УВОД У СТАТИСТИКУ час 13

24. maj ’17.

Пример 1

У сваком статистичком пакету централно место има генератор (псеудо)случајних бројева. Доступан је велики број „тестова“ за проверу да ли су симуларни подаци, које продукује генератор, заиста случајни, у смислу одређеног задатог критеријума. Један такав поступак је тест медијане.

Нека је $(y_1, y_2, \dots, y_n)$ n-торка реалних бројева за које се тврди да потичу из обележја Y апсолутно непрекидног типа, са густином расподеле $f_Y(y)$. Нека је k укупан број y_i-ова који су мањи од вредности медијане обележја Y. Ако је ова n-торка заиста један случајан узорак, очекује се да је разлика вредности $\frac{k}{n}$ и $\frac{1}{2}$ мала. Прецизније, реализовани 95% интервал поверења базиран на $\frac{k}{n}$ требало би да садржи вредност $\frac{1}{2}$.

Предложеним генератором добијен је скуп од 60 бројева (наводно) из $\varepsilon(1)$ расподеле:

```
 [1] 2.29622 0.04421 2.53442 0.13955 0.19739 0.46629 0.73158 1.84987 1.51958 1.06313 0.25776 0.51230 1.07546 1.05048
[15] 0.24797 1.18540 0.13040 1.26298 2.53444 0.72599 0.39077 0.11618 2.28168 0.88736 1.28423 0.30041 0.11628 1.79347
[29] 0.07076 2.14579 0.44492 0.07515 0.96897 1.47463 0.50969 1.46188 0.74412 0.15933 0.09408 1.17598 2.44975 1.59606
[43] 1.18172 0.99204 0.11276 1.84582 2.11751 0.31247 0.84832 2.88264 0.94870 0.91993 0.30825 2.56339 0.38480 0.33656
[57] 0.48050 2.11235 0.03130 1.43758
```

Шта се може закључити применом теста медијане??

Решење:

Медијана обележја $Y \in \varepsilon(1)$ је вредност m_e, таква да важи:

$$P\{Y \leq m_e\} = 0.5 \Rightarrow m_e = \ln 2 \approx 0.69135.$$

Од 60 симулираних бројева њих тачно $k = 25$ је на реалној правој позиционирано лево у односу на медијану. Дакле, за дати узорак $\frac{k}{n} = \frac{5}{12}$.

Нека је са p означена (непозната) вероватноћа да појединачан симулирани број буде мањи од медијане обележја.

Реализован (апроксимативни) 95% интервал поверења за p је:

$$\left[\frac{5}{12} - \frac{z_{0.95}}{\sqrt{60}} \cdot \sqrt{\frac{5}{12} \cdot \frac{7}{12}}, \frac{5}{12} + \frac{z_{0.95}}{\sqrt{60}} \cdot \sqrt{\frac{5}{12} \cdot \frac{7}{12}}\right] \approx (0.203, 0.631).$$

Чињеница да вредност 0.5 припада овом интервалу имплицира да је добијени узорак „прошао“ тест медијане. △

Непараметарски тестови

- Пример 2

Предузимач набавља велики број флуоресцентних сијалица. Од стране произвођача сијалица речено му је да нису све сијалице истог квалитета. Наиме, поступак производње је такав да свака сијалица, независно од осталих, има ниво квалитета A, B, C, D или E (најнижи квалитет) са вероватноћама, редом, 0.15, 0.25, 0.35, 0.20, 0.05. Предузимачу се, међутим, чини да му произвођач испоручује превише сијалица нивоа E, па је одлучио да тестира тврдњу произвођача анализирајући квалитет 105 купљених сијалица које је одабрао на случајан начин.

Анализа узорка сијалица показала је следеће: 10 сијалица има ниво квалитета A, 21 ниво квалитета B, 32 ниво квалитета C, 25 ниво квалитета D, а преостале имају најнижи ниво квалитета E.

Да ли ови подаци, за праг значајности $\alpha = 0.05$, оспоравају тврдњу произвођача??

Решење:

Прво уочити да постоји пет класа које одговарају нивоима квалитета флуоресцентних сијалица. Табела са реализованим, односно очекиваним учестаностима, по класама је:

	A	B	C	D	E	Укупно
m_j	10	21	32	25	17	105
$105 \cdot p_j$	15.75	26.25	36.75	21	5.25	105

Реализована вредност тест статистике је:

$$\chi_U^2 = \sum_{j=1}^{5} \frac{m_j^2}{105 \cdot p_j} - 105 \approx 30.82268.$$

Показати да је:
$\sum_{j=1}^{r} \frac{(M_j - np_j)^2}{np_j} = \sum_{j=1}^{r} \frac{M_j^2}{np_j} - n$ ←

Затим се одређује њена P-вредност (читањем из таблица χ^2 расподеле):

$$P\{T \geq \chi_U^2\}, \qquad \text{где } T \in \chi_4^2,$$

и она је једнака $\approx 3.327 \cdot 10^{-6}$.

Дакле, тврдњу произвођача сијалица би требало одбацити као нетачну за задати праг значајности. △

Пример 3

Дат је случајан узорак обима 40 бројева за које се тврди да потиче из обележја X са густином расподеле $f_X(x) = 6x(1-x)$, $0 \leq x \leq 1$.

```
 [1] 0.18 0.06 0.27 0.58 0.98 0.55 0.24 0.58 0.97 0.36 0.48 0.11 0.59 0.15 0.53 0.29 0.46 0.21 0.39 0.89 0.34 0.09 0.64
[24] 0.52 0.64 0.71 0.56 0.48 0.44 0.40 0.80 0.83 0.02 0.10 0.51 0.43 0.14 0.74 0.75 0.22
```

За праг значајности $\alpha = 0.05$ спровести χ^2-тест сагласности.

Решење:

Подаци прво морају бити груписани у класе. У следећој табели налазе се потребне вредности при једној могућој подели на класе:

	$0 \leq x < 0.2$	$0.2 \leq x < 0.4$	$0.4 \leq x < 0.6$	$0.6 \leq x < 0.8$	$0.8 \leq x < 1$
m_j	8	8	14	5	5
p_j	0.104	0.248	0.296	0.248	0.104
$40 \cdot p_j$	4.16	9.92	11.84	9.92	4.16

При томе p_j је вероватноћа да обележје X узме вредност у j-тој класи, ако је тачна хипотеза H_0, тј. она се израчунава на основу претпостављене густине расподеле.

Приметити да је у две (од предложених пет) класа $np_j < 5$, па се морају извршити извесне корекције поделе на класе.

Решење: (наставак)

Спајањем прве две, односно последње две класе постиже се задовољеност услова за примену χ^2-теста:

	$0 \leq x < 0.4$	$0.4 \leq x < 0.6$	$0.6 \leq x < 1$
m_j	16	14	10
p_j	0.352	0.296	0.352
$40 \cdot p_j$	14.08	11.84	14.08

Реализована вредност тест статистике је:

$$\chi^2_U = \sum_{j=1}^{3} \frac{m_j^2}{40 \cdot p_j} - 40 \approx 1.83814.$$

Критична област је:

$$[5.99146, +\infty),$$

па нема разлога за одбацивање тврдње да узорак потиче из предложене расподеле за задати праг значајности. △

Пример 4

У циљу побољшања заштите на раду анализиран је број несрећа на случајном узорку обима 398 радника, који обављају одређену врсту посла, изложени, при томе, ризику од повреде. Подаци из узорка дати су у следећој табели:

Број несрећа	0	1	2	3	4	5	6	7	8	9	10	11	12
Број радника	14	37	76	70	64	53	31	19	14	9	5	5	1
Укупно													398

Питање од интереса је да ли је број несрећа, које су имали радници, расподељен према закону „ретких догађаја“, тј. у складу са Пуасоновом расподелом. Тестирати ову претпоставку за праг значајности $\alpha = 0.05$.

Решење:

Параметар предложене Пуасонове расподеле је непознат, па би га прво требало оценити ММВ: $\hat{\lambda} = \frac{1549}{389} \approx 3.89$.

Табела са реализованим, односно очекиваним (теоријским) учестаностима, по класама, је:

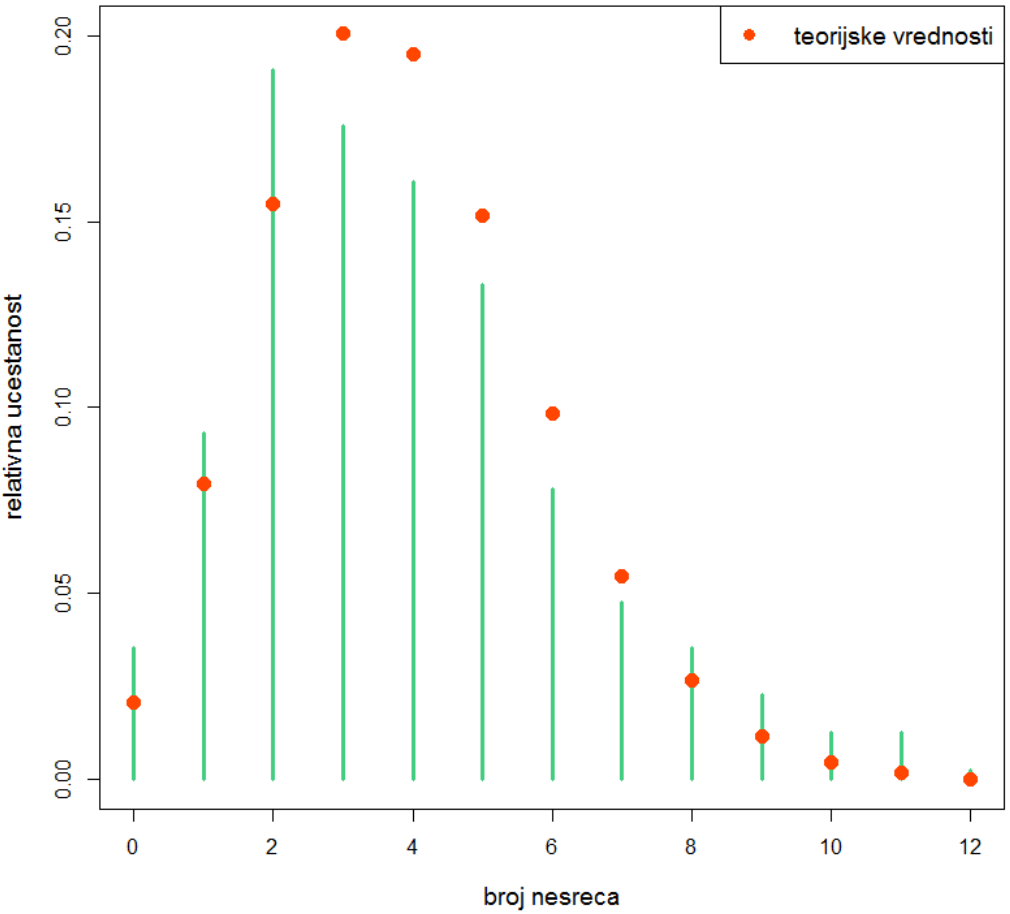
	0	1	2	3	4	5	6	7	8	9	10	11	12 и више
m_j	14	37	76	70	64	53	31	19	14	9	5	5	1
$398 \cdot p_j$	8.14	31.65	61.57	79.83	77.64	60.40	39.16	21.76	10.58	4.57	1.78	0.63	0.29

Решење: (наставак)

Спајањем последње четири класе постиже се задовољеност услова за примену χ^2-теста:

	0	1	2	3	4	5	6	7	8	9 и више
m_j	14	37	76	70	64	53	31	19	14	20
$398 \cdot p_j$	8.14	31.65	61.57	79.83	77.64	60.40	39.16	21.76	10.58	7.27

Dijagram raspodele verovatnoca



Реализована вредност тест статистике је: $\chi^2_U \approx 38.46483$.

Критична област је:

$$[15.57031, +\infty).$$

Претпоставку о Пуасоновој расподели посматраног обележја требало би одбацити за задати праг значајности. △

Пример 5

За податке у следећој табели тврди се да потичу од 144 понављања сл. експеримента, који се састоји у бацању две хомогене коцкице за игру и одређивању збира цифара добијених на њиховим горњим странама.

У табели су дати (наводни) резултати и одговарајуће теоријске вредности:

Збир	2	3	4	5	6	7	8	9	10	11	12
Број експеримената	2	4	10	12	22	29	21	15	14	9	6
Очекивани број експеримената	4	8	12	16	20	24	20	16	12	8	4
Укупно											144

За праг значајности $\alpha = 0.05$ χ^2-тестом испитати да ли су ово могући резултати бацања пара „фер“ коцкица.

Решење:

Након спајања суседних класа (прве две, односно последње две), ради постизања задовољености услова за примену χ^2-теста, добија се реализована вредност тест статистике:

$$\chi^2_U \approx 6.77083.$$

Овде се узима да је критична област двострана:

$$(0, 2.17973] \cup [17.53455, +\infty),$$

па, за дати праг значајности, нема разлога за одбацивање тврдње да су обе коцкице, коришћене у експериментима, хомогене (фер) коцкице. Дакле, добијени узорак има задовољавајући степен случајности, према критеријумима у вези са овим тестом. △

Пример 6

На случајан начин изабран је узорак обима 300 пунолетних грађана једне општине. Они су се изјашњавали на две теме: политичко опредељење и став према увођењу смртне казне. Резултати су дати у следећој табели контингенције:

	Републиканци	Демократе	Неопредељени	Укупно
ЗА	68	56	32	156
ПРОТИВ	52	72	20	144
Укупно	120	128	52	300

За праг значајности $\alpha = 0.05$ тестирати претпоставку да су политичко опредељење и став према смртној казни међусобно независни.

Решење:

Излаз функције (статистички пакет ):

```
        Pearson's Chi-squared test

data:  tab
X-squared = 6.4329, df = 2, p-value = 0.0401
```

Критична област је:

$$[5.99146, +\infty),$$

па би требало одбацити нулту хипотезу о независности посматраних обележја, за задати праг значајности. △

Највише примена табела контингенције у вези је са квалитативним обележјима.

Понекад, међутим, ове табеле представљају врло погодан начин за тестирање независности два обележја X и Y, од којих је бар једно оригинално квантитативно обележје. Тада је потребно прво извршити дискретизацију датог/датих обележја, формирајући интервале у $\mathbb{R}$.

На тај начин конструишу се табеле контингенције у којима су узорачки подаци заправо дати интервално; на основу ње се даље, на уобичајен начин, врши тестирање хипотезе:

H_0(обележја X и Y су независна)).