דף נוסחאות / סיכום

מבוא לסטטיסטיקה
קורס
094480

יש להדפיס במדפסת לייזר
ברזולוציה של (לפחות) DPI 600

דף עזר מספר 1:

תורת הסתברות

הגדרות:

- ניסוי מקרי - הוא ניסוי שאין לדעת בוודאות מראש את תוצאתו אלא קיים מרחב של תוצאות אפשריות.
- מרחב מדגם Ω - הקבוצה הכוללת – אוסף כל התוצאות האפשריות של הניסוי המקרי.
- מאורע פשוט ω - תוצאה בודדת של מרחב המדגם (איבר ב- Ω).
- מאורע A $(B, C, \ldots)$ - אוסף של מאורעות פשוטים (קבוצה חלקית ל- Ω).
- קבוצה חלקית (מוכלת) - $A \subset B$: תת-קבוצה של B (A מוכלת ב- B) אם כל איבר של A הוא איבר של B.
- אומרים שמאורע A קורה אם תוצאת הניסוי (שהיא אחת מנקודות המרחב) שייכת ל-A .
- מאורע משלים למאורע A - A^c - כל האיברים ב- Ω שאינם שייכים ל- A.
- מאורע ריק $\varnothing$ (קבוצה ריקה) - אינו כולל אף איבר.

פעולות בין מאורעות:

1. איחוד - $A \cup B$ - כל האיברים שהם או ב- A או ב- B או בשתי הקבוצות.
2. חיתוך - $A \cap B$ - כל האיברים שהם גם ב- A, וגם ב- B .

אם החיתוך בין הקבוצות הוא ריק, כלומר אין לנו איברים משותפים, אזי הקבוצות נקראות זרות.

מאורע A והמאורע המשלים לו A^c הם מאורעות זרים:
$$A \cap A^c = \varnothing$$
האיחוד בין מאורעות A ו- A^c הוא מרחב המדגם:
$$A \cup A^c = \Omega$$

3. הפרש - $A \setminus B$ - כל האיברים שהם ב- A אך לא ב- B:
$$A \setminus B = A \cap B^c$$

כללי פעולות בין הקבוצות:

חוק החילוף:
$$A \cup B = B \cup A$$
$$A \cap B = B \cap A$$

חוק הפילוג:
$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$
$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

חוקי דה-מורגן:
$$(A \cap B)^c = A^c \cup B^C$$
$$(A \cup B)^c = A^c \cap B^C$$

מבוא לסטטיסטיקה והסתברות
דף עזר 2

משתנים מקריים

משתנה מקרי – נתון מרחב ההסתברות. הפונקציה X = x(w) המתאימה לכל נקודה w במרחב המדגם Ω נקראת **מ"מ** (משתנה מקרי)

משתנה מקרי בדיד - מ"מ X יקרא **מ"מ בדיד** אם הוא מקבל סדרה סופית או בת מניה של ערכים.

פונקציית ההסתברות עבור מ"מ בדיד:

הפונקציה $x \in R,\ P(X = x) = P_x(x)$ נקראת פונקציית ההסתברות עבור מ"מ X.

הערה: $\sum_x P_x(x) = 1$

פונקציית התפלגות מצטברת עבור מ"מ בדיד:

הפונקציה $F_x(x) = \sum_{x_i \le x} P_x(x_i) = P(X \le x)$, המוגדרת על ידי :

(הסתברות שמ"מ X קטן או שווה למספר x), נקראת פונקציית התפלגות מצטברת של מ"מ X בנק' x.

תכונות פונקציית ההתפלגות:

1. $0 \le F_X(x) \le 1$ לכל $x \in \mathbb{R}$.
2. הפונקציה $F_X(x)$ מונוטונית לא יורדת ב- x, כלומר: אם $x_1 < x_2$, אז $F_X(x_1) \le F_X(x_2)$.
3. $\lim_{x_0 \to \infty} F_X(x_0) = 1$, $\lim_{x_0 \to -\infty} F_X(x_0) = 0$.

הערה: אם נתונה $P_X(x)$ ניתן לחשב ממנה את $F_X(x)$.

אם נתונה $F_X(x)$: $P_X(x) = F_X(x) - F_X(x-1)$, כלומר, פונקציית ההסתברות שווה לפונקצית ההתפלגות פחות פונקצית ההתפלגות בנקודה שלפני.

מבוא להסתברות וסטטיסטיקה
דף עזר 3

פונקציית ההתפלגות ברנולי עם פרמטר p

מבצעים ניסוי ברנולי עם הסתברות p להצלחה. X מקבל 1 אם הניסוי הצליח ו- 0 אחרת. אזי :

$X \sim Ber(p)$ ומסמנים : $p_X(x) = \begin{cases} p & x=1 \\ q & x=0 \\ 0 & \text{otherwise} \end{cases}$

פונקציית ההתפלגות גיאומטרית עם פרמטר p

X מ"מ שסופר את מספר הניסויים עד להצלחה הראשונה בסדרת ניסויי ברנולי בלתי תלויים עם הסתברות p להצלחה, אזי :

$X \sim Geo(p)$ ומסמנים : $p_X(x) = \begin{cases} q^{x-1}p & x=1,2,3,... \\ 0 & \text{otherwise} \end{cases}$

אם $X \sim Geo(p)$ (גיאומטרי) אז התוחלת שלו היא : $E(X) = 1/p$

השונות שלו היא: $Var(X) = q/p^2$

תכונת חוסר זיכרון להתפלגות גיאומטרית

אם ידוע כי לא הייתה הצלחה עד לניסיון ה-m אזי : ההסתברות שההצלחה הראשונה תתרחש בניסיון ה-m+k היא : $q^{k-1}p$

מודל גיאומטרי של ניסויים (אחת הדרכים לתאר את ההתפלגות הגיאומטרית)

מבצעים סדרה של ניסויי ברנולי ב"ת עד להצלחה הראשונה. ההסתברות שיהיו בדיוק k ניסיונות, ז"א ההסתברות שההצלחה הראשונה תתרחש בניסיון ה-k היא: $q^{k-1}p$

פונקציית ההתפלגות בינומית עם פרמטרים n,p

X מ"מ שסופר את מספר ההצלחות בn- ניסויי ברנולי **בלתי תלויים** עם הסתברות p להצלחה, אזי :

$X \sim Bin(n,p)$ ומסמנים : $p_X(x) = \begin{cases} \binom{n}{x} p^x q^{n-x} & x=0,1,...,n \\ 0 & \text{otherwise} \end{cases}$

אם X~Bin(n,p) (בינומי) אז התוחלת שלו היא : $E(X) = n \cdot p$

השונות שלו היא: $Var(X) = n \cdot p \cdot q$

מודל בינומי של ניסויים (אחת הדרכים לתאר את ההתפלגות הבינומית)

מבצעים סדרה של n ניסויי ברנולי ב"ת , בעלי הסתברות להצלחה p.

ההסתברות לקבל בדיוק k הצלחות בסדרה היא P(k)

תזכורת:

עבור X בדיד : $F_X(x) = \sum_{t \le x} P(X = t)$

עבור X רציף : $F_X(x) = \int_{-\infty}^{x} f_X(t)dt$

תוחלת של מ"מ רציף:

$E(X) = \int_{-\infty}^{\infty} x \cdot f_X(x)dx$

תכונות התוחלת של מ"מ בדיד נכונות גם עבור מ"מ רציף פרט ל:

$E(g(X)) = \int_{-\infty}^{\infty} g(x) \cdot f_X(x)dx$

שונות של מ"מ רציף:

$Var(X) = E(X^2) - E^2(X)$

התפלגות אחידה רציפה:

סימון : X~U(a,b) (a<b)

$f_X(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & \text{otherwise} \end{cases}$

$E(X) = \frac{a+b}{2}$ $Var(X) = \frac{(b-a)^2}{12}$

התפלגות מעריכית (אקספוננציאלית):

סימון : X~exp(λ) (λ נקרא "הקצב" של התפלגות)

$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \ge 0 \\ 0 & \text{otherwise} \end{cases}$

$E(X) = \frac{1}{\lambda}$ $Var(X) = \frac{1}{\lambda^2}$

$P(X > a) = \int_a^{\infty} f_X(x)dx = \int_a^{\infty} \lambda e^{-\lambda x}\,dx = e^{-\lambda a}$

תכונת **חוסר הזיכרון** של משתנה מקרי אקספוננציאלי :

$P(X > t+s \mid X > t) = P(X > s) = e^{-\lambda s}$

כלומר, אם נתון כי עד רגע t (כולל) לא היה מופע, אזי "מתחילים מהתחלה", ולכן זמן עד המופע הבא הוא exp(λ).

התפלגות נורמלית:

$X \sim N(\mu, \sigma^2)$

$f_X(x) = \frac{1}{\sqrt{2\pi}\cdot\sigma} e^{\frac{(x-\mu)^2}{2\sigma^2}}$ $-\infty < x < \infty$

$E(X) = \mu$

$Var(X) = \sigma^2$

מקרה פרטי של התפלגות נורמלית – מ"מ נורמלי סטנדרטי: $Z \sim N(0,1)$ - מ"מ מתפלג נורמלית עם תוחלת $\mu = 0$ ושונות $\sigma^2 = 1$

חישוב הסתברות בהתפלגות נורמלית (תיקון משתנה נורמלי – מעבר להתפלגות נורמלית סטנדרטית):

אם $X \sim N(\mu, \sigma^2)$, אזי $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$

חישוב הסתברות בהתפלגות נורמלית סטנדרטית:

$P(Z < z) = \phi(z)$

מאחר שההסתברות סימטרית סביב ה-0:
$P(Z \ge z) = 1 - P(Z < z) = 1 - \phi(z) = \phi(-z) = P(Z < -z)$

סימון:
z_α - ערך של Z שמשמאלו שטח (הסתברות) של α

הערה:
$z_\alpha = -z_{1-\alpha}$ מתקיים בגלל הסימטריה של התפלגות נורמלית סטנדרטית

חישוב ערך X כעשר α ידוע ו- $X \sim N(\mu, \sigma^2)$:

$X \sim N(\mu, \sigma^2)$, $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$,

$X_\alpha = \mu + z_\alpha * \sigma$

מבוא להסתברות וסטטיסטיקה
דף עזר 4

נניח כי יש לנו מדגם X_i , $i = 1, \ldots, n$.

1. חציון – הערך שמחצית המקרים קטנים ממנו או שווים לו:
   - עבור $n = 2k+1$ אי-זוגי $MED = X_{(k+1)}$;
   - עבור $n = 2k$ זוגי $MED = \frac{X_{(k)} + X_{(k+1)}}{2}$.
2. רבעון ראשון/תחתון – הערך שרבע מהמקרים קטנים ממנו או שווים לו: $Q_1 = X_{\left(\left\lfloor \frac{n}{4} \right\rfloor + 1\right)}$.
3. רבעון שני – חציון.
4. רבעון שלישי/עליון – הערך ששלושה רבעים מהמקרים קטנים ממנו או שווים לו: $Q_3 = X_{\left(\left\lfloor \frac{3n}{4} \right\rfloor + 1\right)}$.
5. תחום בין-רבעוני – בו מרוכזות 50% מהתצפיות המרכזיות: $IQR = Q_3 - Q_1$.

10. השברון ה- p - הערך שפרופורציה p מהתצפיות קטנות ממנו או שווים לו: $X_p = X_{\left(\lfloor np \rfloor + 1\right)}$.

משתנה מקרי רציף:

פונקציית צפיפות
משתנה מקרי רציף מאופיין ע"י פונקציית צפיפות $f_X(x)$ המקיימת:

1. $f_X(x) \ge 0$
2. $\int_{-\infty}^{\infty} f_X(t)dt = 1$

עבור משתנה מקרי רציף

1. $P(X = x) = 0$ (כלומר, ההסתברות לקבל ערך נקודתי מסויים של מ"מ רציף שווה ל-0)
2. $P(a \le X \le b) = \int_a^b f_X(x)dx$

פונקציית התפלגות מצטברת

יהי X מ"מ. פונק' התפלגות מצטברת של X היא פונקציה $F_X : R \to [0,1]$ המוגדרת ע"י $F_X(x) = P(X \le x)$

תוחלת של מ"מ בדיד

יהי X מ"מ המקבל ערכים $x_1, x_2, \ldots x_n$ בהסתברויות $P_X(x_1), P_X(x_2), \ldots, P_X(x_n)$ בהתאמה,

אזי התוחלת של X מוגדרת ע"י: $E(X) = \sum_{x \in \Omega} x_i P_X(x)$, כאשר Ω היא מרחב המצבים.

תכונות התוחלת:

1. עבור קבוע a , $E(a) = a$
2. $E(aX) = aE(X)$
3. $E(X+a) = E(X) + a$
4. $E(aX+b) = aE(X) + b$
5. $Eg(X) = \sum g(X) \cdot P(X = x)$
6. אם $X_1, \ldots X_k$ סדרה של מ"מ כלשהם אזי: $E(X_1 + X_2 + \ldots X_k) = E(X_1) + \ldots + E(X_n)$

שונות

שונות הנה מדד לפיזור. שונות של מ"מ מסומנת על ידי $Var(X)$ או על ידי $V(X)$.

הגדרה תיאורטית של שונות: $\text{var}(X) = E\left[X - E(X)\right]^2$

נוסחת העבודה: $\boxed{Var(X) = E(X^2) - E^2(X)}$

תכונות של שונות:

1. $Var(X) \ge 0$
2. $Var(a) = 0$
3. $Var(a \cdot X) = a^2 \cdot Var(X)$
4. $Var(a + X) = Var(X)$
5. אם X ו- Y מ"מ **בלתי תלויים** אזי $Var(X+Y) = Var(X) + Var(Y)$
6. אם X ו- Y מ"מ **בלתי תלויים** אזי $Var(X-Y) = Var(X) + Var(Y)$
7. אם X ו- Y מ"מ **תלויים** אזי $Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$
8. אם X ו- Y מ"מ **תלויים** אזי $Var(X-Y) = Var(X) + Var(Y) - 2Cov(X,Y)$

סטיית תקן

$\boxed{\sigma(X) = \sqrt{Var(X)}}$

פונקציית ההתפלגות הגיאומטרית עם פרמטר p

פונקציית ההתפלגות פואסונית עם פרמטר λ

X - מ"מ פואסוני מונה אירועים לאורך זמן :

$X \sim Pois(\lambda)$ ומסמנים : $p_X(x) = \begin{cases} \frac{e^{-\lambda}\lambda^x}{x!} & x = 0,1,2,... \\ 0 & otherwise \end{cases}$

אם $X \sim Pois(\lambda)$ (פואסוני) אז התוחלת שלו היא : $E(X) = \lambda$

השונות שלו היא: $Var(X) = \lambda$

התפלגות אחידה Unif(a,b) עבור מ"מ בדיד

ישנם N ערכים $(a, a+1, a+2, \ldots, b-2, b-1, b)$ עם צעד 1 שאותם מקבל מ"מ בדיד X. ההסתברות לקבל ערך כלשהו מתוך ה-n הערכים האלה היא

$X \sim Unif(a,b)$ ומסמנים : $P_X(x) = \begin{cases} \frac{1}{(b-a+1)} = \frac{1}{N} & x = a, a+1, a+2, \ldots, b-1, b \\ 0 & otherwise \end{cases}$

אם $X \sim Unif(a,b)$ (אחיד) אז התוחלת ושונות שלו היא :

$E(X) = \frac{(a+b)}{2}$

$Var(X) = \frac{(b-a)(b-a+2)}{12}$

במילים אחרות זה מרחב המדגם שווה ההסתברות

מקרה פרטי כאשר X מקבל ערכים (1,2,...N).

$P_X(x) = \begin{cases} \frac{1}{N} & x = 1,2,\ldots N \\ 0 & otherwise \end{cases}$

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

$Var(X) = \frac{(N^2-1)}{12}$

פונקציית ההתפלגות הבינומית עם פרמטרים n,p

פונקציית ההתפלגות ההיפרגיאומטרית עם פרמטרים N,R,n

נתונה אוכלוסיה בת N איברים שמתוכם R מיוחדים. בוחרים מהאוכלוסיה n איברים ללא החזרה.

X – מ"מ המונה את מספר האיברים המיוחדים שנבחרו.

$X \sim HG(N,R,n)$ ומסמנים : $P_X(x) = \begin{cases} \frac{\binom{R}{r}\binom{N-R}{n-r}}{\binom{N}{n}} & 0 \le r \le n \\ 0 & otherwise \end{cases}$

דף עזר 5

משפט הגבול המרכזי

יהיו $X_1,X_2,\ldots,X_n$ מ"מ ב"ת ש"ה עם תוחלת μ ושונות σ^2.

נגדיר $\overline{X}_n=\frac{1}{n}\sum_{i=1}^{n}X_i$, אזי:

$$\frac{\overline{X}_n-\mu}{\sigma/\sqrt{n}}\underset{n\to\infty}{\Rightarrow}N(0,1)$$, ואת אומרת $\overline{X_n}$ בקירוב מתפלג כמו משתנה מקרי מהתפלגות

$$P(\frac{\overline{X}_n-\mu}{\sigma/\sqrt{n}}\le a)\xrightarrow[n\to\infty]{}\Phi(a)$$ כלומר $N(\mu,\frac{\sigma^2}{n})$

באופן שקול אם נגדיר $S_n=\sum_{i=1}^{n}X_i$, אזי: $\frac{S_n-n\mu}{\sqrt{n}\sigma}\underset{n\to\infty}{\Rightarrow}N(0,1)$

S_n מתפלג בקירוב כמו משתנה מקרי נורמלי עם הפרמטרים: $N(n\mu,n\sigma^2)$

קירוב נורמלי להתפלגות בינומית

אם $X\sim Bin(n,p)$, אזי לפי משפט הגבול המרכזי, עבור n מספיק גדול (np>5):

$$X\sim N(np,npq)\Rightarrow Z=\frac{X-np}{\sqrt{npq}}$$

פתרון תרגיל כיתה 8

רווח סמך לשונות:

$$\frac{S^2(n-1)}{\sigma^2}=\frac{\sum(X_i-\overline{X})^2}{\sigma^2}\sim\chi^2(n-1)$$

מכאן: $$\frac{S^2(n-1)}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}\le\sigma^2\le\frac{S^2(n-1)}{\chi^2_{\frac{\alpha}{2}}(n-1)}$$

רווח סמך להפרש תוחלות במדגמים בלתי-תלויים:

$$\overline{X}\sim N\left(\mu_x,\frac{\sigma_x^2}{n_x}\right),\quad \overline{Y}\sim N\left(\mu_y,\frac{\sigma_y^2}{n_y}\right)$$

1. כאשר השונויות של שתי האוכלוסיות ידועות:

$$\overline{X}-\overline{Y}\sim N\left(\mu_x-\mu_y,\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}\right)\Rightarrow\frac{(\overline{X}-\overline{Y})-(\mu_x-\mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}}}\sim N(0,1).$$

ולכן,

$$\mu_x-\mu_y\in\left[\overline{X}-\overline{Y}\pm z_{1-\frac{\alpha}{2}}\sqrt{\left(\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}\right)}\right]$$

2. כאשר שונויות לא ידועות:

נניח: לשתי האוכלוסיות אותה שונות. את השונות נאמוד על-ידי:

$$S_p^2=\frac{(n_x-1)S_x^2+(n_y-1)S_y^2}{n_x+n_y-2},$$

כאשר:

$$S_x^2=\frac{\sum_{i=1}^{n_x}(x_i-\overline{X})^2}{n_x-1}\quad S_y^2=\frac{\sum_{i=1}^{n_y}(y_i-\overline{Y})^2}{n_y-1}$$

מתקיים:

$$\frac{(\overline{X}-\overline{Y})-(\mu_x-\mu_y)}{\sqrt{S_p^2\left(\frac{1}{n_x}+\frac{1}{n_y}\right)}}\sim t_{(n_x+n_y-2)}.$$

ולכן,

$$\mu_x-\mu_y\in\left[\overline{X}-\overline{Y}\pm t^{|-\alpha/2}_{(n_x+n_y-2)}\sqrt{S_p^2\left(\frac{1}{n_x}+\frac{1}{n_y}\right)}\right]$$

- לכל סוג של השערה אלטרנטיבית נקבע **אזור הדחייה** שאם הסטטיסטי המדגמי נופל בו, המסקנה תהיה **לדחות את H_0**. נהוג לסמן אזור דחייה ב- C.
- האזור המשלים נקרא **אזור הקבלה**. אם הסטטיסטי המדגמי נופל בו המסקנה תהיה **לא לדחות את H_0**.

- עבור השערה חד-צדדית ימנית אנו נדחה עבור ערכים גדולים של הסטטיסטי.
- עבור השערה חד-צדדית שמאלית אנו נדחה עבור ערכים קטנים של הסטטיסטי.
- עבור השערה דו-צדדית אנו נדחה עבור ערכים קטנים או גדולים של הסטטיסטי.

ערכים קריטיים לאלטרנטיבות שונות:

מערכת ההשערות:	אזור דחייה:

מבחן השערות על תוחלת μ כאשר השונות ידועה:

$H_0:\mu=\mu_0\quad H_1:\mu>\mu_0$	$Z=\frac{\overline{X}-\mu_0}{\sigma_0/\sqrt{n}}>z_{1-\alpha}$
$H_0:\mu=\mu_0\quad H_1:\mu<\mu_0$	$Z=\frac{\overline{X}-\mu_0}{\sigma_0/\sqrt{n}}<z_{\alpha}$
$H_0:\mu=\mu_0\quad H_1:\mu\neq\mu_0$	$\lvert Z\rvert=\left\lvert\frac{\overline{X}-\mu_0}{\sigma_0/\sqrt{n}}\right\rvert>z_{1-\alpha/2}$

מבחן השערות על תוחלת μ כאשר השונות איננה ידועה:

$H_0:\mu=\mu_0\quad H_1:\mu>\mu_0$	$T=\frac{\overline{X}-\mu_0}{S/\sqrt{n}}>t^{n-1}_{1-\alpha}$
$H_0:\mu=\mu_0\quad H_1:\mu<\mu_0$	$T=\frac{\overline{X}-\mu_0}{S/\sqrt{n}}<t^{n-1}_{\alpha}$
$H_0:\mu=\mu_0\quad H_1:\mu\neq\mu_0$	$\lvert T\rvert=\left\lvert\frac{\overline{X}-\mu_0}{S/\sqrt{n}}\right\rvert>t^{n-1}_{1-\alpha/2}$

דף עזר 6

אמידה נקודתית

אוכלוסייה – אוסף פריטים עליהם מעוניינים במידע מסוים.

מדגם – אוסף חלקי של פריטים מהאוכלוסייה.

פרמטר – מספר קבוע המאפיין את האוכלוסייה (θ).

סטטיסטיקה – ערך הניתן לחישוב מנתוני המדגם.

אמד – סטטיסטיקה המשמש לאמידת פרמטר ($\hat{\theta}$).

אומדן – הערך המספרי של האמד.

אמד $\hat{\theta}$ ייקרא אמד חסר הטיה לפרמטר θ, אם קיים:

$$E(\hat{\theta})=\theta$$

אחרת, הוא אמד מוטה והטיתו ניתנת ע"י:

$$Bias(\hat{\theta})=E(\hat{\theta})-\theta$$

נעדיף אמד חסר הטיה על אמד מוטה.

שגיאה ריבועית ממוצעת (MSE):

$$MSE(\hat{\theta})=E(\hat{\theta}-\theta)^2=Var(\hat{\theta})+[Bias(\hat{\theta})]^2$$

נעדיף את האמד בעל MSE מינימלית.

הערה: אם $\hat{\theta}$ הוא אמד חסר הטיה לפרמטר θ, קיים:

$$MSE(\hat{\theta})=Var(\hat{\theta})\Leftarrow Bias(\hat{\theta})=0$$

סטיית התקן של הממוצע כאשר השונות של X ידועה:

$$SE_{\overline{X}}=\frac{\sigma_X}{\sqrt{n}}$$

סטיית התקן של הממוצע כאשר השונות של X לא ידועה:

$$SE_{\overline{X}}=\frac{S}{\sqrt{n}}$$

כאשר $$S^2=\frac{\sum_{i=1}^{n}(x_i-\overline{X})^2}{n-1}=\frac{\sum_{i=1}^{n}x_i^2-n\overline{X}^2}{n-1}$$

רווח סמך להפרש תוחלות כאשר המדגמים מזווגים:

$$t^{(n-1)}_{\frac{\alpha}{2}}\le\frac{\overline{d}-\delta}{SE(\overline{d})}\le t^{(n-1)}_{1-\frac{\alpha}{2}}$$

כאשר:

$$s_d=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(d_i-\overline{d})^2}\qquad SE(d)=\frac{\overline{s}_d}{\sqrt{n}}$$

מכאן:

$$\delta\in[\overline{d}\pm t^{(n-1)}_{1-\frac{\alpha}{2}}SE(\overline{d})]$$

רווח סמך להפרש פרופורציות:

באותו אופן כמו בהפרש תוחלות:

$$p_1-p_2\in[\hat{p}_1-\hat{p}_2\pm Z_{1-\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}]$$

מבחן השערות על פרופורציה:

$H_0:p=p_0\quad H_1:p>p_0$	$Z=\frac{\hat{p}-p_0}{\sqrt{\frac{p_0q_0}{n}}}>z_{1-\alpha}$
$H_0:p=p_0\quad H_1:p<p_0$	$Z=\frac{\hat{p}-p_0}{\sqrt{\frac{p_0q_0}{n}}}<z_{1-\alpha}$
$H_0:p=p_0\quad H_1:p\neq p_0$	$\lvert Z\rvert=\left\lvert\frac{\hat{p}-p_0}{\sqrt{\frac{p_0q_0}{n}}}\right\rvert>z_{1-\alpha/2}$

P-value value

הגדרה: ההסתברות לקבל תוצאה מדגמית כמו שקיבלנו או יותר מובהקת (קיצונית לכיוון H_1), תחת ההנחה ש H_0 נכונה.

הגדרה נוספת: בהינתן ערכי המדגם, ה- α המקסימלית עבורה עדיין מקבלים את השערת האפס או ה- α המינימלית עבורה עדיין דוחים את השערת האפס.

דף עזר 7

אמידה מרווחית

רמת הסמך (רמת הביטחון) תסומן $1-\alpha$, כאשר α – רמת המובהקות

רווח בר-סמך לתוחלת μ של אוכלוסייה כאשר שונות σ^2 ידועה:

$$\overline{X}\sim N\left(\mu,\frac{\sigma^2}{n}\right)\Rightarrow Z=\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1).$$

ולכן,

$$\mu\in\left[\overline{X}\pm Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

זהו רווח בר-סמך לתוחלת μ ברמת ביטחון $1-\alpha$.

רווח בר-סמך לתוחלת μ של אוכלוסייה כאשר שונות σ^2 אינה ידועה:

נאמוד את השונות בעזרת סטטיסטי S^2 המחושב מהמדגם:

$$S^2=\frac{\sum_{i=1}^{n}(x_i-\overline{X})^2}{n-1}=\frac{\sum_{i=1}^{n}x_i^2-n\overline{X}^2}{n-1}.$$

מתקיים:

$$T=\frac{\overline{X}-\mu}{S/\sqrt{n}}\sim t_{(n-1)},$$

כאשר $(n-1)$ - מספר דרגות חופש – פרמטר בהתפלגות $t_{(n-1)}$ ("ויסטודנט"); n הוא גודל המדגם שעליו מבוסס הממוצע.

ולכן,

$$\mu\in\left[\overline{X}\pm t^{1-\alpha/2}_{n-1}\frac{S}{\sqrt{n}}\right]$$

זהו רווח בר-סמך לתוחלת μ ברמת ביטחון $1-\alpha$.

רווח סמך ל – p - פרופורציה באוכלוסייה:

$$\hat{p}=\frac{x}{n}$$ - הפרופורציה במדגם. x - הנו מספר בעלי התכונה במדגם בגודל n.

עבור n ו – p מספיק גדולים מתקיים בקירוב:

$$\hat{p}\sim N\left(p,\frac{p\cdot(1-p)}{n}\right)\Rightarrow\frac{\hat{p}-p}{\sqrt{\frac{p\cdot(1-p)}{n}}}\sim N(0,1)$$

ולכן-

רווח הסמך:

$$p\in\left[\hat{p}\pm z_{1-\frac{\alpha}{2}}\cdot\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

דף עזר 9

בדיקת השערות:

המטרה: רוצים להחליט על נכונותה או אי – נכונותה של השערה מסוימת לגבי פרמטרים של התפלגויות.

ההחלטה אם לדחות או לא לדחות את ההשערה נעשית על סמך המדגם, ובעזרת כלל הכרעה הקובע לכל תוצאת מדגם אפשרית לדחות או לא לדחות את ההשערה.

השערות:

- **השערת האפס** H_0 - ההשערה אותה מעוניינים לבדוק.
- **ההשערה האלטרנטיבית** H_1 - ההשערה המנוגדת.

האפשרויות צריכות להיות ממצות ומוציאות.

סוגי השערות:

- **השערה פשוטה** – מגדירה באופן חד ערכי את ההתפלגות
- **השערה מורכבת** – כוללת הרבה אפשרויות
- **השערה מורכבת חד-צדדית** (דוגמה: $\begin{cases}H_0:\mu=800\\H_1:\mu<800\end{cases}$)
- **השערה מורכבת דו-צדדית** (דוגמה: $\begin{cases}H_0:\mu=800\\H_1:\mu\neq800\end{cases}$)

ההחלטות האפשריות:

- **דחיית** H_0 - זוהי החלטה האומרת שהשערת האפס איננה נכונה.
- **אי דחיית** H_0 (קבלת H_1) – החלטה האומרת שהשערת האפס איננה ניתנת והיא כנראה נכונה.

טעויות:

α - הסתברות לטעות מסוג ראשון או רמת מובהקות

$\alpha=P(\text{reject the null hypothesis}\mid\text{the null hypothesis is correct})$

β - הסתברות לטעות מסוג שני

$\beta=P(\text{don't reject the null hypothesis}\mid\text{the alternative hypothesis is correct})$

$\pi=1-\beta$ - עוצמת המבחן

לסיכום: $$\pi=P_{H_1}\left\{\text{Reject}\mid H_0\right\}\quad\beta=P_{H_1}\left\{\text{accept}\mid H_0\right\}\quad\alpha=P_{H_0}\left\{\text{Reject}\mid H_0\right\}$$

דף עזר 11

מבחני טיב התאמה:

המטרה לבדוק האם המדגם שייך להתפלגות מסוימת.

$H_0:X\sim F$

$H_1:otherwise$

סטטיסטי המבחן עבור מבחני טיב התאמה הינו:

$$\chi_P^2=\sum_{i=1}^{k}\frac{(n_i-np_{i0})^2}{np_{i0}}=\sum_{i=1}^{k}\frac{(O_i-E_i)^2}{E_i}$$

- n_i - מספר התצפיות בקטגוריה i
- k – מספר הקטגוריות

אזור דחייה ברמת מובהקות α

נדחה את H_0 אם $\chi_P^2>\chi^2_{1-\alpha}(k-l-1)$

- l - מספר הפרמטרים שנאמדו

בדיקת אי תלות:

נלקח מדגם מסוים מתוך אוכלוסייה המסווגת לפי שני משתנים איכותיים (קטגוריאליים).

המטרה היא לבדוק האם שני המשתנים הינם בלתי תלויים.

A – משתנה קטגוריאלי ראשון (שורות)

B – משתנה קטגוריאלי שני (עמודות)

$H_0:P(A_i\cap B_j)=P(A_i)\cdot P(B_j)$

$H_1:otherwise$

סטטיסטי המבחן עבור מבחני אי תלות:

$$\chi_P^2=\sum_{i=1}^{k}\sum\frac{(n_{ij}-n\hat{p}_{i\cdot}\hat{p}_{\cdot j})^2}{n\hat{p}_{i\cdot}\hat{p}_{\cdot j}}=\sum_i\sum_j\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

- $\hat{p}_{i\cdot}=\frac{n_{i\cdot}}{n}$
- $\hat{p}_{\cdot j}=\frac{n_{\cdot j}}{n}$
- $n_{i\cdot}$ - מספר התצפיות בעלות ערך i במשתנה A
- $n_{\cdot j}$ - מספר התצפיות בעלות ערך j במשתנה B

אזור דחייה ברמת מובהקות α

נדחה את H_0 אם $\chi_P^2>\chi^2_{1-\alpha}(r-1)(c-1)$

- r – מספר הקטגוריות במשתנה A
- c – מספר הקטגוריות במשתנה B

דף עזר 12 - רגרסיה לינארית

Y – משתנה תלוי / משתנה מוסבר.

X – משתנה בלתי-תלוי / משתנה מסביר.

לפעמים יש מספר משתנים מסבירים (p משתנים מסבירים) ואז הם מסומנים: $X_1, X_2, ..., X_p$

המטרה רוצים להסביר את Y בעזרת המשתנה/ים המסבירי/ם ולתת תחזיות לציון Y על סמך ציון ה-X.

שלבים:

1. ציור דיאגרמת פיזור: מדיאגרמת הפיזור נלמד האם קיים קשר לינארי בין X ל- Y.
2. מציאת קו הרגרסיה: $Y = aX + b$ כאשר a - שיפוע הקו ו b - חותך הקו
3. חיזוי Y על-פי X ועל-פי קו הרגרסיה.

סימונים:

$$S_{XY} = \sum (X_i - \bar{X})(Y_i - \bar{Y}) = \sum XY - n\bar{X}\bar{Y}$$
$$S_{XX} = S_X{}^2(n-1) = \sum (X_i - \bar{X})^2 = \sum (X^2) - n(\bar{X})^2$$
$$S_{YY} = S_Y{}^2(n-1) = \sum (Y_i - \bar{Y})^2 = \sum (Y^2) - n(\bar{Y})^2$$

שיפוע הקו:

$$a = \frac{S_{XY}}{S_{XX}} = r\sqrt{\frac{S_{YY}}{S_{XX}}}$$

החותך:

$$b = \bar{Y} - a\bar{X}$$

שונות הרגרסיה

$$\hat{\sigma}^2 = \frac{\sum (Y_i - (aX_i + b))^2}{n-2} = \frac{S_{YY} - a^2 S_{XX}}{n-2}$$

מקדם המתאם:

$$r = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)\left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right)}} = \frac{\sum_{i=1}^{n}(x_i \cdot y_i) - n \cdot \bar{x} \cdot \bar{y}}{\sqrt{\left(\sum_{i=1}^{n}x_i^2 - n\bar{x}^2\right)\left(\sum_{i=1}^{n}y_i^2 - n\bar{y}^2\right)}}$$

תכונות מקדם המתאם:

- $-1 \le r \le 1$
- $r = 0$: אין קשר לינארי בין x ל- y.
- $0 < r \le 1$: קשר לינארי חיובי בין x ל- y. (כש- x עולה, y עולה).
- $-1 \le r < 0$: קשר לינארי שלילי בין x ל- y. (כש- x עולה, y יורד).

r^2 - **פרופורציית ה"שונות המוסברת"** על-ידי x (מקדם דטרמיניציה)

$0 \le r^2 \le 1$

קשר בין מקדם הרגרסיה לשיפוע הקו:

$$a = r\sqrt{\frac{S_{YY}}{S_{XX}}} = r \cdot \frac{s_y}{s_x} = r \cdot \frac{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

בדיקת השערות לשיפוע הרגרסיה

$$E(a) = \beta$$
$$Var(a) = \sigma_a{}^2 = \frac{\sigma^2}{S_{XX}}$$
$$\hat{Var}(a) = SE^2(a) = \hat{\sigma}_a{}^2 = \frac{\hat{\sigma}^2}{S_{XX}}$$

H_0: $\beta=0$
H_1: $\beta\neq 0$

$$T = \frac{a-\beta}{\hat{\sigma}_a} = \frac{a}{SE(a)} = \frac{a}{\hat{\sigma}}\sqrt{S_{XX}} \sim t_{n-2}$$
$$T^2 = \left[\frac{a}{SE(a)}\right]^2 = \left[\frac{a}{\hat{\sigma}}\right]^2 S_{XX} \sim F_{1,n-2}$$

רווח סמך עבור a:

$$a - t_{n-2}^{1-\alpha/2}\hat{\sigma}_a \le \beta \le a + t_{n-2}^{1-\alpha/2}\hat{\sigma}_a$$

לוח ניתוח השונות - ANOVA

	SS	df	MS	F
רגרסיה	SS_B	1	$MS_B=SS_B/df_B$	$MS_B/MS_E \sim F_{(1,n-2)}^{1-\alpha}$
שארית	SS_E	n-2	$MS_E=SS_E/df_E=\hat{\sigma}^2$	
סה"כ	$SS_T = SS_B + SS_E = S_{YY}$	n-1		

$$SS_B + SS_E = SS_T = S_{YY} = S_Y{}^2(n-1) = \sum (Y_i - \bar{Y})^2 = \sum (Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2$$
$$SS_E = \sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2 = S_{YY}(1-R^2)$$
$$SS_B = SS_T - SS_E = S_{YY}R^2 = a^2 S_{XX} = \sum (\hat{Y}_i - \bar{Y})^2$$
$$R^2 = \frac{SS_B}{S_{YY}} = \frac{SS_B}{SS_T}$$

אמידת קו רגרסיה:

$$S_{XY} = \sum (X_i - \bar{X})(Y_i - \bar{Y}) = \sum XY - n\bar{X}\bar{Y}$$
$$S_{XX} = S_X{}^2(n-1) = \sum (X_i - \bar{X})^2 = \sum (X^2) - n(\bar{X})^2$$
$$S_{YY} = S_Y{}^2(n-1) = \sum (Y_i - \bar{Y})^2 = SS_T = \hat{\beta}^2 S_{XX} + \sum e_i^2$$
$$\hat{\alpha} \sim N\left(\alpha, \frac{\sigma^2 \sum x^2}{nS_{XX}}\right)$$
$$\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{S_{XX}}\right)$$
$$\hat{\beta} = \frac{S_{XY}}{S_{XX}} = r\sqrt{\frac{S_{YY}}{S_{XX}}} = r\sqrt{\frac{(n-1)S_Y{}^2}{(n-1)S_X{}^2}} = r\frac{S_Y}{S_X}$$
$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$
$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2} = \frac{\sum (Y_i - (\hat{\alpha} + \hat{\beta}X_i))^2}{n-2} = \frac{S_{YY} - \hat{\beta}^2 S_{XX}}{n-2} = \frac{SS_E}{n-2} = MS_E$$

ANOVA

$$SS_T = SS_B + SS_E = S_{YY} = \hat{\beta}^2 S_{XX} + \sum e_i^2 = S_Y{}^2(n-1) = \sum (Y_i - \bar{Y})^2$$
$$SS_B = SS_{reg} = \hat{\beta}^2 S_{XX} = S_{YY} r^2$$
$$SS_E = \sum e_i^2 = S_{YY} - \hat{\beta}^2 S_{XX} = S_{YY}(1-r^2)$$
$$MS_E = \frac{SS_E}{n-2} = \frac{\sum e_i^2}{n-2} = \hat{\sigma}^2$$

$$MS_E = \frac{SS_B}{df_B}$$
$$\frac{SS_B}{SS_T} + \frac{SS_E}{SS_T} = 1 = r^2 + (1-r^2) = \frac{\hat{\beta}^2 S_{XX}}{S_{YY}} + \frac{\sum e_i^2}{S_{YY}}$$
$$F = \frac{MS_B}{MS_E} = \frac{\hat{\beta}^2 S_{XX}}{\hat{\sigma}^2} = T^2 = \left[\frac{\hat{\beta}}{SE(\hat{\beta})}\right]^2 = \frac{r^2(n-2)}{1-r^2}$$

דף עזר 13

התפלגויות משותפות

יהיו X ו Y שני משתנים מקריים המוגדרים על אותו מרחב הסתברות $(X,Y): \Omega \to R^2$:Ω

פונקצית ההסתברות המשותפת של הזוג (X,Y) מוגדרת ע"י:

$$P_{XY}(x,y) = P(X=x, Y=y)$$

אזי פונקצית ההסתברות השולית של מ"מ X ע"י:

$$P_X(x) = \sum_y P_{XY}(x,y)$$

בדומה, פונקצית ההסתברות השולית של מ"מ Y ע"י:

$$P_Y(y) = \sum_x P_{XY}(x,y)$$

פונקצית ההסתברות המותנית של X בהינתן {Y=y} מוגדרת ע"י:

$$P_{X|Y}(x \mid y) = \frac{P_{XY}(x,y)}{P_Y(y)} = \frac{P(X=x, Y=y)}{P(Y=y)} = P(X=x \mid Y=y)$$

פונקצית ההסתברות המותנית של Y בהינתן {X=x} מוגדרת ע"י:

$$P_{Y|X}(y \mid x) = \frac{P_{XY}(x,y)}{P_X(x)} = \frac{P(X=x, Y=y)}{P(X=x)} = P(Y=y \mid X=x)$$

אי תלות

שני משתנים מקריים נקראים בלתי-תלויים אם לכל זוג נקודות (x,y):

$$P_{XY}(x,y) = P_X(x)P_Y(y)$$

אם X ו Y בלתי תלויים אזי:

$$P_{X|Y}(x \mid y) = \frac{P_{XY}(x,y)}{P_Y(y)} = \frac{P(X=x)P(Y=y)}{P(Y=y)} = P(X=x)$$
$$P_{Y|X}(y \mid x) = \frac{P_{XY}(x,y)}{P_X(x)} = \frac{P(X=x)P(Y=y)}{P(X=x)} = P(Y=y)$$

	SS	df	MS	F
רגרסיה	SS_B	1	MS_B	MS_B / MS_E
טעות	SS_E	n-2	MS_E	
סה"כ	SS_T	n-1		

מקדם המתאם - r

$$r^2 = \frac{\hat{\beta}^2 S_{XX}}{S_{YY}} = \frac{SS_B}{SS_T} = \frac{S_{XY}{}^2}{S_{XX}S_{YY}}$$
$$r = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}}$$
$$\rho = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

בדיקת השערות

H0: β=0 H1: β≠0

$$T = \frac{\hat{\beta} - \beta}{\hat{\sigma}_{\hat{\beta}}} = \frac{\hat{\beta}}{SE(\hat{\beta})} = \frac{\hat{\beta}}{\hat{\sigma}}\sqrt{S_{XX}} \sim t_{n-2}^{1-\alpha/2}$$
$$T = \sqrt{\frac{r^2}{1-r^2}(n-2)}$$
$$T^2 = \left[\frac{\hat{\beta}}{SE(\hat{\beta})}\right]^2 = \left[\frac{\hat{\beta}}{\hat{\sigma}}\right]^2 S_{XX} = \frac{MS_B}{MS_E} \sim F_{1,n-2}^{1-\alpha}$$

רווח סמך לשיפוע

$$\hat{\beta} - t_{n-2}^{1-\alpha/2}\hat{\sigma}_{\hat{\beta}} \le \beta \le \hat{\beta} + t_{n-2}^{1-\alpha/2}\hat{\sigma}_{\hat{\beta}}$$

סכום של משתנים מקריים

השונות של סכומם של מ"מ X ו Y נתונה ע"י:

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X+Y)$$

כאשר Cov(X,Y) היא השונות המשותפת של שני המשתנים:

$$Cov(X,Y) = E[(X - E(X)) \cdot (Y - E(Y))] = E(X \cdot Y) - E(X)E(Y)$$
$$E(X \cdot Y) = \sum_x \sum_y (xy) P_{XY}(x,y)$$

תכונות השונות המשותפת:

1. $-\infty \le Cov(X,Y) \le +\infty$
2. $Cov(X,Y) > 0 \Leftrightarrow$ קיים מתאם חיובי בין שני המשתנים
3. $Cov(X,Y) = 0 \Leftrightarrow$ המשתנים בלתי מתואמים
4. $Cov(aX+b, cY+d) = a \cdot c \cdot Cov(X,Y) \Leftrightarrow$
5. $Cov(X,X) = Var(X)$

- מכאן, אם X ו Y בלתי מתואמים אזי $Var(X+Y) = Var(X) + Var(Y)$
- X ו Y בלתי-תלויים $\leftarrow Cov(X,Y) = 0$ (X ו Y בלתי מתואמים). הכיוון ההפוך לא תמיד נכון!

מקדם המתאם:

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

תכונות מקדם המתאם:

1. $-1 \le \rho(X,Y) \le 1$
2. $\rho(X,Y) = 0 \Leftrightarrow$ המשתנים בלתי מתואמים
3. $\rho(X,Y) = 1 \quad Y = aX + b, a > 0$ המשתנים בלתי מתואמים
4. $\rho(X,Y) = -1 \quad Y = aX + b, a < 0$

נוצר ע"י אסף נתן
מבחן אחרון לסיום התואר
! 8.3.2007

בהצלחה לכולם, טוב לדעת שיש
להנדסת חשמל סוף.