72+2

הפקולטה למדעים מדויקים
החוג למדעי המחשב
אוניברסיטת תל אביב

מועד א'
סמסטר א', תשס"ט
12.2.2009

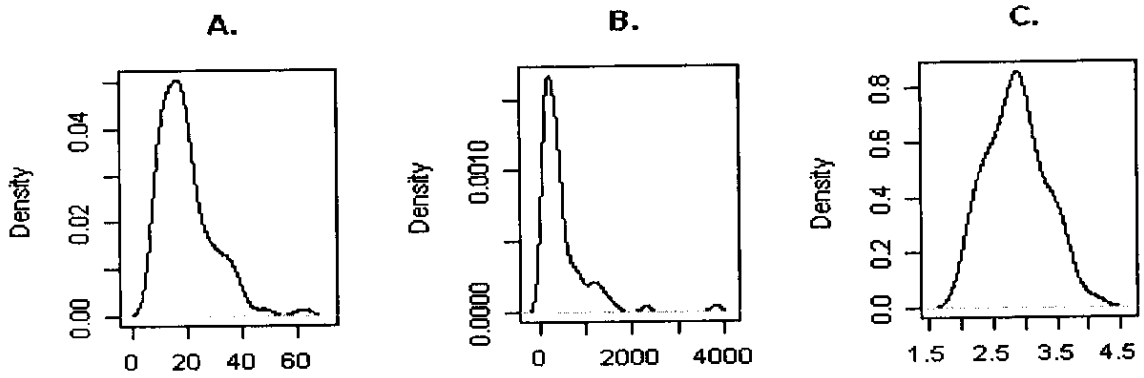
סטטיסטיקה למדעי המחשב
מרצה: ד"ר סהרון רוסט

1	6
2	5
3	33
4	15
5	13
*	2

משך הבחינה 3 שעות. הינכם רשאים להביא עמכם מחשבון בלבד. דפי נוסחאות וטבלאות יסופקו עם הבחינה. עליכם לענות על כל השאלות בגוף השאלון. הקפידו על תשובות קצרות וברורות. המחברת תשמש לחישובים בלבד.
סעיפים המסומנים ב(*) הם סעיפי בונוס והמבחן מסתכם ל100 נקודות בלעדיהם.
בהצלחה!

מס' מחברת: 47

שאלה 1 (10): נתונות שלוש הצגות של מחיריהן של 93 דגמי מכוניות בארה"ב (בדולרים, נכון לשנת 1993). בחלון A מוצגת החלקת חלון של המחירים המקוריים, בB של המחירים אחרי העלאה בריבוע, ובC של המחירים אחרי טרנספורמציית log.



א. אם נתון שערכי תמצית פירסון לנטייה עבור שלושת הגרפים הם 0.25, 1.48, 3.29, שייכו כל אחד מהערכים לאחד מהגרפים והסבירו בקצרה.

יזכור כי תמצית T נקראת אי-שיוך: $T(\bar{x}) \approx 0$ כאשר אין שיוך לאף כיוון, ושיוך סימטרי.
לפיכך: 0.25 - מתאים לגרף C שכן לפי הגרף הנתון כיותר מאין השפעה, וגרף C הוא הקרוב ביותר לסימטריה. ✓
1.48 - מתאים לגרף A שכן הנטיה בו ימינה אחלה מאשר ב-C, וכן לפי הסקאלה ניתן לראות שקיבוי מאשר ב-B. מה שמובן!
3.29 - מתאים לגרף B בו הנטיה ימינה הכאלה בולטת מבין השלושה.

ב. מה ניתן ללמוד על התפלגות מחירי מכוניות מכך שהגרף בC (בסקאלת לוג) נראה כמעט סימטרי, בניגוד לגרף בA?
הדרכה: נסו להימנע משימוש במונח לוג ולדבר במונחים שיהיו מובנים לאדם ברחוב.

ניתן ללמוד שמחירי הרכבים מתפלגים בקירוב אחיד-לוגריתמית, ✗
כלומר היתה שכיחות דומה לרמות שונות בקירוב.
לפיכך ניתן להניח שרוב מחירי המכוניות הם סביב ערך מסוים (הממוצע)
ומעט מאוד גדולים או מאוד יקרים.
גרף A הינו מסיים, שכן בו נראה שישנה נטיה ימינה, כלומר בין ערך

שאלה 2 (15) : בדיקה חדשה וזולה לHIV מפורסמת כ"מדויקת ב99%". העלון המצורף מסביר שהכוונה ש99% מהנשאים מזוהים ככאלה ו99% מהבריאים מזוהים ככאלה.

א. מה הם שני סוגי הטעויות שהמבחן יכול לעשות? כתבו אותן כהסתברויות מותנות.

(1) טעות מסוג I : המבחן עלול לזהות אדם בריא כחולה, כאשר למעשה הוא אולה. ההסתברות המותנית (נסמן אדם כ- X המקבל H לבריא ו- S לחולה ונסמן תוצאת בדיקה כ- Y המקבל H לשלילי ו- S לחיובי):

$$P(Y=S \mid X=H)$$

(2) טעות מסוג II : המבחן עלול לזהות אדם חולה כבריא, כלומר לשלוח אדם חולה הביתה. ההסתברות המותנית המתאים אותה ושלם סימונים יהיה: $P(Y=H \mid X=S)$

ב. אנחנו בוחרים מדגם מקרי של 50 אנשים שהוכרזו נשאים על ידי הבדיקה ובודקים אותם שוב בבדיקה יקרה שדיוקה 100%. אנו מגלים ש45 מתוכם למעשה בריאים. כתבו את האמירה הזו כהסתברות מותנית אמפירית.

נסמן את מספר החולים האמיתי במ"מ X , המקבל ערכים $0 \le k \le 50$ (לצורך).

נסמן את מספר התולים שהבדיקה מגלה במ"מ Y המקבל ערכים $0 \le l \le 50$ (לצורך).

ההסתברות המותנית היא: $P(Y=l \mid X=k)$

כאשר המקרה לה $l=5$ (5 חולים בפועל) ו- $l=50$ (תוצאות הבדיקה).

ג. הניחו ששתי הסתברויות הטעות בסעיף א שוות ל1% בדיוק. האם התוצאה בסעיף ב אפשרית? נמקו בקצרה.

התשובה בסעיף ג' לא יכולה להתקיים, מה שאר הסתברויות הן ההסתברות לדיוק.

למעשה גם בסעיף ב' הבדיקה לא מדויקת, ויש שאנו חולים הנשארים בחיים.

הבעיה של הבדיקה היא, נתיחס את מספר הנשים החיוביות על חולים שהבדיקה לחלים:

50 איש ← 45 בריאים: מהרכב שכזה של 5 חולים ו45 בריאים, וכאשר כולם נשלחים הביתה

נשלח לחולה ונניח שכולם בריאים ש-5 לכן יש לנו 1 חולה.

5 חולים: נניח ואין טעות (למרות שכזה הוא לא) וכל החולים ילכו הביתה. המשך בעמוד הבא (קלר...)

(א) הסברו את משמעות הסיכוי/ההסתברות (5/50) כאמור, ואת הטעות שאתם מוצאים בה.

ד. (*+5 נקודות) בהינתן שני מבחנים א,ב, צרו דוגמא לפרדוקס של סימפסון שבה התניה על תוצאות מבחן א מייצרת אשליה כאילו מבחן ב אינו עוזר לזהות נשאים, בעוד שהמסקנה ללא ההתניה היא שונה.

המשך לסעיף ג':

נניח היתכן 6 אנשים יגיעו בתוצאה לפי הנתונים הנ"ל, ולכן
⇐ לא יתכן חלק מן שצוהו 25 תוצים, שאינו האמיתי של התוצאה בטוחות
הוא כפי שנתנך בהבדלה מינימלי ב-1000 (5 מתוך 50). לכן תוצאה סעיף
ב' לא נכונה.

שאלה 3 (35): חוק הַאבל (Hubble) באסטרונומיה קובע שמהירות התרחקותן v של גלקסיות זו מזו פרופורציונלי למרחקן d. פרמטר היחס ידוע בשם קבוע האבל והוא מסומן בH. בשאלה הזו אנו מבקשים להעריך אותו.
להלן נתונים על שתי גלקסיות (ליתר דיוק, צבירי גלקסיות): מרחקן משביל החלב (הגלקסיה שלנו) ומהירות ההתרחקות שלהן מאיתנו.

צביר	מרחק d (מיליוני שנות אור)	מהירות v (אלפי ק"מ/שניה)
וירגו	22	1.13
ג'מיני	405	23.2

נניח שהמרחקים d_i ידועים בדיוק מראש (דהיינו הם מספרים) ואילו המהירויות v_i הן **תצפיות מקריות** הנמדדות עם רעש: $v_i = Hd_i+\varepsilon_i$ עבור $i=1,2$, כאשר ε_i הן שגיאות בלתי תלויות עם תוחלת אפס.

אנו שוקלים שני אומדים אפשריים עבור הפרמטר H:

$$\hat{H}_1 = \frac{d_1 v_1 + d_2 v_2}{d_1^{\,2} + d_2^{\,2}} \qquad \hat{H}_2 = 0.5 \cdot \left(\frac{v_1}{d_1} + \frac{v_2}{d_2} \right)$$

א. בהנחה שהשגיאות ε_i מתפלגות $N(0,\sigma^2)$ (ולכן $v_i \sim N(Hd_i,\sigma^2)$), כתבו את לוג פונקצית הנראות של H במונחים של d_1, d_2, v_1, v_2.

$$L(H; d_1, d_2, v_1, v_2) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left\{-\frac{(v_1 - Hd_1)^2}{2\sigma^2}\right\} \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left\{-\frac{(v_2 - Hd_2)^2}{2\sigma^2}\right\} = \frac{1}{2\pi\sigma^2} \cdot \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{2}(v_i - Hd_i)^2\right\}$$

$$\Rightarrow \ell(H; d_1, d_2, v_1, v_2) = \log\left(\frac{1}{2\pi\sigma^2}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{2}(v_i - Hd_i)^2 = -\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\cdot\left[v_1^2 - 2Hv_1d_1 + H^2d_1^2 + v_2^2 - 2Hv_2d_2 + H^2d_2^2\right] =$$

$$= -\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left[v_1^2 + v_2^2\right] + H\left[\frac{v_1 d_1 + v_2 d_2}{\sigma^2}\right] - H^2\left[\frac{d_1^2 + d_2^2}{2\sigma^2}\right]$$

ב. גזרו אותו והסיקו שאחד משני האומדים לעיל הוא אומד נראות מקסימלי במצב זה (איזה מהם?).

$$\frac{\partial \ell(H)}{\partial H} = 0 + 0 + \frac{v_1 d_1 + v_2 d_2}{\sigma^2} - \frac{2(d_1^2 + d_2^2)}{2\sigma^2} H$$

$$\frac{\partial \ell(H)}{\partial H} = 0 \iff v_1 d_1 + v_2 d_2 = H(d_1^2 + d_2^2) \iff \boxed{H = \frac{v_1 d_1 + v_2 d_2}{d_1^2 + d_2^2}}$$

(שכן $\sigma^2 > 0$ ו...)

$$\frac{\partial^2 \ell(H)}{\partial H^2} = -\frac{d_1^2 + d_2^2}{\sigma^2} < 0$$

ולכן $\hat{H}_1$ הוא אומד ל-H.

ג. האם האומדים מוטים?

(1) $E(\hat{H}_1) = E\left(\frac{d_1 v_1 + d_2 v_2}{d_1^2 + d_2^2}\right) = \frac{1}{d_1^2 + d_2^2}\cdot\left(d_1 E(v_1) + d_2 E(v_2)\right) = \frac{d_1 \cdot Hd_1 + d_2 Hd_2}{d_1^2 + d_2^2} = \frac{d_1^2 + d_2^2}{d_1^2 + d_2^2} H = H$

$\Rightarrow \hat{H}_1$ אומד חסר הטיה.

(2) $E(\hat{H}_2) = E\left(\frac{1}{2}\left[\frac{v_1}{d_1} + \frac{v_2}{d_2}\right]\right) = \frac{1}{2}\left(\frac{1}{d_1}E(v_1) + \frac{1}{d_2}E(v_1)\right) = \frac{1}{2}\left(\frac{Hd_1}{d_1} + \frac{Hd_2}{d_2}\right) = \frac{2H}{2} = H$

$\Rightarrow \hat{H}_2$ גם כן אומד חסר הטיה.

ד. מהן השונויות שלהם? איזו גדולה יותר?

(1) $\mathrm{Var}(\hat{H}_1) = \mathrm{Var}\left(\frac{d_1 v_1 + d_2 v_2}{d_1^2+d_2^2}\right) = \frac{1}{(d_1^2+d_2^2)^2}\left(d_1^2\,\mathrm{var}(v_1) + d_2^2\,\mathrm{var}(v_2)\right) = \frac{d_1^2+d_2^2}{(d_1^2+d_2^2)^2}\sigma^2 = \frac{\sigma^2}{d_1^2+d_2^2}$

(2) $\mathrm{var}(\hat{H}_2) = \mathrm{var}\left(\frac{1}{2}\left[\frac{v_1}{d_1}+\frac{v_2}{d_2}\right]\right) = \frac{1}{4}\left[\frac{1}{d_1^2}\mathrm{var}(v_1) + \frac{1}{d_2^2}\mathrm{var}(v_2)\right] = \frac{d_1^2+d_2^2}{4d_1^2d_2^2}\sigma^2$

נשאיר את ההפרש: $\mathrm{var}(\hat{H}_2)$ פחות המונה של $\mathrm{var}(\hat{H}_1)$: $0 \overset{?}{<} d_1^4 - 2d_1^2d_2^2 + d_2^4$

עבור $d_1 = d_2$ נקבל שוויון ולכן $d_1 \neq d_2$ נקבל אי־שוויון אשר מוסיף לנו שהמונה של $\mathrm{var}(\hat{H}_1)$ קטן ממנו (ולכן $\hat{H}_1$ עדיף).

ה. על סמך סעיפים א-ד, איזה משני האומדים עדיף? ציינו שתי סיבות לבחירתכם.

משני האומדים הנתונים עדיף $\hat{H}_1$ בהיותו ... ושתי סיבות:

- השונות שלו קטנה יותר מזו של $\hat{H}_2$, כלומר הוא יותר מדויק (פחות ביזור).
- הוא אומד נראות מקסימלי לפרמטר H.

ו. חשבו את הערך של שני האומדים על הנתונים שלנו.
הערה: האומדים הם ביחידות מהירות שכן המרחק נמדד בשנות אור והמהירות בק"מ לשנייה. כדי להגיע למדידה ביחידות מקובלות יש להכפיל בקבועי המרה עצומים, ואנו חוסכים זאת מכם כאן.

$$\hat{H}_1 = \frac{22\cdot 1.13 + 405\cdot 23.2}{22^2+405^2} \cong 0.0572$$

$$\hat{H}_2 = \frac{1}{2}\left(\frac{1.13}{22} + \frac{23.2}{405}\right) \cong 0.0543$$

ז. בהנחה שהקצב אכן קבוע, 1/H הוא הזמן שעבר מאז המפץ הגדול. האם $1/\hat{H}_1$ או $1/\hat{H}_2$ הוא אומד חסר הטיה לזמן זה? האם אחד מהם הוא אומד נראות מקסימלי?

$\hat{H}_1$ ו־$\hat{H}_2$, שניהם ~~שגויים~~ תסרי הטיה של H, ומכאן ש־$\frac{1}{\hat{H}_1}$ ו־$\frac{1}{\hat{H}_2}$ הם אומדנים תסרי הטיה של $\frac{1}{H}$.

- $\frac{1}{\hat{H}_1}$ הוא אומד של $\frac{1}{H}$ כיוון שהפונקציה ממנה הוא מתקבל $\frac{1}{x}$ היא פונקציה על הקטע $(0,\infty)$ ולכן, כיון שמקבלים אי־שוויון (הפוכה), ותהי אומד נראות מקסימלי של H - $\frac{1}{H}$.

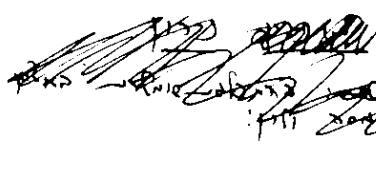
שאלה 4 (20): ציינו נכון/לא נכון ונמקו בקצרה (תשובות ללא נימוק תקף לא יקבלו נקודות):

א. מבחן א ומבחן ב בודקים את אותה השערת **אפס**. אזי אם מבחן א דוחה בכל מקרה שמבחן ב דוחה, אזי מבחן א הוא בעל עצמה גבוהה יותר. לא נכון
(או שווה)

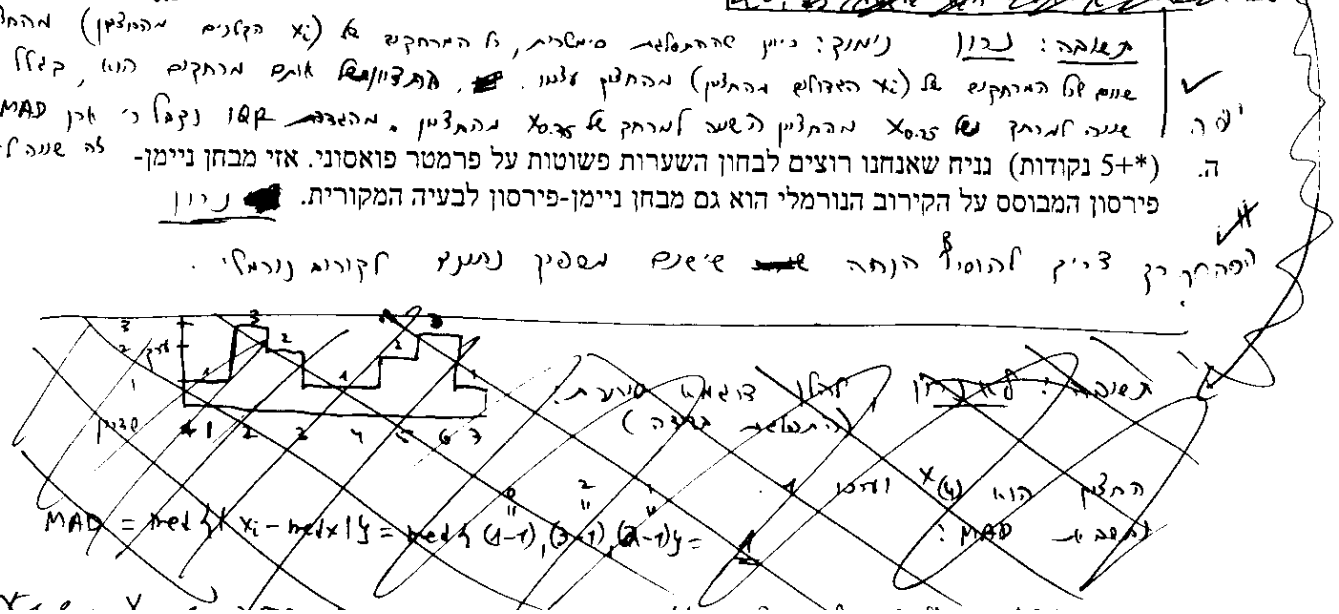
ב. נניח $X \sim N(\mu, \sigma_X^2)$, ואנו עושים הסקה בייזיאנית על התוחלת הנורמלית μ, תוך שימוש בהתפלגות מקדימה על μ של $N(0, \sigma_\mu^2)$. אזי לשונות ההתפלגות המקדימה σ_μ^2 אין השפעה על תוחלת האומד הבייזיאני הא-פוסטריורי. לא נכון

ג. בונים רווח סמך ברמה 95% לתוחלת נורמלית, פעם בהנחת שונות ידועה σ^2 ושימוש בZ, ופעם בהנחת שונות לא ידועה --- אומדים את σ^2 ובונים רווח סמך תוך שימוש בערך t המתאים. אזי רווח הסמך הראשון יהיה קטן יותר. נכון

ד. אם התפלגות היא סימטרית אזי MAD שווה למחצית הטווח הבין רבעוני IQR (בדיוק על ההתפלגות התיאורטית ובערך על מדגם).



(שייתכן עם הפלגן)



שאלה 5 (20): בארץ גיבורי הילדות התמודדו בבחירות המועמדים קיפי ובילבי. בסקר ראשון שנערך בקרב מדגם מקרי של 600 מצביעים, תמכו בקיפי 40% ובבילבי 60% מתוכם. אז הכריזה קיפי מלחמה על ארץ חלאס השכנה, שיושביה הטרידו את גיבורי הילדות. לאחר המלחמה נערך מדגם מקרי נוסף של 600 מצביעים, וכעת תמכו בקיפי 45% מתוכם, והשאר בבילבי.

א. ברצוננו לבחון את השערת האפס שהמלחמה לא השפיעה על התמיכה בקיפי (מול האלטרנטיבה שהעלתה התמיכה). נסחו את ההשערות ובצעו את המבחן הבינומי המתאים ברמה 0.05. מה המסקנה?

יהי p שיעור התמיכה בקיפי לאחר המלחמה:

$H_0: p = 0.4$ (שיעור התמיכה בקיפי בטרם המלחמה (שהוא 0.4 בהתבסס על סקר ו))

$H_1: p > 0.4$

המבחן הבינומי המדויק:

$C_\alpha = [p_0 + z_{\alpha} \cdot \sqrt{\frac{p_0(1-p_0)}{n}}, \infty) = [0.4 + 1.645 \cdot \sqrt{\frac{0.4 \cdot 0.6}{600}}, \infty) = [0.4329, \infty)$

כיוון ש- $\beta = 0.45$ גבוה מדי, ולכן נדחה את H_1,

למרות המסקנה הזו:

ברמת מובהק של 5%, המלחמה הגבירה את שיעור התמיכה בקיפי באופן מובהק (שיעור התמיכה עלה).

ב. מכון הסקרים "מין הצמח" העדיף לבדוק את השערת האפס לעיל באמצעות מבחן חי-בריבוע לטבלת 2*2 המתאימה. בצעו את המבחן והראו שמסקנתו ברמה 0.05 שונה מזו של סעיף א.

נבדוק האם יש קשר בין תמיכה בקיפי/בילבי ובין לפני/אחרי המלחמה:

obs:

	תמיכה בקיפי	תמיכה בבילבי	סה"כ
לפני	240	360	600
אחרי	270	330	600
סה"כ	510	690	1200

נחשב ערכים exp לפי ההתפלגות השולית:

	קיפי	בילבי	סה"כ
לפני	255	345	600
אחרי	255	345	600
סה"כ	510	690	1200

כל אחד מהם גדול מ-5 ולכן ניתן להשתמש במבחן χ^2, במקום במבחן z בסעיף א.

$$\sum \frac{(O_{ij} - e_{ij})^2}{e_{ij}} = 0.882 + 0.652 + 0.882 + 0.652 = 3.069 \sim \chi^2_{1,\alpha}$$

כאשר אנו בודקים במקרה זה: $\chi^2 < 3.84$ ל

(עם $\alpha = 0.05$)

← אפשר להתקבל שקשקוט לכן ניתן שאת H_0 (כלומר שיעור התמיכה לא השתנה)

(המסקנה, כלומר שיעור התמיכה בקיפי לא השתנה ברמה של 95%).

ג. הסבירו בקצרה את שני ההבדלים העיקריים בין המבחנים בסעיפים א וב. אילו מהמבחנים מתאים יותר לסיטואציה המתוארת?
רמז: אחד מההבדלים נוגע להנחות על המודל הסטטיסטי (אילו גדלים ידועים מראש ואילו מקריים) ואחד להשערות הנבדקות.

ד. (*+5 נקודות) נניח כעת ששני המדגמים בסעיף א הם על אותם 600 מצביעים. דונו בקצרה איך ניתן כעת לבחון את השערת האפס במדגם מזווג כזה.
רמז והצעה: כדי לענות נכון על שאלה זו תצטרכו לפתח הרחבות מעבר לדברים שלמדנו בכיתה. לכן מומלץ לנסות אותה רק אם זמנכם בידכם.