, חברות הסקרים הציגו ביצועים גרועים בבחירות הכלליות של 2015 בבריטניה

בחירות, למרות שדגמו אלפי מצביעים פוטנציאליים. חקירה מאוחרת יותר האשימה דגימות לא מייצגות, במיוחד מסקרים טלפוניים: לא רק שטלפונים קוויים היוו את רוב המספרים שהתקשרו, אלא שפחות מ-10% מאלה שהתקשרו השיבו בפועל. זה לא צפוי להיות מדגם מייצג.

מעבר מאוכלוסיית המחקר (שלב 3) לאוכלוסיית היעד (שלב 4): לבסוף, אפילו עם מדידה מושלמת ומדגם אקראי מוקפד, התוצאות עדיין עשויות שלא לשקף את מה שרצינו לחקור מלכתחילה אם לא הצלחנו לשאול את האנשים שבהם אנו מתעניינים במיוחד. אנחנו רוצים שלמחקר שלנו יהיה **תוקף חיצוני**.

דוגמה קיצונית היא כאשר אוכלוסיית היעד שלנו מורכבת מבני אדם, בעוד שהצלחנו לחקור רק בעלי חיים, כגון ההשפעה של כימיקל על עכברים. פחות דרמטי הוא כאשר ניסויים קליניים של תרופות חדשות נערכו רק על גברים בוגרים, אבל התרופה משמשת אז "מחוץ לתווית" על נשים וילדים. היינו רוצים לדעת את ההשפעות על כולם, אבל זה לא יכול להיפתר על ידי ניתוח סטטיסטי בלבד – אנחנו בהכרח צריכים להניח הנחות ולהיות זהירים מאוד.

כשיש לנו את כל הנתונים

למרות שהרעיונות של למידה מנתונים מודגמים בצורה מסודרת על ידי התבוננות בסקרים, למעשה רוב הנתונים המשמשים כיום אינם מבוססים על דגימה אקראית, או למעשה דגימה כלשהי. נתונים שנאספים באופן שגרתי על, למשל, רכישות מקוונות או עסקאות חברתיות, או לניהול מערכת כגון חינוך או שיטור, יכולים להיות בעלי ייעוד מחדש כדי לעזור לנו להבין מה קורה בעולם. במצבים האלה יש לנו את כל הנתונים. מבחינת תהליך האינדוקציה המוצג באיור 3.1, אין פער בין שלבים 2 ו-3 – ה'מדגם' ואוכלוסיית המחקר זהים במהותם. זה אמנם מונע כל חשש לגבי גודל מדגם קטן, אבל בעיות רבות אחרות עדיין יכולות להישאר.

חשבו על השאלה כמה פשיעה יש בבריטניה, ועל השאלה הרגישה מבחינה פוליטית האם היא עולה או יורדת. ישנם שני מקורות נתונים עיקריים – אחד מבוסס סקר ואחד מנהלי. ראשית, סקר הפשיעה עבור אנגליה וויילס הוא פיסת דגימה קלאסית שבה כ -38,000 אנשים נחקרים מדי שנה על חוויות הפשע שלהם. בדיוק כמו סקר המין של Natsal, בעיות יכולות להתעורר

כאשר משתמשים בדוחות עצמם (שלב 1) כדי להסיק מסקנות לגבי חוויותיהם האמיתיות (שלב 2), מכיוון שהמשיבים עשויים שלא לומר את האמת – נניח על פשעי סמים שהם עצמם השתתפו בהם. לאחר מכן יש להניח שהמדגם מייצג את אוכלוסיית הזכאים ולקחת בחשבון את גודלה המצומצם (שלב 2 עד שלב 3), ולבסוף להכיר בכך שתכנון המחקר אינו מגיע לחלק כלשהו מאוכלוסיית היעד הכוללת, כגון העובדה שאף אחד מתחת לגיל 16 או מתגורר במעון קהילתי אינו מוטל בספק (שלב 3 עד שלב 4). עם זאת, עם אזהרות מתאימות, סקר הפשיעה עבור אנגליה וויילס הוא "סטטיסטיקה לאומית ייעודית" ומשמש לניטור מגמות ארוכות טווח.[5]

מקור הנתונים השני כולל דיווחים על פשעים שנרשמו על ידי המשטרה. הדבר נעשה לצרכים מנהליים ואינו מדגם: מכיוון שניתן לספור כל פשע שנרשם במדינה, 'אוכלוסיית המחקר' זהה למדגם. כמובן שאנחנו עדיין צריכים להניח שהנתונים שנרשמו באמת מייצגים את מה שקרה לאותם קורבנות שמדווחים על פשעים (שלב 1 עד שלב 2), אבל הבעיה העיקרית מתרחשת כאשר אנו רוצים לטעון שהנתונים על אוכלוסיית המחקר – אנשים שדיווחו על פשעים – מייצגים את אוכלוסיית היעד של כל הפשעים שבוצעו באנגליה ובוויילס. למרבה הצער, פשע המתועד במשטרה מחמיץ באופן שיטתי מקרים שהמשטרה אינה רושמת כפשע או שלא דווחו על ידי הקורבן; שימוש בסמים לא חוקיים, למשל, ואנשים שבוחרים שלא לדווח על גניבות וונדליזם במקרה שהאזור שלהם סובל מירידת ערך רכוש. כדוגמה קיצונית, לאחר שבנובמבר 2014 מתח דו"ח ביקורת על נוהלי ההקלטה של המשטרה, עלה מספר עבירות המין המתועדות מ-64 אלף בשנת 2014 ל-121 אלף בשנת 2017: כמעט הכפלה בתוך שלוש שנים.

אין זה מפתיע ששני מקורות נתונים שונים אלה יכולים להגיע למסקנות שונות למדי לגבי מגמות: למשל, סקר הפשיעה העריך כי הפשיעה ירדה ב-9% בין 2016 ל-2017, בעוד המשטרה רשמה 13% יותר עבירות. במה עלינו להאמין? לסטטיסטיקאים יש אמון רב יותר בסקר, וחששות לגבי אמינות נתוני הפשיעה המתועדים על ידי המשטרה הובילו אותו לאבד את ייעודו כסטטיסטיקה לאומית בשנת 2014.

כשיש לנו את כל הנתונים, פשוט לייצר סטטיסטיקות שמתארות את מה שנמדד. אבל כאשר אנו רוצים להשתמש בנתונים כדי להסיק מסקנות רחבות יותר על מה שקורה סביבנו, אז האיכות של

הנתונים הופכים להיות בעלי חשיבות עליונה, ועלינו להיות ערניים לסוג של הטיות שיטתיות שיכולות לסכן את אמינותן של טענות כלשהן.

אתרים שלמים מוקדשים לרשימה של הטיות אפשריות שיכולות להתרחש במדע הסטטיסטי, החל מהטיית הקצאה (הבדלים שיטתיים במי מקבל כל אחד משני טיפולים רפואיים מושווים) ועד להטיית מתנדבים (אנשים המתנדבים למחקרים שונים באופן שיטתי מהאוכלוסייה הכללית). רבים מהם הם הגיון בריא למדי, אם כי בפרק 12 נראה כמה דרכים מעודנות יותר שבהן סטטיסטיקה יכולה להיעשות רע. אך ראשית עלינו לשקול דרכים לתאר את מטרתנו הסופית – אוכלוסיית היעד.

"העקומה בצורת פעמון"

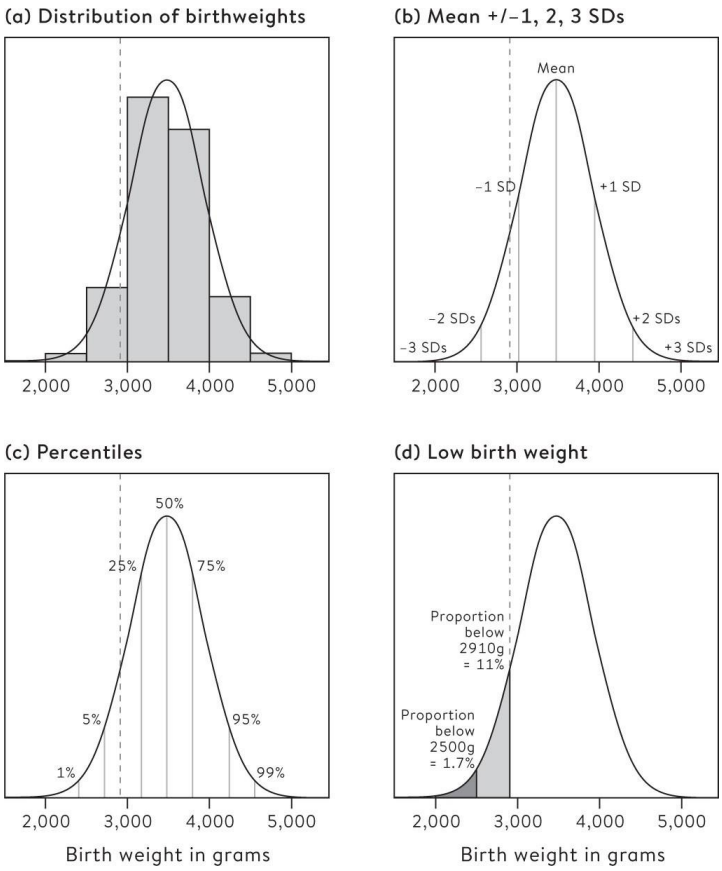
חברה בארה"ב ילדה זה עתה תינוק במשקל 2.91 ק"ג. נאמר לה שזה מתחת לממוצע, והיא מודאגת. האם המשקל נמוך באופן חריג?

כבר דנו במושג התפלגות נתונים – הדפוס שהנתונים יוצרים, המכונה לעתים התפלגות אמפירית או מדגם. בשלב הבא עלינו להתמודד עם הרעיון של **התפלגות אוכלוסייה** – הדפוס בכל קבוצת האינטרסים.

חשבו על אישה אמריקאית שזה עתה ילדה. אנו עשויים לחשוב על התינוק שלה כאילו נלקח, כמעין מדגם של אדם אחד בלבד, מכלל אוכלוסיית התינוקות שנולדו לאחרונה לנשים לבנות לא היספניות בארה"ב (הגזע שלה חשוב, שכן משקלי לידה מדווחים עבור גזעים שונים). התפלגות האוכלוסייה היא הדפוס שנוצר על ידי משקלי הלידה של כל התינוקות הללו, שאנו יכולים לקבל מהדו"ח של המערכת הלאומית לסטטיסטיקה חיונית של ארה"ב על משקלם של למעלה ממיליון תינוקות שנולדו במועד מלא בארה"ב בשנת 2013 לנשים לבנות שאינן היספניות – אם כי זה לא כל מערך הלידות העכשוויות, זה מדגם כל כך גדול שאנחנו יכולים לקחת את זה כאוכלוסייה.[6] משקלי לידה אלה מדווחים רק כמספרים בקבוצות המשתרעות על פני 500 גרם, ומוצגים באיור 3.2(a).

משקלו של התינוק של חברך מצוין כקו של 2,190 גרם, וניתן להשתמש במיקומו בחלוקה כדי להעריך אם משקלו "יוצא דופן". צורתה של התפלגות זו חשובה. מידות כגון משקל, הכנסה, גובה וכן הלאה יכולות, לפחות באופן עקרוני, להיות דקיקות באותה מידה

כרצונכם, ולכן ניתן להחשיב ככמויות "רציפות" שהתפלגות האוכלוסייה שלהן חלקה. הדוגמה הקלאסית היא "עקומת הפעמון", או **ההתפלגות הנורמלית**, שנחקרה לראשונה בפירוט על ידי קרל פרידריך גאוס בשנת 1809 בהקשר של טעויות מדידה באסטרונומיה ובמדידות.[*] התיאוריה מראה כי ההתפלגות הנורמלית צפויה להתרחש עבור תופעות המונעות על ידי מספר רב של השפעות קטנות, למשל תכונה פיזיקלית מורכבת שאינה מושפעת מגנים מעטים בלבד. משקל לידה, כאשר מסתכלים על קבוצה אתנית אחת ותקופת הריון אחת, עשוי להיחשב לתכונה כזו, ואיור 3.2(a) מראה עקומה נורמלית עם אותו ממוצע וסטיית תקן כמו המשקולות המתועדות. העקומה הנורמלית החלקה וההיסטוגרמה קרובות באופן מספק, ולתכונות מורכבות אחרות כגון גובה ומיומנויות קוגניטיביות יש גם התפלגות אוכלוסייה נורמלית בקירוב. תופעות אחרות, פחות טבעיות, עשויות להיות בעלות התפלגות אוכלוסייה שאינה נורמלית במובהק ולעתים קרובות מתאפיינות בזנב ימני ארוך, הכנסה היא דוגמה קלאסית.



תרשים 3.2

(א) התפלגות משקלי הלידה של 1,096,277 ילדים של נשים לבנות לא-היספניות בארה"ב בשנת 2013, שנולדו בשבועות 39-40 להריון, עם עקומה נורמלית עם ממוצע וסטיית תקן זהים למשקלי הרישום באוכלוסייה. A

תינוק במשקל 2,910 גרם מוצג כקו מקווקו. (ב) הממוצע ± 1, 2, 3 סטיות תקן (SDs) עבור העקומה הנורמלית. (ג) אחוזונים של העקומה הנורמלית. (ד) שיעור התינוקות במשקל לידה נמוך (אזור מוצל כהה), ותינוקות פחות מ-2,910 גרם (אזור מוצל בהיר).

ההתפלגות הנורמלית מאופיינת בממוצע שלה, או **בציפייה**, וסטיית התקן שלה, שכפי שראינו היא מדד להתפשטות – לעקומה המתאימה ביותר באיור 3.2(a) יש ממוצע של 3,480 גרם וסטיית תקן של 462 גרם (1 ליברות). אנו רואים שניתן ליישם את המדדים המשמשים לסיכום מערכי נתונים בפרק 2 גם כתיאורים של אוכלוסייה – ההבדל הוא שמונחים כגון ממוצע וסטיית תקן ידועים כסטטיסטיקה כאשר מתארים קבוצת נתונים, **ופרמטרים** כאשר מתארים אוכלוסייה. זהו הישג מרשים להיות מסוגל לסכם מעל 1,000,000 מדידות (כלומר, מעל מיליון לידות) רק על ידי שתי כמויות אלה.

יתרון גדול של הנחת צורה נורמלית עבור הפצה הוא כי כמויות חשובות רבות ניתן פשוט להשיג טבלאות או תוכנה. לדוגמה, איור 3.2(b) מציג את מיקום הממוצע וסטיות תקן 1, 2 ו-3 בכל צד של הממוצע. מהתכונות המתמטיות של ההתפלגות הנורמלית, אנו יודעים שכ-95% מהאוכלוסייה ייכללו במרווח הזמן הנתון בממוצע ± שתי סטיות תקן, ו-99.8% במרכז ± שלוש סטיות תקן. התינוק של החבר שלך נמצא בסביבות 1.2 סטיות תקן מתחת לממוצע - זה ידוע גם בשם **ציון Z שלה,** אשר פשוט מודד כמה סטיות תקן נקודת נתונים היא מהממוצע.

הממוצע וסטיית התקן יכולים לשמש כתיאורי סיכום עבור (רוב) ההתפלגויות האחרות, אך גם אמצעים אחרים עשויים להיות שימושיים. תרשים 3.2 (ג) מציג **אחוזונים** נבחרים המחושבים מתוך העקומה הנורמלית: לדוגמה, האחוזון ה-50 הוא החציון, הנקודה שמפצלת את האוכלוסייה לשניים ושניתן לומר שהיא משקלו של תינוק "ממוצע" – זה זהה לממוצע במקרה של התפלגות סימטרית כמו העקומה הנורמלית. האחוזון ה-25 (3,167 גרם) הוא המשקל שתחתיו שוכבים 25% מהתינוקות – האחוזונים ה-25 וה-75 (3,791 גרם) ידועים **כרביעונים**, והמרחק ביניהם (624 גרם), המכונה הטווח הבין-רבעוני, הוא מדד להתפשטות ההתפלגות. שוב, אלה בדיוק אותם סיכומים שבהם נעשה שימוש בפרק 2, אבל כאן הם מתייחסים לאוכלוסיות ולא למדגמים.

התינוק של חברכם נמצא באחוזון ה-11, מה שאומר ש-11% מהתינוקות שנולדו לנשים לבנות שאינן היספניות ישקלו פחות – איור 3.2(d) מראה את ה-11% האלה כאזור מוצל אפור בהיר. אחוזונים במשקל לידה

יש חשיבות מעשית, שכן משקל התינוק של חברך יהיה במעקב ביחס לגדילה הצפויה לתינוקות באחוזון 11,-* וירידה באחוזון התינוק עשויה להיות סיבה לדאגה.

מסיבות רפואיות ולא סטטיסטיות, תינוקות מתחת ל-2,500 גרם נחשבים ל"משקל לידה נמוך", ותינוקות מתחת ל-1,500 גרם נחשבים ל"משקל לידה נמוך מאוד". תרשים 3.2(d) מראה כי היינו מצפים ש-1.7% מהתינוקות בקבוצה זו יהיו במשקל לידה נמוך – למעשה המספר האמיתי היה 14,170 (1.3%), בהתאמה קרובה לתחזית מהעקומה הנורמלית. נציין כי בקבוצה מסוימת זו של לידות מלאות לאימהות לבנות שאינן היספניות יש שיעור קטן מאוד של משקלי לידה נמוכים – השיעור הכולל של כל הלידות בארה"ב בשנת 2013 היה 8%, בעוד שהשיעור בקרב נשים שחורות היה 13%, הבדל ניכר בין הגזעים.

אולי הלקח החשוב ביותר מדוגמה זו הוא שהאזור המוצל באפור כהה באיור 3.2(d) ממלא שני תפקידים:

א. הוא מייצג את *חלקה* של אוכלוסייה זו של תינוקות במשקל לידה נמוך.
ב. זוהי גם *ההסתברות* שתינוק שנבחר באופן אקראי בשנת 2013 שוקל פחות מ -2,500 גרם.

כך שניתן לחשוב על אוכלוסייה כעל קבוצה פיזית של פרטים, אך גם כמספקת את **התפלגות ההסתברות** לתצפית אקראית. פרשנות כפולה זו תהיה בסיסית כאשר נגיע להיסק סטטיסטי פורמלי יותר.

כמובן שבמקרה זה אנו יודעים את הצורה והפרמטרים של האוכלוסייה, ולכן אנו יכולים לומר משהו הן על הפרופורציות באוכלוסייה, והן על הסיכויים להתרחשות אירועים שונים לתצפית אקראית. אבל כל העניין של הפרק הזה הוא שאנחנו בדרך כלל לא יודעים על אוכלוסיות, ולכן רוצים לעקוב אחר התהליך האינדוקטיבי וללכת הפוך, מנתונים לאוכלוסייה. ראינו שהמדדים הסטנדרטיים של ממוצע, חציון, מצב וכן הלאה, שפיתחנו עבור דגימות, מתרחבים לאוכלוסיות שלמות – אבל ההבדל הוא שאנחנו לא יודעים מה הם. וזה האתגר העומד בפנינו בפרק הבא.

מהי האוכלוסייה?

שלבי האינדוקציה המתוארים לעיל עובדים היטב עם סקרים מתוכננים, אך ניתוח סטטיסטי רב אינו מתאים בקלות למסגרת זו. ראינו כי, במיוחד כאשר משתמשים ברשומות מנהליות כגון דוחות משטרה על פשע, ייתכן שיש לנו את כל הנתונים האפשריים. אבל למרות שאין דגימה, הרעיון של אוכלוסייה בסיסית עדיין יכול להיות בעל ערך.

שקול את נתוני ניתוחי הלב של ילדים בפרק 1. יצאנו מנקודת הנחה נועזת למדי שאין בעיות מדידה – במילים אחרות, יש לנו אוסף שלם של הניתוחים ושל ניצולים ל-30 יום בכל בית חולים. אז הידע שלנו על המדגם (שלב 2) הוא מושלם.

אבל מהי אוכלוסיית המחקר? יש לנו נתונים על כל הילדים ועל כל בתי החולים, ולכן אין קבוצה גדולה יותר שממנה הם נדגמו. למרות הרעיון של אוכלוסייה הוא הציג בדרך כלל די כלאחר יד לתוך קורסים סטטיסטיקה, דוגמה זו מראה שזה רעיון מסובך ומתוחכם כי שווה לחקור בפירוט מסוים, כמו הרבה רעיונות חשובים לבנות על הרעיון הזה.

ישנם שלושה סוגים של אוכלוסיות שמהן ניתן לשאוב מדגם, בין אם הנתונים מגיעים מאנשים, עסקאות, עצים או כל דבר אחר.

- אוכלוסייה פשוטו *כמשמעו*. זוהי קבוצה הניתנת לזיהוי, למשל כאשר אנו בוחרים אדם באופן אקראי בעת ההצבעה. או שאולי יש קבוצה של אנשים שאפשר למדוד, ולמרות שאנחנו לא באמת בוחרים קבוצה באקראי, יש לנו נתונים ממתנדבים. לדוגמה, אנו עשויים לשקול את האנשים שניחשו את מספר פולי הג'לי כמדגם מאוכלוסיית כל החנונים המתמטיים שצופים בסרטוני YouTube.
- אוכלוסייה וירטואלית . לעתים קרובות אנו מבצעים מדידות באמצעות מכשיר, כגון מדידת לחץ דם של מישהו או מדידת זיהום אוויר. אנו יודעים שתמיד נוכל לבצע מדידות נוספות ולקבל תשובה מעט שונה, כפי שתדעו אם אי פעם ביצעתם מדידות לחץ דם חוזרות. הקרבה של הקריאות המרובות תלויה בדיוק המכשיר וביציבות הנסיבות – אנו עשויים לחשוב על כך כעל הסקת תצפיות מאוכלוסייה וירטואלית של כל המדידות שניתן לבצע אם היה לנו מספיק זמן.

- אוכלוסייה *מטאפורית* , כאשר אין אוכלוסייה גדולה יותר. זהו מושג יוצא דופן. כאן אנו פועלים *כפי* שהנתונים נלקחו מאוכלוסייה כלשהי באקראי, אבל ברור שזה לא – כמו במקרה של ילדים שעברו ניתוח לב: לא עשינו שום דגימה, יש לנו את כל הנתונים, ואין יותר מה לאסוף. חישבו על מספר מקרי הרצח המתרחשים מדי שנה, על תוצאות הבדיקה של כיתה מסוימת, או על נתונים על כל מדינות העולם – אף אחד מאלה לא יכול להיחשב כמדגם מאוכלוסייה אמיתית.

הרעיון של אוכלוסייה מטאפורית הוא מאתגר, ואולי מוטב לחשוב על מה שראינו כאילו נשאב ממרחב דמיוני כלשהו של אפשרויות. לדוגמה, ההיסטוריה של העולם היא מה שהיא, אבל אנחנו יכולים לדמיין את ההיסטוריה מתנהלת אחרת, ובמקרה הגענו רק לאחד ממצבים אפשריים אלה של העולם. קבוצה זו של כל ההיסטוריות האלטרנטיביות יכולה להיחשב אוכלוסייה מטאפורית. כדי להיות יותר קונקרטיים, כשהסתכלנו על ניתוחי לב בילדות בבריטניה בין 2012 ל-2015, היו לנו את כל הנתונים על ניתוחים באותן שנים וידענו כמה מקרי מוות וכמה ניצולים היו. עם זאת, אנו יכולים לדמיין היסטוריות נגדיות שבהן אנשים שונים היו עשויים לשרוד, באמצעות נסיבות בלתי צפויות שאנו נוטים לכנות "מקריות".

זה צריך להיות ברור כי מעט מאוד יישומים של מדע סטטיסטי באמת כרוך דגימה אקראית פשוטו כמשמעו, וכי זה נפוץ יותר ויותר לקבל את כל הנתונים הזמינים פוטנציאלית. אף על פי כן, חשוב מאוד להחזיק ברעיון של אוכלוסייה דמיונית שממנה נלקח ה"דגימה" שלנו, שכן אז נוכל להשתמש בכל הטכניקות המתמטיות שפותחו לדגימה מאוכלוסיות אמיתיות.

באופן אישי, אני די אוהב להתנהג כאילו כל מה שקורה סביבנו הוא תוצאה של בחירה אקראית כלשהי מתוך כל הדברים האפשריים שיכולים לקרות. זה תלוי בנו אם נבחר להאמין שזה באמת מקרי, אם זה רצון של אל או אלים, או כל תיאוריה אחרת של סיבתיות: זה לא משנה את המתמטיקה. זוהי רק אחת הדרישות המותחות את המוח ללמידה מנתונים.

תקציר

- הסקה אינדוקטיבית דורשת עבודה מהנתונים שלנו, דרך מדגם המחקר ואוכלוסיית המחקר, לאוכלוסיית יעד.
- בעיות והטיות יכולות לצוץ בכל שלב של נתיב זה.
- הדרך הטובה ביותר להתקדם ממדגם לאוכלוסיית המחקר היא לצייר מדגם אקראי.
- ניתן לחשוב על אוכלוסייה כעל קבוצה של פרטים, אך גם כמספקת את התפלגות ההסתברות לתצפית אקראית השאובה מאותה אוכלוסייה.
- ניתן לסכם אוכלוסיות באמצעות פרמטרים המשקפים את הנתונים המסכמים של נתוני המדגם.
- לעתים קרובות הנתונים אינם עולים כמדגם מאוכלוסייה מילולית. כשיש לנו את כל הנתונים שיש, אז אנחנו יכולים לדמיין אותם שאובים מאוכלוסייה מטאפורית של אירועים שהיו יכולים להתרחש, אבל לא התרחשו.