אורט בראודה
המכללה האקדמית להנדסה
ORT Braude College

המחלקה למתמטיקה
Department of Mathematics

פרויקט מסכם לתואר בוגר במדעים (B.Sc)
במתמטיקה שימושית

שיטת הבדיקה הקבוצתית

מעיין צדוק

The Group Test Method

Maayan Tzadok

תוכן עניינים

1 הקדמה

בדיקה של קבוצת אנשים גדולה לאיתור "פגומים" היא תהליך יקר ומייגע.
זיהוי של אנשים "פגומים" באוכלוסיה גדולה טופל לראשונה ע"י דורפמן ויושם על פרויקט בקנה מידה גדול, שבו שירות לבריאות הציבור בארה"ב רצה לאתר את כל הגברים המועמדים לגיוס שנגועים בעגבת בשנת 1943.

רוברט דורפמן (1916 – 2002) היה פרופסור לכלכלה פוליטית באוניברסיטת הרווארד וסטטיסטיקאי. דורפמן תרם רבות לתחומי הכלכלה, בדיקות קבוצתיות ושיטות הצפנה.
מאמרו 'The Detection of Defective Members of Large Populations' [1] הוא אבן דרך בתחום הבדיקה הקבוצתית.

באופן כללי הבעיה היא כזו: נדרש לבצע בדיקת דם לצורך זיהוי מחלה מסויימת למספר גדול של אנשים - N.
שכיחות המחלה באוכלוסיה היא p, כלומר p היא ההסתברות שאדם מסוים נגוע במחלה.
לאחר שדגימות הדם נלקחות מהנבדקים, ניתן לבצע את בדיקת הדם בשתי דרכים:

1. כל אדם יכול להיבדק בנפרד, במקרה זה נדרשות N בדיקות.

2. מחלקים את הדגימות לקבוצות בגודל n: ניתן לערבב דגימות מ-n אנשים לדגימה אחת ולבדוק את התערובת.
   אם התוצאה שמתקבלת שלילית, אף אחד מ-n האנשים באותה קבוצה אינו נגוע, ואין צורך לבצע בדיקות נוספות.
   אם התוצאה חיובית, יש לפחות נגוע אחד בקבוצה ונדרש לבצע בדיקה אינדיבידואלית לכל אחד מ-n האנשים. סה"כ נבצע במקרה זה $n+1$ בדיקות לקבוצה.
   שיטה זו נקראת שיטת דורפמן. ברור באופן אינטואטיבי ששיטה זו תביא לחיסכון במספר הבדיקות שבוצעו, במיוחד כאשר p מאוד קטן:
   במקרה ש-p קטן, שכיחות המחלה קטנה ואז הסיכוי שבקבוצה בגודל n יהיה אדם נגוע גם קטן ולכן כאשר נבצע בדיקה לתערובת הדגימות של n

אנשים הסיכוי לקבל תוצאה שלילית גבוה יותר, ואז נבצע בדיקה אחת ל-n אנשים במקום n בדיקות, וקיבלנו חיסכון גדול בכמות הבדיקות שנבצע.

השאלות שעולות לגבי שיטה זו הן:

- מה היקף החיסכון בכמות הבדיקות שנבצע?
- מה הגודל היעיל ביותר לקבוצה?
- איך ניתן לשפר את שיטה זו על מנת להביא לחיסכון גדול יותר במספר הבדיקות?
- האם קיימים חסרונות לשיטה? ואם כן, האם ניתן למנוע אותם?

בשאלות אלה נעסוק בפרוייקט.

בפרק הראשון נתאר את השיטה, ננסח אותה ונקבל ביטוי המתאר את הקשר בין גודל הקבוצה שנבחר לתוחלת מספר הבדיקות שנבצע, נחקור ביטוי זה וננסה למקסם את החיסכון במספר הבדיקות.
בנוסף נציג דרך משופרת שמגדילה את החיסכון בכמות הבדיקות שנבצע.
בפרק השני נציג שיטה נוספת לביצוע בדיקות קבוצתיות ונבצע השוואה בין השיטות.

2 שיטת דורפמן

2.1 תאור השיטה

נחלק אוכלוסיה בגודל N לקבוצות בגודל n. לכל קבוצה נערבב את דגימות הדם של כל חברי הקבוצה לתערובת אחת. נבצע בדיקה על התערובת.
אם התוצאה שהתקבלה מבדיקת התערובת חיובית, נבצע בדיקה פרטנית לדגימה של כל אחד מחברי הקבוצה.
אם התוצאה שהתקבלה מבדיקת התערובת שלילית, נסיק שאף אחד מחברי הקבוצה אינו נגוע ואין צורך בבדיקות נוספות.
שיטה זו הוצגה לראשונה במאמר [1] ברשימת המקורות.

2.2 ניתוח השיטה

נגדיר:
p - שכיחות המחלה באוכלוסיה, כך שבדגימה אקראית הסיכוי למציאת חולה היא p.
$q = 1 - p$ - ההסתברות שאדם שנבחר באקראי לא יהיה נגוע.
נבחין ש- $(1-p)^n$ היא ההסתברות שבקבוצה של n אנשים שנבחרה באקראי, אף אחד מהם אינו נגוע, לכן:
$p' = 1 - (1-p)^n = 1 - q^n$ - ההסתברות שבקבוצה של n אנשים שנבחרה באקראי לפחות אדם אחד נגוע.

מספר הקבוצות בגודל n באוכלוסיה בגודל N הוא: $\dfrac{N}{n}$
לכן המספר הצפוי של קבוצות נגועות בגודל n באוכלוסיה בגודל N עם שיעור שכיחות p הוא: $\dfrac{p' \cdot N}{n}$

נסמן T- תוחלת מספר הבדיקות שיש לעשות בשיטת הבדיקה הקבוצתית של דורפמן. אז:

$$T = \frac{N}{n} + n \cdot \frac{N}{n} \cdot p'$$

כאשר:

$\frac{N}{n}$ מציין את מספר הבדיקות הקבוצתיות שיעשו.

$n \cdot \frac{N}{n} \cdot p'$ מציין את מספר הקבוצות שנמצאו נגועות.

$\frac{N}{n} \cdot p'$ מציין את מספר הבדיקות האינדיבידואלית שנבצע עבור כל הקבוצות שנמצאו נגועות.

נגדיר C - תוחלת מספר הבדיקות לאדם

$$C = \frac{T}{N} = \frac{1}{n} + p' = \frac{n+1}{n} - (1-p)^n$$

(במעבר האחרון הצבתי את הביטוי עבור p')

C מוגדר כיחס בין מספר הבדיקות בשיטת דורפמן לבין מספר הבדיקות בשיטה הרגילה (בדיקה אינדיבידואלית לכל אדם).

לכן $S = 1 - C$ יהיה היקף החיסכון שניתן להשיג ע"י שיטת דורפמן והוא תלוי בגודל הקבוצה ובשיעור השכיחות.

2.3 חקירת הפונקציה $C(n)$

$C(n)$ מבטאת את היחס בין מספר הבדיקות הנדרש ע"י שיטת דורפמן לבין מספר הבדיקות בשיטה הרגילה או במילים אחרות מספר הבדיקות הצפוי לאדם.
נחקור את הפונקציה הזו כדי לדעת מתי יעיל להשתמש בשיטת דורפמן ומתי עדיף לבצע בדיקות אינדיבידואליות, וכן כדי למצוא מהו הגודל הקבוצה-n האופטימלי עבור ערכים שונים של p בהנחה ש-p ידוע ונתון.

האינטואיציה היא שככל שהמחלה שכיחה יותר, כלומר ערכו של p גדול יותר גודל הקבוצה האופטימלי יהיה גדול קטן יותר. נסביר את האינטואיציה:
אם המחלה מאוד נפוצה וניקח קבוצות גדולות אז הסיכוי שבכל אחת מהקבוצות יהיה נגוע גדול, כי המחלה שכיחה, ואז נקבל תוצאה חיובית ברוב הבדיקות הקבוצתיות שנבצע ונצטרך לבצע בדיקה פרטנית לכל אחת מקבוצות אלה, ובצורה כזו אנו עלולים להגדיל את כמות הבדיקות או לקבל חיסכון מאוד קטן בכמות הבדיקות.
מצד שני, אם גודל הקבוצה יהיה קטן יותר הסיכוי שיהיה נגוע בקבוצה קטן, וכך נקבל עבור יותר קבוצות תוצאה שלילית ונוכל לחסוך יותר בדיקות פרטניות וכתוצאה מכך לקבל חיסכון גדול יותר במספר הבדיקות הכולל שנבצע.

נתבונן בפונקציה:

$$C(n) = \frac{n+1}{n} - (1-p)^n$$

כפי שראינו בסעיף 2.2, פונקציה זו מבטאת את היחס בין מספר הבדיקות הנדרש ע"י שיטת דורפמן לבין מספר הבדיקות בשיטה הרגילה, לכן כדי ששיטת הבדיקה הקבוצתית תהיה יעילה נרצה לקבל את הערך המינימלי עבור $C(n)$ ונרצה שהוא יהיה קטן מ-1 על מנת לקבל חיסכון בכמות הבדיקות שנצטרך לבצע.
יש לציין שלמרות ש-n הוא מספר חיובי ושלם המציין גודל קבוצה, את הניתוח עבור $C(n)$ נבצע עבור n רציף ונקח ערך שלם של התוצאה.

נרצה לחקור את הפונקציה $C(n)$ כדי לדעת עבור איזה ערך של n נקבל חיסכון מקסימלי במספר הבדיקות. תחילה נבחן את התנהגות הפונקציה כאשר n שואף ל-0 וכאשר n שואף ל-n.

לשם כך נחשב את הגבולות הבאים: $\lim_{n\to\infty} C(n)$, $\lim_{n\to 0} C(n)$

$$\lim_{n\to 0} C(n) = \lim_{n\to 0} \underbrace{\frac{n+1}{n}}_{\downarrow \atop \infty} - \underbrace{(1-p)^n}_{\downarrow \atop 1} = \infty$$

$$\lim_{n\to\infty} C(n) = \lim_{n\to\infty} \underbrace{\frac{n+1}{n}}_{\downarrow \atop 1} - \underbrace{(\overbrace{1-p}^{<1})^n}_{\downarrow \atop 0} = 1$$

מתוצאות אלה לא ניתן להסיק מסקנה גורפת לגבי יעילות שיטת דורפמן, אבל מה שניתן להסיק הוא שהמצב שבו n שואף ל-0 לא טוב, כיוון שאנו מקבלים שערך הפוקנציה $C(n)$ שואף ל-∞ ואנו רוצים ערך כמה שיותר קטן, אך ממילא מצב כזה לא יתכן כי n הוא מספר שלם.

כאשר n שואף ל-∞ מקבלים שהפונקציה שואפת ל-1, אך לא ניתן לדעת האם הפונקציה שואפת ל-1 מלמטה או מלמעלה. לכן נשאלת השאלה האם מתישהו הפונקציה מקבלת ערך קטן מ-1 או שתמיד היא נמצא מעלה הישר $C=1$? ואם כן, עבור אילו ערכים של n זה קורה ועבור איזה ערך של n נקבל את $C(n)$ המינימלי? על מנת לענות על שאלות אלה נצרך להתבונן בנגזרת של $C(n)$ ולחקור את התנהגותה.

נבדוק האם יש לפונקציה זו מינימום בקטע $(0,\infty)$. לשם כך נגזור את $C(n)$:

$$C'(n) = -\frac{1}{n^2} - \ln(1-p)\cdot(1-p)^n$$

כעת נרצה לדעת האם קיים n שעבורו הנגזרת מתאפסת. אם נשווה את הביטוי המבטא את הנגזרת לאפס נקבל משוואה שלא ניתן לפתור אנליטית.

לכן נצטרך לחקור את התנהגות הפונקציה $C'(n)$. לשם כך נחשב תחילה את הגבול של הנגזרת כאשר $n\to 0$

$$\lim_{n\to 0} C'(n) = \lim_{n\to 0} -\frac{1}{n^2} - \ln(1-p)\cdot(1-p)^n = -\lim_{n\to 0} \underbrace{\frac{1}{n^2}}_{\downarrow \atop \infty} - \lim_{n\to 0} \ln(1-p)\cdot\underbrace{(\underbrace{1-p}_{\vee \atop 0})^n}_{0<p<1 \atop \downarrow \atop 1} = -\infty$$

נראה שעבור $n \to \infty$, $C'(n) \to 0$ מהכיוון השלילי. כלומר, נראה שהנגזרת $C'(n)$ שלילית החל מ-n מסוים:

$$\lim_{n\to\infty} C'(n) = \lim_{n\to\infty} -\underbrace{\frac{1}{n^2}}_{\substack{\downarrow\\0}} - \ln(1-p)\cdot \underbrace{(\underbrace{1-p}_{0<1-p<1})^n}_{\substack{\downarrow\\0}} = 0$$

כדי לחקור את הסימן של $C'(n)$ עבור n גדול, נכפיל את הביטוי שבתוך הגבול ב-n^2 ונחשב את הגבול שמתקבל:

$$\lim_{n\to\infty} n^2\cdot C'(n = \lim_{n\to\infty} -1-n^2\cdot\ln(1-p)\cdot(1-p)^n = -1-\underbrace{\lim_{n\to\infty} n^2\cdot\ln(1-p)\cdot(1-p)^n}_{(*)=0} = -1$$

$(*)$

$$\lim\nolimits_{n\to\infty} \underbrace{n^2}_{\substack{\downarrow\\\infty}} \cdot \ln(1-p)\cdot \underbrace{(1-p)^n}_{\substack{\downarrow\\0}} \quad [\infty\cdot 0] =$$

$$\lim\nolimits_{n\to\infty} \frac{n^2\cdot\ln(1-p)}{(1-p)^{-n}} \quad \left[\frac{\infty}{\infty}\right] \underset{l'Hopital}{=} \lim_{n\to\infty} \frac{2n\cdot\ln(1-p)}{-\ln(1-p)\cdot(1-p)^{-n}} \left[\frac{\infty}{\infty}\right] \underset{l'Hopital}{=}$$

$$\lim\nolimits_{n\to\infty} \frac{2\ln(1-p)}{(\ln(1-p))^2\cdot(1-p)^{-n}} = \frac{\overbrace{2\ln(1-p)}^{<0}}{\underbrace{\ln(1-p))^2}_{>0}} \underbrace{\lim_{n\to\infty} (\underbrace{1-p}_{0<1-p<1})^n}_{\substack{\downarrow\\0}} = 0$$

אז קיבלנו ש:

$$\lim_{n\to\infty} n^2\cdot C'(n) = -1$$

הביטוי $-\frac{1}{n^2}$ שלילי, לכן $C'(n)$ שואף לאפס מהכיוון השלילי כאשר $n \to \infty$.
קיבלנו ש:

$\lim_{n\to 0} C'(n) = -\infty$

$\lim_{n\to\infty} C'(n) = 0$

עבור n גדול מתקיים $C'(n) < 0$

כעת, כדי לדעת האם $C'(n)$ מתאפסת נבדוק האם קיימים ערכים של $n > 0$ שעבורם $C'(n) \overset{?}{>} 0$. אם קיימים ערכים כאלה, אז לפי משפט ערך הביניים $C'(n)$ מתאפסת בנקודה כלשהי.

$$\begin{aligned}
C'(n) &= -\underbrace{\frac{1}{n^2}}_{>0\ \forall n} - \underbrace{\ln(1-p)}_{\le 0} \cdot \underbrace{(1-p)^n}_{>0\ \forall n} \overset{?}{>} 0 \\
&\iff -\frac{1}{n^2} > \ln(1-p)\cdot(1-p)^n \\
&\iff -1 > n^2\ln(1-p)\cdot(1-p)^n
\end{aligned}$$

לפי התבוננות בשרטוטים במחשב בגרף של $C'(n)$ זה לא מתקיים תמיד. (נראה שרטוטים במהשך)

כלומר, עבור ערכים מסוימים של p נקבל שאף ערך של n לא מקיים את האי שיוויון הזה ומפה נסיק שהנגזרת לא מתאפסת (נבחן מקרה זה בהמשך).

בנוסף כפי שנראה כעת, קיים ערך מסויים של p שהוא הערך הקריטי, כלומר שעבור כל ערך של p שגדול ממנו לא יהיה אף ערך של n שמקיים את אי השיוויון הנ"ל. נבדוק עבור אילו ערכי p קיים n כך ש: $C'(n) > 0$.

$$C'(n) = -\frac{1}{n^2} - \ln(1-p)\cdot(1-p)^n > 0$$

$$\iff n^2\cdot(1-p)^n > \frac{1}{\ln\left(\frac{1}{1-p}\right)} \tag{$\star$}$$

נסמן: $h(n) = n^2\cdot(1-p)^n$

באגף ימין של אי השיוויון יש ביטוי שתלוי רק ב-p ובאגף שמאל הפונקציה $h(n)$.

אגף שמאל צריך להיות גדול מאגף ימין עבור איזשהו ערך של n אז נגדיל את אגף שמאל עד למקסימום האפשרי. כלומר, נקח את הערך של n שעבורו הפונקציה $h(n)$ מקבלת את הערך המקסימלי שלה ונציב אותו באי שיוויון ואז נקבל אי שיוויון עם המשתנה p בלבד שיתקיים אם ורק אם קיים ערך של n שעבורו $C'(n) > 0$.

נחפש מקסימום לפונקציה $h(n)$.

$$\begin{aligned}
&h'(n) = 2n \cdot (1-p)^n + n^2 \cdot \ln(1-p) \cdot (1-p)^n = 0 \quad / : (1-p)^n \\
&2n + n^2 \cdot \ln(1-p) = 0 \quad / : n \\
&2 + n \cdot \ln(1-p) = 0 \\
&n = \frac{-2}{\ln(1-p)}
\end{aligned}$$

נראה שהערך שהתקבל הוא מקסימום גלובלי:
נבדוק את סימן הנגזרת מימין ומשמאל לנקודה שבה הנגזרת מתאפסת.
פונקציה הנגזרת:

$$h'(n) = (1-p)^n \cdot \left(2n + n^2 \cdot \ln(1-p)\right)$$

נקח ערך של n שקטן מהערך שבו הנגזרת מתאפסת ונבדוק שהנגזרת חיובית עבור ערך זה:

$$\frac{-1}{\ln(1-p)} < \frac{-2}{\ln(1-p)}$$

נחשב:

$$\begin{aligned}
h'\left(\frac{-1}{\ln(1-p)}\right) &= (1-p)^{\frac{-1}{\ln(1-p)}} \left(2 \cdot \frac{-1}{\ln(1-p)} + \left(\frac{-1}{\ln(1-p)}\right)^2 \cdot \ln(1-p)\right) \\
&= (1-p)^{\frac{-1}{\ln(1-p)}} \left(\frac{-2}{\ln(1-p)} + \frac{1}{\ln(1-p)}\right) \\
&= \underbrace{(1-p)^{\frac{-1}{\ln(1-p)}}}_{>0} \cdot \underbrace{\frac{-1}{\ln(1-p)}}_{>0} > 0
\end{aligned}$$

בנוסף, נקח ערך של n שגדול מהערך שבו הנגזרת מתאפסת ונבדוק שהנגזרת שלילית עבור ערך זה:

$$\frac{-3}{\ln(1-p)} > \frac{-2}{\ln(1-p)}$$

נחשב:

$$\begin{aligned} h'\left(\frac{-3}{\ln(1-p)}\right) &= (1-p)^{\frac{-3}{\ln(1-p)}}\left(2\cdot\frac{-3}{\ln(1-p)}+\left(\frac{-3}{\ln(1-p)}\right)^2\cdot\ln(1-p)\right) \\ &= (1-p)^{\frac{-3}{\ln(1-p)}}\left(\frac{-6}{\ln(1-p)}+\frac{9}{\ln(1-p)}\right) \\ &= \underbrace{(1-p)^{\frac{-3}{\ln(1-p)}}}_{>0}\cdot\underbrace{\frac{3}{\ln(1-p)}}_{<0}<0 \end{aligned}$$

כעת, אנו יודעים שהפונקציה עולה ואח"כ יורדת, ומזה נובע שהפונקציה מקבלת מקסימום גלובלי בנקודה בה מתאפסת הנגזרת.

נציב את ה-n שקיבלנו באי-שוויון $\star$

$$\frac{4}{(\ln(1-p))^2}\cdot(1-p)^{\frac{-2}{\ln(1-p)}} > \frac{1}{\ln\left(\frac{1}{1-p}\right)}$$

$$\frac{4}{(\ln(1-p))^2}\cdot e^{\frac{-2}{\ln(1-p)}\cdot\ln(1-p)} > \frac{1}{-\ln(1-p)}$$

$$\frac{4}{(\ln(1-p))^2}\cdot e^{-2} > \frac{1}{-\ln(1-p)} \qquad /\cdot(\ln(1-p)<0)^2$$

$$\begin{aligned} &4\cdot e^{-2} > -\ln(1-p) \\ &-4\cdot e^{-2} < \ln(1-p) \\ &e^{-4\cdot e^{-2}} < 1-p \\ &p < 1-e^{-4\cdot e^{-2}} \implies p<p^*\approx 0.418 \end{aligned}$$

קיבלנו ערך קריטי p^* של p, כך שכאשר $p<p^*$ קיים n כך ש: $C'(n)>0$, כלומר לפי משפט ערך הביניים של קושי הנגזרת מתאפסת באיזושהי נקודה. ז"א יש נקודת שמועמדות להיות מינימום/מקסימום. (מה שמעניין אותנו זה מינימום כי נרצה שמספר הבדיקות יהיה מינמלי).

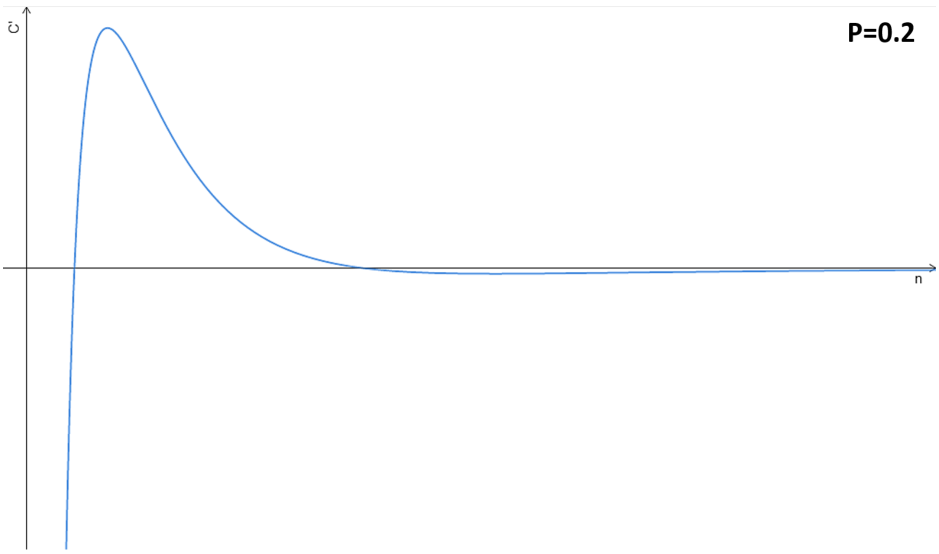
<u>נסכם:</u>

קיבלנו שעבור כל p מתקיים:

$$\lim_{n\to 0} C'(n) = -\infty$$
$$\lim_{n\to \infty} C'(n) = 0^-$$

עבור $p < p^*$ יש ערכים של n שעבורם מתקיים $C'(n) > 0$

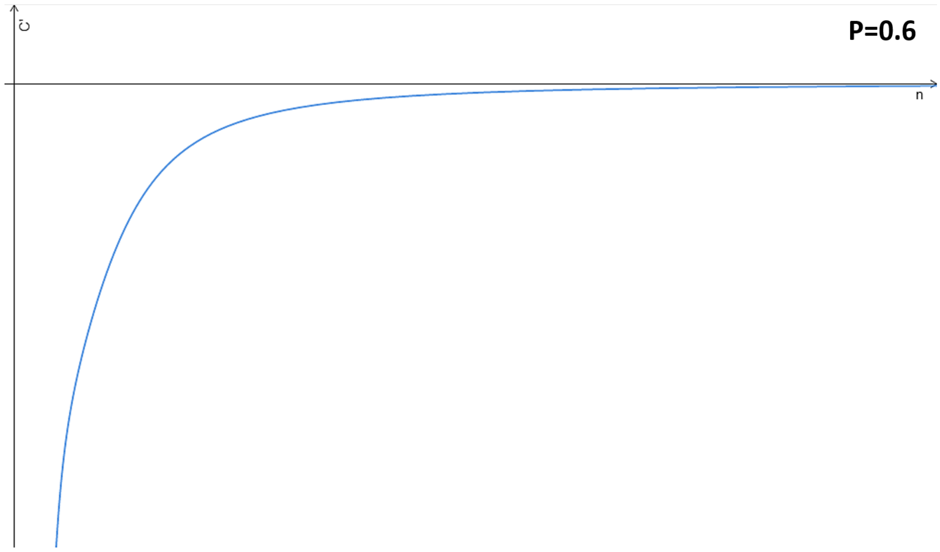
לכן הגרף של הנגזרת $C'(n)$ מהצורה:



עבור n-ים מאוד קטנים הנגזרת שאופת ל-$-\infty$, עבור n-ים מאוד גדולים הנגזרת שואפת ל-0 מהכיוון השלילי ובאיזשהו מקום הנגזרת חיובית ולכן חייבת לחצות את ציר האפס על פי משפט ערך הביניים, כלומר באיזשהו מקום מתקיים $C'(n) = 0$.

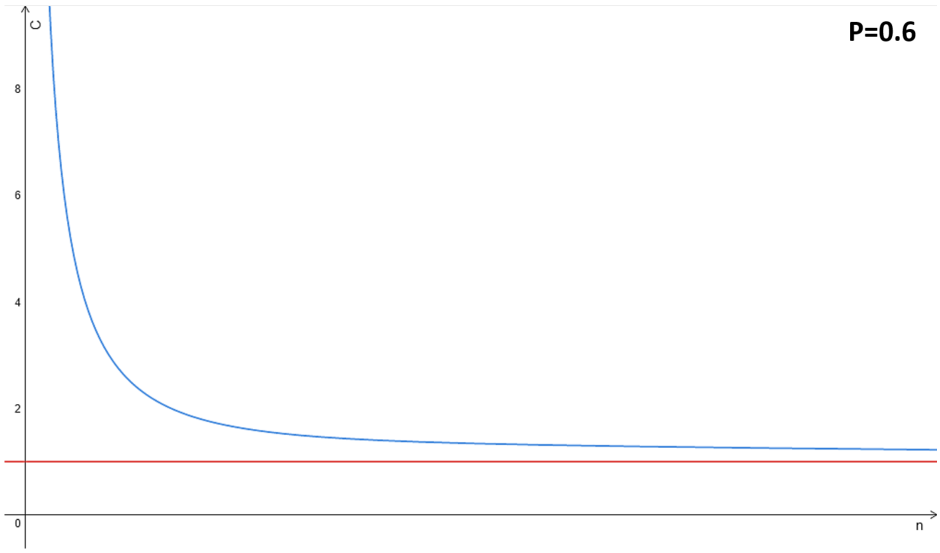
עבור $p > p^*$ אין ערכים של n שעבורם מתקיים $C'(n) > 0$

כלומר גרף הנגזרת תמיד מתחת לציר האפס ואף פעם לא חותך אותו ולכן אין נקודת קיצון וגרף הנגזרת $C'(n)$ מהצורה:



בעזרת המסקנות שקיבלנו לגבי הפונקציה $C'(n)$ נבדוק את צורת הפונקציה $C(n)$:

עבור $p > p^*$ $C'(n)$ לא מתאפסת, מכאן אין ל-$C(n)$ נקודות קיצון והגרף שלה יהיה פונקציה יורדת:



אם הנגזרת לא מתאפסת וידוע ש-$C'(n)$ שואפת ל-$-\infty$ כאשר $n \to 0$ אז $C'(n)$ קטנה מאפס עבור כל ערך של n, כלומר השיפוע תמיד שלילי ולכן הפונקציה $C(n)$

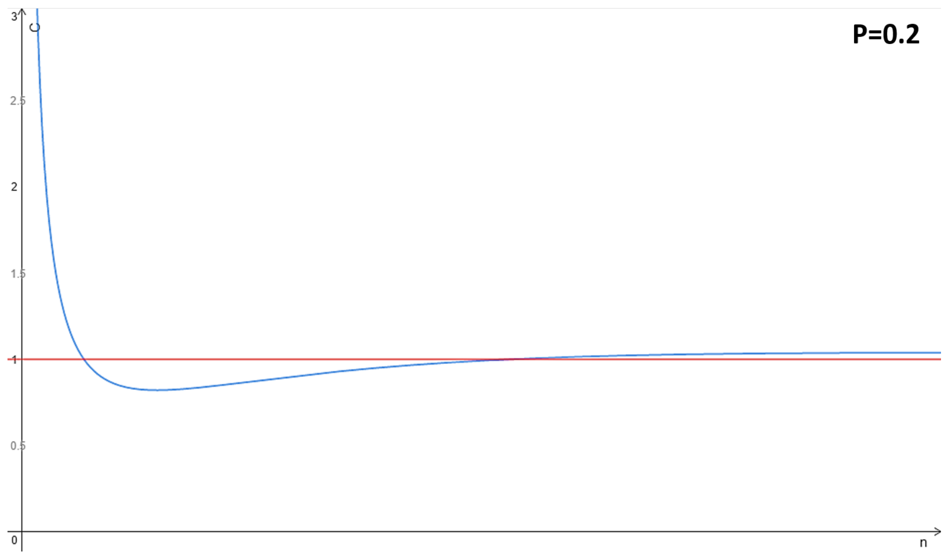
היא פונקציה מונוטונית יורדת בקטע $(0,\infty)$ וכיוון שקיבלנו

$$\lim_{n\to 0} C(n) = \infty$$
$$\lim_{n\to \infty} C(n) = 1$$

לא קיים ערך של n שעבורו $C(n) < 1$, כלומר לא משנה איזה גודל קבוצה נבחר, שיטת דורפמן לא יעילה ולא מביאה לחיסכון בכמות הבדיקות כי תמיד מתקיים $C(n) > 1$ ומכאן שהחיסכון $S = 1 - C$ יהיה תמיד שלילי.

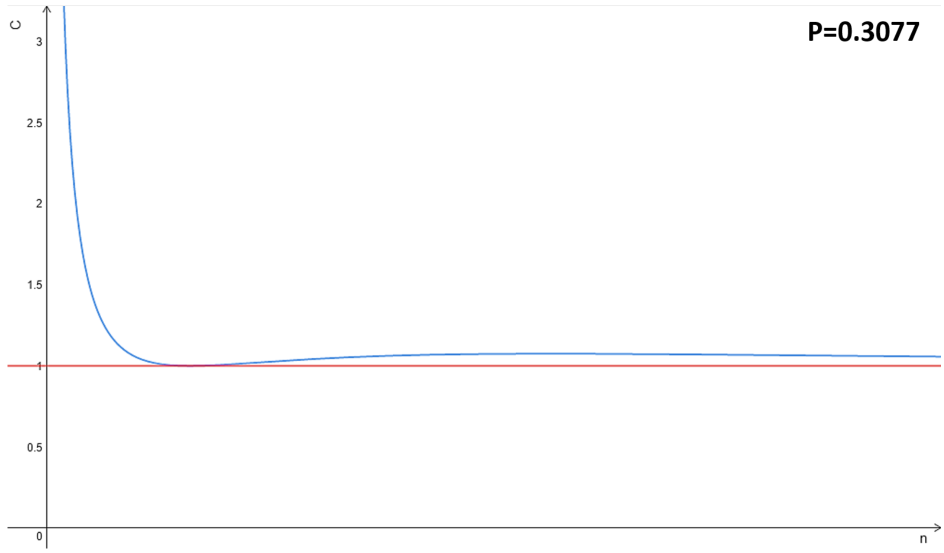
עבור $p < p^*$ הגרף של $C(n)$ יכול לקבל את אחת מבין שלושת הצורות הבאות:

I הגרף יורד מתחת ל-1 כלומר קיים מינימום ל-$C(n)$ שקטן מ-1.
במקרה זה שיטת הבדיקה הקבוצתית יעילה כיוון שקיימים n-ים שעבורם $C(n) < 1$, ואז נקבל שהחיסכון $S = 1 - C$ חיובי, כלומר חסכנו בכמות הבדיקות לעומת בדיקה אינדיבידואלית לכל אדם באוכלוסייה.

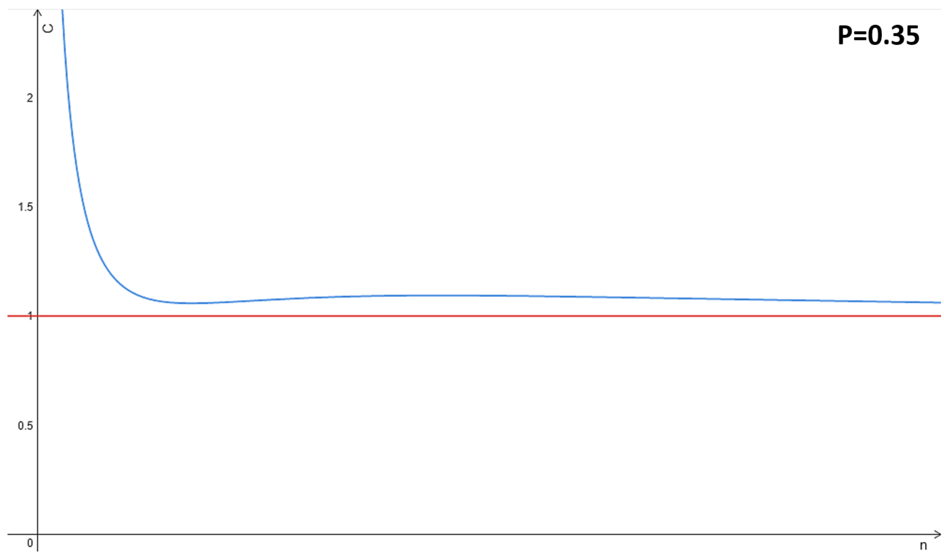


נציין שכדי שיהיה ערך של n שיתן חיסכון בכמות הבדיקות, חייב להיות אי-זשהו ערך שלם בקטע שבו הפונקציה יורדת מתחת לישר $C(n) = 1$, אחרת אין גודל קבוצה n שיתן חיסכון בכמות הבדיקות.

II הגרף כולו נמצא מעל הישר $C(n)=1$ ונוגע בו בנקודה אחת וזו גם נקודת המינימום. במקרה הזה שיטת הבדיקה הקבוצתית לא מועילה כי תמיד מתקיים $C(n)\geq 1$ מכאן שהחיסכון $S=1-C$ יהיה שלילי או שווה לאפס.



III הגרף כולו נמצא מעל הישר $C(n)=1$ ושואף אליו כאשר $n\to\infty$. בנוסף הפונקציה מקבלת את נקודות הקיצון שלה מעל הישר $C(n)=1$.
גם במקרה הזה שיטת הבדיקה הקבוצתית לא מועילה כי תמיד מתקיים $C(n)>1$ ולכן החיסכון יהיה תמיד שלילי.



נחפש את הערך הקריטי של p שעבורו המינימום שמתקבל שווה ל-1.

$$C'(n) = 0 \iff -\frac{1}{n^2} - \ln(1-p)\cdot(1-p)^n = 0 \tag{1}$$

$$C(n) = 1 \iff \frac{n+1}{n} - (1-p)^n = 1 \tag{2}$$

כלומר אנחנו מחפשים מתי אנחנו נמצאים במקרה II.

מ- (1) נקבל:

$$(1-p)^n = -\frac{1}{n^2\cdot\ln(1-p)}$$

נציב את $(1-p)^n$ ב-(2) ונקבל:

$$\frac{n+1}{n} + \frac{1}{n^2\cdot\ln(1-p)} = 1$$
$$n\cdot(n+1)\cdot\ln(1-p) + 1 = n^2\cdot\ln(1-p)$$
$$n^2\cdot\ln(1-p) + n\cdot\ln(1-p) + 1 = n^2\cdot\ln(1-p) \quad / : \ln(1-p)$$
$$n = -\frac{1}{\ln(1-p)}$$

קיבלנו שבנקודה הקריטית: $n^* = -\dfrac{1}{\ln(1-p)}$

נציב את n^* ב- (1):

$$-\frac{1}{\frac{1}{(\ln(1-p))^2}} - \ln(1-p)\cdot(1-p)^{-\frac{1}{\ln(1-p)}} = 0$$

$$-(\ln(1-p))^2 - \ln(1-p)\cdot(1-p)^{-\frac{1}{\ln(1-p)}} = 0 \quad / : -\ln(1-p)$$

$$\ln(1-p) + (1-p)^{-\frac{1}{\ln(1-p)}} = 0$$

$$(1-p)^{-\frac{1}{\ln(1-p)}} = -\ln(1-p)$$

$$\left(e^{\ln(1-p)}\right)^{-\frac{1}{\ln(1-p)}} = -\ln(1-p)$$

$$e^{-\frac{1}{\ln(1-p)}\cdot\ln(1-p)} = \ln(1-p)$$

$$e^{-1} = -\ln(1-p) \quad\implies\quad \ln(1-p) = -e^{-1}$$

$$1-p = e^{-\frac{1}{e}} \quad\implies\quad p^{**} = 1-e^{-\frac{1}{e}} \approx 0.30779$$

אז קיבלו ש:

גרף I מתקבל עבור: $p < p^{**}$

גרף II מתקבל עבור: $p = p^{**}$

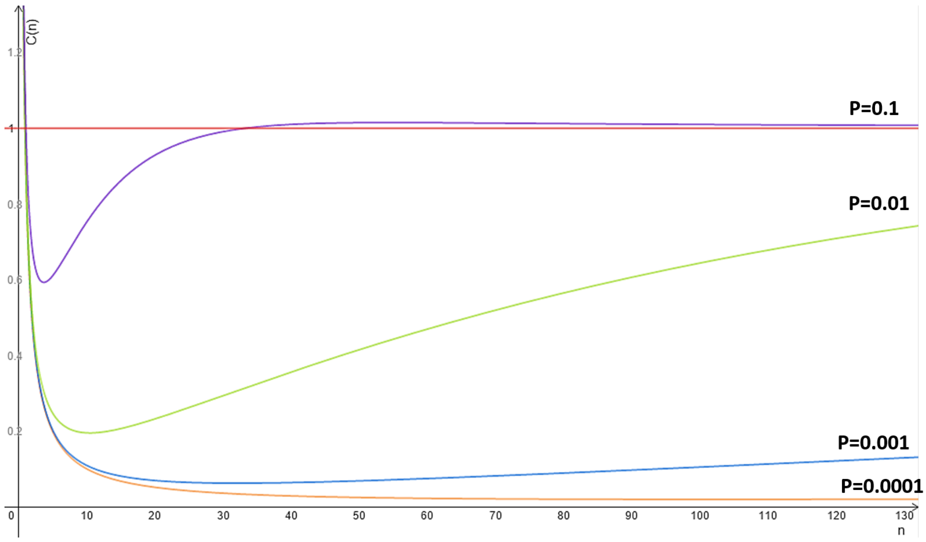
גרף III מתקבל עבור: $p^{**} < p < p^*$

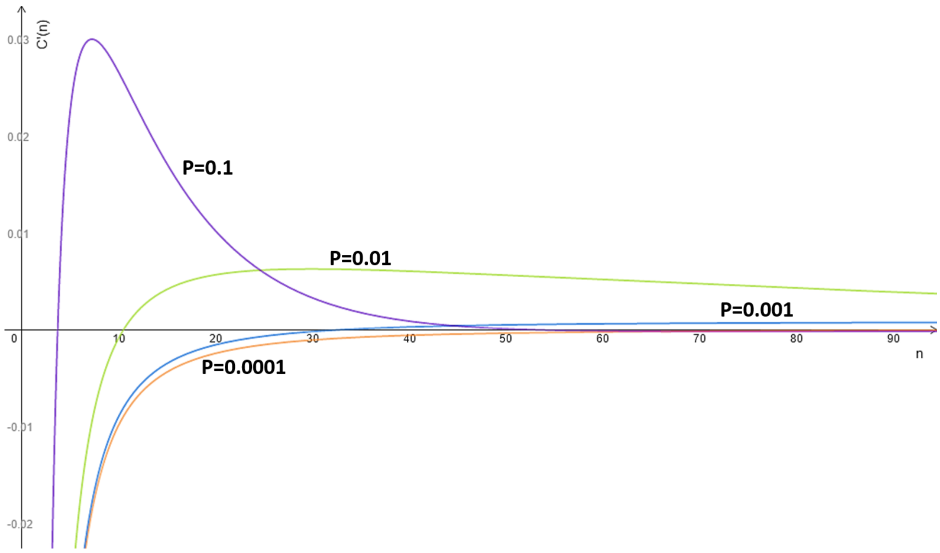
אנו רוצים להיות במקרה של גרף I כדי ששיטת הבדיקה הקבוצתית תהיה יעילה, כיוון שאם המינימום גדול מ-1 אז בבדיקה הקבוצתית נבצע יותר בדיקות מאשר בבדיקה האינדיבידואלית ואם הוא שווה ל-1 אז נשארנו עם אותה כמות בדיקות כמו בבדיקה האינדיבידואלית, כלומר שיטת הבדיקה הקבוצתית לא הועילה ולא הזיקה.

לסיכום, אם יותר מ 30.779% מהאוכלוסיה נגועים במחלה אז לא יעיל להשתמש בשיטת הבדיקה הקבוצתית, כמובן שבמציאות לרוב המחלות שכיחות נמוכה מזו

ולכן השיטה כן רלוונטית עבורן.

נשרטט את הגרף $C(n)$ וגרף הנגזרת שלו עבור ערכים שונים של p המקיימים $p < p^{**}$.





ניתן לראות שככל שערכו של p גדול יותר גודל הקבוצה המביא את ערכו של C למינימום קטן יותר. זאת אומרת, נדרשת קבוצה קטנה יותר כאשר שכיחות המחלה גדולה יותר, דבר המתיישב עם האינטואיציה.

ערכי n האופטימלים ושיעור החיסכון המתקבלים עבור ערכי p שמופיעים בגרפים מופיעים בטבלה הבאה:

p	n	C	S
0.0001	101	0.02	0.98
0.001	32	0.0628	0.9372
0.01	11	0.1956	0.8044
0.1	4	0.5939	0.4061

2.4 קירוב לגודל הקבוצה האופטימלי - n

בסעיף זה, ננסה למצוא קירוב לגודל הקבוצה האופטימלי בשיטת דורפמן, שמביא לחיסכון המקסימלי בכמות הבדיקות (או לחילופין ל-C מינימלי).
נרצה למצוא קירוב ל-n האופטימלי מכיוון שבהינתן p מסויים ורוצים למצוא מהו n האופטימלי צריך לפתור את המשוואה הבאה:

$$C'(n) = -\frac{1}{n^2} - \ln(1-p) \cdot (1-p)^n = 0$$

משוואה זו ניתן לפתור רק בשיטה נומרית ואנו רוצים לקבל ביטוי פשוט יותר לחישוב n.
בשיטה שנפתח למציאת קירוב ל-n האופטימלי הוצעה לראשונה במאמר של Finucan [2].

נזכיר ש: $p \cdot n$ הוא תוחלת מספר הנפשות הנגועות בכל קבוצה.

נמצא קירוב ל-n האופטימלי עבור p קטן.
נשתמש <u>בהנחה</u>: אין יותר מנגוע אחד בקבוצה.
הנחה זו סבירה כאשר p קטן כי אז $p \cdot n$ קטן.

מספר האנשים הנגועים מתוך כלל האוכלוסיה הוא $N \cdot p$ ולכן מספר הקבוצות הנגועות הוא $N \cdot p$ כי הנחנו שאין יותר מנגוע אחד בקבוצה.
נכפיל את $N \cdot p$ בגודל הקבוצה n ונקבל: $N \cdot p \cdot n$ - מספר הבדיקות הפרטניות בשלב הבא.
לכן תוחלת מספר הבדיקות הכולל שנבצע הוא:

$$\underbrace{\frac{N}{n}}_{\text{מספר הקבוצות}} + \underbrace{N \cdot p \cdot n}_{\text{מספר בדיקות פרטניות}}$$

לכן מספר הבדיקות לאדם:

$$C = \frac{\frac{N}{n} + N \cdot p \cdot n}{N} = \frac{1}{n} + p \cdot n$$

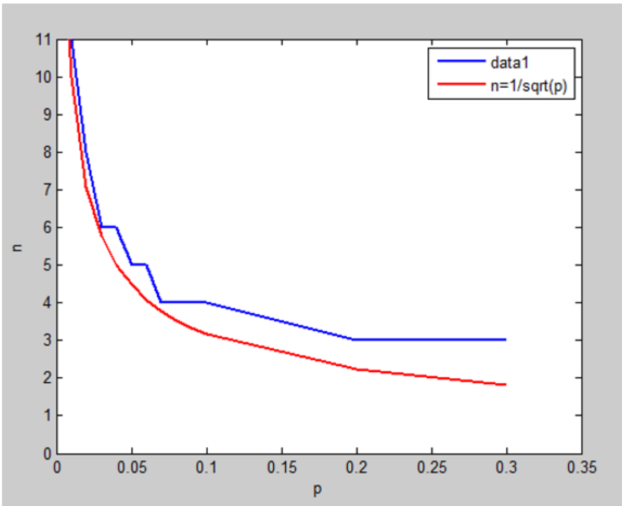
קיבלנו ביטוי פשוט יותר ל-$C(n)$ לעומת הביטוי המקורי, אם כי זה ביטוי מקורב.
נוכל לגזור את הביטוי הזה ולמצוא את הערך של n שמביא את C למינימום.

$$C' = -\frac{1}{n^2} + p = 0$$
$$\implies n = \frac{1}{\sqrt{p}}$$

זה הקירוב לגודל הקבוצה האופטימלי.
נציב את ה-n שקיבלנו ב-$C(n)$ ונקבל ביטוי ל-C כפונקציה של p:

$$C = 1 + \sqrt{p} - (1-p)^{\frac{1}{\sqrt{p}}}$$

נתבונן בגרף הבא ונראה עד כמה הקירוב שקיבלנו ל-n טוב:



בציר האופקי מופיעים ערכים שונים של p ובציר האנכי מופיעים הערכים המתאימים של n.
בגרף הכחול n חושב ע"י פתרון של המשוואה:

$$C'(n) = -\frac{1}{n^2} - \ln(1-p) \cdot (1-p)^n = 0$$

בגרף האדום n חושב ע"י:

$$n = \frac{1}{\sqrt{p}}$$

ניתן לראות את הפערים בין הערכים האופטימלים של n גם בטבלה הבאה:

p	n	~n
0.0001	101	100
0.0002	71	70.7107
0.0003	58	57.735
0.0004	51	50
0.0005	45	44.7214
0.0006	41	40.8248
0.0007	38	37.7964
0.0008	36	35.3553
0.0009	34	33.3333
0.001	32	31.6228
0.002	23	22.3607
0.003	19	18.2574
0.004	16	15.8114
0.005	15	14.1421
0.006	13	12.9099
0.007	12	11.9523
0.008	12	11.1803
0.009	11	10.5409
0.01	11	10
0.02	8	7.0711
0.03	6	5.7735
0.04	6	5
0.05	5	4.4721
0.06	5	4.0825
0.07	4	3.7796
0.08	4	3.5355
0.09	4	3.3333
0.1	4	3.1623
0.2	3	2.2361
0.3	3	1.8257

2.5 חקירת התפלגות מספר הבדיקות

לאחר שמצאנו את גודל הקבוצה האופטימלי, נוכל לחשב את תוחלת מספר הבדיקות שנדרש לבצע בבדיקה הקבוצתית כפי שהוצג בפרקים קודמים.
כעת נרצה לדעת מהי סטיית התקן מממוצע מספר הבדיקות שקיבלנו.

מספר הבדיקות שיש לבצע היא אקראית כיוון שמספר הבדיקות תלוי בהסתברות שבקבוצה מסוימת יהיה אדם נגוע. אם נמצא אדם נגוע מתבצעת בדיקה פרטנית לכל חברי הקבוצה. אם לא נמצא נגוע, מספיקה רק הבדיקה הקבוצתית.

נגדיר את המ"מ X- מציין את מספר הבדיקות בשיטת הבדיקה הקבוצתית של דורפמן.
נזכיר ש-N מציין את מספר הנבדקים ו-p מציין את הסיכוי שהאדם הנבחר חולה.
למען הפשטות נניח ש- $\frac{N}{n}$ שלם.
נרצה למצוא את ההתפלגות של X, כלומר לחשב: $P(X=k)=?$
נשים לב ש:
הערך המינמלי ש-k יכול לקבל הוא 0.
הערך המקסימלי ש-k יכול לקבל הוא $N+\frac{N}{n}$ - זהו המקרה שבו עבור כל קבוצה נקבל תוצאה חיובית, כלומר שיש לפחות חולה אחד בקבוצה, ואז נצטרך לבדוק אינדיבידואלית כל אחד מהקבוצה. אז סך הכל יצא שגם ביצענו בדיקה עבור כל אחת מהקבוצות וגם בדקנו אינדיבידואלית כל אחד מ- N הנבדקים.

נתבונן בכמה מקרים:

I ההסתברות לכך ש: $X<\frac{N}{n}$ היא 0 כיוון ש- $\frac{N}{n}$ זה מספר הקבוצות, ולא יתכן שנבצע פחות בדיקות ממספר הקבוצות.

II ההסתברות לכך ש: $X=\frac{N}{n}$ שווה להסתברות שעבור כל קבוצה נקבל תוצאה שלילית (כלומר שאין בה חולים) וזה שווה להסתברות שכל N הנבדקים לא נגועים כלומר ההסתברות היא: $(1-p)^N$

III נחשב את ההסתברות לכך ש: $X=k$ כאשר $k>\frac{N}{n}$.
מספר הבדיקות הקבוצתיות שמבצעים בשיטת דורפמן הוא: $\frac{N}{n}$

נגדיר:

X - מ"מ המציין את מספר הבדיקות בשיטת הבדיקה הקבוצתית.

Y - מ"מ המציין את מספר הבדיקות הקבוצתית שיצאו חיוביות.

$$Y \sim \mathrm{B}\left(\frac{N}{n}, p'\right)$$

כלומר Y מתפלג בינומית כאשר:

p' היא ההסתברות ל"הצלחה" בניסוי ברנולי בודד, כלומר שהקבוצה תהיה נגועה.

$\frac{N}{n}$ מספר ניסויי ברנולי שהתבצעו, כלומר מספר הקבוצות שנבדקו.

Z - מ"מ המצין את מספר הבדיקות שנבצע בשלב השני.

$$Z = n \cdot Y$$

אז:

$$X = \frac{N}{n} + Z = \frac{N}{n} + n \cdot Y$$

$$P(X=k) = P\left(\frac{N}{n} + n \cdot Y = k\right) = P\left(n \cdot Y = k - \frac{N}{n}\right) = P\left(Y = \frac{k - \frac{N}{n}}{n}\right)$$

לסיכום, קיבלנו שההתפלגות של X היא:

$$P(X=k) = \begin{cases} 0 & k < \frac{N}{n} \\ P\left(Y = \left[\frac{k - \frac{N}{n}}{n}\right]\right) & k \geq \frac{N}{n} \end{cases}$$

נוכל לחשב גם את התוחלת והשונות של המ"מ X:

$$E(X) = E\left(\tfrac{N}{n} + n \cdot Y\right) = \tfrac{N}{n} + n \cdot E(Y) = \tfrac{N}{n} + n \cdot \tfrac{N}{n} \cdot p' = \tfrac{N}{n} + N \cdot p'$$

$$= \tfrac{N}{n} + N \cdot (1 - (1-p)^n)$$

אותה תוצאה שכבר התקבלה בסעיף 2.2.

$$V(X) = V\left(\tfrac{N}{n} + n \cdot Y\right) = V\left(\tfrac{N}{n}\right) + n^2 \cdot V(Y) = n^2 \cdot \tfrac{N}{n} \cdot p'(1-p')$$
$$= n \cdot N \cdot (1 - (1-p)^n)\,(1-p)^n$$

דוגמה מספרית:

עבור אוכלוסיה בגודל $N = 100,000$ ומחלה עם שיעור שכיחות $p = \dfrac{1}{1000}$

נקבל שגודל הקבוצה האופטימלי הוא: $n = \dfrac{1}{\sqrt{0.001}} \approx 32$

נחשב את התוחלת וסטיית התקן:

$$E(X) = \frac{100,000}{32} + 100,000 \cdot (1 - (1 - 0.001)^{32}) \approx 6276$$

$$\sigma = \sqrt{V(X)} = \sqrt{32 \cdot 100,000 \cdot (1 - (1 - 0.001)^{32})\,(1 - 0.001)^{32}} \approx 312$$

נוכל לקרב את המ"מ הבינומי ע"י התפלגות נורמלית כיוון שבאוכלוסיות גדולות נקבל ש- $\dfrac{N}{n}$ גדול מספיק, נקבל:

$$X \sim \mathrm{N}\left(\underbrace{\frac{N}{n} \cdot p'}_{\mu},\ \underbrace{\frac{N}{n} \cdot p'(1 - p')}_{\sigma^2}\right)$$

לכן ב-95% מהמקרים נקבל שמספר הבדיקות שנבצע יהיה בטווח:

$$[\mu - 1.96\sigma, \mu + 1.96\sigma]$$

במקרה שלנו הטווח שמתקבל הוא:

$$[6276 - 1.96 \cdot 312, 6276 + 1.96 \cdot 312] = [5664.48, 6887.52]$$

2.6 שיפור שיטת דורפמן

ניתן להקטין את הערך של C, כלומר להקטין את מספר הבדיקות שנבצע ע"י שיפור שנוסיף לשיטת דורפמן.

רעיון זה תואר במאמר של Finucan [2] .

בשלב הראשון: מבצעים חלוקה של האוכלוסיה לקבוצות בגודל n_1 ומבצעים בדיקה לתערובת הדגימות של כל קבוצה. כעת, לפי דורפמן מבצעים בדיקה פרטנית לכל אחת מהדגימות בקבוצות שנמצאו נגועות. השיפור שנוסיף בא לידי ביטוי לאחר שהנגוע הראשון נמצא בקבוצה. ברגע שהוא נמצא משעים את שאר הבדיקות הפרטניות של הקבוצה, שמים אותן בצד וממתינים עם הבדיקה שלהן. כך נעשה בכל קבוצה, נבצע בדיקות פרטניות עד שנגיע לנגוע הראשון ואת שאר הבדיקות שנותרו לבצע נשים בצד.

בשלב השני: נקח את הדגימות שצברנו מכל הקבוצות ונבצע להן בדיקה קבוצתית לפי שיטת דורפמן עם גודל קבוצה חדש n_2 שנבחר בצורה אופטימלית.

גם פה נשתמש לצורך הניתוח ב הנחה מסעיף 2.4 האומרת: יש נגוע אחד לכל היותר בכל קבוצה, כלומר ערכו של p קטן.

הסיבה לשימוש בהנחה ואינטואיציה לשימוש בשיטה זו:

במידה וקיבלנו תוצאה חיובית עבור קבוצה מסוימת אנו נאלצים לבצע בדיקה פרטנית לכל אדם בקבוצה, אם אנו מניחים שערכו של p קטן אז כמות האנשים הנגועים בכל קבוצה גם תהיה קטנה. נניח שיש נגוע אחד בקבוצה, אז ברגע שמצאנו אותו על פי ההנחה כל הבדיקות הפרטניות האחרות שנותר לבצע אמורות לצאת שליליות, אז אם נצבור את בדיקות אלה לאחר שמצאנו את הנגוע הראשון בכל קבוצה ונפעיל את שיטת דורפמן על כל הבדיקות שצברנו הסיכוי שנקבל תוצאה חיובית עבור הקבוצות החדשות קטן ובצורה כזו נחסוך חלק גדול מהבדיקות הפרטניות שנצטרך לבצע.

נחשב את C שמתקבל מהתהליך המתואר לעיל, נסמנו ב- $C_*(n_1,n_2)$.
כאמור לעיל אם קיבלנו עבור קבוצה תוצאה חיובית נבצע בדיקות פרטניות עד אשר נמצא את הנגוע הראשון ואת שאר הבדיקות נצבור, מכאן שמספר הבדיקות הפרטניות שנבצע בכל קבוצה שנמצאה נגועה תלוי במיקום הנגוע בקבוצה. בנוסף גם מספר הבדיקות שנבצע בשלב השני תלוי באותו פרמטר.
לכן נרצה למצוא ביטוי המתאר כמה בדיקות פרטניות נצטרך לבצע בממוצע עבור קבוצה שנמצאה נגועה.
לשם כך נגדיר: X - מ"מ המציין את מספר הבדיקות שנבצע בקבוצה מסויימת שנמצאה נגועה, או לחילופין המקום של הנגוע בקבוצה שנמצאה נגועה.
בהנחה שיש נגוע אחד בקבוצה הסיכוי שהוא יהיה במקום ה-x הוא קבוע ושווה ל- $\frac{1}{n_1}$.
נחשב את התוחלת של מספר הבדיקות שנבצע בקבוצה מסויימת שנמצאה נגועה:

$$E(X)=1\cdot\frac{1}{n_1}+2\cdot\frac{1}{n_1}+3\cdot\frac{1}{n_1}+\cdots+n_1\cdot\frac{1}{n_1}=\sum_{i=1}^{n_1} i\cdot\frac{1}{n_1}=\frac{1}{n_1}\frac{(n_1+1)n_1}{2}$$

$$E(X)=\frac{n_1+1}{2}$$

i מציין את מיקום הנבדק בסדרת הבדיקות הפרטניות עבור קבוצה מסויימת בגודל n_1 שנמצאה נגועה, וכל מיקום אפשרי מכפילים בהסתברות שהאדם שנמצא במיקום זה הוא הנגוע.
בצורה זו מקבלים את $E(X)$ - מספר הבדיקות שנעשה בממוצע בכל קבוצה.
כעת נוכל לחשב את C_*

$$C_*(n_1,n_2)=\overbrace{\underbrace{\frac{N}{n_1}}_{\text{מספר הקבוצות}}+\underbrace{N\cdot p}_{\text{מספר הקבוצות הנגועות}}\cdot\underbrace{\frac{n_1+1}{2}}_{E(X)}}^{\text{מספר הבדיקות שנעשה בשלב הראשון}}+\overbrace{\frac{N'}{n_2}+N'\cdot p\cdot n_2}^{\text{מספר הבדיקות שנעשה בשלב השני}}$$

כאשר N' שווה למספר האנשים שצוברים מכל הקבוצות הנגועות.

$$N'=\underbrace{N\cdot p\cdot n_1}_{\text{כל האנשים בקבוצות הנגועות}}-\underbrace{N\cdot p\cdot\frac{n_1+1}{2}}_{\text{האנשים מהקבוצות הנגועות שכבר נבדקו}}=N\cdot p\cdot\frac{n_1-1}{2}$$

נציב את N' בביטוי של C_*:

$$C_*(n_1,n_2)=\frac{N}{n_1}+N\cdot p\cdot\frac{n_1+1}{2}+\frac{N\cdot p}{n_2}\cdot\frac{n_1-1}{2}+N\cdot p^2\cdot\frac{n_1-1}{2}\cdot n_2$$

כעת נרצה למצוא מינימום ל- $C_*(n_1, n_2)$, כלומר נרצה למצוא ביטויים לערך של n_1, ו-n_2 שעבורם נקבל מספר בדיקות מינימלי. לשם כך נגזור את הפונקציה לפי כל אחד מהמשתנים ונשווה ל-0, ונבודד מתוך מערכת המשוואות שנקבל את n_1, ו-n_2.

נגזור לפי n_1 ונשווה לאפס

$$\frac{\partial C_*}{\partial n_1} = -\frac{N}{n_1^2} + \frac{N \cdot p}{2} + \frac{N \cdot p}{2n_2} + \frac{N \cdot n_2 \cdot p^2}{2} = 0$$

$$\frac{2}{n_1^2} = p + \frac{p}{n_2} + n_2 \cdot p^2$$

$$n_1^2 = \frac{2}{p + \frac{p}{n_2} + n_2 \cdot p^2}$$

$$n_1 = \sqrt{\frac{2n_2}{n_2 \cdot p + p + n_2^2 \cdot p^2}}$$

נגזור לפי n_2 ונשווה לאפס

$$\frac{\partial C_*}{\partial n_2} = -\frac{N \cdot p}{n_2^2} \cdot \frac{n_1 - 1}{2} + N \cdot p^2 \cdot \frac{n_1 - 1}{2} = 0$$

$$\frac{-(n_1 - 1)}{n_2^2} + p \cdot (n_1 - 1) = 0$$

$$p \cdot n_2^2 \cdot (n_1 - 1) = n_1 - 1$$

$$p \cdot n^2 = 1$$

$$n_2^* = \frac{1}{\sqrt{p}}$$

נציב את n_2^* בביטוי שקיבלנו ל-n_1 ונקבל:

$$n_1^* = \sqrt{\frac{2}{p + 2p\sqrt{p}}}$$

נציב את n_1^*, n_2^* חזרה ב-C^* ונקבל ביטוי ל-C^* האופטימלי:

$$C^* = N \cdot \sqrt{\frac{p + 2p\sqrt{p}}{2}} + \frac{N \cdot p}{2} \cdot \left(\sqrt{\frac{2}{p + 2p\sqrt{p}}} + 1\right) + N \cdot p \cdot \sqrt{p} \left(\sqrt{\frac{2}{p + 2p\sqrt{p}}} - 1\right)$$

כעת נשווה בין שימוש בשיטת דורפמן הרגילה לשיטת דורפמן המשופרת ע"י
דוגמה מספרית: נניח ששכיחות המחלה באוכלוסיה היא $p = 0.01$

<u>חישוב על פי שיטת דורפמן הרגילה:</u>

תחילה נחשב את n על פי הנוסחה שקיבלנו לקירוב n

$$n = \frac{1}{\sqrt{p}} = \frac{1}{\sqrt{0.01}} = 10$$

כעת נחשב את תוחלת מספר הבדיקות לאדם

$$C = \frac{\frac{1}{N} + N \cdot p \cdot n}{N} = \frac{1}{n} + p \cdot n = \frac{1}{10} + 0.01 \cdot 10 = \mathbf{0.2}$$

<u>חישוב על פי השיטה החדשה:</u>

נחשב את גדלי הקבוצות n_1 ו-n_2 האופטימליים על פי הנוסחאות שהתקבלו:

$$n_1^* = 12.9 \approx 13$$
$$n_2^* = 10$$

כעת נחשב את תוחלת מספר הבדיקות לאדם

$$\frac{C_*(n_1, n_2)}{N} = \frac{1}{n_1} + p \cdot \frac{n_1 + 1}{2} + \frac{p}{n_2} \cdot \frac{n_1 - 1}{2} + p^2 \cdot \frac{n_1 - 1}{2} \cdot n_2$$
$$\frac{C_*(13, 10)}{N} = \frac{1}{13} + 0.01 \cdot \frac{14}{2} + \frac{0.01}{10} \cdot \frac{12}{2} + 0.01^2 \cdot \frac{12}{2} \cdot 10 = \mathbf{0.1589}$$

הקטנו את תוחלת כמות הבדיקות שנבצע ב 0.0411, **שיפור של** 20%.

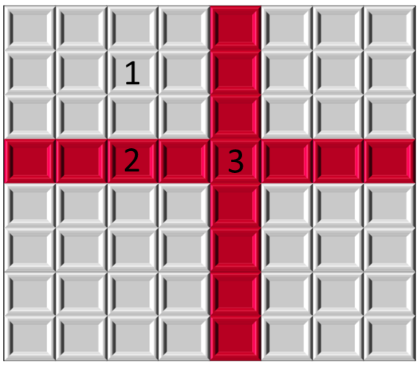
3 שיטת מערך ריבועי

3.1 תאור השיטה

שיטה זו הוצעה לראשונה במאמר של Phatarfod [3]. במעבדה, דגימות דם לעיתים קרובות מונחות במגש מרובע ($n \times n$) ושיטת מערך ריבועי מנצלת את הסדר הזה. כל אחת מ-n השורות יוצרת תערובת דגימות שכל אחת מהן נבדקת, אותו הדבר עושים גם ל-n העמודות. לאחר מכן מבצעים בדיקה פרטנית עבור כל צומת שבה קיבלנו תוצאה חיובית גם בבדיקה הקבוצתית של השורה וגם בבדיקה הקבוצתית של העמודה.

שיטה זו נקראת SA1 והוצעה במאמר של Phatarfod [3].

נמחיש זאת באיור הבא:



באיור מתואר מגש ובו בכל ריבוע דגימת דם שונה, העמודה והשורה המסומנות באדום הן קבוצות שעבורן נמצאה תוצאה חיובית.

נניח לצורך ההסבר שבשאר הקבוצות התקבלה תוצאה שלילית ושיש נגוע אחד בקבוצה (נמצא בצומת המסומנת ב-3).

אם צומת מסוים נמצא בשורה שהתקבלה בה תוצאה שלילית ובעמודה שהתקבלה בה תוצאה שלילית (למשל צומת מספר 1 באיור), אז כמובן שצומת זה אינו נגוע.

אם צומת מסוים נמצא בשורה שהתקבלה בה תוצאה חיובית ובעמודה שהתקבלה בה תוצאה שלילית (למשל צומת מספר 2 באיור), אז בהכרח צומת זה לא נגוע ולא צריך לבצע לו בדיקה פרטנית, כי אם הוא היה נגוע אז הוא היה "מזהם" את תערובת דגימות הדם של העמודה שבו הוא נמצא.

אם צומת מסוים נמצא בשורה ובעמודה שהתקבלו בהן תוצאות חיוביות (למשל

צומת מספר 3 באיור) אז יתכן והוא נגוע וכדי לבדוק זאת צריך לבצע בדיקה פרטנית לצומת זה.
אם הצומת אכן נגוע ודגימת הדם של צומת זה מעורבבת עם דגימות הדם של כל חבריו לשורה, גם תערובת דגימות הדם של הקבוצה בהכרח תהיה נגועה. אותו הדבר גם לעמודה שבה צומת זה נמצא.
יתכן גם שצומת זה אינו נגוע ובמקרה בשתי הקבוצות בהן הוא נמצא בתערובת (שורה ועמודה) יש חבר קבוצה אחר נגוע. (מקרה אינו אפשרי באיור שלעיל).

לסיכום, הבדיקה הפרטנית מתבצעת רק למי שנמצא בהצטלבות של שורה ועמודה שבהן התקבלה תוצאה חיובית, כי רק צומת כזה "מועמד" להיות נגוע.

נחשב את T_{SA1} - מספר הבדיקות הכולל הצפוי בשיטת SA1.
תחילה נגדיר:
X - מ"מ המציין את מספר הבדיקות הכולל שנצטרך לבצע בשיטת SA1.
R_i - המאורע ששורה i חיובית
C_j - המאורע שעמודה j חיובית
$i, j = 1, 2, \ldots n$
נגדיר גם את משתנה האינדיקטור הבא:

$$I_{i,j} = \begin{cases} 1 & R_i \cap C_j \\ 0 & \text{אחרת} \end{cases}$$

$R_i \cap C_j$ - מציין שגם הבדיקה של השורה ה-i וגם הבדיקה של העמודה ה-j יצאו חיוביות.
אז מספר הבדיקות שנצטרך לבצע יהיה:

$$X = 2n + \sum_{i=1}^{n} \sum_{j=1}^{n} I_{i,j}$$

כלומר, תחילה מבצעים בדיקות קבוצתיות ל-n השורות ול-n העמודות, סך הכל $2n$ בדיקות. ומוסיפים את מספר המקרים שעבורם קיבלנו תוצאה חיובית בשורה ובעמודה כי להם נצטרך לבצע שוב בדיקה פרטנית.

נחשב את תוחלת מספר הבדיקות שנצטרך לבצע:

$$E(X) = E\left(2n + \sum_{i=1}^{n}\sum_{j=1}^{n} I_{i,j}\right) = E(2n) + E\left(\sum_{i=1}^{n}\sum_{j=1}^{n} I_{i,j}\right) = 2n + \sum_{i=1}^{n}\sum_{j=1}^{n} E(I_{i,j})$$

עבור $I_{i,j}$ מסויימים :

$$E(I_{i,j}) = 1 \cdot P(I_{i,j} = 1) + 0 \cdot P(I_{i,j} = 0) = P(I_{i,j} = 1)$$

נציב בביטוי של התוחלת ונקבל:

$$E(X) = 2n + \sum_{i=1}^{n}\sum_{j=1}^{n} P(I_{i,j} = 1)$$

אז המספר הכולל הצפוי של הבדיקות הנדרשות הוא:

$$T_{\text{SA1}} = 2n + \sum_{i=1}^{n}\sum_{j=1}^{n} P(R_i \cap C_j)$$

$P(R_i \cap C_j)$ - מציין את ההסתברות שגם הבדיקה של השורה ה-i וגם הבדיקה של העמודה ה-j יצאו חיוביות.

$$P(R_i \cap C_j) = 1 - P((R_i \cap C_j)^c) = 1 - P(R_i^c \cup C_j^c)$$

$P(R_i^c \cup C_i^c)$ - ההסתברות שהבדיקה של השורה ה-i או הבדיקה של העמודה ה-j שלילית (או שניהם).

לפי עקרון ההכלה וההפרדה:

$$\begin{aligned} P(R_i^c \cup C_j^c) &= \underbrace{P(R_i^c)}_{\text{כל איברי השורה שליליים}} + \underbrace{P(R_j^c)}_{\text{כל איברי העמודה שליליים}} - \underbrace{P(R_i^c \cap C_j^c)}_{\text{החסרת מקרים שנספרו פעמיים}} \\ &= q^n + q^n + q^{2n-1} \end{aligned}$$

לכן נקבל:

$$P(R_i \cap C_j) = 1 - P(R_i^c \cup C_j^c) = 1 - 2q^n + q^{2n-1}$$

נציב את מה שהתקבל בביטוי של T_{SA1} ונקבל:

$$T_{\text{SA1}} = 2n + \sum_{i=1}^{n}\sum_{j=1}^{n}(1 - 2q^n + q^{2n-1}) = 2n + n^2 \cdot (1 - 2q^n + q^{2n-1})$$

נחשב את מספר הבדיקות הצפוי לאדם:

$$C_{\mathrm{SA1}}(n)=\frac{T_{\mathrm{SA1}}}{n^2}=\frac{2}{n}+1-2(1-p)^n+(1-p)^{2n-1}$$

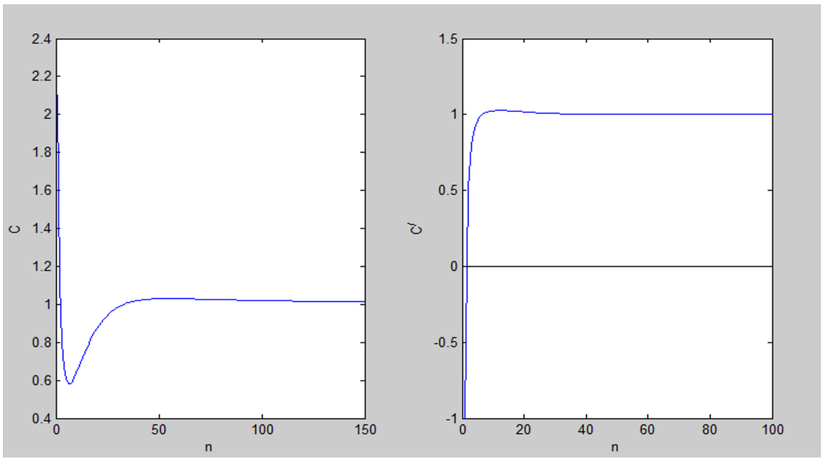
על מנת למצוא את ה-n האופטימלי עבורו שיטת הבדיקה הקבוצתית יעילה יש לגזור את הפונקציה שהתקבלה ולהשוות אותה לאפס.

$$C'_{\mathrm{SA1}}(n)=-\frac{2}{n^2}-2\cdot\ln(1-p)\cdot(1-p)^n+2\cdot\ln(1-p)\cdot(1-p)^{2n-1}=0$$

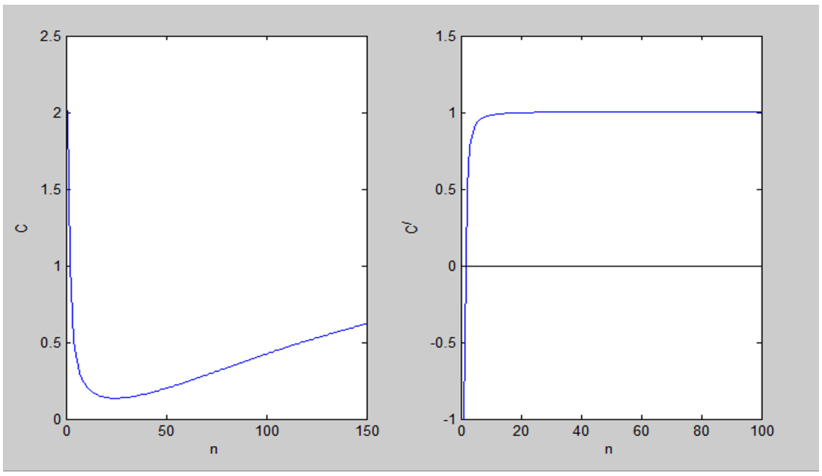
משוואה זו לא ניתן לפתור אנליטית.

נשרטט את הגרף של C_{SA1} ואת גרף הנגזרת עבור ערכים שונים של p

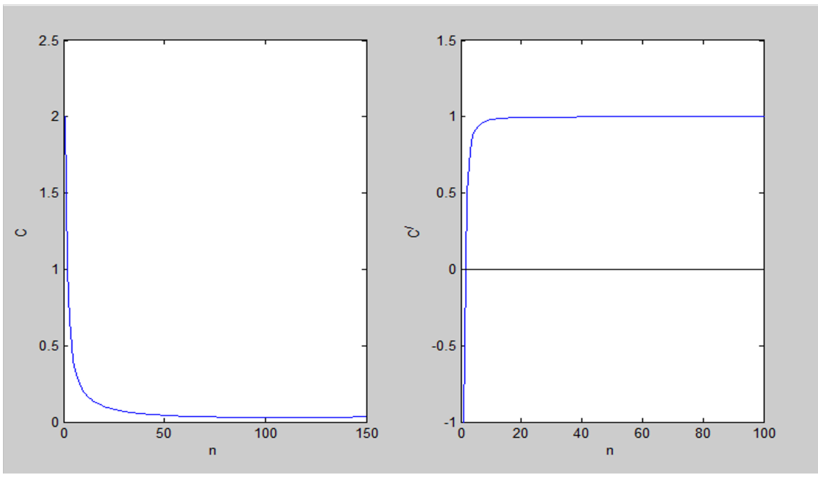
P=0.1



P=0.01



P=0.001



על פי השרטוטים ניתן לראות שגרף הנגזרת חותך את ציר ה-x וקיימת נקודת מינימום. בנוסף, נקודה זו נמצאת מתחת לישר $C_{\text{SA1}} = 1$, מכאן שנקודת המינימום היא רלוונטית, כלומר נקבל חיסכון בכמות הבדיקות בשיטת הבדיקה הקבוצתית.

ע"י פתרון נומרי של המשוואה $C'_{\text{SA1}} = 0$ נוכל למצוא את הגודל הקבוצה - n שיתן לנו את הערך המינמלי של תוחלת מספר הבדיקות C_{SA1}.

3.2 השוואה בין שיטת דורפמן לשיטת SA1

נבצע השוואה בין שיטת דורפמן לשיטת SA1 מתי עדיף להשתמש בכל אחת מהשיטות.

ע"י חישובים ב-Matlab התקבלו שתי הטבלאות הבאות:

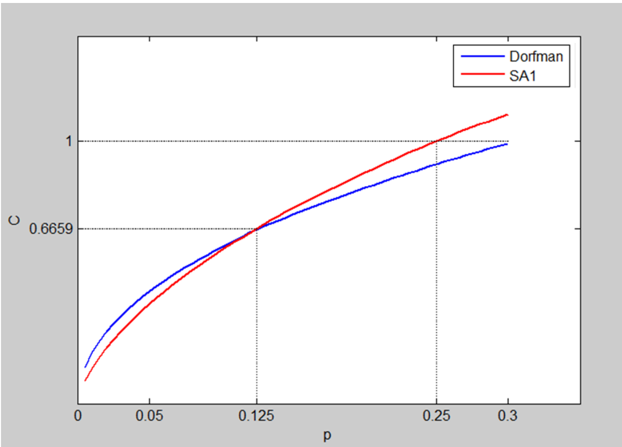
שיטת דורפמן			
p	n	C	S
0.005	15	0.1391	0.8609
0.01	11	0.1956	0.8044
0.015	9	0.2383	0.7617
0.02	8	0.2742	0.7258
0.025	7	0.3053	0.6947
0.03	6	0.3337	0.6663
0.035	6	0.3591	0.6409
0.04	6	0.3839	0.6161
0.045	5	0.4056	0.5944
0.05	5	0.4262	0.5738
0.055	5	0.4464	0.5536
0.06	5	0.4661	0.5339
0.065	4	0.4857	0.5143
0.07	4	0.5019	0.4981
0.075	4	0.5179	0.4821
0.08	4	0.5336	0.4664
0.085	4	0.5491	0.4509
0.09	4	0.5643	0.4357
0.095	4	0.5792	0.4208
0.1	4	0.5939	0.4061
0.105	4	0.6084	0.3916
0.11	4	0.6226	0.3774
0.115	4	0.6366	0.3634
0.12	3	0.6519	0.3481
0.125	3	0.6634	0.3366
0.13	3	0.6748	0.3252
0.135	3	0.6861	0.3139
0.14	3	0.6973	0.3027
0.145	3	0.7083	0.2917
0.15	3	0.7192	0.2808
0.155	3	0.73	0.27
0.16	3	0.7406	0.2594
0.165	3	0.7512	0.2488
0.17	3	0.7615	0.2385
0.175	3	0.7718	0.2282
0.18	3	0.782	0.218
0.185	3	0.792	0.208
0.19	3	0.8019	0.1981
0.195	3	0.8117	0.1883
0.2	3	0.8213	0.1787
0.205	3	0.8309	0.1691
0.21	3	0.8403	0.1597
0.215	3	0.8496	0.1504
0.22	3	0.8588	0.1412
0.225	3	0.8678	0.1322
0.23	3	0.8768	0.1232
0.235	3	0.8856	0.1144
0.24	3	0.8944	0.1056
0.245	3	0.903	0.097
0.25	3	0.9115	0.0885
0.255	3	0.9198	0.0802
0.26	3	0.9281	0.0719
0.265	3	0.9363	0.0637
0.27	3	0.9443	0.0557
0.275	3	0.9523	0.0477
0.28	3	0.9601	0.0399
0.285	3	0.9678	0.0322
0.29	3	0.9754	0.0246
0.295	3	0.9829	0.0171
0.3	3	0.9903	0.0097

שיטת SA 1			
p	n	C	S
0.005	38	0.0861	0.9139
0.01	25	0.1355	0.8645
0.015	19	0.1761	0.8239
0.02	16	0.212	0.788
0.025	14	0.2445	0.7555
0.03	13	0.2748	0.7252
0.035	12	0.3031	0.6969
0.04	11	0.3297	0.6703
0.045	10	0.3549	0.6451
0.05	9	0.3798	0.6202
0.055	9	0.4024	0.5976
0.06	9	0.4255	0.5745
0.065	8	0.4467	0.5533
0.07	8	0.4675	0.5325
0.075	8	0.4886	0.5114
0.08	7	0.5083	0.4917
0.085	7	0.5269	0.4731
0.09	7	0.5456	0.4544
0.095	7	0.5645	0.4355
0.1	7	0.5833	0.4167
0.105	6	0.6005	0.3995
0.11	6	0.6169	0.3831
0.115	6	0.6332	0.3668
0.12	6	0.6496	0.3504
0.125	6	0.6659	0.3341
0.13	6	0.6822	0.3178
0.135	6	0.6984	0.3016
0.14	6	0.7145	0.2855
0.145	5	0.7304	0.2696
0.15	5	0.7442	0.2558
0.155	5	0.758	0.242
0.16	5	0.7718	0.2282
0.165	5	0.7855	0.2145
0.17	5	0.7991	0.2009
0.175	5	0.8127	0.1873
0.18	5	0.8261	0.1739
0.185	5	0.8395	0.1605
0.19	5	0.8527	0.1473
0.195	5	0.8659	0.1341
0.2	5	0.8789	0.1211
0.205	5	0.8917	0.1083
0.21	5	0.9044	0.0956
0.215	5	0.917	0.083
0.22	5	0.9294	0.0706
0.225	5	0.9417	0.0583
0.23	5	0.9538	0.0462
0.235	5	0.9657	0.0343
0.24	5	0.9775	0.0225
0.245	4	0.99	0.01
0.25	4	1.0007	-0.0007
0.255	4	1.0113	-0.0113
0.26	4	1.0218	-0.0218
0.265	4	1.0322	-0.0322
0.27	4	1.0425	-0.0425
0.275	4	1.0527	-0.0527
0.28	4	1.0628	-0.0628
0.285	4	1.0728	-0.0728
0.29	4	1.0827	-0.0827
0.295	4	1.0925	-0.0925
0.3	4	1.1022	-0.1022

* בשיטת SA1 עבור כל ערך של p בוצע חישוב של הערך האופטימלי של n על ידי פתרון נומרי של המשוואה $C'_{\text{SA1}} = 0$.

ניתן לראות מהטבלה שעבור $p < 0.125$ שיטת SA1 עדיפה על שיטת דורפמן וככל ש-p קטן יותר החיסכון בשיטת SA1 גדול יותר.
עבור $p = 0.125$ מתקבל אותו C עבור שתי השיטות (בטבלה הערכים של C לא שווים מדויק כי הוא חושב עם n מעוגל).
עבור $p > 0.125$ שיטת דורפמן עדיפה על שיטת SA1 וככל ש-p גדול יותר החיסכון בשיטת דורפמן גדול יותר.
החל מ-$p = 0.25$ החיסכון שלילי ושיטת SA1 לא יעילה כלל, כלומר נבצע בה יותר בדיקות מאשר אם נבצע בדיקה פרטנית לכל אדם מהאוכלוסיה.

ניתן לראות את התוצאות האלה גם בגרף הבא:



4 סיכום ומסקנות

תחילה הצגנו את השיטה הבסיסית של דורפמן לבדיקה הקבוצתית.
הצגנו את הנוסחה עבור $C(n)$ - תוחלת מספר הבדיקות לאדם התלויה בשכיחות המחלה -p ובגודל הקבוצה - n.
בהנתן p מסוים, ניתן לבחור גודל קבוצה n ולחשב את כמות הבדיקות שנחסכו יחסית לבדיקה פרטנית של כל האוכלוסיה: $S = 1 - C$

לאחר שקיבלנו את הנוסחה ל-$C(n)$ חקרנו את הפונקציה על מנת למצוא מהו גודל הקבוצה האופטימלי שיש לבחור בהנתן p מסוים המביא את C למינימום, כלומר שנבצע כמה שפחות בדיקות וכתוצאה מכך נקבל את החיסכון המקסימלי.
מצאנו שתמיד קיים מינימום כזה, קיבלנו $p^* = 0.148$ קריטי כך ש:
עבור $p > p^*$ לא קיים מינימום, הפונקציה $C(n)$ מונוטונית יורדת ונמצאת כולה מעל לישר $C(n) = 1$, כלומר מתקבל חיסכון שלילי. לכן במקרה זה שיטת דורפמן לא יעילה לכל גודל קבוצה שנבחר.
עבור $p < p^*$ קיים מינימום אבל שיטת דורפמן לא בהכרח יעילה, היא תהיה יעילה רק אם נקודת המינימום מתקבלת מתחת לישר $C(n) = 1$. מצאנו $p^{**} = 0.30779$ קריטי, כך ש:
עבור $p^{**} < p < p^*$ שיטת דורפמן לא יעילה כלל כי המינימום מתקבל מעל לישר $C(n) = 1$.
עבור $p = p^{**}$ החיסכון שווה לאפס כי המינימום מתקבל על הישר $C(n) = 1$.
ועבור $p < p^{**}$ המינימום מתקבל מתחת לישר $C(n) = 1$, לכן יש חיסכון במספר הבדיקות שנבצע ושיטת הבדיקה הקבוצתית יעילה.

שכיחות $p^{**} = 0.30779$ היא שכיחות מאוד גבוהה למחלה, במציאות לרוב השכיחויות הרבה יותר נמוכות כך שברוב המקרים שיטת דורפמן יעילה.

בנוסף, הגענו למסקנה שככל שערכו של p גדול יותר, גודל הקבוצה האופטימלי קטן יותר, כיוון שאם נקח קבוצה גדולה כאשר השכיחות גדולה, הסיכוי מאוד גדול שנקבל שרוב הקבוצות או כולן נגועות ואז נצטרך לבצע לכולן בדיקה פרטנית, וכתוצאה מכך נקבל שלא חסכנו כלום ואף יתכן שהגדלנו את כמות הבדיקות.

כיוון שאת המשוואה $C'(n) = 0$ ניתן לפתור רק בשיטה נומרית, הצגנו דרך לקרב את הערך האופטימלי של n. בצורה זו נוכל לחשב את n האופטימלי במהירות ובקלות. בהינתן p מסוים על ידי $n = \frac{1}{\sqrt{p}}$.
ראינו שקירוב זה מאוד טוב עבור ערכים של p בטווח הרלוונטי $p < p^{**}$.

חקרנו איך מספר הבדיקות מתפלג, מצאנו את תוחלת מספר הבדיקות ושונות מספר הבדיקות שבאמצעותה חישבנו את סטיית התקן. בצורה כזו בהינתן p מסוים ולאחר שנחשב את n האופטימלי, נוכל לחשב כמה בדיקות נצטרך לבצע בממוצע ומה השגיאה שיכולה להתקבל, כלומר כמה יתכן שנסטה מהממוצע ימינה או שמאלה. באמצעות זה נוכל להחליט האם למרות האקראיות השיטה נותנת חיסכון במספר הבדיקות שנבצע.

הצגנו שיפור לשיטה הבסיסית של דורפמן שיכול לתת חיסכון גדול יותר במספר הבדיקות שנבצע. השיפור מתבטא בכך שהוספנו עוד שלב לשיטת הבסיסית. ז"א בשלב הראשון מבצעים עדיין חלוקה לקבוצות ומבצעים בדיקה קבוצתית אבל לקבוצות הנגועות לא מבצעים בדיקה פרטנית רגילה אלא מבצעים בדיקה פרטנית עד שנתקל בנגוע הראשון ואת שאר חברי הקבוצה שלא הספקנו לבדוק נצרף לחברי הקבוצות האחרות שלא הספקנו לבדוק. בשלב השני נבצע על ה"אוכלוסיה" החדשה שהתקבלה בדיקה לפי שיטת דורפמן הבסיסית עם
גודל קבוצה מתאים.
ניתן גם להרחיב שיפור זה לשלושה או ארבעה שלבים ואף יותר, ז"א שעל האוכלוסיה החדשה שהתקבלה בשלב השני נפעיל שוב את השיטה המשופרת ולא את שיטת דורפמן הבסיסית.

הצגנו את שיטת מערך ריבועי המתבססת על שיטת דורפמן. בשיטה זו בדיקות הדם נמצאות במגש מרובע $(n \times n)$ ומשתמשים בשיטת דורפמן תוך ניצול סדר הדגימות במגש. כלומר, מתייחסים לכל עמודה ולכל שורה כקבוצה ולכל אחת מהקבוצות מבצעים בדיקת דם. אם עבור דגימה מסוימת שנמצאת במגש גם הבדיקה של השורה שלה יצאה חיובית וגם הבדיקה של העמודה שלה יצאה חיובית דגימת הדם הזו חשודה כנגועה ומבצעים לה בדיקה פרטנית.
חישבנו לשיטה זו את תוחלת מספר הבדיקות שנבצע - T_{SA1} וחישבנו את הנגזרת

של ביטוי זה והתבוננו בשרטוטים שלהם עבור ערכים שונים של p, ראינו שקיים מינימום עבור p-ים אלו ונקודת המינימום הזו נותנת חיסכון חיובי.
בנוסף השוונו בין שיטת דורפמן הבסיסית לשיטת מערך ריבועי וראינו שקיים $p = 0.125$ כך ש:
עבור $p < 0.125$ שיטת SA1 עדיפה על שיטת דורפמן וככל ש-p קטן יותר החיסכון בשיטת SA1 גדול יותר.
עבור $p > 0.125$ שיטת דורפמן עדיפה על שיטת SA1 וככל ש-p גדול יותר החיסכון בשיטת דורפמן גדול יותר.
מלבד שיטת מערך ריבועי יש מגוון ווריאציות נוספות לשימוש בשיטת דורפמן.

דוגמאות לשימושים בשיטת הבדיקה הקבוצתית:

1. על מנת למנוע את התפשטות מחלת האיידס, הצלב האדום האמריקאי החל לבדוק את דמם של כל התורמי הדם, אבל באזורים עניים בעולם זה לא היה אפשרי, כיוון שאין מספיק כסף כדי לרכוש ערכות בדיקה לכולם. ההשלכות האנושיות כתוצאה מכן הן הרסניות.
פרופסור סטפנוס זניאוס התעניין בכך ויחד עם פרופסור לורנס ויין פיתחו דרך מדוייקת לבצע בדיקה במחיר זול יותר, ע"י בדיקת של קבוצות של דגימות דם.
אם למשל מבצעים בדיקה לקבוצה של 10 דגימות והתוצאה שלילית, אז נחסכה עלות של 9 דגימות. אם התוצאה חיובית, יש צורך בבדיקה פרטנית לכל אחת מדגימות הדם בקבוצה.
הרעיון הזה של איחוד דגימות דם לא היה חדש אבל בדיקת הוירוס שגורם לאיידס הופכת את העניין לקצת יותר מסובך.
כאשר מבצעים בדיקת דם ל- HIV זה לא מתבצע בתהליך פשוט, צריך להשאיר את דגימת הדם בצלחת ואז מוסיפים אנזימים. אנזימים אלה משפיעים על צבע הדם שנבדק, על פי הצבע שמתקבל קובעים את ריכוז הנוגדנים ל-HIV. אם התוצאה עולה על הסף הקריטי האדם נגוע ב- HIV.
הבעיה היא שכאשר מערבבים מספר דגימות דם, כמות הדם היא גדולה מאוד ויכולה לדלל את הנוגדנים ולשבש את תוצאת הבדיקה.

לשם כך זאינוס וויין בנו מודל הקובע מהו גודל הקבוצה האופטימלי כך שלא ישנה את תוצאת הבדיקה, כלומר מודל זה קובע כמה צריך לדלל דגימה נגועה כדי לקבל תוצאה שגויה ע"י בדיקת פרמטרים מסויימים וביניהם p - שכיחות המחלה באוכלוסיה.

2. גילוי תרופות, מערבבים מספר תרופות שונות יחד ובאמצעות שיטת הבדיקה הקבוצתית בודקים האם יש יחסים סינרגטים בין התרכובות השונות. יחסים סינרגטים בין תרופות ניתן לגלות רק באמצעות מחקר שבו מערבבים מספר תרופות יחד. לכן מערבבים מספר תרופות יחד ויוצר-ים תרכובת. בכל פעם שבודקים תרכובת, בודקים אם היא עומדת במספר דרישות.

   אם כן, התרכובת הזו נקראת תרכובת מובילה (מקביל לקבוצה נגועה בהקשר של בדיקות הדם) ובמקרה זה נמשיך לבדוק את היחסים בין התר-ופות המרכיבות את התרכובת.

   אחרת, לא נמשיך לבדוק את היחסים בין התרופות המרכיבות את התרכובת. בדרך זו ניתן לגלות טיפולים משולבים במחלות קשות, ז"א ששילוב נטילת מספר תרופות שונות יכול לטפל במחלה בצורה יותר טובה.

ראינו בפרויקט את תרומתה של המתמטיקה לבריאות הציבור ולחיסכון במשאבים. לצורך פיתוח הביטוי המתמטי המבטא את תוחלת מספר הבדיקות לאדם השתמשנו בכלים מהסתברות.

לצורך חקירת הביטוי ובדיקה האם אכן הצלחנו לקבל שיטה המביאה לשיפור המצב הקיים השתמשנו בכלים מחשבון אינפיניטסימלי, אנליזה נומרית, ואופטימזציה.

באופן כללי, המתמטיקה תורמת רבות לאנושות במגוון תחומים. לרוב התרומה הזו מתבצעת "מאוחרי הקלעים" כך שהרבה אנשים לא מודעים לכך. לא פעם נשאלתי "למה מתמטיקה שימושית? למה זה שימושי?" אז בפרויקט זה הצגתי שימוש אחד מיני רבים של המתמטיקה.

5 נספחים

תוכניות ששימשו לשרטוט הגרפים ב- Matlab:

שרטוט גרף $C(n)$ עבור p מסוים:

```
clear all;
close all;
nn=100;
n=0:0.001:nn;
p=0.15;
C=(n+1)./n-(1-p).^n;
 plot(n,C,'m','LineWidth',1.8)
hold on
plot(0:0.001:nn,1,'b','LineWidth',3)
plot(0:0.001:nn,0,'k','LineWidth',3)
axis([0,20,-5,15])
xlabel('n')
ylabel('C(n)')
set(gca,'YTick',[0 1]);
set(gca,'XTick',[0]);
```

שרטוט גרף $C'(n)$ עבור p מסוים:

```
clear all;
close all;
nn=100;
n=0:0.001:nn;
p=0.3;
Ctag=-1./n.^2-log(1-p)*(1-p).^n;
plot(n,Ctag,'m','LineWidth',1.8)
hold on
plot(0:0.001:nn,0,'k','LineWidth',3)
axis([0,25,-0.4,0.1])
xlabel('n')
ylabel('C^,(n)')
set(gca,'YTick',[0]);
set(gca,'XTick',[0]);
```

שרטוט גרף $C(n)$ ו- $C'(n)$ עבור p-ים שונים:

```
close all;

nn=600;
n=0:nn;
p=[0.0001,0.001,0.01,0.1];

for i=1:length(p)
  C=(n+1)./n-(1-p(i)).^n;
  figure(1)
  hold all;
   plot(n,C);
   Ctag=-1./n.^2-log(1-p(i))*(1-p(i)).^n;
    figure(2)
    plot(n,Ctag);
    hold all;
end

 figure(1)
xlabel('n')
ylabel('C(n)')
plot(0:0.001:nn,1,'k','LineWidth',3)
 figure(2)
xlabel('n')
ylabel('C^,(n)')
plot(0:0.001:nn,0,'k','LineWidth',3)
axis([0,100,-1,0.4])
```

שרטוט גרף המשווה בין n המתקבל מפתרון נומרי של המשוואה $C'(n) = 0$ לבין n המקורב :

```
clear all;
close all;

p=0.0001:0.0001:0.3;
n0=5;
for i=1:length(p)
root(i)=fzero(@(n) -1./n.^2-log(1-p(i))*(1-p(i)).^n, n0);
root(i)=round(root(i));
c(i)=(root(i)+1)/root(i)-(1-p(i))^root(i);
s(i)=1-c(i);
end

plot(p, root,'LineWidth',2);
xlabel('p');
ylabel('n');
hold on
plot(p,1./sqrt(p),'r','LineWidth',2);
axis([0 0.35 0 120])
```

שרטוט גרף T_{SA1} ו- T'_{SA1} עבור p מסוים:

```
close all;

p=0.1;
nn=300;
n=0:nn;

  c=2./n+1-2*(1-p).^n+(1-p).^(2*n-1);
 subplot(1,2,1);
  plot(n,c)
  xlabel('n')
  ylabel('C')

   ctag=-2./n.^2+1-2*log(1-p)*(1-p).^n+2*log(1-p)*(1-p).^(2*n-1);
  subplot(1,2,2);
   plot(n,ctag)
  xlabel('n')
  ylabel('C^/')
  hold on
  plot(0:0.001:nn,0,'k')
  axis([0 200 -1 1.5]);
```

שרטוט גרף המשווה בין שיטת דורפמן לשיטת מערך ריבועי:

```
clear all;
close all;

p=0:0.005:0.3;
n0=5;

for i=1:length(p)
root(i)=fzero(@(n) -1./n.^2-log(1-p(i))*(1-p(i)).^n, n0,optimset('TolX',1e-10));
root(i)=round(root(i));
c1(i)=(root(i)+1)/root(i)-(1-p(i))^root(i);
s1(i)=1-c1(i);
end

for i=1:length(p)
root(i)=fzero(@(n) -2./n.^2-2*log(1-p(i))*(1-p(i)).^n+2*log(1-p(i))*(1-p(i)).^(2*n-1), n0,optimset('TolX',1e-10));
root(i)=round(root(i));
c2(i)=2/root(i)+1-2*(1-p(i))^root(i)+(1-p(i))^(2*root(i)-1);
s2(i)=1-c2(i);
end

plot(p,c1, 'LineWidth',1.3)
hold on
plot(p,c2,'r','LineWidth',1.3)
plot(0.125,0:0.01:0.6659,'-k','LineWidth',1.3)
plot(0:0.001:0.125,0.6659,'-k','LineWidth',1.3)
plot(0:0.001:0.3,1,'-k','LineWidth',1.3)
plot(0.25,0:0.01:1.007,'-k','LineWidth',1.3)
set(gca,'XTick',[0 0.05 0.125 0.25 0.3]);
set(gca,'YTick',[0.6659 1]);
xlabel('p')
ylabel('C')
```

יצירת טבלה המשווה את הערכים המתקבלים משתי השיטות:

```
clear all;
close all;

p=0:0.005:0.3;
n0=5;
for i=1:length(p)
root(i)=fzero(@(n) -2./n.^2-2*log(1-p(i))*(1-p(i)).^n+2*log(1-p(i))*(1-p(i)).^(2*n-1), n0,optimset('TolX',1e-10));
root(i)=round(root(i));
c(i)=2/root(i)+1-2*(1-p(i))^root(i)+(1-p(i))^(2*root(i)-1);
s(i)=1-c(i);
end

disp('      p      n      cost    save');
disp([p' root' c' s']);
```

רשימת מקורות

[1] Dorfman, Robert. "The detection of defective members of large populations." The Annals of Mathematical Statistics 14.4 (1943): 436-440.

[2] Finucan, H. M. "The blood testing problem." Applied Statistics (1964): 43-50.

[3] Phatarfod, R. M., and Aidan Sudbury. "The use of a square array scheme in blood testing." Statistics in Medicine 13.22 (1994): 2337-2343.

[4] A Statistical Solution to Testing the Blood Supply for HIV. September 1, 1997—by Barbara Buell. http://stanford.io/1oNma2f

[5] Hughes-Oliver, Jacqueline. "Pooling experiments for blood screening and drug discovery." Screening: Methods for Experimentation in Industry, Drug Discovery, and Genetics, Springer New York (2006): 48-68.