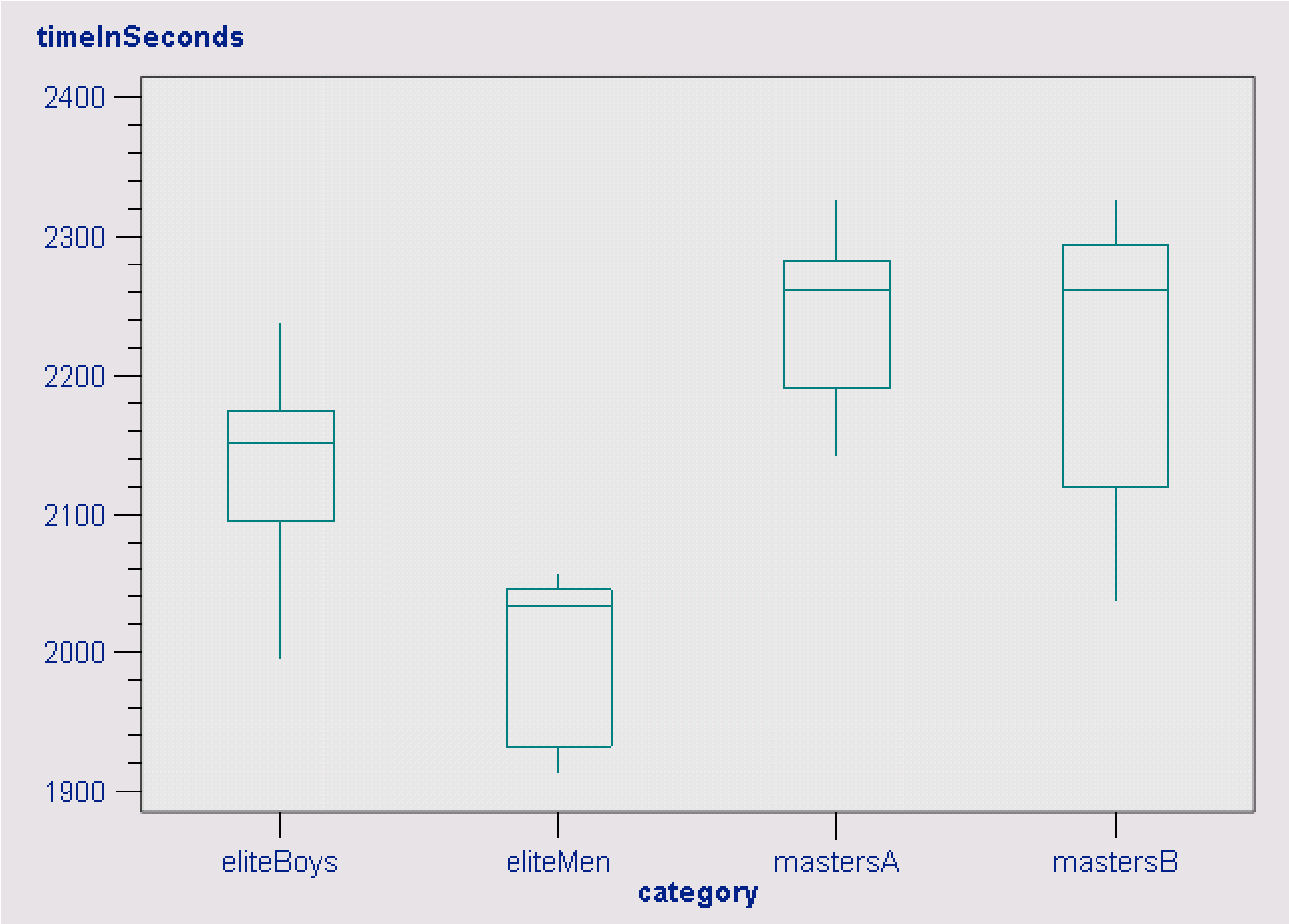
פרק ב-9

שיטות א-פרמטריות להשוואת תוחלות: Kruskal-Wallis.

אלטרנטיבה א-פרמטרית לניתוח שונות

- רוצים לבדוק אם התוחלות של a אוכלוסיות (טיפולים) הן זהות/שונות.
- כאשר לא ניתן להניח נורמאליות של הנתונים אז ייתכן ונרצה להשתמש בשיטה א-פרמטרית.
- זהו מבחן דרגות (בדומה למבחן Mann-Whitney אשר נלמד בפרק א-3).
- ההנחה היחידה שיש להניח היא שהתצפיות בלתי תלויות.
- השפעת תצפיות חריגות (outliers) מופחתת.

דוגמא: אליפות ישראל ברכיבת נגד השעון באופני כביש - 2006



ראשית נדרג את הנתונים

$$\begin{array}{ccccc} y_{11} & y_{12} & y_{13} & \cdots & y_{1n_1} \\ y_{21} & y_{22} & \cdots & y_{2n_2} & \\ \vdots & \vdots & & & \vdots \\ y_{a1} & y_{a2} & \cdots & & y_{an_a} \end{array}$$

דרוג כל הנתונים וסידור
שלהם בחזרה בתוך כל
קבוצת טיפול (כל שורה)

$$\begin{array}{ccccc} R_{11} & R_{12} & R_{13} & \cdots & R_{1n_1} \\ R_{21} & R_{22} & \cdots & R_{2n_2} & \\ \vdots & \vdots & & & \vdots \\ R_{a1} & R_{a2} & \cdots & & R_{an_a} \end{array}$$

נחשב סכום
וממוצע דרגות
עבור כל קבוצת
טיפול

$$R_{i\bullet} = \sum_{j=1}^{n_i} R_{ij}$$

$$\overline{R}_{i\bullet} = \frac{\sum_{j=1}^{n_i} R_{ij}}{n_i}$$

במקרה של שוויונות (ties) יש לרשום את הממוצע של
השוויונות ב R (כמו ב Mann-Whitney).

הערה: ניתן לבקש מ SAS לבצע דירוג

	timeInSeconds	category	rank_timeInSeco
1	1914	eliteMen	1
2	1930	eliteMen	2
3	1932	eliteMen	3
4	2012	eliteMen	5
5	2029	eliteMen	6
6	2039	eliteMen	8
7	2042	eliteMen	9
8	2046	eliteMen	10
9	2052	eliteMen	11
10	2058	eliteMen	12
11	1995	eliteBoys	4
12	2076	eliteBoys	14
13	2095	eliteBoys	15
14	2119	eliteBoys	16.5
15	2141	eliteBoys	18
16	2160	eliteBoys	21
17	2164	eliteBoys	22
18	2174	eliteBoys	23
19	2233	eliteBoys	26
20	2238	eliteBoys	27
21	2142	mastersA	19
22	2182	mastersA	24
23	2191	mastersA	25
24	2243	mastersA	29
25	2248	mastersA	30
26	2275	mastersA	31
27	2282	mastersA	33
28	2283	mastersA	34.5
29	2292	mastersA	36
30	2327	mastersA	40
31	2037	mastersB	7
32	2067	mastersB	13
33	2119	mastersB	16.5
34	2143	mastersB	20
35	2242	mastersB	28
36	2280	mastersB	32
37	2283	mastersB	34.5
38	2295	mastersB	37
39	2300	mastersB	38
40	2326	mastersB	39

ב SAS Enterprise V2.0 ניתן לדרג נתונים
ע"י Data->Rank

דוגמא לטיפול בשוויון:

…	15	16	17	18	…
…	2095	2119	2119	2141	…

$$\frac{16+17}{2}=16.5$$

השערות המבחן

- H0: התוחלת של כל הטיפולים שווה.
- H1: קיימים טיפולים בעלי תוחלת שונה.

תחת H0:
$$\overline{R}_{1\bullet} \approx \overline{R}_{2\bullet} \approx \ldots \approx \overline{R}_{a\bullet} \approx \overline{R}_{\bullet\bullet}$$

- נבנה סטטיסטי הבודק זאת המבוסס על ריבועי מרחקים של ממוצעי הדרגות לכל טיפול מהממוצע הכללי.

אינטואיציה לסטטיסטי המבחן

לצורך המחשה נניח שהמדגמים מאוזנים:

$$\sum_{i=1}^{a}(\overline{R}_{i\bullet}-\overline{R}_{\bullet\bullet})^2=\sum_{i=1}^{a}\overline{R}_{i\bullet}^{\,2}-2\overline{R}_{\bullet\bullet}\sum_{i=1}^{a}\overline{R}_{i\bullet}+a\overline{R}_{\bullet\bullet}^{\,2}$$

$$=\sum_{i=1}^{a}\overline{R}_{i\bullet}^{\,2}-2\left(\sum_{i=1}^{a}\frac{R_{i\bullet}}{an}\right)a\left(\sum_{i=1}^{a}\frac{R_{i\bullet}}{an}\right)+a\left(\sum_{i=1}^{a}\frac{R_{i\bullet}}{an}\right)^2$$

$$=\sum_{i=1}^{a}\overline{R}_{i\bullet}^{\,2}-a\overline{R}_{\bullet\bullet}^{\,2}$$

$$\boxed{\overline{R}_{\bullet\bullet}=\frac{\sum_{i=1}^{a}\sum_{j=1}^{n}R_{ij}}{N}=\frac{1+2+\ldots+an}{an}=\frac{an(an+1)/2}{an}=\frac{an+1}{2}=\frac{N+1}{2}}$$

$$=\sum_{i=1}^{a}\frac{R_{i\bullet}^{\,2}}{n^2}-a\frac{(an+1)^2}{4}=\frac{1}{n}\left[\sum_{i=1}^{a}\frac{R_{i\bullet}^{\,2}}{n}-\frac{N(N+1)^2}{4}\right]$$

בערך כך צריך להראות סטטיסטי המבחן

סטטיסטי המבחן: H

$$H = \frac{1}{S^2}\left[\sum_{i=1}^{a}\frac{R_{i\bullet}^{\ 2}}{n_i} - \frac{N(N+1)^2}{4}\right]$$

השונות המדגמית של הדרגות

$$S^2 = \frac{1}{N-1}\left[\sum_{i=1}^{a}\sum_{j=1}^{n_i} R_{ij}^{\ 2} - \frac{N(N+1)^2}{4}\right]$$

תזכורת (לשונות מדגמית):

$$S^2(x_1,\ldots,x_m) = \frac{1}{m-1}\sum_{j=1}^{m}(x_j - \bar{x})^2 = \ldots$$

$$= \frac{1}{m-1}\left[\sum_{j=1}^{m} x_j^{\ 2} - m\bar{x}^2\right]$$

בקרוב מתקיים:

$$H \overset{H_0}{\sim} \chi^2_{a-1}$$

שונות הדרגות וסטטיסטי המבחן כאשר כל התצפיות יחודיות (no ties)

במידה וכל התצפיות ייחודיות אז הדרגות הן: $1, 2, \ldots, N$

במקרה זה: $$S^2 = \frac{1}{N-1}\left[\sum_{j=1}^{N} j^2 - \frac{N(N+1)^2}{4}\right]$$

תזכורת: $$\sum_{i=1}^{N} i^2 = \frac{N(N+1)(2N+1)}{6}$$

ולכן מתקיים: $$S^2 = \frac{1}{N-1}\left[\frac{N(N+1)(2N+1)}{6} - \frac{N(N+1)^2}{4}\right] = \ldots = \frac{N(N+1)}{12}$$

$$H = \frac{12}{N(N+1)}\sum_{i=1}^{a}\frac{R_{i\bullet}^{\ 2}}{n_i} - 3(N+1)$$

כלל החלטה

- תחת H0 $H \sim \chi^2_{a-1}$ בקירוב.
- אז נדחה H0 אם: $H > \chi^2_{1-\alpha,a-1}$

ביצוע ב SAS

- הביצוע ע"י SAS-Enterprise הוא בדומה לביצוע מבחן Mann-Whitney.
  - Analysis->Anova-Non-Parametric One Way Anova
  - בחירת Wilcoxon בלבד.
- זה מפעיל את PROC NPAR1WAY אשר מזהה שיש יותר מ -2 טיפולים ולכן מבצע Kruskal-Wallis במקום Mann-Whitney.

תוצאות

Wilcoxon Scores (Rank Sums) for Variable timeInSeconds Classified by Variable category					
category	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
eliteMen	10	67.00	205.0	32.012618	6.700
eliteBoys	10	186.50	205.0	32.012618	18.650
mastersA	10	301.50	205.0	32.012618	30.150
mastersB	10	265.00	205.0	32.012618	26.500
Average scores were used for ties.					

Kruskal-Wallis Test	
Chi-Square	23.6375
DF	3
Pr > Chi-Square	<.0001

הערות

- ניתן להשתמש בWilcoxon Rank Sums לטובת ביצוע ניתוח Post-Hoc.
- ניתן לבצע ניתוח שונות חד-כווני על הדרגות. ביצוע טרנספורמציה זו על הנתונים וביצוע ניתוח שונות חד-כווני מניב תוצאות דומות למבחן Kruskal – Wallis.