מודלים סטטיסטיים ויישומיהם 52518 תשע"ח – פתרון בוחן בית 2

להגשה עד 20.12.17 בשעה 23:55

1. (20 נקודות) סטטיסטי המבחן קולמוגורוב-סמירנוב-ליליפורס מוגדר באופן

$$D=\max_{t\in[-\infty,\infty]}\left\{\left|\hat{F}_n(t)-\Phi\left(\frac{t-\bar{X}}{s}\right)\right|\right\}$$

יהי מדגם $X=x_1,\dots,x_n \underset{iid}{\sim} F$ כאשר F הינה פונקציית התפלגות רציפה ועולה ממש. נגדיר את המשפחה של ההתפלגויות הנורמליות באופן הבא:

$$N=\left\{F:F(t)=\Phi\left(\frac{t-\mu}{\sigma}\right)\right\},\qquad \mu\in\mathbb{R},\sigma\in\mathbb{R}^+$$

הראו כי עבור בדיקת ההשערות $H_0: F\in N,\ H_1: F\notin N$, התפלגות של D (סטטיסטי קולמוגורוב-סמירנוב-ליליפורס) תחת השערת האפס היא Distribution Free, כלומר אינה תלויה בפונקציה ההתפלגות F.

$$D=\max_{t\in[-\infty,\infty]}\left\{\left|\hat{F}_n(t)-\Phi\left(\frac{t-\bar{X}}{s}\right)\right|\right\}$$

נציב את $\hat{F}_n(t)=\frac{1}{n}\sum_{i=1}^n I_{\{X_i\le t\}}$:

$$D=\max_{t\in[-\infty,\infty]}\left\{\left|\frac{1}{n}\sum_{i=1}^n I_{\{X_i\le t\}}-\Phi\left(\frac{t-\bar{X}}{s}\right)\right|\right\}$$

נגדיר $u=\frac{t-\bar{X}_n}{s}$, אז $t=\bar{X}_n+us$. כיוון שזוהי טרנספורמציה לינארית על t, המקסימום על t הוא כמו המקסימום על u ונוכל לרשום:

$$D=\max_{t\in[-\infty,\infty]}\left\{\left|\frac{1}{n}\sum_{i=1}^n I_{\{X_i\le t\}}-\Phi\left(\frac{t-\bar{X}}{s}\right)\right|\right\}=\max_{u\in[-\infty,\infty]}\left\{\left|\frac{1}{n}\sum_{i=1}^n I_{\{X_i\le \bar{X}_n+us\}}-\Phi(u)\right|\right\}$$

$$=\max_{u\in[-\infty,\infty]}\left\{\left|\frac{1}{n}\sum_{i=1}^n I_{\left\{\frac{X_i-\bar{X}_n}{s}\le u\right\}}-\Phi(u)\right|\right\}$$

נגדיר $Z_i\sim N(0,1)\rightarrow Z_i=\frac{X_i-\mu}{\sigma}$ ונקבל $X_i=\mu+\sigma Z_i$, מכאן כי $\bar{X}_n=\mu+\sigma\bar{Z}_n$ ולכן $X_i-\bar{X}_n=\sigma(Z_i-\bar{Z}_n)$. נסמן $s_Z^2=\frac{1}{n-1}\sum_{i=1}^n(Z_i-\bar{Z}_n)^2$ אז $s=\sigma s_Z$ ונקבל

$$\frac{X_i-\bar{X}_n}{s}=\frac{\sigma(Z_i-\bar{Z}_n)}{\sigma s_Z}=\frac{(Z_i-\bar{Z}_n)}{s_Z}$$

ובסך הכל

$$D=\max_{u\in[-\infty,\infty]}\left\{\left|\frac{1}{n}\sum_{i=1}^n I_{\left\{\frac{(Z_i-\bar{Z}_n)}{s_Z}\le u\right\}}-\Phi(u)\right|\right\}$$

2. (30 נקודות) מצורף קובץ נתונים qn2.txt ממחקר על התפתחות תינוקות. הקבוצות מוגדרות באופן הבא:
   1) נולד פג עם משקל לידה נמוך
   2) נולד פג עם משקל לידה בינוני
   3) נולד באופן תקין

מדד התוצאה הוא מדד התפתחות של התינוקות, רוצים לבחון את ההשערה $H_0:\mu_1=\mu_2=\mu_3$.

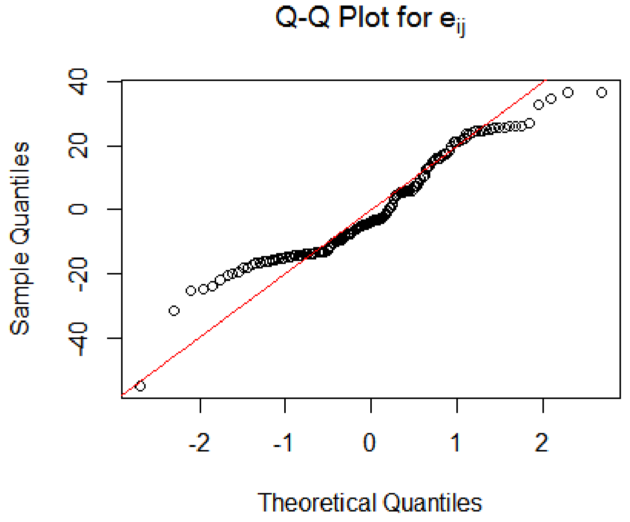
ערכו את המבחנים המתאימים לבדיקת הנחות המודל והסיקו אילו מהן נשמרות ואילו נדחות. במידה ויש הפרות של ההנחות, נקטו בצעדים המתאימים לטפל בעניין. בנו רווחי סמך בו-זמניים עבור אוסף הקונטרסטים הבא:

$$\psi_1=\mu_2-\mu_1,\qquad \psi_2=\mu_3-\mu_1,\qquad \psi_3=\mu_3-\mu_2,\qquad \psi_4=\frac{\mu_2+\mu_3}{2}-\mu_1$$

במודל ניתוח שונות חד-כווני יש לנו שתי הנחות:

1. השגיאה מפולגת נורמלית $(y_{ij}-\mu_i=\epsilon_{ij}\sim N(0,\sigma_i^2))$
2. שוויון שונויות $(\sigma_1=\sigma_2=\dots=\sigma_I)$

עבור הנתונים בקובץ qn2.txt, נביט ב-QQ-Plot של השאריות וכן נערוך מבחן קולמוגורוב-סמירנוב:




```
Lilliefors (Kolmogorov-Smirnov) normality test

data:  e
D = 0.11279, p-value = 0.0001636
```

כמעט בכל רמת מובהקות, נוכל להכריע כי הנתונים אינם מפולגים נורמלית.

כיוון שגדלי הקבוצות מקיימים $n_1 = 32, n_2 = 71, n_3 = 37$ נוכל לקרב לפי CLT באופן $\bar{y}_{i.} \sim N\left(\mu_i, \frac{\sigma_i^2}{n_i}\right)$ ובשל כך נוכל להשתמש במבחן Levene לשוויון שונויות (למרות שהוא מניח $\frac{\epsilon_{ij}}{\sigma_i} \sim N(0,1)$):

```
Levene's Test for Homogeneity of Variance (center = "mean")
       Df F value Pr(>F)
group   2  0.2442 0.7837
      137
```

מתוצאה זו, לא נדחה את השערת שוויון השונויות ולכן נערוך את מבחן Welch בהתאם:

```
        One-way analysis of means

data:  y and G
F = 7.6979, num df = 2, denom df = 137, p-value = 0.0006788
```

לפי תוצאות אלו, נדחה את השערת שוויון התוחלות. גם לפי מבחן Kruskal-Wallis נקבל את אותה מסקנה:

```
 Kruskal-Wallis k-sample test.

Number of samples:  3
Sample sizes:  32, 71, 37
Number of ties: 42

Null Hypothesis: All samples come from a common population.

Based on Nsim = 1e+07 simulations

  test statistic  asympt. P-value     sim. P-Value
       11.370000         0.003389         0.002990
```

כיוון שדחינו את ההנחה על נורמליות השגיאות בנתונים, לא נוכל להשתמש בשיטות שפה וטוקי (טוקי מיועדת להפרשים זוגיים, שפה משתמשת במבחן F); לא מצאנו קשר מהצורה $\sigma_i \approx c\mu_i^\alpha$ ולכן גם לא מצאנו טרנספורמציה מתאימה לנתונים; כל שנותר לנו הוא רווחי סמך בשיטה המדויקת:

$$\Psi = \begin{cases} \mu_2 - \mu_1 \\ \mu_3 - \mu_1 \\ \mu_3 - \mu_2 \\ \dfrac{\mu_2 + \mu_3}{2} - \mu_1 \end{cases} \rightarrow C = \begin{pmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \\ 0 & -1 & 1 \\ -1 & 0.5 & 0.5 \end{pmatrix}, \hat{\Psi} = C\hat{\beta} = \begin{cases} \bar{y}_{2.} - \bar{y}_{1.} \\ \bar{y}_{3.} - \bar{y}_{1.} \\ \bar{y}_{3.} - \bar{y}_{2.} \\ \dfrac{\bar{y}_{2.} - \bar{y}_{3.}}{2} - \bar{y}_{1.} \end{cases}$$

$$R = \Omega^{-1} H \Omega^{-1} \rightarrow \hat{\psi}\left(c^{(k)}\right) \pm t^*_{N-I,1-\alpha} \cdot s \cdot \sqrt{V\left(C^{(k)}\right)}$$

$$S(c^{(1)}) = [3.7838, 18.5242]$$
$$S(c^{(2)}) = [6.7803, 23.4929]$$
$$S(c^{(3)}) = [-3.0360, 11.0011]$$
$$S(c^{(4)}) = [6.0913, 20.1993]$$

ובסך הכל נוכל לדחות לחלוטין את ההשערה $H_0: \mu_1 = \mu_2 = \mu_3$. קוד R:

```
library(car)
library(kSamples)
library(nortest)
library(mvtnorm)
D <- read.table("qn2.txt")
G <- factor(D$V1)
y <- D$V2
N <- length(y)
I <- nlevels(G)
mu.hat <- ave(y, G)
e <- y - mu.hat
# Q-Q plot
qqnorm(e, main = expression("Q-Q Plot for"~e[ij]))
qqline(e, col = 2)
# K-S
lillie.test(e)
#levene
leveneTest(y ~ G, center = "mean")
# Welch
oneway.test(y~G, var.equal = T)
# K-W
y1 <- y[G==1]
y2 <- y[G==2]
y3 <- y[G==3]
qn.test(y1, y2, y3, test = "KW", method = "simulated", Nsim = 1e7)
# CIs
alpha <- 0.1
C <- matrix(c(-1, -1, 0, -1, 1, 0, -1, 0.5, 0, 1, 1, 0.5), ncol = 3)
K <- nrow(C)
mu_i.hat <- tapply(y, G, mean)
n_i <- tapply(y, G, length)
s <- sqrt(sum(e^2)/(N - I))
psi.hat <- C %*% mu_i.hat
V <- diag(s^2 / n_i)
cov.hat <- C %*% V %*% t(C)
H <- cov.hat / s^2
V_i <- sqrt(diag(H))
Omega <- diag(V_i)
R <- solve(Omega) %*% H %*% solve(Omega)
t.crit <- qmvt(p = 1 - alpha, corr = R, tail = "both", df = (N - I))$quantile
w <- t.crit * s * V_i
CI <- cbind(psi.hat - w, psi.hat + w)
```

3. (30 נקודות) בקורס מסוים יש 9 תלמידים. 4 תלמידים למדו ברצינות במהלך הסמסטר והבינו תרגילים באופן שוטף, 5 נוספים לא למדו כלל במהלך הסמסטר ולמדו רק לקראת המבחן. הציונים במבחן הסופי הם:

לומדים באופן שוטף	87	80	77	90	
לומדים רק למבחן	70	75	69	82	85

רוצים לבדוק את ההשערות הבאות:

H_0 – לימוד שוטף לא מצביע על ציון סופי גבוה יותר בקורס זה.

H_1 - לימוד שוטף מצביע על ציון סופי גבוה יותר בקורס זה.

כיוון שזהו קורס חדש לחלוטין, אין ידע על התפלגות הציונים ונדרש שימוש במבחן לא-פרמטרי.

א. איזה מבחן יתאים לבעיה זו?

ב. מה רמת המובהקות של מבחן הדוחה את H_0 אם סכום הדרגות הרלוונטי גדול או שווה ל-26? יש לחשב את רמת המובהקות במדויק, ללא שימוש בקירוב הנורמלי.

ג. האם, לפי המבחן שבנית בסעיף ב', עליך לדחות את H_0?

א. כיוון שנדרש מבחן לא-פרמטרי להשוואה בין קבוצות ולא ידועה לנו התפלגות הנתונים, נוכל להשתמש במבחן סכום דרגות של ווילקוקסון.

ב. יש $\binom{9}{4}$ קומבינציות שונות לסטטיסטי הדרגות, אנחנו מחפשים את אלו בהן סכום הדרגות של קבוצת הלומדים הוא לכל הפחות 26, יש בדיוק 12 כאלו:

$$w = 26: (9,8,7,2), (9,8,6,3), (9,8,5,4), (9,7,6,4), (8,7,6,5)$$
$$w = 27: (9,8,7,3), (9,8,6,4), (9,7,6,5)$$
$$w = 28: (9,8,7,4), (9,8,6,5)$$
$$w = 29: (9,8,7,5)$$
$$w = 30: (9,8,7,6)$$

(את מספר הקומבינציות ניתן היה למצוא באמצעות שאלה 3 בתרגיל 4)

כלומר רמת המובהקות של המבחן היא $P(W_s \geq 26) = \frac{12}{\binom{9}{4}} = \frac{12}{126} = 0.0952$

ג. הדרגות של קבוצת הלומדים הן $(9,4,5,8)$, סכומן הוא 26 ולכן לפי המבחן של סעיף ב' עלינו לדחות את H_0 ברמת מובהקות של 0.0952.

4. (20 נקודות) נניח מודל ANOVA דו-כווני עם $I = J = 2$ וממוצעי קבוצות באופן הבא:

$$\mu_{11} = \theta$$
$$\mu_{12} = \theta + \Delta_B$$
$$\mu_{21} = \theta + \Delta_A$$
$$\mu_{22} = \theta + \Delta_A + \Delta_B + \psi$$

ע"י שימוש בפרמטריזציה $\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$ ובמשקלות $\pi_1 = \pi_2 = \tau_1 = \tau_2 = \frac{1}{2}$, מצאו ביטויים עבור הרכיבים $\mu, \alpha_1, \alpha_2, \beta_1, \beta_2, \gamma_{11}, \gamma_{12}, \gamma_{21}, \gamma_{22}$.

$$\mu_{11} = \mu + \alpha_1 + \beta_1 + \gamma_{11}$$
$$\mu_{12} = \mu + \alpha_1 + \beta_2 + \gamma_{12}$$
$$\mu_{21} = \mu + \alpha_2 + \beta_1 + \gamma_{21}$$
$$\mu_{22} = \mu + \alpha_2 + \beta_2 + \gamma_{22}$$

מאילוצים על הפרמטרים ומהמשקלות השווים נקבל $\alpha_2 = -\alpha_1$ וכן $\beta_2 = -\beta_1$, כמו גם $\gamma_{21} = -\gamma_{11}$, $\gamma_{22} = -\gamma_{12}$ וכן $\gamma_{12} = -\gamma_{11}$, כלומר $\gamma_{22} = \gamma_{11}$.

$$\mu = \sum_i \sum_j \pi_i \tau_j \mu_{ij} = \frac{1}{4}\big(\theta + (\theta + \Delta_B) + (\theta + \Delta_A) + (\theta + \Delta_A + \Delta_B + \psi)\big) = \theta + \frac{\Delta_A}{2} + \frac{\Delta_B}{2} + \frac{\psi}{4}$$

$$\mu_{12} - \mu_{11} = \beta_2 - \beta_1 + \gamma_{12} - \gamma_{11} = -2\beta_1 - 2\gamma_{11} = \Delta_B \rightarrow -\beta_1 - \gamma_{11} = \frac{\Delta_B}{2}$$

$$\mu_{11} = \mu + \alpha_1 + \beta_1 + \gamma_{11} \rightarrow \alpha_1 = \mu_{11} - \mu - \beta_1 - \gamma_{11} = \theta - \theta - \frac{\Delta_A}{2} - \frac{\Delta_B}{2} - \frac{\psi}{4} + \frac{\Delta_B}{2} = -\frac{\Delta_A}{2} - \frac{\psi}{4}$$

$$\mu_{11} + \mu_{21} = 2\mu + 2\beta_1 = 2\theta + \Delta_A \rightarrow \beta_1 = \theta + \frac{\Delta_A}{2} - \mu = \theta + \frac{\Delta_A}{2} - \theta - \frac{\Delta_A}{2} - \frac{\Delta_B}{2} - \frac{\psi}{4} = -\frac{\Delta_B}{2} - \frac{\psi}{4}$$

$$\mu_{11} + \mu_{22} = 2\mu + 2\gamma_{11} = 2\theta + \Delta_A + \Delta_B + \psi \to \gamma_{11} = \theta + \frac{\Delta_A}{2} + \frac{\Delta_B}{2} + \frac{\psi}{2} - \mu$$
$$= \theta + \frac{\Delta_A}{2} + \frac{\Delta_B}{2} + \frac{\psi}{2} - \theta - \frac{\Delta_A}{2} - \frac{\Delta_B}{2} - \frac{\psi}{4} = \frac{\psi}{4}$$

ובסך הכל נקבל:

$$\mu = \theta + \frac{\Delta_A}{2} + \frac{\Delta_B}{2} + \frac{\psi}{4}$$
$$\alpha_1 = -\frac{\Delta_A}{2} - \frac{\psi}{4}$$
$$\alpha_2 = \frac{\Delta_A}{2} + \frac{\psi}{4}$$
$$\beta_1 = -\frac{\Delta_B}{2} - \frac{\psi}{4}$$
$$\beta_2 = \frac{\Delta_B}{2} + \frac{\psi}{4}$$
$$\gamma_{11} = \frac{\psi}{4}$$
$$\gamma_{12} = -\frac{\psi}{4}$$
$$\gamma_{21} = -\frac{\psi}{4}$$
$$\gamma_{22} = \frac{\psi}{4}$$