EX2

roi hezkiyahu

4 3 2022

שאלה 1

נתון מודל ניתוח שונות דו כיווני

$$Y_{ijk} = \mu_{ij} + \epsilon_{ijk}. \quad \epsilon_{ijk} \sim N(0, \sigma^2) \quad (\epsilon_{ijk}\ iid)$$

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

$$i = 1, ..., I,\ j = 1, ..., J, k = 1, ..., n$$

א. הוכיחו דרך ההגדרה של התוחלות כי במודל ניתוח שונות דו כיווני מתקיים $\beta_j = \mu_{.j} - \mu$ (מתוך מצגת 2 שקף 4).

ב. הוכיחו דרך ההגדרה של התוחלות כי במודל ניתוח שונות דו כיווני מתקיים $\gamma_{ij} = \mu_{ij} - \mu_{i.} - \mu_{.j} + \mu$ (מתוך מצגת 2 שקף 4).

ג. הוכיחו כי במודל ניתוח שונות דו כיווני $E(MSA) = \sigma^2 + nJ\frac{\sum_{i=1}^{I}\alpha_i^2}{I-1}$ (מתוך מצגת 2 שקף 13) . (ההוכחה שכתבתם בתרגיל 1 שאלה 2 יכולה לעזור).

ד. הוכיחו כי במודל ניתוח שונות דו כיווני $\sum_{i=1}^{I}\sum_{j=1}^{J}(\overline{Y}_{i..} - \overline{Y})(\overline{Y}_{ij.} - \overline{Y}_{i..} - \overline{Y}_{.j.} + \overline{Y}) = 0$ (מתוך מצגת 2 שקף 11).

Q1

a

$$Reminder\ of\ assumtions:$$

$$(1) \sum_{i=1}^{I} \alpha_i = 0$$

$$(2) \sum_{j=1}^{J} \beta_j = 0$$

$$(3) \sum_{i=1}^{I} \gamma_{ij} = 0 \ \ \forall j$$

$$(4) \sum_{j=1}^{J} \gamma_{ij} = 0 \ \ \forall i$$

$$\mu_{.j} - \mu = \frac{1}{nI}\sum_{i=1}^{I}\sum_{k=1}^{n}(\mu_{ij}) - \mu = \frac{1}{nI}\sum_{i=1}^{I}\sum_{k=1}^{n}(\mu + \alpha_i + \beta_j + \gamma_{ij}) - \mu = \frac{1}{nI}nI\mu + \frac{n\sum_{i=1}^{I}\alpha_i}{nI} + \frac{nI\beta_j}{nI} + \frac{n\sum_{i=1}^{I}\gamma_{ij}}{nI} - \mu =$$

$$= \mu + \frac{\sum_{i=1}^{I}\alpha_i}{I} + \beta_j + \frac{\sum_{i=1}^{I}\gamma_{ij}}{I} - \mu := A + \beta_j + C$$

$$(1), (3) \Rightarrow A = C = 0 \Rightarrow \mu_{.j} - \mu = \beta_j$$

b

$$\gamma_{ij} = \mu_{ij} - \mu - \alpha_i - \beta_j = \mu_{ij} - \mu - (\mu_{.j} - \mu) - (\mu_{i.} - \mu) = \mu_{ij} + 2\mu - \mu - \mu_{i.} - \mu_{.j} = \mu_{ij} - \mu_{i.} - \mu_{.j} + \mu$$

c

$$MSA = \frac{SSA}{I-1} = \frac{nJ\sum_{i=1}^{I}(\bar{Y}_{i..} - \bar{Y})^2}{I-1}$$

$$\bar{Y}_{i..} - \bar{Y} = \frac{1}{nJ}\sum_{j=1}^{J}\sum_{k=1}^{n} Y_{ijk} - \frac{1}{N}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n} Y_{ijk} = \frac{1}{nJ}\sum_{j=1}^{J}\sum_{k=1}^{n}(\mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}) - \frac{1}{N}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n}(\mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk})$$

$$= \mu + \alpha_i + 0 + 0 + \bar{\varepsilon}_{i..} - (\mu + 0 + 0 + 0 + \bar{\varepsilon}) = \alpha_i - (\bar{\varepsilon} + \bar{\varepsilon}_{i..}) \Rightarrow (\bar{Y}_{i..} - \bar{Y})^2 = \alpha_i^2 + 2\alpha_i(\bar{\varepsilon}_{i..} - \bar{\varepsilon}) + (\bar{\varepsilon}_{i..} - \bar{\varepsilon})^2$$

$$E(MSA) = E(\frac{nJ\sum_{i=1}^{I}(\alpha_i^2 + 2\alpha_i(\bar{\varepsilon}_{i..} - \bar{\varepsilon}) + (\bar{\varepsilon}_{i..} - \bar{\varepsilon})^2)}{I-1}) = \frac{nJ\sum_{i=1}^{I}\alpha_i^2}{I-1} + \frac{2nJ\sum_{i=1}\alpha_i E(\bar{\varepsilon}_{i..} - \bar{\varepsilon})}{I-1} + \frac{E(nJ\sum_{i=1}^{I}(\bar{\varepsilon}_{i..} - \bar{\varepsilon})^2)}{I-1}$$

$$(5)\frac{2nJ\sum_{i=1}^{I}\alpha_i E(\bar{\varepsilon}_{i..} - \bar{\varepsilon})}{I-1} = \frac{nJ\sum_{i=1}^{I}\alpha_i(-E(\bar{\varepsilon}) + E(\bar{\varepsilon}_{i..}))}{I-1} = 0$$

$$(6)nJ\sum_{i=1}^{I}\frac{E((\bar{\varepsilon}_{i..} - \bar{\varepsilon})^2)}{I-1} = nJ\sum_{i=1}^{I}\frac{V(\bar{\varepsilon}_{i..} - nJ\sum_{i=1}^{I}\bar{\varepsilon})}{I-1} + nJ\sum_{i=1}^{I}\frac{(E(\bar{\varepsilon}_{i..} - \bar{\varepsilon}))^2}{I-1} = nJ\sum_{i=1}^{I}\frac{V(\bar{\varepsilon}_{i..} - \bar{\varepsilon})}{I-1} =$$

$$= \frac{nJ}{I-1}\sum_{i=1}^{I}(\frac{\sigma^2}{nJ} - \frac{\sigma^2}{N}) = \frac{nJ}{I-1}\frac{(I-1)\sigma^2}{nJ} = \sigma^2$$

$$E(MSA) = \frac{nJ\sum_{i=1}^{I}\alpha_i^2}{I-1} + (5) + (6) = \frac{nJ\sum_{i=1}^{I}\alpha_i^2}{I-1} + \sigma^2$$

d

$$(7)\,\bar{Y} = \frac{1}{N}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n} Y_{ijk} = \frac{1}{N}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n}\mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} = \mu + \bar{\varepsilon}$$

$$(8)\,\bar{Y}_{i..} = \frac{1}{nJ}\sum_{j=1}^{J}\sum_{k=1}^{n} Y_{ijk} = \frac{1}{nJ}\sum_{j=1}^{J}\sum_{k=1}^{n}\mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} = \mu + \alpha_i + \bar{\varepsilon}_{i..}$$

$$(9)\,\bar{Y}_{.j.} = \frac{1}{nI}\sum_{i=1}^{I}\sum_{k=1}^{n} Y_{ijk} = \frac{1}{nI}\sum_{i=1}^{I}\sum_{k=1}^{n}\mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} = \mu + \beta_j + \bar{\varepsilon}_{.j.}$$

$$(10)\,\bar{Y}_{ij.} = \frac{1}{n}\sum_{k=1}^{n} Y_{ijk} = \frac{1}{n}\sum_{k=1}^{n}\mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \bar{\varepsilon}_{ij.}$$

$$(11)\,\bar{Y}_{i..} - \bar{Y} = \mu + \alpha_i + \bar{\varepsilon}_{i..} - \mu - \bar{\varepsilon} = \alpha_i + \bar{\varepsilon}_{i..} - \bar{\varepsilon}$$

$$(12)\,\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y} = \gamma_{ij}$$

$$(13)\ \sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{Y}_{i..} - \bar{Y})(\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}) = \sum_{i=1}^{I}\sum_{j=1}^{J}(11)(12) = \sum_{i=1}^{I}\sum_{j=1}^{J}\gamma_{ij}(\alpha_i + \bar{\varepsilon}_{i..} - \bar{\varepsilon}) = \sum_{i=1}^{I}(0(\alpha_i + \bar{\varepsilon}_{i..} - \bar{\varepsilon})) = 0$$

שאלה 2

בשאלה זו תכתבו סימולציה. המטרה של הסימולציה היא להדגים מדוע צריך לקחת בחשבון השוואות מרובות כשמבצעים ניתוח פוסט הוק.

לשם כך, נייצר נתונים תחת השערת האפס מספר רב של פעמים ובכל פעם נבדוק האם ניתן לדחות את השערת האפס תחת $\alpha = 0.05$.

א. כתבו קוד היוצר מדגם של 120 תצפיות ולהן שני משתנים -

- משתנה מסביר בשם treatment המשייך כל תצפית לאחת מ4 קבוצות הטיפול A,B,C,D כך שהקבוצות בגודל זהה (n=30).
- משתנה מוסבר בשם outcome הנדגם מהתפלגות נורמלית סטנדרטית לכל תצפית (לא משנה באיזה קבוצת טיפול היא, כלומר התוחלת בכל הקבוצות זהה).

בסעיף זה נבדוק האם יש הבדל בין קבוצה A לקבוצה D.
חיזרו על התהליך של יצירת המדגם הזה 1,000 פעמים, ובכל פעם בצעו את בדיקת ההשערות הבאה: בידקו האם יש עדות להבדל בין קבוצה A לקבוצה D ברמת ביטחון של 95%.
מה פרופורציית הפעמים בהן דחיתם את השערת האפס? האם זה מתאים למה שהייתם מצפים? הסבירו.

ב. השתמשו שוב בקוד היוצר מדגם של 120 תצפיות עם שני המשתנים outcome ו treatment.

בסעיף זה נשווה בכל מדגם את הקבוצה בעלת הממוצע המינימלי לקבוצה בעלת הממוצע המקסימלי.
חיזרו על התהליך של יצירת המדגם הזה 1,000 פעמים, ובכל פעם בצעו את בדיקת ההשערות הבאה: בידקו האם יש עדות להבדל בין הקבוצה שלה יש את הממוצע המינימלי במדגם זה, לבין הקבוצה לה יש את הממוצע המקסימלי במדגם זה. בידקו זאת ברמת ביטחון של 95%.
מה פרופורציית הפעמים בהן דחיתם את השערת האפס? האם זה מתאים למה שהייתם מצפים? הסבירו.

ג. מה המסקנה מסעיפים א - ב?

Q2

a

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------------------------------------------------------------- tidy
verse 1.3.0 --
```

```
## <U+221A> ggplot2 3.3.2     <U+221A> purrr   0.3.4
## <U+221A> tibble  3.0.3     <U+221A> dplyr   1.0.2
## <U+221A> tidyr   1.1.2     <U+221A> stringr 1.4.0
## <U+221A> readr   1.3.1     <U+221A> forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------------------------------------------------------------------------ tidyverse_
conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(glue)
```

```
##
## Attaching package: 'glue'
```

```
## The following object is masked from 'package:dplyr':
##
##     collapse
```

```
options(dplyr.summarise.inform = FALSE)
n <-  120
B <- 1000
alpha <- 0.05
treat <- c(rep("A",30),rep("B",30),rep("C",30),rep("D",30))
get_t <- function(n,treat){
  outcome <- rnorm(n)
  tbl <- tibble(outcome = outcome,treat = treat)
  A <- tbl %>%
    filter(treat == "A")%>%
    select(outcome)
  D <- tbl %>%
    filter(treat == "D")%>%
    select(outcome)
  return(t.test(A,D,var.equal = T)$p.value)
}
p_values <- map2_dbl(rep(n,B),replicate(B,treat,F),get_t)
rejected <- p_values <= alpha
mean_reject <- mean(rejected)
glue("we rejected H_0 {round(mean_reject*100,2)} % of the time")
```

```
## we rejected H_0 6 % of the time
```

התוצאה נראית הגיונית שכן נצפה ל5 אחוז דחייה

b

```
get_t2 <- function(n,treat){
  outcome <- rnorm(n)
  tbl <- tibble(outcome = outcome,treat = treat)
  group_means <- tbl %>%
    group_by(treat)%>%
    summarize(mean=mean(outcome))
  min_group_let <- group_means %>%
    filter(mean == min(mean)) %>%
    select(treat)
  max_group_let <- group_means %>%
    filter(mean == max(mean)) %>%
    select(treat)
  Max_group <- tbl %>%
    filter(treat == max_group_let$treat)%>%
    select(outcome)
  Min_group <- tbl %>%
    filter(treat == min_group_let$treat)%>%
    select(outcome)
  return(t.test(Max_group,Min_group,var.equal = T)$p.value)
}
p_values2 <- map2_dbl(rep(n,B),replicate(B,treat,F),get_t2)
rejected2 <- p_values2 <= alpha
mean_reject2 <- mean(rejected2)
glue("we rejected H_0 {round(mean_reject2*100,2)} % of the time")
```

```
## we rejected H_0 18.7 % of the time
```

זה לא תואם את המצופה, נצפה שנדחה את השערת ה0 כ5 אחוז מהפעמים אך מכיוון ואנו מכניסים כאן בחירה אנו פוגעים בשמירה על אלפא

c

לא ניתן לבצע בחירה לאחר הסתכלות בדאטא ואז בדיקת השערות, עלינו לבצע בדיקת השערות מרובה ולבצע תיאום לפי וואליו על מנת לשמור על האלפא הרצוי