Задатак 1. Један лек помаже у лечењу неке болести у 80% случајева. Нови лек за ту болест је помогао у 250 од 300 случајева. Са прагом значајности 0.03 тестирати хипотезу да је нови лек ефикаснији од старог.

Решење. Нека је обележје X такво да има вредност 1 на јединкама којима је помогао нови лек, а 0 на јединкама којима није помогао. Расподела обележја X је тада:

$$X : \begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix},$$

где је p вероватноћа да нови лек помогне у лечењу.

Добијен је узорак обима 300 такав да је за 250 јединки одабраних у узорак вредност обележја X једнака 1. Дакле, ако је реализовани узорак $(x_1, ..., x_{300})$, важи да:

$$\sum_{k=1}^{300} x_k = 250, \;\; x_k \in \{0, 1\}, \;\; k = 1, ..., 300.$$

На основу добијеног узорка, желимо да закључимо да ли је нови лек ефикаснији од старог. Сматрамо да је нови лек ефикаснији ако је $p > 0.8$, односно ако је вероватноћа да ће пацијенту помоћи нови лек већа него код старог лека. Можемо да тестирамо $H_0(p = 0.8)$ против $H_1(p > 0.8)$. Ако на основу узорка добијемо да треба одбацити нулту хипотезу, закључићемо да је нови лек ефикаснији.

Критичну област формирамо на следећи начин:

$$W = \left\{ \sum_{k=1}^{300} x_k \geq c \right\}.$$

Из задатог прага значајности можемо одредити c тако да важи

$$P_{H_0} \left\{ \sum_{k=1}^{300} X_k \geq c \right\} = 0.03.$$

При H_0:

$$X_k : \begin{pmatrix} 0 & 1 \\ 0.2 & 0.8 \end{pmatrix},$$

према томе $E(X_k) = 0.8$ и $D(X_k) = 0.8(1 - 0.8) = 0.16, \;\; k = 1, ..., 300.$

Обим узорка је $n = 300 > 30$, случајне величине $X_1, ..., X_n$ су независне и једнако расподељене са коначним очекивањем и дисперзијом, па можемо искористити централну граничну теорему да добијемо расподелу тест статистике при H_0

$$\frac{\sum\limits_{k=1}^{300} X_k - E(\sum\limits_{k=1}^{300} X_k)}{\sqrt{D(\sum\limits_{k=1}^{300} X_k)}} \xrightarrow{D} X^* \in \mathcal{N}(0, 1), \;\; n \to \infty.$$

Сада можемо да израчунамо c. При H_0:

$$E\left(\sum_{k=1}^{300} X_k\right) = 300 \cdot 0.8 = 240,$$
$$D\left(\sum_{k=1}^{300} X_k\right) = 300 \cdot 0.16 = 48,$$

па:

$$P_{H_0}\left\{\frac{\sum_{k=1}^{300} X_k - E\left(\sum_{k=1}^{300} X_k\right)}{\sqrt{D\left(\sum_{k=1}^{300} X_k\right)}} \geq \frac{c-240}{\sqrt{48}}\right\} = 0.03.$$

Дакле,

$$\frac{c-240}{4\sqrt{3}} = \Phi^{-1}(0.97) = 1.88$$
$$\implies c = 4\sqrt{3}\cdot 1.88 + 240 = 253.03.$$

Критична област је:

$$W = \left\{\sum_{k=1}^{300} x_k \geq 253.03\right\}.$$

Проверавамо да ли узорак упада у критичну област, односно видимо да је $250 < 253.03$, па узорак не упада у критичну област. Дакле, на основу овог узорка нећемо одбацити H_0, па закључујемо да нови лек није ефикаснији од старог. ✓

Задатак 2. Број пријављених кандидата за упис на Математицхки факултет на студијски програм Информатика има нормалну $\mathcal{N}(m_1, \sigma^2)$ расподелу, а на програм Математика нормалну $\mathcal{N}(m_2, 0.45\sigma^2)$ расподелу, где је $\sigma > 1$ и познато. На случајном узорку од 12 година добијено је да је просечан број пријављених кандидата на студијском програму Информатика 294.6, док је на случајном узорку од 10 година добијено да је просечан број пријављених кандидата на студијском програму Математика 191.3. Тестира се хипотеза да је разлика у средњим вредностима броја пријављених кандидата на студијским програмима Информатика и Математика једнака 100, против алтернативе да је разлика већа од 100. Формиран је тест такав да је вероватноћа тачног прихватања нулте хипотезе 7 пута већа од вероватноће погрешног одбацивања нулте хипотезе. Такође, моћ предложеног теста, при алтернативи да је разлика у средњим вредностима броја пријављених кандидата на студијским програмима Информатика и Математика једнака $105 - 0.25\sigma$, износи 0.85. Одредити који закључак треба донети на основу датих узорака.

Решење. Нека су X и Y обележја које представља број пријављених кандидата на студијске програме Информатика и Математика, редом. Тада је $X \in \mathcal{N}(m_1, \sigma^2)$ и $Y \in \mathcal{N}(m_2, 0.45\sigma^2)$, те је на основу датог узорка $\bar{x}_{12} = 294.6$ и $\bar{y}_{10} = 191.3$. Потребно је да тестирамо $H_0(m_1 - m_2 = 100)$ против $H_1(m_1 - m_2 > 100)$, па посматрамо критичну област

$$W = \{\bar{x}_{12} - \bar{y}_{10} \geq c\}.$$

Како су случајне величине $\bar{X}_{12}$ и $\bar{Y}_{10}$ независне и $\bar{X}_{12} \in \mathcal{N}\left(m_1, \frac{\sigma^2}{12}\right)$ и $\bar{Y}_{10} \in \mathcal{N}\left(m_2, \frac{0.45\sigma^2}{10}\right)$, то је $\bar{X}_{12} - \bar{Y}_{10} \in \mathcal{N}(m_1 - m_2, 0.128\sigma^2)$. Са једне стране имамо да је

$$P_{H_0}\{X \notin W\} = 7P_{H_0}\{X \in W\} \quad \implies \quad 1 - P_{H_0}\{X \in W\} = 7P_{H_0}\{X \in W\}$$

па је

$$P_{H_0}\left\{\frac{\bar{X}_{12} - \bar{Y}_{10} - 100}{\sqrt{0.128\sigma^2}} \geq \frac{c-100}{\sqrt{0.128\sigma^2}}\right\} = 0.125,$$
$$\frac{c-100}{\sqrt{0.128\sigma^2}} = \Phi^{-1}(0.875) = 1.15,$$

$$c = 100 + 0.412\sigma. \tag{1}$$

Са друге стране, при $H_1(m_1 - m_2 = 105 - 0.25\sigma)$, моћ теста је

$$P_{H_1}\{X \in W\} = 0.85,$$

$$P_{H_1}\left\{\frac{\bar{X}_{12} - \bar{Y}_{10} - 105 + 0.25\sigma}{\sqrt{0.128\sigma^2}} \geq \frac{c - 105 + 0.25\sigma}{\sqrt{0.128\sigma^2}}\right\} = 0.125,$$

$$\frac{c - 105 + 0.25\sigma}{\sqrt{0.128\sigma^2}} = \Phi^{-1}(0.15) = -1.04,$$

$$c = 105 - 0.622\sigma. \tag{2}$$

Из (1) и (2) следи да је

$$100 + 0.412\sigma = 105 - 0.622\sigma$$

тј.

$$\sigma = 4.836.$$

Сада је $c = 100 + 0.412\sigma = 100 + 0.412 \cdot 4.836 = 101.992$, па је критична област

$$W = \{\bar{x}_{12} - \bar{y}_{10} \geq 101.992\}.$$

На основу података из задатка имамо да је $\bar{x}_{12} - \bar{y}_{10} = 103.3$, па реализовани узорци упадају у критичну област и одбацујемо нулту хипотезу. ✓

Непараметарски тестови

До сада смо се бавили параметарским проблемима, где смо претпостављали да је расподела обележја од интереса одређена до на непознати параметар. Дакле, ако је обележје од интереса X и X има расподелу $F(\cdot, \theta)$, претпостављали смо да знамо тип расподеле F, али не знамо вредност параметра θ (који може бити и вектор). Код непараметарских проблема, не знамо ни тип расподеле F. На основу узорка, код непараметарског тестирања хипотеза, тестирамо да ли је расподела обележја X једнака некој конкретној расподели, тј. тестирамо

$$H_0(F = F_0) \quad \text{против} \quad H_1(F \neq F_0).$$

Пирсонов χ^2 тест

Пирсонов χ^2 тест је један од тестова за тестирање непараметарских хипотеза. Тестирање се спроводи на следећи начин:

- Скуп вредности које узима случајна величина X (узорачки простор) се дели на r дисјунктних подскупова - $S_1, ..., S_r$.
- Ако постоје непознати параметри, оцењују се методом максималне веродостојности.
- Нека је M_k број елемената из узорка обима n који су упали у S_k. Тада је $n = \sum_{k=1}^{r} M_k$.
- Нека је $p_k = P_{H_0}\{X \in S_k\}$. Тада, при H_0, $M_k \in \mathcal{B}(n, p_k)$ расподелу, $k = 1, ..., r$.
- Тест статистика која се користи је $\chi_0^2 = \sum_{k=1}^{r} \frac{(M_k - np_k)^2}{np_k}$ и при H_0 важи да $\chi_0^2 \in \chi^2_{r-s-1}$, где је s број непознатих параметара, оцењених методом максималне веродостојности.

- Критична област је облика $W = \{\chi_0^2 \geq c\}$, где c рачунамо из задатог нивоа значајности теста α и познате расподеле тест статистике при H_0:

$$P_{H_0}\{\chi_0^2 \geq c\} = \alpha.$$

Напомена: Увек прво проверити да ли узорак који је доступан у задатку покрива целу област дефинисаности расподеле која се тестира. Уколико не покрива, треба додати недостајуће вредности и доделити им фреквенције појављивања елемената из узорка једнаке 0. Такође, мора бити $np_k \geq 5$, за свако $k = 1, ..., r$. Ако ово није случај, спајамо класе са најмањим np_k, све док не добијемо да овај услов важи.

Задатак 3. Из популације чије је обележје X извучен је узорак:

X_k	1	2	3	4	5	$\geq$6
M_k	45	30	15	6	2	2

Са прагом значајности 0.05 тестирати хипотезу да обележје X има закон расподеле $P\{X = k\} = \frac{1}{2^k},\ k \in \mathbb{N}$.

Решење. Имамо 6 класа, којима су покривене све вредности које може да узме случајна величина X. За почетак, рачунамо вредности np_k, $k = 1, ..., 6$, да видимо да ли ћемо морати да спојимо неке класе (ако је негде $np_k < 5$).

Обим узорка је $n = 45 + 30 + 15 + 6 + 2 + 2 = 100$.

Рачунамо вероватноће $p_k = P_{H_0}\{X \in S_k\}$, где S_k означава k-ту класу:

$$p_k = P_{H_0}\{X = k\} = \frac{1}{2^k},\ \ k = 1, 2, 3, 4, 5,$$

$$p_6 = P_{H_0}\{X \geq 6\} = 1 - P_{H_0}\{X \leq 5\} = 1 - \sum_{k=1}^{5} p_k.$$

Добијају се следеће вредности:

X_k	1	2	3	4	5	$\geq$6
M_k	45	30	15	6	2	2
p_k	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{32}$
np_k	50	25	12.5	6.25	3.125	3.125

Пошто је $np_5 < 5$ и $np_6 < 5$, спојићемо пету и шесту класу:

X_k	1	2	3	4	$\geq$5
M_k	45	30	15	6	4
np_k	50	25	12.5	6.25	6.25

Сада можемо да пређемо на тестирање. Тест статистика $\chi_0^2 = \sum_{k=1}^{5} \frac{(M_k - np_k)^2}{np_k}$ при H_0 има χ^2_{5-0-1} расподелу, јер имамо пет класа и нисмо вршили оцењивање параметара. Критична област је облика:

$$W = \{\chi_0^2 \geq c\}.$$

Вредност константе c налазимо из задатог нивоа значајности и познате расподеле тест статистике при H_0, односно тако да важи

$$P_{H_0}\{\chi_0^2 \geq c\} = 0.05.$$

Дакле,

$$c = \chi^2_{4;0.05} = 9.488.$$

Критична област је облика

$$W = \{\chi_0^2 \geq 9.488\}.$$

Рачунамо вредност тест статистике на добијеном узорку:

$$\begin{aligned}\chi_0^2 &= \frac{(45-50)^2}{50} + \frac{(30-25)^2}{25} + \frac{(15-12.5)^2}{12.5} + \frac{(6-6.25)^2}{6.25} + \frac{(4-6.25)^2}{6.25} \\ &= 0.5 + 1 + 0.5 + 0.01 + 0.81 = 2.82.\end{aligned}$$

Проверавамо да ли узорак упада у критичну област:

$$2.82 < 9.488,$$

па закључујемо да узорак не упада у критичну област, те не одбацујемо H_0. ✓

Задатак 4**.** Из популације чије је обележје X извучен је узорак:

I_k	[0,1]	[1.5,2.5]	(2.5,3.5]	(3.5,5]
M_k	52	35	9	4

Са прагом значајности 0.02 тестирати хипотезу да обележје X има експоненцијалну расподелу.

Решење. Наведене класе не покривају читав узорачки простор, па морамо да направимо нову табелу у којој ћемо да покријемо све:

I_k	[0,1]	(1,1.5)	[1.5,2.5]	(2.5,3.5]	(3.5,5]	(5,∞)
M_k	52	0	35	9	4	0

Обим узорка је $n = 52 + 35 + 9 + 4 = 100$.

За почетак треба да израчунамо np_k, $k = 1, ..., 6$. Рачунамо p_k по формули

$$p_k = P_{H_0}\{X \in I_k\} = \int_{I_k} \lambda e^{-\lambda x} dx.$$

Параметар λ је непознат, па не можемо израчунати тачне вероватноће, али можемо оценити λ методом максималне веродостојности па добити процене вероватноћа. Оцењујемо λ методом максималне веродостојности:

$$\begin{aligned}f(x;\lambda) &= \lambda e^{-\lambda x}, x \geq 0 \\ L(\lambda) &= \prod_{k=1}^{n} f(x_k, \lambda) = \lambda^n e^{-\lambda \sum\limits_{k=1}^{n} x_k} \\ \ln L(\lambda) &= n \ln \lambda - \lambda \sum_{k=1}^{n} x_k \\ \frac{\partial \ln L(\lambda)}{\partial \lambda} &= \frac{n}{\lambda} - \sum_{k=1}^{n} x_k = 0 \\ \frac{n}{\lambda} &= \sum_{k=1}^{n} x_k \implies \lambda = \frac{n}{\sum\limits_{k=1}^{n} x_k} = \frac{1}{\overline{x}_n}\end{aligned}$$

Још треба да проверимо да ли је максимум. Пошто је $\frac{\partial \ln L(\lambda)}{\partial \lambda} > 0$, ако је $\lambda < \frac{1}{\overline{x}_n}$, а $\frac{\partial \ln L(\lambda)}{d\partial \lambda} < 0$, ако је $\lambda > \frac{1}{\overline{x}_n}$. Дакле, функција $\log L(\lambda)$ расте до $\frac{1}{\overline{x}_n}$, а после опада. Закључујемо да је оцена максималне веродостојности:

$$\widehat{\lambda} = \frac{1}{\overline{X}_n}.$$

Желимо да добијемо оцену λ на реализованом узорку. Узорачку средину можемо оценити са:

$$\overline{x}_n = \frac{0.5 \cdot 52 + 2 \cdot 35 + 3 \cdot 9 + 4.25 \cdot 4}{100} = 1.4,$$

па на основу овог узорка добијамо да је

$$\widehat{\lambda} = \frac{1}{1.4} = 0.71.$$

Сада можемо да израчунамо вероватноће p_k:

$$p_1 = P\{X \in I_1\} = \int_0^1 0.71e^{-0.71x}dx = 0.71\frac{e^{-0.71x}}{-0.71}\Big|_0^1 = 1 - e^{-0.71} = 0.508,$$
$$p_2 = P\{X \in I_2\} = \int_1^{1.5} 0.71e^{-0.71x}dx = e^{-0.71} - e^{-0.71\cdot 1.5} = 0.147,$$
$$p_3 = P\{X \in I_3\} = \int_{1.5}^{2.5} 0.71e^{-0.71x}dx = e^{-0.71\cdot 1.5} - e^{-0.71\cdot 2.5} = 0.175,$$
$$p_4 = P\{X \in I_4\} = \int_{2.5}^{3.5} 0.71e^{-0.71x}dx = e^{-0.71\cdot 2.5} - e^{-0.71\cdot 3.5} = 0.086,$$
$$p_5 = P\{X \in I_5\} = \int_{3.5}^{5} 0.71e^{-0.71x}dx = e^{-0.71\cdot 3.5} - e^{-0.71\cdot 5} = 0.055,$$
$$p_6 = 1 - p_1 - p_2 - p_3 - p_4 - p_5 = 0.029.$$

Последњу вероватноћу коју рачунамо би увек требало рачунати одузимањем претходно израчунатих вероватноћа од 1, због заокруживања, као што је и урађено.

Дакле,

I_k	[0,1]	(1,1.5)	[1.5,2.5]	(2.5,3.5]	(3.5,5]	(5,∞)
M_k	52	0	35	9	4	0
p_k	0.508	0.147	0.175	0.086	0.055	0.029
np_k	50.8	14.7	17.5	8.6	5.5	2.9

Видимо да је $np_6 < 5$, па ћемо шесту класу спојити са класом која има најмање np_k, односно петом класом. Добијамо нову табелу:

I_k	[0,1]	(1,1.5)	[1.5,2.5]	(2.5,3.5]	(3.5,∞)
M_k	52	0	35	9	4
np_k	50.8	14.7	17.5	8.6	8.4

Критична област је облика:

$$W = \{\chi_0^2 \geq c\},$$

где c рачунамо из једнакости

$$P_{H_0}\{\chi_0^2 \geq c\} = 0.02.$$

При H_0, пошто имамо пет класа после спајања и оценили смо један параметар методом максималне веродостојности, $\chi_0^2 \in \chi^2_{5-1-1}$. Према томе,

$$c = \chi^2_{3;0.02} = 9.837,$$

па је критична област

$$W = \{\chi_0^2 \geq 9.837\}.$$

Рачунамо вредност тест статистике на добијеном узорку:

$$\begin{aligned}\chi_0^2 &= \sum_{k=1}^{5} \frac{(M_k - np_k)^2}{np_k} \\ &= \frac{(52-50.8)^2}{50.8} + \frac{(0-14.7)^2}{14.7} + \frac{(35-17.5)^2}{17.5} + \frac{(9-8.6)^2}{8.6} + \frac{(4-8.4)^2}{8.4} \\ &= 0.028 + 32.2 + 17.5 + 0.019 + 2.305 = 34.552.\end{aligned}$$

Пошто је

$$34.552 > 9.837,$$

узорак упада у критичну област, па одбацујемо нулту хипотезу. ✓