

סוגיות נוספות בניתוח שונות דו-כווני.

מה נעשה בפרק זה?

- נדון (בקצרה) במגוון נושאים הקשורים לניתוח שונות דו-כווני.
 - נתונים לא מאוזנים.
 - בדיקת הנחות המודל.
 - השוואות מרובות.
 - שני מקרים פרטיים:
 - מודל שבו מניחים שאין אינטראקציה.
 - מודל עם תצפית אחת עבור קומבינציה של טיפולים.

מה עושים כאשר הנתונים אינם מאוזנים...

נתונים לא מאוזנים

- ניתוח שונות דו-כווני הרבה פעמים מיושם עבור נתונים אשר מקורם בניסוי. במקרה זה הניסוי מתוכנן מראש ודואגים לרוב לאסוף מספר זהה של תצפיות עבור כל קומבינציה של טיפולים. אבל...
- לפעמים מקור הנתונים הוא אחר (Observational data), ובמקרים כאלו לרוב הנתונים אינם מאוזנים.
- לפעמים מאבדים/פוסלים תצפיות והנתונים "הופכים" ללא מאוזנים.
- לפעמים לא ניתן לבצע ניסוי מאוזן ועדיף לבצע ניסוי לא מאוזן.
- הבעיה: עבור נתונים לא מאוזנים פרוק סכום הריבועים לא מתקיים (ודרגות החופש אינן מסתכמות) ולכן (באופן כללי) לא ניתן לבצע ניתוח שונות דו-כווני באופן ישיר ומדויק. (זאת בניגוד לניתוח שונות חד-כווני אשר עדיין מדויק במקרה הלא מאוזן).

סימונים

מספר תצפיות עבור כל קומבינציה:

$$\begin{array}{cccc} n_{11} & n_{12} & \cdots & n_{1b} \\ n_{21} & \cdot & \cdot & \vdots \\ \vdots & \cdot & n_{ik} & \vdots \\ n_{a1} & \cdots & \cdots & n_{ab} \end{array}$$

$$\longrightarrow n_{i\cdot} = \sum_{k=1}^b n_{ik} \quad i = 1, \dots, a$$

$$\begin{array}{l} n_{\cdot k} = \sum_{i=1}^a n_{ik} \\ k = 1, \dots, b \end{array}$$

$$n_{\cdot\cdot} = \sum_{i=1}^a n_{i\cdot} = \sum_{k=1}^b n_{\cdot k} = \sum_{i=1}^a \sum_{k=1}^b n_{ik}$$

מקרה עם פתרון מדויק: מספר תצפיות פרופורציונאלי

$$n_{ik} = \frac{n_{i\cdot} \cdot n_{\cdot k}}{n_{\cdot\cdot}} \quad \text{נאמר שמספר התצפיות הוא פרופורציונאלי אם:}$$

באופן כללי כל שתי עמודות וכל שתי שורות הן פרופורציונאליות

מה זה אומר?

נסתכל על עמודות k_1 ו k_2 :

קבלנו שעמודה k_1
ועמודה k_2
פרופורציונאליות

$$\frac{n_{1k_1}}{n_{1k_2}} = \frac{n_{2k_1}}{n_{2k_2}} = \dots = \frac{n_{ak_1}}{n_{ak_2}}$$

$$\begin{array}{l} \frac{n_{1k_1}}{n_{1k_2}} = \frac{n_{\cdot k_1}}{n_{\cdot k_2}} \\ \frac{n_{2k_1}}{n_{2k_2}} = \frac{n_{\cdot k_1}}{n_{\cdot k_2}} \\ \vdots \\ \frac{n_{ak_1}}{n_{ak_2}} = \frac{n_{\cdot k_1}}{n_{\cdot k_2}} \end{array}$$

$$\begin{array}{l} n_{1k_2} = n_{1\cdot} \frac{n_{\cdot k_2}}{n_{\cdot\cdot}} \\ n_{2k_2} = n_{2\cdot} \frac{n_{\cdot k_2}}{n_{\cdot\cdot}} \\ \vdots \\ n_{ak_2} = n_{a\cdot} \frac{n_{\cdot k_2}}{n_{\cdot\cdot}} \end{array}$$

$$\begin{array}{l} n_{1k_1} = n_{1\cdot} \frac{n_{\cdot k_1}}{n_{\cdot\cdot}} \\ n_{2k_1} = n_{2\cdot} \frac{n_{\cdot k_1}}{n_{\cdot\cdot}} \\ \vdots \\ n_{ak_1} = n_{a\cdot} \frac{n_{\cdot k_1}}{n_{\cdot\cdot}} \end{array}$$

נוסחאות חישוב

- כאשר מספר התצפיות פרופורציונאלי, ניתן לבצע ניתוח שונות דו-כווני באופן מדויק.

- צריך לעדכן את נוסחאות החישוב:

$$SS_{Total} = \sum_{i=1}^a \sum_{k=1}^b \sum_{j=1}^{n_{ik}} y_{ikj}^2 - \frac{y_{\dots}^2}{n_{\dots}}$$

$$SS_A = \sum_{i=1}^a \frac{y_{i\bullet}^2}{n_{i\bullet}} - \frac{y_{\dots}^2}{n_{\dots}}$$

$$SS_B = \sum_{k=1}^b \frac{y_{\bullet k}^2}{n_{\bullet k}} - \frac{y_{\dots}^2}{n_{\dots}}$$

$$SS_{AB} = \sum_{i=1}^a \sum_{k=1}^b \frac{y_{ik\bullet}^2}{n_{ik\bullet}} - \frac{y_{\dots}^2}{n_{\dots}} - SS_A - SS_B$$

$$SS_E = SS_{Total} - SS_A - SS_B - SS_{AB} = \sum_{i=1}^a \sum_{k=1}^b \sum_{j=1}^{n_{ik}} y_{ikj}^2 - \sum_{i=1}^a \sum_{k=1}^b \frac{y_{ik\bullet}^2}{n_{ik\bullet}}$$

דוגמא

- בחנות ביגוד מעוניינים לבחון כיצד הגורמים הבאים משפיעים על כמות המכירות (בשקלים) הממוצעת לשעה:
 - אמצע שבוע (D) או סוף שבוע (E).
 - אדם המכירות הוא גבר (M) או אדם המכירות הוא אישה (W).
- החנות עובדת 4 ימים באמצע השבוע ויומיים בסוף השבוע. באמצע השבוע ישנן 4 משמרות של גברים ו 8 משמרות של נשים. בסוף השבוע ישנן 2 משמרות של גברים ו 4 משמרות של נשים (בכל משמרת יש אדם מכירות יחיד).
- להלן הנתונים:

	<i>D</i>	<i>E</i>
<i>M</i>	535,345,1245,419	2352,1552
<i>W</i>	688,854,868,1010, 1442,647,673,1235	1245,1023,2451,1045

האם מספר התצפיות פרופורציונאלי? ←

שיטות נוספות

- כאשר הנתונים אינם מאוזנים הם גם לרוב לא יהיו פרופורציונאליים.
- ניתן ל"אזן" את הנתונים ע"י מספר שיטות:
 - שערך תצפיות חסרות (ע"י ממוצע האוכלוסייה) - מתאים כאשר יש תצפיות בודדות חסרות.
 - הורדת עודף תצפיות ע"י בחירה אקראית - מתאים כאשר יש עודף של תצפיות בודדות - אבל זהו "חטא סטטיסטי".
 - PROC GLM מבצע ניתוח מדויק ע"י מודל רגרסיה... כאן יש ארבע סוגי סכומי ריבועים (Type I, II, III, IV).
 - Type I – "Sequential" (תלוי בסדר המשתנים במודל).
 - "Extra" – Type III

בדיקת הנחות המודל... בדומה לניתוח שונות חד-כווני

השוואות מרובות ... בדומה לניתוח שונות חד-כווני

מודל ללא אינטראקציה

לפעמים ניתן להניח שאין אינטראקציה...

$$\underbrace{SS_{Total}}_{abn-1} = \underbrace{SS_A}_{a-1} + \underbrace{SS_B}_{b-1} + \underbrace{SS_E}_{\underbrace{abn-a-b+1}_{ab(n-1)+(a-1)(b-1)}}$$

$$y_{ikj} = \underbrace{\mu + \tau_i + \beta_k}_{\mu_{ik}} + \tilde{\varepsilon}_{ikj}$$

$$i = 1, \dots, a$$

$$k = 1, \dots, b$$

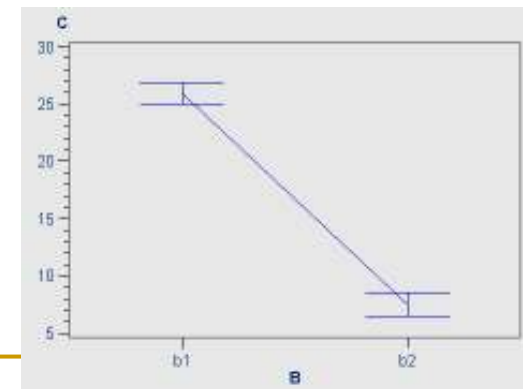
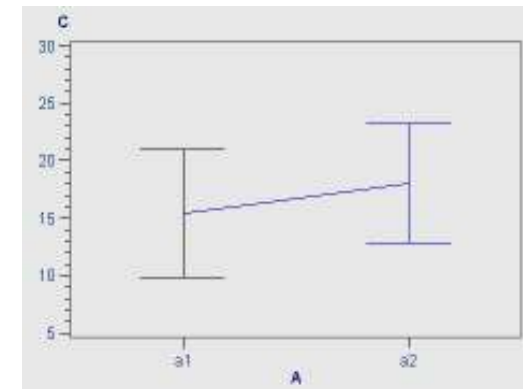
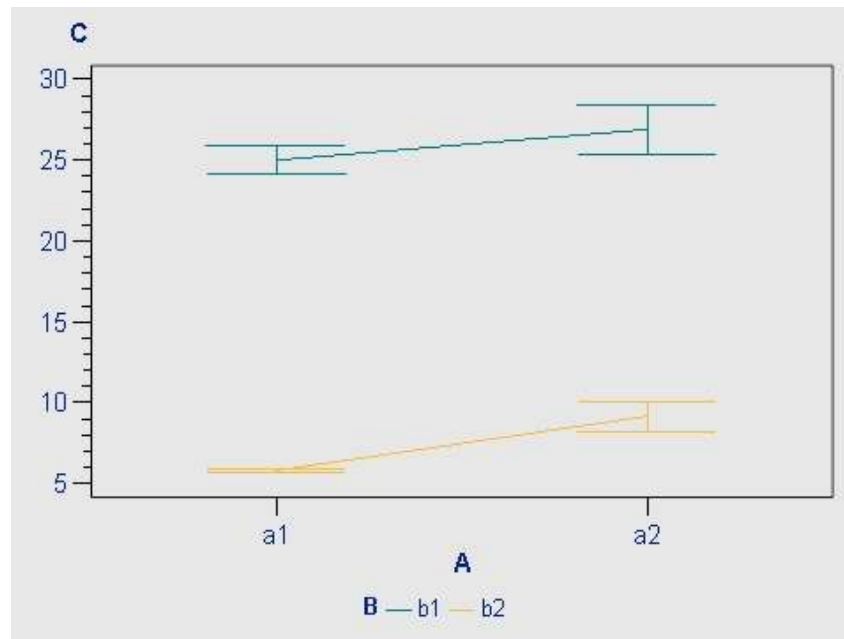
$$j = 1, \dots, n$$

$$\sum_{i=1}^a \tau_i = 0, \quad \sum_{k=1}^b \beta_k = 0$$

$$\tilde{\varepsilon}_{ikj} \sim NID(0, \sigma^2)$$

דוגמא

	▲ A	▲ B	● C
1	a1	b1	24.1
2	a1	b1	25.9
3	a1	b2	5.7
4	a1	b2	5.9
5	a2	b1	25.4
6	a2	b1	28.4
7	a2	b2	8.2
8	a2	b2	10.1



מודלים סטטיסטים ב' ארתור צ'ירגייב, יוני נצרתי

המשך דוגמא:

CLASS A B;
MODEL C= A B / SS3;

CLASS A B;
MODEL C= A B A*B / SS3;

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	696.4325000	348.2162500	193.53	<.0001
Error	5	8.9962500	1.7992500		
Corrected Total	7	705.4287500			

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	697.4837500	232.4945833	117.05	0.0002
Error	4	7.9450000	1.9862500		
Corrected Total	7	705.4287500			

R-Square	Coeff Var	Root MSE	C Mean
0.987247	8.026096	1.341361	16.71250

R-Square	Coeff Var	Root MSE	C Mean
0.988737	8.432873	1.409344	16.71250

Source	DF	Type III SS	Mean Square	F Value	Pr > F
A	1	13.7812500	13.7812500	7.66	0.0395
B	1	682.6512500	682.6512500	379.41	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
A	1	13.7812500	13.7812500	6.94	0.0579
B	1	682.6512500	682.6512500	343.69	<.0001
A*B	1	1.0512500	1.0512500	0.53	0.5072

מודל עם תצפית אחת עבור כל קומבינציה של טיפולים.

מה קורה כאשר יש תצפית אחת עבור כל קומבינציה של טיפולים?

המודל

$$y_{ik} = \underbrace{\mu + \tau_i + \beta_k + (\tau\beta)_{ik}}_{\mu_{ik}} + \tilde{\varepsilon}_{ik}$$

$$i = 1, \dots, a$$

$$k = 1, \dots, b$$

$$\sum_{i=1}^a \tau_i = 0, \quad \sum_{k=1}^b \beta_k = 0$$

$$\sum_{i=1}^a (\tau\beta)_{ik} = 0, \quad \sum_{k=1}^b (\tau\beta)_{ik} = 0$$

$$\tilde{\varepsilon}_{ik} \sim NID(0, \sigma^2)$$

סכום הריבועים מתפרק כך:

$$\underbrace{SS_{Total}}_{ab-1} = \underbrace{SS_A}_{a-1} + \underbrace{SS_B}_{b-1} + \underbrace{SS_{Residual}}_{(a-1)(b-1)}$$

SSResidual יכול לתפקד כ SSe או SSab

$$E[MS_A] = E\left[\frac{SS_A}{a-1}\right] = \sigma^2 + \frac{b \sum_{i=1}^a \tau_i^2}{a-1}$$

$$E[MS_B] = E\left[\frac{SS_B}{b-1}\right] = \sigma^2 + \frac{a \sum_{k=1}^b \beta_k^2}{b-1}$$

אמדים לשונות...

$$E[MS_{Residual}] = E\left[\frac{SS_{Residual}}{(a-1)(b-1)}\right] = \sigma^2 + \frac{\sum_{i=1}^a \sum_{k=1}^b (\tau\beta)_{ik}^2}{(a-1)(b-1)}$$

רואים שלא ניתן לעמוד את השונות בקלות. ז"א לא ניתן לבצע מבחנים לגבי הגורמים A ו B כל עוד שיש אינטראקציה.

צריך לוותר על האינטראקציה...

- אם כך, כאשר יש תצפית בודדת עבור כל קומבינציה נצטרך לוותר על האינטראקציה מהמודל.
- האם זו הנחה סבירה?
- ניתן להחליט על כך על פי המבחן הבא....

מבחן של Tukey לקיום אינטראקציה
(עבור ניתוח שונות דו-כווני עם תצפית אחת עבור כל קומבינציה)

$$(\tau\beta)_{ik} = \gamma\tau_i\beta_k$$

$$SS_N = \frac{\left[\sum_{i=1}^a \sum_{k=1}^b y_{ik} y_{i\cdot} y_{\cdot k} - y_{\cdot\cdot} (SS_A + SS_B + \frac{y_{\cdot\cdot}^2}{ab}) \right]^2}{ab SS_A SS_B}$$

לא בחומר

$$SS_{Error} = SS_{Residual} - SS_N$$

$$F_0 = \frac{SS_N}{SS_{Error} / ((a-1)(b-1)-1)}$$

$$F_0 > F_{1-\alpha, 1, (a-1)(b-1)-1}$$

המודל המתקבל

$$y_{ik} = \underbrace{\mu + \tau_i + \beta_k}_{\mu_{ik}} + \tilde{\varepsilon}_{ik}$$

$$i = 1, \dots, a$$

$$k = 1, \dots, b$$

$$\tilde{\varepsilon}_{ik} \sim NID(0, \sigma^2)$$

בפרק הבא נראה כיצד באמצעות מודל זה (ניתוח שונות עם תצפית בודדת וללא אינטראקציה) ניתן לבצע ניתוח של מבחנים עם בלוקים (הכללה של מבחנים מזווגים).