

Анализа података у биологији

Марија Цупарић

Тринаести час



- ★ Претходно разматране процедуре (t —тестови, линеарна регресија, дисперзиона анализа) засноване су на претпоставци о нормалној расподели обележја
- ★ Уколико нисмо сигурну у претпоставку о нормалности, боље је користити процедуру која је не подразумева
- ★ Таква тестирања спадају у домен непараметарске статистике
- ★ Размотрићемо
 - ★ непараметарски тестови који су аналогни раније описаним тестовима
 - ★ тестови сагласности са расподелом



- ★ Имамо узорак $X_1, \dots, X_n$ из расподеле која има непознату медијану M и желимо да тестирамо у вези са неном вредношћу.
- ★ Нулта хипотеза је $H_0 : M = M_0$, а алтернатива једна од $H_1 : M < M_0$, $H_1 : M > M_0$ или $H_1 : M \neq M_0$
- ★ Од сваког елемента узорка одузмемо M_0 и пребројимо колико има позитивних (q^+), а колико негативних разлика (q^-).
- ★ Уколико имамо неку нулу, сматрамо је позитивном ако је $H_1 : M < M_0$, а негативном ако је $H_1 : M > M_0$
- ★ Случајна променљива Q^+ је број позитивних разлика (у општем узорку), а Q^- број негативних разлика.
- ★ Обе статистике Q^+ и Q^- имају биномну расподелу с параметрима n и $\frac{1}{2}$



Ако је алтернатива

$M < M_0$ тада је p —вредност теста вероватноћа да је Q^+ мање од q^+ ,
колико их има у нашем узорку, тј. $p = P\{Q^+ \leq q^+\}$,

$M > M_0$ тада је p —вредност теста вероватноћа да је Q^- мање од q^- ,
колико их има у нашем узорку, тј. $p = P\{Q^- \leq q^-\}$,

$M \neq M_0$ тада је p —вредност теста двострука вредност мање од ове две
вероватноће

За рачунање вредности p користи се биномна расподела, односно
функција у R-у: `pbinom(q,n,0.5)`

Пример Златни правоугаоник је правоугаоник код којег однос ширине (ω) и дужине (l) је златни пресек, што је приближно 0.618. У многим културама овакав правоугаоник је присутан у уметности и архитектури. Приликом проучавања културе племена Шошони проучаване су корпе са истканим правоугаонцима. Нека X представља количник ширине и дужине истканог материјала. Узет је узорак од 20 количника ширина и дужина материјала

0.553	0.570	0.576	0.601	0.606	0.606	0.609	0.611	0.615	0.628
0.654	0.662	0.668	0.670	0.672	0.690	0.693	0.749	0.844	0.933

Да ли се може рећи да ови симболи имају облик златног правоугаоника?

Желимо да тестирамо $H_0 : M = 0.618$ против $H_1 : M \neq 0.618$. Приметимо да је 9 елемената узорка мање од 0.618 ($q^- = 9$) и 11 елемената веће од 0.618 ($q^+ = 11$). p -вредност теста је

$$p = 2 \min\{P\{Q^+ \leq q^+\}, P\{Q^- \leq q^-\}\} = 2P\{Q^- \leq 9\} = 0.8238.$$

Дакле, нема основа да се одбаци нулта хипотеза.



R

```
zlatni_presek<-c( 0.553 , 0.570, 0.576, 0.601, 0.606, 0.606, 0.609,
 0.611, 0.615, 0.628, 0.654, 0.662, 0.668, 0.670, 0.672, 0.690,
 0.693, 0.749, 0.844, 0.933)

library(BSDA)
SIGN.test(zlatni_presek,md=0.618)
```

One-sample Sign-Test

```
data: zlatni_presek
s = 11, p-value = 0.8238
alternative hypothesis: true median is not equal to 0.618
95 percent confidence interval:
 0.6063494 0.6717671
sample estimates:
median of x
 0.641
```

Achieved and Interpolated Confidence Intervals:

	Conf.Level	L.E.pt	U.E.pt
Lower Achieved CI	0.8847	0.6090	0.6700
Interpolated CI	0.9500	0.6063	0.6718
Upper Achieved CI	0.9586	0.6060	0.6720

- ★ Уколико имамо индиција да је расподела симетрична (а није нормална), можемо користити Вилкоксонов тест означених рангова
- ★ Тестира се нулта хипотеза $H_0 : M = M_0$, где је M_0 претпостављена вредност медијане расподеле
- ★ Формирају се разлике $D_1 = X_1 - M_0, D_2 = X_2 - M_0, \dots, D_n = X_n - M_0$
- ★ Апсолутне вредности $|D_i|$ поређају се по величини од најмање до највеће и свакој се додели ранг од 1 до n .
- ★ Ако има једнаких елемената међу $|D_i|$, онда им се додељује средња вредност њихових рангова (нпр. ако су први и други једнаки, онда добијају ранг по 1.5)
- ★ Тест статистике су

$$W_- = \sum_{\text{по негативним } D_i} R_i \text{ или } W_+ = \sum_{\text{по позитивним } D_i} R_i$$

- ★ p -вредност се рачуна коришћењем Вилкоксонове расподеле означених рангова



Ако је алтернатива

$M < M_0$ онда се посматра статистика W_+

$M > M_0$ онда се посматра статистика W_-

$M \neq M_0$ онда посматрамо мању од W_- и W_+

p –вредност теста за вредност тест статистике w одређујемо функцијом
`psignrank(w, n)`

Пример

- ★ Ако је $n = 18$, а тестирамо против алтернативе $M > M_0$ и добијемо $W_- = 35$, тада је p –вредност једнака `psignrank(35, 18) = 0.013`
- ★ Ако је $n = 21$, а тестирамо против алтернативе $M \neq M_0$ и добијемо $W_+ = 85$ и $W_- = 146$, тада је p –вредност, узимајући у обзир мању статистику W_+ , једнака `2 · psignrank(85, 21) = 0.30`

Постоје таблице тачне расподеле тест статистике, а за велике вредности n може се користити нормална апроксимације.

Пример Проценат неписмених меѓу белцима у Америци 1969. године је била 0.7. Сматра се да је просечан проценат неписмених у великим градовима (преко пола милиона становника) већи. Да би се то испитало, изабрано је 20 градова и изведено је истраживање које је дало следеће податке: 0.6, 0.5, 0.62, 1.7, 0.75, 1.0, 0.69, 0.8, 0.8, 0.57, 0.9, 1.5, 0.95, 0.53, 1.1, 1.2, 2.0, 0.65, 0.79, 0.61. Претпостављајући симетричност расподеле, испитати наведену претпоставку.

Тестира се $H_0 : M = 0.7$ против $H_1 : M > 0.7$. Пошто је симетрија око 0.7, одузимамо ту вредност и правимо варијациони низ: -0.1 -0.2 -0.08 1.0 0.05 0.3 -0.01 0.1 0.1 -0.13 0.2 0.8 0.25 -0.17 0.4 0.5 1.3 -0.05 0.09 -0.09

$ D_i $	0.01	0.05	0.05	0.08	0.09	0.09	0.1	0.1	0.1	0.13
знак	-	-	+	-	-	+	-	+	+	-
R_i	1	2.5	2.5	4	5.5	5.5	8	8	8	10
$ D_i $	0.17	0.2	0.2	0.25	0.3	0.4	0.5	0.8	1	1.3
знак	-	-	+	+	+	+	+	+	+	+
R_i	11	12.5	12.5	14	15	16	17	18	19	20

Пошто је $w_- = 54.5$, p -вредност овог теста је $\text{psignrank}(54.5, 20) = 0.032$, па на нивоу значајности 0.05 закључујемо да је медијални проценат неписмених у великим градовим већи него на националном нивоу.



```
procenat.pismenih<-c(0.6,1.0,0.9,1.2,0.5,0.69, 1.5,2.0,0.62, 0.8,  
0.95, 0.65,1.7,0.8,0.53,0.79,0.75,0.57,1.1,0.61)  
wilcox.test(procenat.pismenih,mu=0.7,alternative="g")
```

Wilcoxon signed rank test with continuity correction

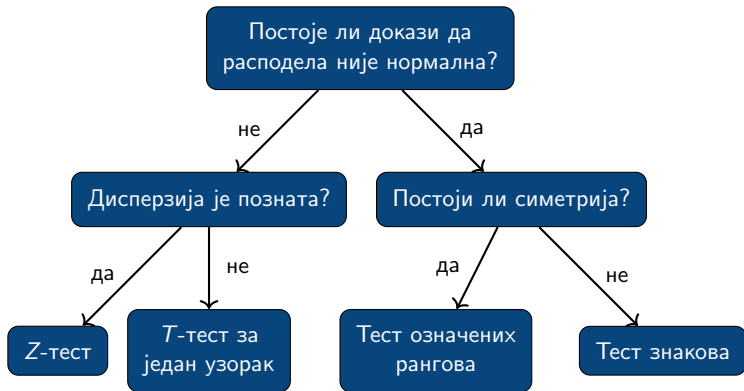
data: procenat.pismenih
V = 158, p-value = 0.02499
alternative hypothesis: true location is greater than 0.7

Warning message:

In wilcox.test.default(procenat.pismenih, mu = 0.7, alternative = "g") :
cannot compute exact p-value with ties



Схема бирања за тестирање хипотезе о параметру положаја



- ★ Тест знакова и тест означених рангова можемо применити и у случају два спарена узорка $X_1, \dots, X_n$ и $Y_1, \dots, Y_n$
- ★ Желимо да тестирамо да су просечне (медијалне) вредности једнаке за ове две променљиве ($H_0 : M_X = M_Y$), а не претпостављамо нормалну расподелу
- ★ Формирамо разлике $D_1 = X_1 - Y_1, \dots, D_n = X_n - Y_n$ и добијемо нови узорак $D_1, \dots, D_n$
- ★ Тестирамо хипотезу да је $M_D = 0$ против неке од алтернатива

претпостављамо симетрију: тест означених рангова

не претпостављамо симетрију: тест знакова



Пример Произвођач сапуна жели да покаже како његов нови сапут за лице не исушује кожу. Узет је узорак од 10 жена и свака је две недеље једном у два дана прала једну половину лица тим сапуном, а другу сапуном конкурентског произвођача. На крају је мерена сувоћа лица (на скали од 1 до 10, мањи број - већа сувоћа). Добијени су следећи резултати

нови сапун	5.0	4.3	7.3	2.1	9.8	6.9	10.0	1.5	8.2	7.3
конкурентски сапун	6.1	4.5	6.0	2.0	7.5	8.0	9.2	1.0	8.0	6.9

Тестирати да ли је нови сапун бољи од конкурентског.

Нека је X обележје повезано са новим сапуном, а Y обележје повезано са конкурентским сапуном. Тестира се $H_0 : M_D = 0$ против $H_1 : M_D > 0$, где је M_D медијана обележја $D = X - Y$. Пошто је

$D = X - Y \mid -1.1 \quad -0.2 \quad 1.3 \quad 0.1 \quad 2.3 \quad -1.1 \quad 0.8 \quad 0.5 \quad 0.2 \quad 0.4$
то је $q^- = 3$. Пошто је $Q^- \in \mathcal{B}(10, 0.5)$, p -вредност теста је

$$p = P_{H_0}\{Q^- \leq 3\} = 0.1719,$$

одакле закључујемо да немамо доказа да је сапун тог произвођача бољи од конкурентског. Лажна реклама могла имати озбиљне последице, те је не би требало правити.



```

    sapun.nas<-c(5.0,4.3,7.3,2.1,9.8,6.9,10.0,1.5,8.2,7.3)
    sapun.njihov<-c(6.1,4.5,6.0,2.0,7.5,8.0,9.2,1.0,8.0,6.9)
    SIGN.test(sapun.nas,sapun.njihov,alternative="g")
    
```

Dependent-samples Sign-Test

```

data:  sapun.nas and sapun.njihov
S = 7, p-value = 0.1719
alternative hypothesis: true median difference is greater than 0
95 percent confidence interval:
 -0.296      Inf
sample estimates:
median of x-y
      0.3
    
```

Achieved and Interpolated Confidence Intervals:

	Conf.Level	L.E.pt	U.E.pt
Lower Achieved CI	0.9453	-0.200	Inf
Interpolated CI	0.9500	-0.296	Inf
Upper Achieved CI	0.9893	-1.100	Inf



Пример Фармацеутска компанија има два метода за производњу лека за лечење уједа пчеле. Сматра се да метод A даје мање успешне резултате од метода B . Да би се ово испитало узет је узорак од 12 пацијената на којима су примењена оба третмана. Добијено је

метод A	1.5	1.4	1.4	1.0	1.1	0.9	1.3	1.2	1.1	0.9	0.7	1.8
метод B	2.0	1.8	0.7	1.3	1.2	1.5	1.1	0.9	1.5	1.7	0.9	0.9

Претпостављајући симетрију разлике ових података, испитати наведену тврдњу.

Тестира се $H_0 : M_D = 0$ против $H_1 : M_D < 0$, где је M_D медијана обележја $D = X - Y$ (X је метод A , Y метод B). Формирамо следећу табелу

D_i	-0.5	-0.4	0.7	-0.3	-0.1	-0.6	0.2	0.3	-0.4	-0.8	-0.2	0.9
$ D_{(i)} $	0.1	0.2	0.2	0.3	0.3	0.4	0.4	0.5	0.6	0.7	0.8	0.9
знак	-	-	+	-	+	-	-	-	-	+	-	+
R_i	1	2.5	2.5	4.5	4.5	6.5	6.5	8	9	10	11	12

Вредност статистике је $W^+ = 29$. p -вредност теста је 0.23, те закључујемо да нема довољно доказа да метода A даје ниже мере квалитета лека.



```
metod.A<-c(1.5,1.4,1.4,1.0,1.1,0.9,1.3,1.2,1.1,0.9,0.7,1.8)
metod.B<-c(2.0,1.8,0.7,1.3,1.2,1.5,1.1,0.9,1.5,1.7,0.9,0.9)
wilcox.test(metod.A,metod.B,paired = TRUE,alternative="l")
```

Wilcoxon signed rank exact test

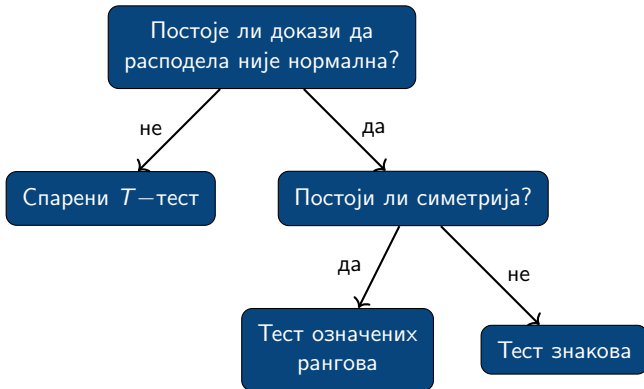
data: metod.A and metod.B

V = 28, p-value = 0.2119

alternative hypothesis: true location shift is less than 0



Схема бирања за случај спарених узорака



- ★ У случају да желимо да упоредимо медијане код два независна узорка користимо Вилкоксонов тест збира рангова
- ★ Нека је један узорак $X_1, \dots, X_m$, а други $Y_1, \dots, Y_n$ и нека је $m \leq n$
- ★ Тестирамо нулту хипотезу $H_0 : M_X = M_Y$
- ★ Сваком елементу доделимо ранг који би имао у обједињеном узорку
- ★ Статистика W_m је збир рангова елемената из мањег узорка (обима m) умањена за $\frac{m(m+1)}{2}$
- ★ p -вредност се рачуна коришћењем Вилкоксонове расподеле збира рангова



Ако је алтернатива

$M_X < M_Y$ онда p –вредност рачунамо као $P\{W_m < w_m\}$

$M_X > M_Y$ онда p –вредност рачунамо као $P\{W_m > w_m\}$

$M_X \neq M_Y$ онда p –вредност рачунамо као
 $2 \min\{P\{W_m > w_m\}, P\{W_m < w_m\}\}$

Пример

- ★ Ако је $m = 18, n = 21$, а тестирамо против алтернативе $M_X > M_Y$ и добијемо $W_m = 254$, тада је p –вредност једнака $1 - \text{pwilcox}(254, 18, 21) = 0.032$
- ★ Ако је $m = 21, n = 21$, а тестирамо против алтернативе $M \neq M_0$ и добијемо $W_m = 174$, тада је p –вредност једнака $2 \cdot \text{pwilcox}(174, 21, 21) = 0.25$

Пример Осигуравајућа компанија тврди да агенти који су прошли специјални тренинг имају више успеха у успостављању односа са потенцијалним клијентима. Да би се ово тестирало, случајно су изабрана 22 млада агента и од њих је 10 похађало специјални тренинг. Затим су оцењивани приликом пробних интервјуа с клијентима. Добијени су следећи резултати (просечна оцена са три интервјуа)

X	8.1	7.9	9.0	4.3	7.0	9.1	7.2	8.0	3.1	9.0		
Y	9.1	6.3	2.5	6.0	0.0	2.0	7.0	9.0	5.5	9.7	5.1	1.0

Испитати тврдњу компаније.

Сортирани целокупан узорак и одговарајући рангови су дати у табели:

узорак	0.0	1.0	2.0	2.5	3.1	4.3	5.1	5.5	6.0	6.3	7.0
популација	Y	Y	Y	Y	X	X	Y	Y	Y	Y	X
ранг	1	2	3	4	5	6	7	8	9	10	11.5
узорак	7.0	7.2	7.9	8.0	8.1	9.0	9.0	9.0	9.1	9.1	9.7
популација	Y	X	X	X	X	X	X	Y	Y	X	Y
ранг	11.5	13	14	15	16	18	18	18	20.5	20.5	22

Тестира се хипотеза $H_0 : M_X = M_Y$ против $H_1 : M_X > M_Y$. Реализована вредност тест статистике је $w_X = 137$, па је $w_m = 82$. p -вредност теста је 0.07. Према томе немамо доказа да се обуком постиже бољи резултат.



```
trenirani<-c(8.1,7.9,9.0,4.3,7.0,9.1,7.2,8.0,9.0,3.1)
kontrolni<-c(9.1,6.3,2.5,6.0,0.0,2.0,7.0,5.5,1.0,9.0,9.7,5.1)
wilcox.test(trenirani,kontrolni,alternative="g")
```

Wilcoxon rank sum test with continuity correction

data: trenirani and kontrolni

W = 82, p-value = 0.07779

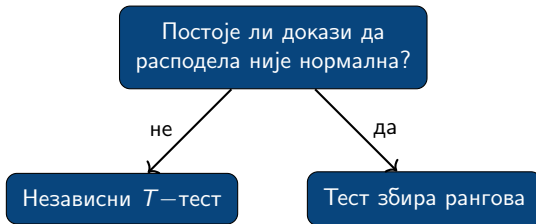
alternative hypothesis: true location shift is greater than 0

Warning message:

In wilcox.test.default(trenirani, kontrolni, alternative = "g") :
cannot compute exact p-value with ties



Схема бирања за случај независних узорака



Случај више узорака

- ★ Имамо k група и желимо да тестирамо да су просечне вредности неког обележја у свим групама једнаке
- ★ Уколико претпостављамо нормалну расподелу обележја, користимо једнофакторску дисперзиону анализу (ако су узорци независни) или блок дизајн (ако су зависни)
- ★ Ако имамо доказе да расподела обележја није нормална, онда ако су узорци

независни користимо Крускал-Валисов тест

зависни користимо Фридманов тест



Крускал-Валисов тест

- ★ Имамо k група и независне узорке у свакој од њих обима n_i (укупно n)
- ★ Тестирамо хипотезу да су им медијане једнаке $H_0 : M_1 = \dots = M_k$
- ★ Одредимо рангове свих елемената узорка и нека је R_i збир рангова у i -тој групи
- ★ Тест статистика је

$$H = \frac{12}{n(n+1)} \sum_{i=1}^k \frac{R_i^2}{n_i} - 3(n+1)$$

- ★ H има χ^2 расподелу са параметром $k - 1$
- ★ p -вредност теста је површина χ_{k-1}^2 расподеле десно од h_0 , вредности коју је H узело у узорку.



Пример Упоређује се ниво загађености три велике реке. Количина загађења мерена је на по пет места у свакој реци. Добијени су подаци (у заградама је ранг сваког елемента узорка)

прва река	друга река	трећа река
2.7 (13)	2.9 (14)	0.6 (1)
1.4 (4)	2.4 (11.5)	1.2 (2.5)
2.0 (8)	3.7 (15)	1.5 (5)
1.2 (2.5)	1.6 (6)	1.7 (7)
2.1 (9.5)	2.4 (11.5)	2.1 (9.5)

Из табеле добијамо да су збирови рангова $R_1 = 37$, $R_2 = 58$, $R_3 = 25$. Вредност тест статистике у овом узорку је

$$h_0 = \frac{12}{15 \cdot 16} \left(\frac{37^2}{5} + \frac{58^2}{5} + \frac{25^2}{5} \right) - 3 \cdot 16 = 5.58.$$

На основу χ^2_2 расподеле добијамо да је p -вредност теста 0.06. Закључак доносимо у зависности од нивоа значајности α . Ако је $\alpha = 0.05$, онда нема довољно доказа о различитом нивоу загађености река.





```
zagadjenost<-data.frame(zagadjenje=c(2.7,2.9,0.6,1.4,2.4,1.2,  
  2.0,3.7, 1.5, 1.2, 1.6,1.7,2.1,2.4,2.1),  
  reka=c(1,2,3,1,2,3,1,2,3,1,2,3,1,2,3))  
kruskal.test(zagadjenje~reka,zagadjenost)
```

Kruskal-Wallis rank sum test

data: zagadjenje by reka

Kruskal-Wallis chi-squared = 5.6101, df = 2, p-value = 0.06051



- ★ Имамо k група и b блокова, укупно kb елемената (исто као код блок дизајна)
- ★ Тестирамо хипотезу да су им медијане једнаке $H_0 : M_1 = \dots = M_k$
- ★ Одредимо рангове елемената у оквиру својих блокова и нека је R_i збир рангова у i -тој групи
- ★ Тест статистика је

$$S = \frac{12}{bk(k+1)} \sum_{i=1}^k \left(R_i - \frac{b(k+1)}{2} \right)^2$$

- ★ S има χ^2 расподелу с параметром $k - 1$
- ★ p -вредност теста је површина χ^2_{k-1} расподеле десно од s_0 , вредности коју је S узело у узорку



Пример Упоређује се квалитет три врсте кочница за бицикле. Сматра се да и марка бицикла има утицај на перформансе кочница, па је узето шест познатијих марки и формирано 6 блокова. Свака кочница тестирана је на сваком бициклу и мерен је број недеља коришћења пре првог сервиса. Добијени су подаци (у заградама је ранг по блоку)

марка бицикла	врста кочнице		
	<i>A</i>	<i>B</i>	<i>C</i>
<i>S</i>	5.2 (2)	7.3 (3)	3.0 (1)
<i>V</i>	6.8 (1)	8.9 (3)	7.5 (2)
<i>JH</i>	6.3 (2.5)	6.3 (2.5)	6.0 (1)
<i>R</i>	13.0 (1.5)	14.8 (3)	13.0 (1.5)
<i>C</i>	12.8 (2.5)	12.8 (2.5)	11.0 (1)
<i>Ra</i>	15.0 (2)	15.2 (3)	14.5 (1)

Из табеле добијамо да су збирови рангова $R_1 = 11.5$, $R_2 = 17$, $R_3 = 7.5$. Вредност тест статистике у овом узорку је

$$s_0 = \frac{12}{6 \cdot 3 \cdot 4} \left((11.5 - 12)^2 + (17 - 12)^2 + (7.5 - 12)^2 \right) = 7.58.$$

На основу χ^2_2 расподеле добијамо је p -вредност теста 0.02, што значи да закључујемо да постоји разлика између ових врста кочница. Судећи по томе што кочница *B* има највећи ранг, препоручује се њена употреба.



```
bicikle<-data.frame(rok.trajanja=c(5.2,7.3,3.0,6.8,8.9,7.5, 6.3,
6.3, 6.0,13.0,14.8,13.0,12.8,12.8,11.0,15.0,15.2,14.5),
blok=c(1,1,1,2,2,2,3,3,3,4,4,4,5,5,5,6,6,6),
vrsta=c(1,2,3,1,2,3,1,2,3,1,2,3,1,2,3,1,2,3))
friedman.test(rok.trajanja~vrsta|blok,bicikle)
```

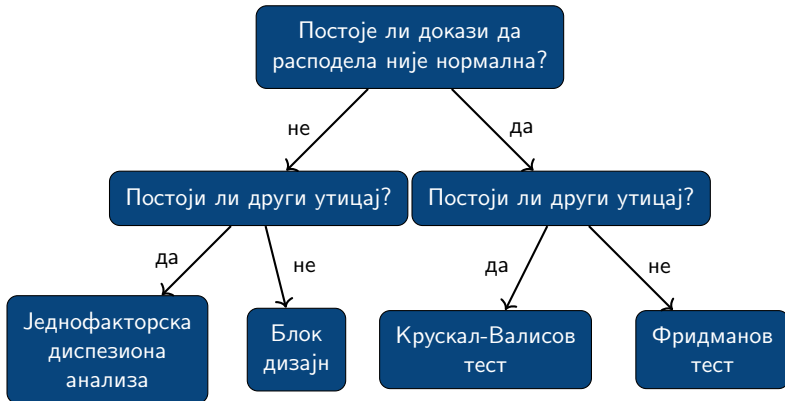
Friedman rank sum test

data: rok.trajanja and vrsta and blok

Friedman chi-squared = 8.6667, df = 2, p-value = 0.01312



Схема бирања теста за случај више узорака



- ★ Проверавају хипотезу о статистичком моделу
- ★ Нека је X обележје са функцијом расподеле F
- ★ Тестирамо $H_0 : F = F_0$, где је F_0 претпостављена распоела
- ★ Размотрићемо само тестове засноване на емпиријској функцији расподела
- ★ Претпоставимо да је F_0 апсолутно непрекидна функција расподеле и да не зависи од непознатих параметара



- ★ Тест статистике

$$D_n = \sup_x |F_n(x) - F_0(x)|,$$

где је $F_n(x) = \frac{1}{n} \sum_{i=1}^n I\{X_i \leq x\}$

- ★ p -вредност теста је површина десно од вредности d_n коју је статистика D_n узела у узорку
- ★ У R-у се користи функција `ks.test`

Крамер-фон Мизесов тест

- ★ Тест статистике

$$\omega_n^2 = \int_{-\infty}^{\infty} (F_n(x) - F_0(x))^2 dF_0(x),$$

где је $F_n(x) = \frac{1}{n} \sum_{i=1}^n I\{X_i \leq x\}$

- ★ Велике вредности тест статистике ω_n^2 говоре у прилог несагласности расподеле F_n са F_0
- ★ У R-у се користи функција `cvm.test` из пакета `goftest`

Пример



```
x<-c(1.48,0.5,1.62,1.76,1.34,0.64,0.78,2.18,2.32,1.9,2.38,0.92,
      1.06,2.04,1.2)
ks.test(x, "pnorm", 1.5, 0.6)
library(goftest)
cvm.test(x, "pnorm", 1.5, 0.6)
```

Exact one-sample kolmogorov-smirnov test

```
data: x
D = 0.10166, p-value = 0.9933
alternative hypothesis: two-sided
```

Cramer-von Mises test of goodness-of-fit
 Null hypothesis: Normal distribution
 Parameters assumed to be fixed

```
data: x
omega2 = 0.03224, p-value = 0.9724
```