חקירת המאפיינים הסטטיסטיים של אומדן צפיפות הקרנל הכפול

הרולד שיפ

תַקצִיר

מניע: ישנן דוגמאות רבות באפידמיולוגיה של נתונים המתרחשים כנקודות שמופיעות במרחב. דוגמה אחת היא כתובת הבית של אנשים בשכונה, ובמיוחד, אם יש להם מצב רפואי מסוים. נתונים אלה מאפשרים לאדם לחשב את שיעור השכיחות של מצב זה. לאחרונה, טכניקה הידועה בשם צפיפות הקרנל הכפול שימשה להערכת פונקציות שכיחות.

החוקרים נוטים לאחרונה להשתמש צפיפות הקרנל הכפול בשל אובדן נתונים עקב סיכום של נתונים באיזורים גיאוגרפיים כמו ערים או שכונות. צבירה זו נחשבה נחוצה הן לצמצום השפעות השונות על הטעויות היחסיות של התוצאות והן לשמירה על הפרטיות של אנשים שחלו במחלה.

מחקר זה בוחן חלק מהמאפיינים הסטטיסטיים של צפיפות הקרנל הכפול. בפרט, אנו מנתחים באופן אמפירי כיצד גורמים שונים של האוכלוסייה והפצת השכיחות משפיעים על הדיוק של צפיפות הקרנל הכפולה.

התרומה של מחקר זה תהיה לענות על שאלות המחקר הבאות:

1. כיצד הדיוק של צפיפות הקרנל הכפול מושפע מגורמים שונים?
2. עד כמה מדויק צפיפות הקרנל הכפולה באופן גלובאלי, על אזור המחקר בכללותו, וגם באופן נקודתי?

תיאוריה ושיטה: סוג המחקר שאנו מעוניינים בו כרוך באירועים של מחלה כרונית. המדד הנפוץ של תדירות בשימוש בספרות נקרא שיעור ההיארעות. עבור אוכלוסייה מסוימת, תקופת זמן ומערכת של תקריות, אנו יכולים לחשב את שיעור השכיחות הכולל על ידי לקיחת המספר הכולל של האירועים וחלוקת האוכלוסייה כולה. עם זאת, אנו עשויים להשוות את שיעור ההיארעות במקומות שונים בתוך שטח. לשם כך, אנו מחשבים את שיעור השכיחות לכל נקודה.

היחידה הבסיסית של המחקר שלנו היא הניסוי, סדרה של סימולציות מונטה קרלו שבוצעו עם אותה קונפיגורציה הראשונית מאותן המדידות. כל ניסוי מופעל עם קבוצה קבועה של פרמטרים. אנו מפעילים קבוצה של ניסויים, משתנה פרמטר אחד בכל פעם, כדי לבחון את ההשפעה של פרמטר זה על הדיוק של צפיפות הליבה כפולה.

כדי לחשב את צפיפות הקרנל הכפולה, יש לקבוע פרמטר הידוע כרוחב הפס, ודיוק צפיפות הליבה הכפולה תלוי בה במידה רבה. בכל ניסוי השתמשנו בשתי טכניקות, כלל אצבע של סילברמן ו- Least Squares Cross Validation - כדי לבחור את רוחב הפס. אנו משווים את התוצאות של שתי הטכניקות הללו.

על מנת לתאר את הדיוק של צפיפות הקרנל הכפולה כשיטה להערכת פונקצית הסיכון האמיתית, λ אנו משתמשים במספר מדידות דיוק. בפרט, עבור כל ניסוי אנו מודדים ממוצע מרובע שגיאה משולבת, ממוצע שגיאה מוחלטת משולבת, שגיאת supremum, הטית השיא, סחיפת השיא, הטית centroid וסחיפת centroid.

תוצאות: השגנו תוצאות המראות כי הגדלת המספר הצפוי של תקלות מקטינה את טעות האמידה של צפיפות הקרנל הכפולה, הן לטעויות גלובליות והן מבחינת טעויות נקודתיות. כמו כן, נצפה כי הגדלת התפשטות פונקציית הסיכון (שיעור) גם מפחיתה את הערכת השגיאה. כמו כן, ראינו כי הגדלת מספר האירועים הצפוי במקביל לגודל האוכלוסייה צמצמה את טעות האמידה, וכי העובדה שאוכלוסייה לא אחידה הגדילה את טעות האמידה. ברוב המקרים, התוצאות שלנו היו דומות הן ברוחבי פס שנבחרו בשיטת סילברמן והן בשיטת CV.

מסקנה: בדוגמאות שחקרנו, מצאנו כי סטטיסטית, צפיפות הקרנל הכפול יכולה לתת קירוב טוב של הסיכון האמיתי או פונקציית שיעור ההיארעות. זה כך במונחים של הדיוק הכללי של שער בכל נקודה נתונה, כמו גם עבור המיקום של נקודת השיא. התוצאות שלנו לא הראו הבדל גדול בין תוכניות סילברמן ובין תוכניות CV לאימות רוחב פס, למעט בקביעת מיקום השיא. שתי התכניות בחרו רוחב פס בין 5% ל -20% מגודל אזור המחקר. מאחר שהדיוק משתפר עם מספר התצפיות, יש להגביר את מסגרת הזמן של המחקר כאשר מספר התצפיות בשנה מסוימת קטן מ -50. במקרים שבהם צפיפות האוכלוסין משתנה במידה רבה על פני שטח מחקר, צפיפות הקרנל היא פחות מדויקת מאשר אוכלוסיות יותר אחידות.