אופרה אהרון   בס״ד

סיכום סטטיסטיקה - הגדרות

* שכיחות יחסית $\frac{f(x)}{n}$ * שכיחות מצטברת $F(x)=\sum f(t)$ * שכיחות יחסית מצטברת $\frac{F(x)}{n}$
* נקודת אמצע של מחלקה: ממוצע של גבול עליון ותחתון
* רוחב המחלקה: ההפרש של גבול עליון ותחתון (בגבולות אמיתיים).

מדדי מרכז
* שכיח - Mode - הערך שמופיע הכי הרבה פעמים.
* אמצע הטווח - Midrange - ממוצע בין הערך הגבוה לנמוך.
* חציון - Median - הערך שנמצא באמצע כאשר מסדרים את הערכים בסדר עולה.
* ממוצע - Mean - סכום הערכים חלקי הכמות

· חציון: $L_0+(L_1-L_0)\cdot\frac{n/2-F(x_{m-1})}{f(x_m)}$ = Median (n = מס' התצפיות הכולל, x_m = מחלקת החציון, L_0, L_1 = גבול המחלקה תחתון/עליון).
- במצב כזה של תצפיות. כאשר יש תצפיות חריגות ונמצאות במחלקות שונות - החציון הוא המדד המועדף.

מדדי פיזור
* יחס השונות - מס' התצפיות שאינן במחלקת השכיח.
* טווח - ההפרש בין הערך המקסימלי למינימלי.
* ממוצע הסטיות מהממוצע - סכום המרחקים מהממוצע $\frac{\sum|x_i-\bar{x}|}{n}$
* שונות $s^2=\frac{\sum|x_i-\bar{x}|^2}{n}\cdot f(n)$ * סטיית תקן $s=\sqrt{s^2}=\sqrt{\frac{\sum(x_i-\bar{x})^2}{n}\cdot f(n)}$
* אחוזון p - הערך הx כך שp אחוזים מערכי ההתפלגות נמצאים מתחתיו.
* טווח בין רבעוני - נחלק ל4 רבעים ונסתכל על $x_{75}-x_{25}$ שזה הב' האמצעיים של הנתונים.
* ציון תקן - מספר מיקום יחסי של תצפית יחידה בהתפלגות שלה $z_i=\frac{x_i-\bar{x}}{s}$

הערה! יש לשים את כל התצפיות בצורה לינארית (מספר/כפל בקבוע) - הממוצע, החציון והשכיח גם ישתנו. הסטיית תקן תשתנה כפול הקבוע בערך מוחלט.
- אמצע הטווח והממוצע מושפעים מערכים קיצוניים.
- השכיח והחציון נשענים יותר על שכיחות התצפיות השונות, הממוצע ואמצע הטווח נשענים על ערכיהם.

* סטטיסטי - פונק' של תצפיות המדגם $\hat{\theta}=f(x_1,\dots,x_n)$ - נקרא גם אומד.
- אומדן - הערך עצמו אחרי שחישבנו את האומד.
- הטיה - הטיה של אומד זה התוחלת של האומד פחות האומד עצמו $bias(\hat{\theta})=E(\hat{\theta})-\theta$
- אומד $\hat{\theta}$ הוא אומד חסר הטיה אם $E(\hat{\theta})=\theta$ - אם התוחלת שווה לאומד עצמו = ההטיה שווה 0.
- חוסר הטיה - הערכים שהאומד יכול לקבל יהיו קרובים למה שאנו רוצים למצוא.
- אם התוחלת ידועה השונות הם אומד חסר הטיה עבור σ^2, אם התוחלת אינה ידועה נבחר $\frac{\sum(x_i-\bar{x})^2}{n-1}$
- כאשר נתונים לנו שני אומדים חסרי הטיה נבחר באומד עם השונות המינימלית יותר. (ממוצע - אומד חסר הטיה לתוחלת).

* סטטיסטי = תצפיות = מדגם = $\bar{x}, s$ * פרמטר = הסקית = אוכלוסיה = μ, σ
- התכנסות - ככל שהמדגם גדל יותר, הערכים של האומד יתקרבו יותר ויותר לתוחלת.

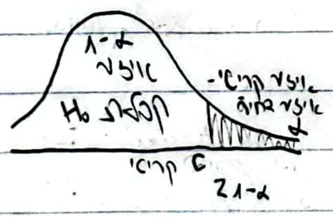
סטטיסטיקה - סיכום

* שיטת הנראות המירבית - אנחנו נרצה לדעת מי מהערכים שנבחר לפרמטר יתן בגלל ההסתברות בצורה הגבוהה לקבל את תוצאות המדגם הנצפות.
נגדיר פונק' $L(x_1 ... x_n, \theta)$ שמקבלת את הנתונים והפרמטר ומחזירה מכפלה של הצפיפות.
נחפש את הפונק' L שמחזירה ערך מקסימלי, כדי למצוא מקסימום נחשב את $\ln L(x,\theta)$
נגזור לפי θ, (ואם מבקשים) נחשב נגזרת שניה ונבדוק שהיא שלילית (שזה הערך מקסימלי).

* התפלגות T
$T=\frac{Z}{\sqrt{\frac{V}{k}}}\sim t_k$ $\qquad$ $T=\frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}}\sim t_{n-1}$
כאשר $V\sim\chi^2_k$, $Z\sim N(0,1)$

* רווח סמך - הרעיון הוא לקבוע רווח (אינטרוול) כך שההסתברות שמכיל בתוכו את הפרמטר נמצא מקבל מאוד. אם נוכל מבין שאומר הוא לא מדגם כשהוא לא על כל האוכלוסיה.
אנו רוצים לתת "מרווח בטחון" כלומר טווח שבו נוכל לומר כמה בוודאות שהפרמטר נמצא.
- רמת הבטחון / הסמך $1-\alpha$ (ההסתברות שהפרמטר אכן בטווח שמצאנו).
- משפט הגבול המרכזי, יש $X\sim N(\mu,\sigma^2)$ ו-σ ידועה אז: $P(\bar{X}-c<\mu<\bar{X}+c)=1-\alpha$
- אורך רווח הסמך $= 2c$ $\qquad 2\cdot Z_{1-\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}}$ (ככל שn גדל המרווח יקטן).
- כאשר מגדילים את רמת הבטחון אנחנו מגדילים את רווח הסמך
רווח סמך לפרופורציה: $V(\hat{p})=\frac{pq}{n}$, $E(\hat{p})=p$, לפי משפט הגבול המרכזי: $\frac{\hat{p}-p}{\sqrt{\frac{pq}{n}}}\sim N(0,1)$
- α וβ משפיעים זה את זה, לכן צריך למצוא מבחן ששומר את שניהם מינימליים.
- טעות באומדן = ההפרש המקסימלי בין $\hat{p}$ לp = חצי רווח הסמך (c).
- α - רמת המובהקות, ההסתברות לטעות באומדן.

* בדיקת השערות סטטיסטיות:
השערת האפס - מה שהיה ידוע עד היום, השערה אלטרנטיבית - החידוש צריך לשכנע את הקהל.
- α היא ההסתברות לטעות בהחלטת החוקר לדחות את H_0
כאשר H_0 נכונה. α קובע את רמת המובהקות.
ככל שα קטנה יותר - התוצאה מובהקת יותר.
- c מייצג את הערך הקריטי המבדיל בין אזור קבלה לדחיה.



$C=\mu_0+Z_{1-\alpha}\cdot\frac{\sigma}{\sqrt{n}}$
הטעויות האפשריות:
I טעות מסוג ראשון - לדחות את השערת האפס למרות שהיא נכונה. ההסתברות לטעות בו היא α.
II טעות מסוג שני - לקבל את השערת האפס למרות שההשערה האלטרנטיבית נכונה. ההסתברות β.
$\beta=P(H_0|H_1)$, $\alpha=P(H_1|H_0)$
- $1-\beta$ * עוצמת המבחן (דחיית H_0 בצדק) $\qquad$ $1-\alpha$ * רמת הבטחון (אי דחיית H_0 בצדק).

סטטיסטיקה - סיכום

* מובהקות - P-value מובהקות התוצאה, ה-α המינימלית שעבורה נדחה את H_0.
כאשר $\alpha > P\text{-value}$ נדחה את H_0.

- במבחן ימני חישוב ה-PV הוא: $PV = 1-\Phi(z_0)$, במבחן שמאלי חישוב ה-PV: $PV=\Phi(z_0)$
- במבחן דו צדדי חישוב הערך הוא: $PV = 2\cdot(1-\Phi(|z_0|))$

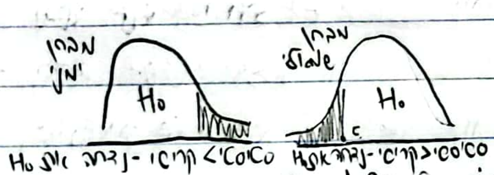
חישוב מובהקות p value: א) נחשב את ציון התקן $z=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$ ב) במבחן ימני $PV=\alpha_{min}=1-\Phi(z_0)$

* הפרש תוחלות - (מדגמים בזוגות) - משווים בין ממוצע של אוכלוסיות כאשר לא שונויות ידועות ויש הפרש.

* מבחן על יחס שונויות - נבחר את s_1^2 להיות השונות הגדולה יותר.
מבחן 2 כיווני: $F_{n_1-1,n_2-1;1-\frac{\alpha}{2}}=\frac{1}{F_{n_2-1,n_1-1,\frac{\alpha}{2}}}$

* התפלגות χ^2 - בודקים את ההשערה מתאימה להתפלגות מסוימת. $\chi^2=\sum\frac{(O_i-E_i)^2}{E_i}$ (סטטיסטי)
כאשר: O_i - מס' התצפיות בפועל, E_i - מס' התצפיות שאמורות להתקיים לפי מודל מסוים.
אם $\chi^2_{\text{סטטיסטי}} > \chi^2_{\alpha,k-1\ \text{קריטי}}$ נדחה את H_0

לזכור!



* תמיד כשמשרטטים יש לתת את הקריטי, ואז הסטטיסטי
גדול מימין או נחליט לדחות את H_0 (במבחן שמאלי זה הפוך).

* בתוכנת SPSS הפלט מראה לנו את הסטטיסטי ו/עם את המובהקות P-value.

* בהתפלגות פואסון יש לנו נתונה ה-λ וצריך לחשב אותה במקרה זה נוריד ב דרגות חופש. (כמו אם נחשב ממוצע, שונות וכו').

* כאשר $n\leq5$ נאחד נתונים בטבלה לשונות את הדרגות חופש בהתאם.

* ניתוח שונויות - בודקים הבדלים של קבוצות מסוימות באוכלוסיה.
ככל שהשונות בתוך הקבוצות קטנה והשונות בין הקבוצות גדולה או נדחה את H_0.

ממוצע הכללי SST	בין הקבוצות SSB	בתוך הקבוצה SSW
$n-1$	$k-1$	$n-k$ מס' התצפיות, מס' הקבוצות

$SST = SSB + SSW$ $\quad F=\frac{MSB}{MSW}$ $\quad MSB=\frac{SSB}{k-1}$ $\quad MSW=\frac{SSW}{n-k}$

כלומר $SST=(n-1)S^2$

ככל ש-F_e גדול יותר כך סביר יותר שההבדל בין הקבוצות במדגם משקף הבדל בין הקבוצות באוכלוסיה.

$SST=\sum y_{ij}^2-\frac{T^2}{N}$ $\quad SSB=\sum\frac{T_i^2}{n_i}-\frac{T^2}{N}$ כאשר סכ T_i, מס' איברים בכל n_i

המשך סיכום -

* רגרסיה לינארית - רוצים לבדוק האם קיים קשר כלשהו בין 2 משתנים.
כלומר אנחנו מחפשים את העוצמה והכיוון של הקשר בין 2 משתנים.
המשוואה $Y = \hat{b}x + \hat{a} + \varepsilon$ (נרצה ש-ε יהיה כמה שיותר קטן. המקדם של x הוא b!!!).
כאשר y תלוי ב-x, x בלתי תלוי; השערת האפס היא שאין קשר בין המשתנים
- סך הסטיות של התצפיות מהישר הוא 0. סך ריבועי הסטיות הוא מינימלי.
- הסטטיסטי:

$$F_{סטטיסטי} = \frac{SSR/1}{SSE/n-2} = \frac{(n-2)SSR}{SSE} = \frac{(n-2)\cdot R^2}{1-R^2}$$

נקבל את H_0 כאשר $F_{סטטיסטי} < F_{\alpha(1,n-2)}$

ניבוי/תחזיה $r^2 = R^2$; $\frac{SSE}{SST}$ = השונות הבלתי מוסברת ; $\frac{SSR}{SST} = R^2$ השונות המוסברת

r - מקדם המתאם של פירסון

כאשר $b \neq 0$ אז יש קשר בין x ל-y, כלומר הרגרסיה מובהקת.

$$\underbrace{\sum(y_i-\bar{y})^2}_{\text{סה"כ הפיזור}} = \underbrace{\sum(y_i-\hat{y}_i)}_{\text{הפיזור הבלתי מוסבר ע"י הרגרסיה}} + \underbrace{\sum(\hat{y}_i-\bar{y})^2}_{\text{הפיזור המוסבר ע"י הרגרסיה}}$$

$$SST = SSE + SSR$$

מתקיים $-1 \le r \le 1$ כאשר $|r|$ קרוב ל-1 זה סימן שעוצמת הקשר גבוהה.

* הסבר לפלט spss:
- בפלט של רגרסיה השורה הראשונה זה a (1 constant) (הדרגת חופש היא 1)
השורה השנייה זה b (x) (הדרגת חופש היא n-2). לפי המשוואה! $\hat{Y} = bx + a$
- ה-sig מראה לנו את ה-p-value שזה ה-α המינימלי שעבורו נדחה את H_0,
כלומר כאשר $\alpha > p\text{-value}$ נדחה את השערת האפס.
- ניתוח פלט של ניתוח שונות:

		Sum Of Squers	df (דרגות החופש)	Mean Square (ממוצעים, $\frac{SS}{df}$)	F (הסטטיסטי)	Sig (P-value)
SSB	Betwin group (בין הממוצעים של מדגם לממוצע הכללי.)	225.75	k-1 3	75.25	.745	.541
SSW	within groups (בין כל ערך לממוצע המדגם שלו)	1616.8	N-k 16	101.05	$F = \frac{\frac{225.75}{3}}{\frac{1616.8}{16}} = \frac{75.25}{101.05}$	
SST	Total (בין כל ערך לממוצע הכללי)	1842.55	N-1 19	—		

נשים לב: N - גודל המדגם, k - מס' הקבוצות, מתקיים השוויון: $SST = SSW + SSB$

$$F_{סטטיסטי} = \frac{\frac{\text{Sum Of Squers SSB}}{\text{df SSB}}}{\frac{\text{Sum Of Squers SSW}}{\text{df SSW}}} = \frac{\text{Mean Square SSB}}{\text{Mean Square SSW}}$$

- כאשר יש לנו פלט ולא ניתן להסיק מסקנה לפי הפלט (α = p-value) אז נבדוק לפי ה-F הקריטי.
- לא ניתן להסיק מנתוני מדגם על הפרופורציה כולה כיוון ש-$\bar{x}$, s הם רק של מדגם והם שונים מהסטיית התקן/השונות/התוחלת (μ) הכלליות.
- התוכנה נותנת לנו מבחן דו כיווני, במבחן חד כיווני נחלק את ה-sig ב-2.

סטטיסטיקה - דברים שכדאי לזכור

- רמת מובהקות = אלפא α , מובהקות התוצאה = p-value
- אלפא וביתא מושפעים זה מזה וכאשר מקטינים אחד מגדילים את השני, לכן צריך למצוא מקום שישמור את שניהם מינימליים.
- ברירת המחדל כאשר $\hat{p}$ לא ידוע היא לקחת $\hat{p}=0.5$ (תוצאה הכי מחמירה).
- אם מקטינים את אורך הרווח פי a, צריך להגדיל את המדגם פי a^2
- טעות האומדן = חצי רווח הסמך (המרחק המקסימלי בין $\hat{p}$ ל-p).
- כאשר מוסיפים מס' קבוע לממוצע, החציון והשכיח משתנים בהתאם לאותו מספר, הס"ת תקן לא משתנה. כאשר מוסיפים אחוז מעלים לכולם הממוצע, החציון והשכיח גדלים באותו אחוז. הס"ת תקן משתנה.
- רווח הסמך תלוי בגודל המדגם ולא בממוצע.
- בהתפלגות t כל פרמטר שנחשב ידוע מהתוצאות הנתונות יוסיף לנו 1 (כלומר נצטרך להוריד דרגת חופש נוספת). לדוג' בהתפלגות פואסון אם לא נתונה למדא אז שצריך לחשב אותה (אז שנתון שחשבנו) אז נוריד 2 בדרגות חופש (במקום 1).
- בהתפלגות כי (χ^2) כאשר $n<5$ נאחד נתונים בטבלה ונשנה את הדרגות חופש בהתאם. E_i חייב להיות גדול או שווה 5, אחרת נחבר טווחים.
- $\hat{p}$ מציין את המצב שיצא לנו, P_0 זה הממוצע לפי השערת ה-H_0.
- בהיסטוגרמה ציר הx זה הנתונים, ציר הy זה הצפיפות שזה $f(x)$/רוחב המחלקה שכיחות....

בס"ד

איך ניגשים לשאלה:

* בשאלות של רגרסיה - סוכמים את כל הx-ים ואת כל הy-ים
וכן עושים סכימה של מכפלת x-ים עם y-ים, ועוד סכימה של x-ים בריבוע.
ואז מציבים בנוסחאות. מוצאים את $\hat{b}$ ולאחר מכן באמצעות המשוואה $\hat{a}=\bar{y}-\hat{b}\bar{x}$
מוצאים את a ואז מוצאים את משוואת הרגרסיה, לזכור המקדם של x הוא b ולא a.
בדר"כ השאלה היא להציב נתון בx ולבדוק את התוצאה.
* בשאלות של ניתוח שונות - עושים טבלה עם כל הנתונים, ומחשבים את n_i
ו-T_i (n_i מבטא את מס' התצפיות, T_i מבטא את סכומם). וכן מחשבים את K, N, T
(כאשר k-זה מס' המדגמים, N- מס' התצפיות, T- סכום הערכים). וכמובן מציבים
את הנתונים בנוסחאות. משתמשים בכך ש $SST=SSB+SSW$ וכן נחשב את הF.
* בשאלות של בדיקת טיב התאמה - מחשבים את הE_i, יש לזכור שכאשר הוא קטן
מ-5 מאחדים עמודות בטבלה וכן מורידים דרגת חופש נוספת. בנוסף כאשר לא נתון
מחשבים נתון כמו λ וכד' אז גם מורידים דרגת חופש נוספת (בנוסף לדרגה
שמורידים בכל מקרה). מציבים את הנתונים בנוסחה ומשווים את הχ^2 שיצא לנו
עם הχ^2 הקריטי ודוחים/מקבלים השערה.
* בשאלות של התפלגות מצב ואוכלוסיה - שצריך להשלים נתונים חסרים, לזכור!
קודם כל לזכור שהסתברויות אמורים להיות שוות את כל התשובות. בשכיחות יחסית מצטברת
$F(x)/n$ המצטבר הוא 1.
* בשאלות שנתונה פונקציית צפיפות - צריך לחשב אומד נראות מרבית לפרמטר.
נרשום פונקציית נראות לפרמטר לפי מה שנתון $L(\theta)=\Pi(x_1,x_2...x_n,\theta)$
לאחר מכן מפעילים ln על המשוואה כדי לפשט את הפונקציה ואז גוזרים אותה
לפי המשתנה (במקרה שלנו θ). יש רוצים לבדוק נקודת מקסימום נגזור שוב ונראה
שהנגזרת שלילית. כדי לבדוק שאומד הוא חסר הטיה מראים $E[\hat{\theta}]=\theta$
* בשאלות של רמת בטחון/רווח סמך - מציבים לפי הנוסחה המתאימה, צריך לזכור
שאורך רווח הסמך $=2c=2\cdot Z_{1-\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}}$
* בשאלות שצריך לחשב את רמת מובהקות (α) - נשתמש בכך ש $\alpha=P(H_1|H_0)$
ונציב בהתאם לאמור. בהתפלגות נורמלית נרמול: $P(\bar{x}>x_0)=P\left(\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}>\frac{x_0-\mu_i}{\sigma_i/\sqrt{n_i}}\right)$
בהתאם לנתוני השאלה.

סיכום סטטיסטיקה - איך ניגשים לשאלות.

* בשאלות של לוח שכיחויות: חשוב נתונים לפי האמצעים גבולות אמיתיים!

תזכורות כיצד מחשבים:

- טווח: $R = X_{max} - X_{min}$ • אמצע טווח: $MR = \frac{1}{2}[X_{max} - X_{min}]$ (מושפעים מערכי קיצון!).
- טווח בין רבעוני: $C_{75} - C_{25}$ (טווח רבעוני וחציון לא מושפעים מערכי קיצון).
- במחלקות שונות רוחב השכיח לא רלוונטי.
- כאשר צריך לשרטט היסטוגרמה ציר הX מורכב מגבולות אמיתיים, ציר הY עד $f(x)/\Delta x$ כלומר שכיחות המחלקה לחלק לרוחב, וכך לצבע שכיחות ($f(x)$) = שטח.
- כאשר צריך לשרטט מצולע, כל מחלקה מיוצגת ע"י אמצע המחלקה וציר הY הוא $f(x)$ (שטח השכיחות).
- n (גודל המדגם) שווה לסכום השכיחויות. כלומר סכום העמודה $f(x)$.
- כאשר השכיח מוגדר ניתן להסיק שהמחלקות שוות רוחב.
- כדי לחשב ציון תקן צריך לחשב ממוצע וסטיית תקן (בנוסחאות). ואז להציב בנוסחת ציון תקן את הנתונים.
- סטיית תקן $= S^2$, שונות $= S = \sqrt{S^2}$, כלומר סטיית התקן היא שורש השונות.

* בשאלות של חי"ב התאמה - לשים לב!!

- כאשר מחשבים את ערכי ה Expected לא לשכוח להכפיל ב n.
- כאשר $n < 5$ לשים לב לאחד קטגוריות ולשנות את ה df בהתאם.
- כאשר נתון אומדן או שמחשבים משהו ע"ב להוריד דרגת חופש נוספת.

* לשים לב כאשר מבקשים תחום של $\bar{X}$ בבדיקת השערות להשוות לפי הבדיקה והתחום הוא $\mu_0 \pm \frac{\sigma}{\sqrt{n}} \cdot Z_{1-\frac{\alpha}{2}} < \bar{X}$

- מרכז רווח בר סמך לפרופורציה הוא הפרופורציה המדגמית. ($\frac{f(x)}{n}$)
- מרכז רווח בר סמך לתוחלת הוא ממוצע המדגם.

* לא לשכוח שבנוסחת המחושבת נותן מציבים את $f(x)$ ו $F(x)$ (שכיחות/שכיחות מצטברת) ולכן לפעמים צריך להכפיל או לחלק ב n. (אם נתון $\frac{f(x)}{n}$ וכו').