21.12.06 (1)

סטטיסטיקה למדעי המחשב - שיעור 8

רווחי סמך

$$\forall \alpha.\ P(\theta \in [S(\bar{X}) - \varepsilon_1,\ S(\bar{X}) + \varepsilon_2]) \geq 1-\alpha$$

$$\Updownarrow$$

$$\forall \alpha.\ P_r(S(\bar{X}) \in [\theta - \varepsilon_2,\ \theta + \varepsilon_1]) \geq 1-\alpha$$

רווח סמך לתוחלת (רמ"ל) עם שונות ידועה

נשתמש בשימוש $\hat{\mu} = \bar{X}$, רמ"ס $1-\alpha$:

$$\theta \in [\bar{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}},\ \bar{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}]$$

במקרה והשונות אינה ידועה

אומדן שונות ע"י השימוש $\hat{\sigma}^2 = \frac{1}{n-1} \sum (X_i - \bar{X})^2$

ואז רווח הסמך: $\theta \in [\bar{X} - t_{n-1,1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}},\ \bar{X} + t_{n-1,1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}}]$

נזכר ב התפלגות t מה הגדרה של Gosset: $t_{n-1} \sim \frac{N(0,1)}{\sqrt{\chi^2_{n-1}/n-1}}$

(אם בפנים, אנו רואים את התפלגות t גם מול ההתפלגות הנורמלית. ניתן לה שכבל ש ברבים המדגם עולות התפלגות t ברשת ויאה ויאה כמו ההתפלגות הנורמלית.

הבהרה משכנע שעבר

שנה לפרמטר בינומי

$$\log(\mathcal{L}(p)) = \sum x_i \log p + (n - \sum x_i) \log(1-p)$$

שנה:

$$\frac{\partial}{\partial p} = \frac{\sum x_i (1-p) - (k - \sum x_i) p}{p(1-p)} \Rightarrow \hat{p} = \frac{\sum x_i}{n} \qquad (\sum x_i = n_1 \text{ נסמן})$$

(מקס) שגרה מקסימום

$$\frac{\partial^2 \mathcal{L}(p)}{\partial^2 p} = \frac{-np(1-p) - n_1 + np + 2n_1 p - 2np^2}{p^2(1-p)^2 > 0} \Rightarrow$$

$$\Rightarrow -np^2 + 2n_1 p - n_1 \overset{p>0}{<} 0 \Rightarrow$$

$$\Delta = 4n^2 - 4n \cdot n_1 \leq 0$$

הנגזרת תמיד שלילית

רווח סמך לפרמטר בינומי

$X \sim Bernoulli(p)$, יהי $\bar{X} = (X_1, \dots, X_n)$ מדגם מקרי.

נבקש לעצב רמ"ס עם הפרפורציה השמאלית $\hat{p} = \frac{\sum x_i}{n}$ (נזכר $\sum x_i \sim Bin(n,p)$)

2 בעיות בחישוב:

בעיה I: רווח סמך מדויק עם כל הנתונים הבינומיים

$$P_r(p \in [\hat{p} - \varepsilon_1, \hat{p} + \varepsilon_2]) = P_r(\hat{p} \in [p - \varepsilon_2, p + \varepsilon_1]) = \sum_{k=\underbrace{(p-\varepsilon_1)n}_{\text{נתון}}}^{(p+\varepsilon_2)n} \underbrace{P_r(\sum X_i = k)}_{P_r(\hat{p} = \frac{k}{n})} = \sum_{k=(p-\varepsilon_1)n}^{(p+\varepsilon_2)n} \binom{n}{k} p^k (1-p)^{n-k} \underset{\text{רמ"ס מדויק}}{\geq} 1-\alpha$$

עם בעיה לא מבטיחה!

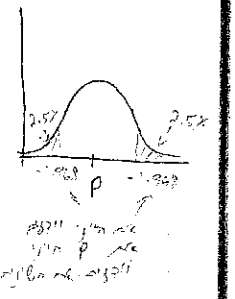
u.multinet.co.il

סטטיסטיקה למדעי המשפט- שיעור 8

נושא II: להשתמש בקירוב הנורמלי.

אמרנו שגם מספר ההצלחות והשונות שנמדד מן, אז ניתן להגיד $\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$

או $\hat{p} \sim N(p, \frac{p(1-p)}{n})$.



בעיה: השונות מכילה את הפרמטר הלא ידוע p, אשר איתו אנו מנסים לאמוד.

פתרון א': נחליף את p ב- $\hat{p}$ בשונות.

נסמן: $SD(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$ סטיית תקן

$SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ Standard Error

ונקבל: $\frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \approx N(0,1)$

אין לנו בעיה אלא להשתמש בהתפלגות הנורמלית.

אם הקירוב בסימן שם מדבר שמה "לפרק". סימן כזה מול ב דומות של איך קוראים: א. קרוב נורמלי לבינומי. ב. שימוש בשאר ל-p בלי p בחישוב.

לכן, נקבל:

$$P_r\left(p \in \left[\hat{p} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]\right) \approx 1-\alpha$$

דוגמא

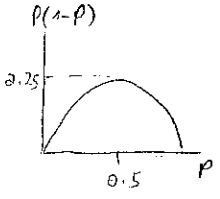
p: אחוז הסטודנטים הטבעונים, $n=25$, $\hat{p}=0.6$ רוצים $\alpha=0.05$.

רווח 95% ל- p: $\left[0.6 - 1.96\sqrt{\frac{0.6\cdot 0.4}{25}},\ 0.6 + 1.96\sqrt{\frac{0.6\cdot 0.4}{25}}\right]$

$= [0.41, 0.79]$

(אם) שרצינו שנשתמש בגרסה הישנה, אפשר לעשות רווחים הסתברותיים. אחר להגיד שקיבלנו רווח כמה מקיפים הפרמטר (מבנה בתוכו). לה"ל נכון לגבי שיש סיכוי של, עפ"י הפרמטר אמיתי בתוכו.

פתרון ב': להחליף את $p(1-p)$ במקסימום הבטוח שלו: $p(1-p) \le 0.25$



רווח סמך בשיטה הזו: $P_r\left(p \in \left[\hat{p} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{4n}},\ \hat{p} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{4n}}\right]\right) \ge 1-\alpha$

כדי כמו נמוכה הקירוב תיאורטי לבינומי, אינטרוולים לא שווים הסיכוי של הרווח שלנו הוא לפחות $1-\alpha$. זאת אומרת שאתה מבטיח לבניית רווח סמך עבור p.

בדוגמא: (קודם הנלון $[0.41, 0.79]$). $n=25$, $\hat{p}=0.6$.

$\left[0.6 - 1.96\sqrt{\frac{1}{100}},\ 0.6 + 1.96\sqrt{\frac{1}{4\cdot 25}}\right] = [0.4, 0.8]$

מגדיל הרווח מגיע כי $\hat{p}$ קרוב מגיע לחצי.

אם הנקודה $\hat{p}=0.6$ הוא $\hat{p}=0.9$ אבל הוא כבר גדול הסמך הוא $\sqrt{\frac{0.25}{0.16}} = \frac{5}{4}$.

u.multinet.co.il

הרעיון של שגיאה ל. השמרנית חשוב ומופיע במיוחד בסקרים פוליטיים ובתחומים:

1. לפני עריכת הסקר ניתן לקבוע כמה נדגמים דרוש לשגיאה כדי להגיע לרמה הרצויה של רווח הסמך.

2. בהנ"ל p השאיר קבוע $\frac{1}{2}$, ולכן השמרנית לא יורה מכי גדולה.

(*) מקבלים לקבוע $\pm Z_{0.975}\cdot\frac{1}{2\sqrt{n}}=\pm\frac{1}{\sqrt{n}}$ בתור טווח הגבולות השמרנית במדגם ע"י הטרכה שמרנית של רווח המאפיין שמרני.

כלומר, לפני מדגם ה מדגם ניתן לחשב את הטבלה:

n	שגיאה
100	$\pm 10\%$
400	$\pm 5\%$
600	$\pm 4\%$
1100	$\pm 3\%$

אינטרוולים לפי אמת מרבי שונות

דוגמה: גובה סטודנטים, תלוי לפי הסטודנטים שגבהם נמוך מ-170 ס"מ.

$$f(\mu)=\theta=P_r(X\le 170)=\Phi\left(\frac{\mu-170}{\sigma}\right)$$

נתון: שונות $\sigma^2=16$ ידוע, $\bar{X}=175$, $n=25$.

אז θ - $\hat{\theta}_{MLE}=f(\hat{\mu}_{MLE})=\Phi\left(\frac{170-175}{4}\right)=\Phi\left(-\frac{5}{4}\right)=0.016$

מטרה: נתון פרמטר θ, לפי מדד $S(\bar{X})$, ברמת סמך $1-\alpha$.

$$I=[S(\bar{X})-\varepsilon_1,\ S(\bar{X})+\varepsilon_2]$$

אם $f(\theta)$ הוא מתפלג מונוטונית (גדל ע"י ירידת) מן θ אזי $f(I)$ הוא לפי ברמת סמך $1-\alpha$ עבור $f(\theta)$.

בדוגמתנו הסטודנטים: לפי עבור μ $[173.4,176.6]=[\bar{X}-1.96\frac{4}{5},\ \bar{X}+1.96\frac{4}{5}]$

לפי עבור $\Phi\left(\frac{\mu-170}{4}\right)$: $\left[\Phi\left(\frac{170-(\bar{X}-1.96\frac{4}{5})}{4}\right),\ \Phi\left(\frac{170-(\bar{X}+1.96\frac{4}{5})}{4}\right)\right]=$ (Φ פונקציה יורדת מונוטונית)

$=\left[\Phi\left(\frac{-6.4}{4}\right),\ \Phi\left(\frac{-3.6}{4}\right)\right]=[0.069,\ 0.198]$ ← רווח סמך לאחוז סטודנטים גבוהה מהגבוה מ-170.

הוכחת הטענה:

לפי הנתון: $P_r(\theta\in\underbrace{[S-\varepsilon_1,\ S+\varepsilon_2]}_{I})\ge 1-\alpha$

אם f מונוטונית עולה אזי:

$$S-\varepsilon_1\le\theta\le S+\varepsilon_2 \iff f(S-\varepsilon_1)\le f(\theta)\le f(S+\varepsilon_2)$$

אם f מונוטונית יורדת אזי

$$S+\varepsilon_1\le\theta\le S+\varepsilon_2 \iff f(S+\varepsilon_2)\le f(\theta)\le f(S-\varepsilon_1)$$

לכן המאורעות שקולים, $P_r(f(\theta)\in f(I))\ge 1-\alpha$.

מסקנה

סטטיסטיקה למדעי החברה - שיעור 8

דרך נוספת ("הדרך הרגילה") לחישוב רווח הסטטיסטים מתחת לר- פב 1:

- (סעיף) על השגיאה הנורמלי בינומי.
- נחנה רווח בינוי על סיכוי שיעור כזה.

נניח לדוגמא ל סטטיסטים מתחת לר- פב במדגם. לה יוצר $\hat{p} = \frac{\text{מספר מתחת}}{25} = 0.16$.

רווח סמך 95%: $\left[\hat{p} - 1.96 \cdot \frac{\sqrt{\hat{p}(1-\hat{p})}}{5}, \hat{p} + 1.96 \frac{\sqrt{\hat{p}(1-\hat{p})}}{5}\right] = [0.019, 0.303]$

סקרים ושקרים

קולגה ל': הטיה במדגם

בארה"ב 1936 בחירות.

מועמד א': רוזוולט (הנשיא הנוכחי)
מועמד ב': לנדון.

רוב הציבור: רוזוולט ינצח.

העיתון Literary digest עשה סקר: שלח מכתבים לקוראים, ביקש העיתון לשלוח הבעת רצון.

⇐ מדגם בגודל 2.4 מיליון. הנסקרים בסיכוי לנדון יותר. (נניח $\hat{p} = 0.8$)
בפועל: רוזוולט ניצח, נכון לקולות.

אין הבדל בסקר הזה ובסקרים שם רק שכבות אוכלוסיה - קוראי העיתון.

הבעיה: הוכלו הנתונים בעל המנויים על השאלון בפרט, לא יוכלו את כל הבוחרים
⇐ מכתבים מאוד לא מקרי.

בחירות 1996 (הסקר לג נחוה)

נובמבר 1995: רבין רצח, פרס ראש הממשלה.

סקר יוניר 1996:

פרס	46.5%	
נתניהו	35.8%	
שחר	12.5%	לא יכולים / לא הוכרע
		לא יצא

⇐ הרוב מקבלים את הבחירות.

טבה הבחירות: לפרס בפועל: פרס מוביל ב- 3%: פרס 51.5%, נתניהו 48.5% שיעור השגיאה 3%, מספר הנבחרים 1355.

אין נשוא לכך שיעור השגיאה הוא 3% בהשוואה לכל מועמד, לא ביחסים.

p: שיעור תומכי בפרס, מהו ההפרש בתומכים: $2p-1 = p-(1-p)$.

$2p-1$ טרנס מונוטוני של p, לכן רווח להפרש כולל מרווח לשיעור התומכים.

⇐ רווח סמך ל- 95% תומכי פרס $[48.5\%, 54.5\%]$
רווח להפרש $[-3\%, +9\%]$.

u.multinet.co.il

סטטיסטיקה למדעי המחשב - שיעור 8

האם ההחלטה להקים חנות נכונה?

$X \sim$ לא התיוו אז לא מבלים.

ידוע היסטורית ש- 75% מהלקוחות מבלים יין. נניח 8% תנאי ו-19% פרס.
$\Leftarrow$ 49.4% פרס מול 43.8% תנאי.

אם (נרמה) מחקר לא גילה שמלצים נתן 53% פרס מול 46% תנאי.
כלומר סה"כ שגיאה של הקרן יותר מ- 3% לא ניתן להכריע מי יעדיף.

בדיקת / בחינת השערות Hypothesis testing

מה למדנו בהסקה בקירוב?

1. לאמוד פרמטר.
2. לקבוע רווח סמך.

הבעיה השונה היא לבדוק השערות על פרמטרים.

דוגמא: סקר בחירות, שואלים אלפי בוחרים מי מועמד p במועמד ל. מול מועמד כ.

השאלה הראשונה שמעניינת: האם $p > \frac{1}{2}$ בבוודאי? האם $p < \frac{1}{2}$ בבוודאי?
או האם לא ניתן לפסול $p = \frac{1}{2}$?

(בדרך כלל)
לשם כך נגדיר "השערת הלא" שיש הסבירה שניתן הגיע לתוצאות כדי להפריך אותה, ולא נפריך אותה רק אם נוכל בוודאות מהתוצאות שהשערת הלא אינה נכונה.

במקרה של בחירות: $H_0: p = \frac{1}{2}$, $H_A: p \neq \frac{1}{2}$ (ניתן לבדוק מי יותר)
אלטרנטיבה

דוגמא:

1. סיסטרילבת אומרת שמעמדת מרגלית שבני תינוקות יקר מאוד יותר מסרטן.

נניח שעדיין הסרטן בשלבים כלה הוא p_0 ידוע. נסמן ב- p שיעור הסרטן בקרב תינוקות ידוע.

נסמן, $H_0: p = p_0$, $H_A: p < p_0$. (ניתן לסמן את $H_0: p \geq p_0$)

2. מחקר על מחלה חדשה לגידם.

נניח מספר תגי T לא מחלה מתפלג $X \sim N(\mu, \sigma^2)$.
רוצים לבדוק האם המחלה מלרת. נניח שגם המחלה $X \sim N(\mu_1, \sigma^2)$

ואז $H_0: \mu_0 = \mu_1$ (המחלה לא משפיעה)

$H_A: \begin{cases} \mu_0 \neq \mu_1 & \text{(משפיעה בדרך כלשהי)} \\ \mu_0 > \mu_1 & \text{(מלרת)} \end{cases}$

סטטיסטיקה למדעי המחשב - שיעור 8

הגדרת מבחן סטטיסטי:

נניח שיש לנו השערת 0 $H_0: \theta = \theta_0$

מבחן סטטיסטי ברמת α מוגדר ע"י גזירה כמה $C_\alpha \subset \mathbb{R}$ כך ש- $Pr_{H_0}(S(\bar{X}) \in C_\alpha) \le \alpha$

המבחן: אם $S(X) \in C_\alpha$ נדחה את H_0, אחרת לא נדחה (נקבל) את H_0.

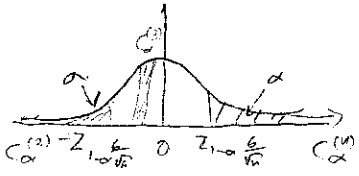
קיימת הרבה הגדרות שאפשריות ל- C_α בהגדרה נבחר בזמן ע"פ **הלא מטיבה** שמעניינת אותנו.

דוגמא: $X \sim N(\mu, \sigma^2)$ (σ^2 ידוע).

$H_0: \mu = 0$, $H_A: \mu > 0$.

נשתמש באומד $\bar{X}$, מה המתאים תחת H_0?

$\bar{X} \overset{H_0}{\sim} N(0, \frac{\sigma^2}{n})$



u.multinet.co.il