ПРИРОДНО–МАТЕМАТИЧКИ ФАКУЛТЕТ
УНИВЕРЗИТЕТА У НИШУ

Проф. др БИЉАНА ПОПОВИЋ

МАТЕМАТИЧКА СТАТИСТИКА

Друго издање

Серија: уџбеници

НИШ, 2015

Аутор: Др Биљана Поповић, редовни професор
Природно–математичког факултета у Нишу
МАТЕМАТИЧКА СТАТИСТИКА
Рецензенти:
Др Загорка Лозанов–Црвенковић, редовни професор Природно–математичког факултета у Новом Саду
Др Мила Стојаковић, редовни професор Факултета техничких наука у Новом Саду
Техничко уређивање: Аутор

Одлуком Наставно–научног већа Природно–математичког факултета у Нишу, број 687/1–01 од 09.09.2009. године, рукопис је одобрен за штампу као универзитетски уџбеник.

Издавач: Природно–математички факултет, Ниш
Штампа: Атлантис, Ниш
Тираж: 100 примерака

CIP-Каталогизација у публикацији
Народна библиотека Србије, Београд

519.2(075.8)

ПОПОВИЋ, Биљана, 1954-
Математичка статистика/ Биљана Поповић. – 2. изд. – Ниш :
Природно–математички факултет Универзитета, 2015 (Ниш : Атлантис). – VII,
250 стр. : граф. прикази, табеле; 25 cm. – (Серија Уџбеници /
Природно–математички факултет, Ниш)

Тираж 100. – О аутору: стр. 245–246. – Библиографија: стр. 241–243. –
Регистар.

ISBN 978–86–6275–045–7

а) Математичка статистика
COBISS.SR–ID 217636364

Најдражима:

Милици, Предрагу и Миодрагу

Садржај

Предговор другом издању

Ауторка овом издању претенциозно даје епитет "исправљено" свесна, међутим, да нема објављене књиге у другом издању у којој се не може наћи понека грешка. Дакле, све уочене грешке у првом издању ове књиге су исправљене. Ауторка се искрено захваљује свима који су јој на грешке у првом издању указали.

Осим овога, у овом издању, у односу на претходно, је промењен редослед неких тема у првој глави, али не и садржај.

Ниш, 2015. године Аутор

Предговор првом издању

Уџбеник **Математичка статистика** је, пре свега, намењен студентима Основних академских студија Одсека за математику и информатику Природно–математичког факултета у Нишу. У том погледу, ова књига у потпуности покрива теоријски садржај истоименог предмета предвиђен програмом.

За коришћење уџбеника у целости неопходно је опште математичко знање до нивоа теорије мера и интеграла, међутим, чињеница је да се његов већи део може читати и користити већ са стеченим знањем класичног курса Математике који обухвата вишеструке интеграле. Штавише, за добар део читања и коришћења ове књиге неопходно је само знање из математике које укључује познавање једноструких интеграла.

Посебно математичко знање којим се подразумева да располажу студенти, корисници уџбеника у целини, је стандардни курс савремене Теорије вероватноћа.

Књига је итекако погодна за оне који само примењују статистику у другим научним дисциплинама због великог броја урађених примера који се односе на најразличитија подручја науке, почев од медицине, па преко психологије до технике.

Аутор се захваљује рецензентима на корисним саветима.

Ниш, јуна 2009. године Аутор

Глава 1

Увод

Математичка статистика је примењена математичка дисциплина сродна теорији вероватноће. Базира се на питањима и методама теорије вероватноће, али решава своје специфичне (проблеме) задатке својим методама. (Свака математичка теорија се развија у оквиру неког модела који описује одређени круг реалних појава чијим се проучавањем и бави дата теорија.)

У теорији вероватноће се полази од претпоставке да је познат простор вероватноћа $(\Omega, \mathcal{F}, P)$, где је Ω скуп свих елементарних исхода, $\mathcal{F}$ је σ-алгебра на скупу Ω а P је вероватноћа.

Вероватноћа P, у практичним проблемима које треба решавати, није у потпуности позната. У већини случајева се претпоставља да $P \in \mathcal{P}$, где је $\mathcal{P} = \{P\}$ фамилија вероватноћа. Такви практични проблеми називају се статистичким моделима.

Дакле, за разлику од модела теорије вероватноћа, статистички модел је $(\Omega, \mathcal{F}, \mathcal{P})$.

Пример 1. (**Шема Бернулија**) Обавља се n независних опита у којима се реализује 0 или 1 са вероватноћама редом $1 - p = q$ и p, $0 \leq p \leq 1$. Исход овог експеримента је

$$\Omega = \{\omega : \omega = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n), \varepsilon_i = 0, 1\}.$$

При томе је вероватноћа појединог елементарног исхода

$$P(\omega) = p^{\sum \varepsilon_i} q^{n - \sum \varepsilon_i}.$$

Ако вероватноћа p није претходно позната, означићемо је са θ и ту ознаку ћемо надаље користити за сваки непознати параметар.

У том случају једина информација коју имамо о параметру овог примера је да је $\theta \in \Theta = [0,1]$. Тачније, имамо једино информацију да расподела вероватноћа којом овај експеримент описујемо припада фамилији $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$, где је $P_\theta = \theta^{\sum \varepsilon_i}(1-\theta)^{n-\sum \varepsilon_i}$. $\triangle$

У претходном примеру је дефинисан један статистички модел, дакле модел који у себи садржи неку врсту неодређености. Задатак математичке статистике је да се коришћењем информације добијене посматрањем исхода експеримента, дакле статистичких података, смањи та неодређеност, односно да се, што је могуће тачније, изврши избор $P \in \mathcal{P}$.

Математичка статистика је наука о статистичком закључивању. Статистичко закључивање подразумева решавање задатака обрнутих од оних које решава теорија вероватноће: она утврђује структуру статистичких модела према резултатима спроведених посматрања, дакле, одређује простор вероватноћа на основу експеримента. При томе посматрања не могу бити произвољна. Наиме, она морају бити еквивалентна статистичком експерименту:

– може се понављати произвољан број пута под истим условима,

– унапред је дефинисано шта се региструје у експерименту при чему су познати сви исходи и

– исход појединачног експеримента није унапред познат.

За прве свесне покушаје дефинисања и примене статистичког закључивања узимају се пописи становништва које су спроводили владари још неколико векова пре наше ере ради утврђивања броја војних поданика или пореских обвезника. Заснивање статистике као науке везује се за појаву школе "политичких аритметичара" у Енглеској у $XVII$ веку. По некима, дело "Natural and Political Observations upon the Bills of Mortality", које је написао Дж. Грант (J. Graunt) и објавио 1622. године, означава почетак статистике као науке. Дуго времена је статистика сматрана научним методом за проучавање друштвених наука. Међутим, математичари који су неминовно били укључени у конституисање, формално дефинисање, и постали одговорни за развој статистичког метода закључивања, одговорни су и за почетак примене статистике у природним наукама. Ту идеју међу првима је прихватио енглески биолог Галтон (Sir

Francis Galton, 1822-1911), који је применио статистички метод у истраживањима у биологији. Теоријски допринос развоју математичке статистике дао је међу првима швајцарски математичар Јакоб Бернули (Jacob Bernoulli, 1654-1705) дефинишући и образлажући закон великих бројева у свом делу "Ars conjectandi". Крупан корак у том правцу дао је и француски астроном и математичар Лаплас (Pierre Simon, Marquis de Laplace, 1749-1827). Познато је његово дело "Théorie analytique de probabilités". Буран развој математичке статистике као теоријске дисциплине у XX веку омогућен је, пре свега, развојем теорије вероватноћа у овом периоду.

Глава 2

Теорија узорака

2.1 Основни појмови

Статистички експеримент се изводи над елементима неког скупа на којима се посматра једно или више заједничких својстава.

Дефиниција 1. *Популација* или *генерални скуп* или *основни скуп* је скуп елемената чија се заједничка својства изучавају статистичким методима. Популација се симболички бележи са Ω, а њен елемент са ω. $\diamondsuit$

Дефиниција 2. *Обележје* је заједничко својство елемената једне популације (које се испитује). Обележје може бити квантитативно (нумеричко) или квалитативно (атрибутивно). $\diamondsuit$

При извођењу статистичког експеримента полази се од претпоставке да се том приликом реализују неки случајни догађаји. Дакле, претпоставља се да се исход експеримента може приказати случајном величином X. Уколико је експеримент понављан n пута, исход се представља случајним вектором $\mathbf{X} = (X_1, X_2, \ldots, X_n)$. При проучавању овог случајног вектора пожељно је познавати његову расподелу. С тим у вези рећи ћемо да треба одредити густину расподеле обележја, а надаље ћемо то појаснити. Овде ће се користити термин густина расподеле у уопштеном значењу, тј. везиваће се и за случајне променљиве дискретног типа.

Пример 2. За случајну променљиву са биномном расподелом

$\mathcal{B}(1,p)$, казаћемо да има густину расподеле

$$f(x) = \begin{cases} p^x(1-p)^{1-x}, & x = 0,1 \\ 0, & x \neq 0,1 \end{cases} . \quad \triangle$$

У савременој литератури се све чешће користи термин *функција масе* за ову функцију када је у питању случајна променљива дискретног типа.

Нека је Y случајна променљива дефинисана као функција случајних променљивих

$$X_1, X_2, \ldots, X_n,$$

тј. нека је $Y = u(X_1, X_2, \ldots, X_n)$. Одређивање густине расподеле ове случајне променљиве на основу познавања заједничке густине расподеле вектора случајних променљивих $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, у ознаци $f(x_1, x_2, \ldots, x_n)$, $(x_1, x_2, \ldots, x_n) \in R^n$, је један од задатака математичке статистике. Сам случајни вектор $\mathbf{X}$ и функције од његових компонената су окосница математичке статистике.

Дефиниција 3. *Узорак* је део популације на коме се испитује посматрано обележје. Број елемената у узорку се назива *обим* или *величина узорка.* $\diamondsuit$

Дакле, на узорку обима n се спроводи статистички експеримент. Исход тог експеримента ће бити вектор $\mathbf{X} = (X_1, \ldots, X_n)$, који је по својим карактеристикама случајна променљива. Вектор $\mathbf{X}$ још зовемо *случајним узорком* за разлику од његове *реализоване вредности* по обављеном експерименту.

Дефиниција 4. Вектор $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ који представља реализацију вектора $\mathbf{X}$ по обављеном експерименту зовемо *реализовани узорак.* $\diamondsuit$

У даљем тексту ће се под узорком подразумевати случајни узорак.

Детаљније о узорку и начинима за избор узорака биће речи надаље.

Популација има нешто шири смисао од извесног догађаја у теорији вероватноће, док је обележје нешто шири појам од појма случајне променљиве. Наиме, извесан догађај је скуп свих могућих елементарних исхода једног експеримента, при чему се подразумевају различити исходи. Популација је, међутим, скуп свих елемената на којима се посматра неко својство (скуп људи, скуп сијалица, део тла, итд.). Дефинишимо функцију из скупа Ω, популације, у скуп који чине категорије једног својства. Прецизније, на скупу Ω се дефинише релација еквиваленције: "два елемента популације су у релацији ако су им једнаке вредности обележја које се на елементима популације посматра". Том релацијом се врши разбијање скупа Ω на класе еквиваленције, односно, дефинише се фактор скуп. Релацију еквиваленције је могуће увести у популацију само до на скуп мере нула. Класе еквиваленције су категорије, те се најпре дефинише пресликавање популације на фактор скуп једном функцијом тако што се сваком елементу популације придружује његова класа еквиваленције. Из последњег скупа је могуће дефинисати нову функцију са вредностима у скупу реалних бројева, R, која је, заправо, случајна променљива. Композиција ових функција је *обележје*. У том смислу се може говорити о расподели обележја посредством расподеле овако дефинисане случајне променљиве, те ће се и обележје, као и случајна променљива, означавати великим словом латинице са краја абецеде, X,Y,Z,. . . .

У вези са уопштењем појма густине расподеле сматраће се да свако обележје има своју густину расподеле.

Пример 3. За популацију ћемо узети студенте Природно-математичког факултета у Нишу. Нека је обележје које посматрамо на тој популацији "образовни профил". У овом моменту ћемо посматрати само основни профил, тј. математика, информатика, физика, хемија, биологија, географија. Ових 6 категорија би чиниле разбијање скупа Ω. Дакле, студенти истог образовног профила би чинили једну класу еквиваленције. Надаље бисмо сваком образовном профилу придружили број (код), рецимо нека су то природни бројеви од 1 до 6. Тиме би била дефинисана случајна променљива. (Вероватноћа да поједини студент истовремено студира два образовна профила је мала, те се може сматрати да је такав догађај вероватноће нула и такви се студенти могу изоставити из даљег посматрања.) $\triangle$

Са гледишта математичке статистике дато обележје X је потпуно одређено ако је одређена његова расподела, $P\{X \in S\}$, где је $S \in \mathcal{B}_1$, а $(R,\mathcal{B}_1,P)$ фазни простор. То је истовремено и један од главних проблема којима се бави математичка статистика: одређивање расподеле обележја. При томе је могуће да унапред није позната фамилија допустивих расподела или да је она позната, а да из ње треба направити прави избор оценом вредности непознатих параметара који у расподели фигуришу. Дакле, основни проблем статистичког закључивања је да на основу статистичког експеримента нешто закључи о расподели обележја.

Закључивање о расподели обележја врши се на основу изабраног узорка. Отуда је важно да изабрани узорак буде репрезентативан, тј. да буде такав да се са довољном тачношћу закључак о расподели посматраног обележја добијеној на узорку може да екстраполује на читаву популацију.

2.2 Сређивање и приказивање реализованих узорака

Експериментални подаци се ради статистичке обраде представљају на два основна начина: *таблично* и *графички*. Таблични метод даје податке сређене у облику табеле. Графички метод те табеле илуструје пригодним скицама, картама, графиконима...

2.2.1 Таблични метод приказа података – организовање база података

Таблице квантитативних обележја

Низ добијених података поређаних у растућем поретку (рангираних) даје варијациони низ обележја. Он пружа полазну основу за даља разматрања у вези са расподелом.

Пример 4. У 20 одељења нижих разреда основне школе регистрован је број ученика са натпросечним способностима: 5, 6, 8, 10, 9, 8, 4, 7, 7, 3, 6, 4, 8, 7, 6, 6, 5, 3, 6, 6. Варијациони низ узорка је: 3, 3, 4, 4, 5, 5, 6, 6, 6, 6, 6, 6, 7, 7, 7, 8, 8, 8, 9, 10. За одређивање расподеле обележја користи се следећа табела:

Број ученика са натпросечним способностима	3	4	5	6	7	8	9	10
$f = k$	2	2	2	6	3	3	1	1
$f^* = k/n$	0,1	0,1	0,1	0,3	0,15	0,15	0,05	0,05
$f^*_{\%} = k/n$ [%]	10	10	10	30	15	15	5	5
$\Sigma f \equiv n_x$	2	4	6	12	15	18	19	20
$\Sigma f^* = n_x/n$	0,1	0,2	0,3	0,6	0,75	0,9	0,95	1
$\Sigma f^*_{\%} = n_x/n$ [%]	10	20	30	60	75	90	95	100

У табели су коришћене ознаке: k–број одељења са посматраним бројем натпросечних ученика, f–апсолутна учестаност, f^*–релативна учестаност, $f^*_{\%}$–процентуална учестаност, n_x–број одељења са не више од x натпросечних ученика, Σf–збирна (кумулативна), учестаност Σf^*–збирна релативна учестаност и $\Sigma f^*_{\%}$–збирна процентуална учестаност. $\triangle$

Приметимо да је број $\frac{n_x}{n}$, заправо, реализована вредност емпиријске функције расподеле за задато $x \in R$.

Код обележја апсолутно непрекидног типа или дискретних обележја са великим бројем различитих вредности, подаци се сређују по унапред одабраним интервалима. Број и распоред тих интервала зависи од броја података и самог обележја.

Пример 5. Бележене су минималне јачине струје које представљају праг осетљивости једног мишића 60 посматраних пацијената и добијени су следећи резултати:

Ред.бр.	Јачина (mA)	Ред.бр.	Јачина (mA)	Ред.бр.	Јачина (mA)
1	7,80	21	6,23	41	6,36
2	9,28	22	7,27	42	5,98
3	8,70	23	6,98	43	5,16
4	5,30	24	4,84	44	11,40
5	5,63	25	10,53	45	8,59
6	6,54	26	8,00	46	8,12
7	7,80	27	7,28	47	10,30
8	7,73	28	9,16	48	10,80
9	6,76	29	5,02	49	11,87
10	12,06	30	8,08	50	11,62
11	16,44	31	3,95	51	11,34
12	9,55	32	6,77	52	9,50
13	3,71	33	5,24	53	6,43
14	8,97	34	6,32	54	6,21
15	7,38	35	9,64	55	10,42
16	5,02	36	10,97	56	7,71
17	5,18	37	8,79	57	6,55
18	7,51	38	7,93	58	7,33
19	4,92	39	7,91	59	7,25
20	4,82	40	14,52	60	4,92

Строгог правила за избор броја и дужине интервала нема, али је могуће управљање по формули која препоручује k интервала, где је

$$k \geq 1 + 3,322 \log_{10} n = 1 + \log_2 n = 1 + \frac{\ln n}{\ln 2}$$

за обим узорка n. Међутим, не препоручује се више од $5 \cdot \log_{10} n$ интервала, тј.

$$k \leq 5 \cdot \log_{10} n = 5 \cdot \frac{\ln n}{\ln 10}\,.$$

За разматрани пример, обим узорка је $n = 60$, па је доња граница броја интервала једнака $k = 1 + \log_2 60 \approx 7$, а горња граница броја интервала је $k \leq 5 \cdot \log_{10} 60 \approx 9$. Значи, може се узети 7, 8 или 9 интервала.

Број интервала k се може одредити и на један од следећих начина: $k \approx \sqrt{n}$, $k \approx 2\sqrt[3]{n}$ или $k \approx 5 \log_{10} n$.

Без обзира на начин одређивања броја интервала, дужине интервала се одређују на следећи начин. Одређују се најмања x_{min} и највећа x_{max} вредност у реализованом узорку, а затим се дужина интервала рачуна по формули:

$$h = \frac{x_{max} - x_{min}}{k}\,,$$

при чему се води рачуна да су границе интервала једноставне за рад (цели бројеви, бројеви дељиви са 5 и сл.).

Узмимо да је број интервала 8. Затим ћемо одредити најмању и највећу вредност реализованог узорка. Оне су редом 3,71 и 16,44. Тада је дужина интервала једнака

$$h = \frac{16,44 - 3,71}{8} = 1,591\,.$$

Како добијени број није погодан за рад, то се може узети други погоднији број, рецимо 2. У доњој табели одређени су интервали, апсолутне, релативне, збирне и збирне релативне учестаности минималних струја разматраног низа:

интервал	средина интервала	f	$f^* = \frac{f}{n}$	$f^*_{\%}$	Σf	Σf^*	$\Sigma f^*_{\%}$
[2,4)	3	2	0,03	3	2	0,03	3
[4,6)	5	12	0,20	20	14	0,23	23
[6,8)	7	22	0,36	36	36	0,59	59
[8,10)	9	12	0,20	12	48	0,79	79
[10,12)	11	9	0,15	15	57	0,94	94
[12,14)	13	1	0,02	2	58	0,96	96
[14,16)	15	1	0,02	2	59	0,98	98
[16,18]	17	1	0,02	2	60	1,00	100

$\triangle$

Пример 6. 50 студената је полагало испит из статистике и добијени су следећи резултати по броју освојених поена од могућих 100: 17, 73, 85, 43, 36, 21, 0, 35, 50, 32, 75, 21, 78, 41, 92, 70, 80, 84, 55, 42, 79, 45, 98, 62, 46, 45, 79, 2, 17, 19, 49, 42, 32, 6, 8, 39, 4, 28, 48, 86, 26, 60, 92, 15, 85, 26, 14, 69, 55, 94. Подаци се могу средити на следећи начин.

Укупно има $n = 50$ података. Најмањи дозвољени број интервала је $k = 1 + 3.322\log_{10} 50 = 6.64 \approx 7$, када се број заокружи. Највећи дозвољени број интервала је $5 \cdot \log_{10} 50 = 8.49 \approx 8$. Тако се може радити са 7 или 8 интервала.

Нека је број интервала $k = 7$. Најмања и највећа вредност реализованог узорка су $x_{min} = 0$ и $x_{max} = 98$, тако да је дужина интервала

$$h = \frac{98 - 0}{7} = 14.$$

Сада се подаци групишу по интервалима: $[0,14)$, $[14,28)$, $[28,42)$, $[42,56)$, $[56,70)$, $[70,84)$ и $[84,98]$, и добија се следећа табела:

Број бодова	$[0,14)$	$[14,28)$	$[28,42)$	$[42,56)$	$[56,70)$	$[70,84)$	$[84,98]$
Број студената (k)	5	9	7	11	3	7	8
$f^* = k/n$	0,1	0,18	0,14	0,22	0,06	0,14	0,16
$f^*_{\%} = k/n$ [%]	10	18	14	22	6	14	16
$\Sigma f \equiv n_x$	5	14	21	32	35	42	50
$\Sigma f^* = n_x/n$	0,1	0,28	0,42	0,64	0,7	0,84	1
$\Sigma f^*_{\%} = n_x/n$ [%]	10	28	42	64	70	84	100

△

Интервали **не морају** бити једнаких дужина, што препоручује са̂мо конкретно обележје.

Најчешће, средине интервала репрезентују реализоване вредности обележја код израчунавања реализованих вредности статистика (о чему ће још бити речи). Као средина интервала $[a,b)$, али такође и интервала $[a,b]$ користи се број $\frac{b+a}{2}$. Ово отуда што је код обележја апсолутно непрекидног типа вероватноћа реализације поједине тачке са реалне праве једнака 0.

Таблице квалитативних обележја

У случају квалитативног обележја може се такође сачинити табела.

Пример 7. Тестом за проверу моторних способности је мерен ниво способности ученика једног одељења и добијени резултати су сврстани у три категорије: низак (н), средњи (с) и висок (в) ниво способности. У одељењу је регистрован следећи низ података: н, н, с, в, с, с, с, н, в, в, с, с, с, н, в, в, в, с, в, н, н, с, в, с. На основу низа реализација добијена је табела

	ниво моторних способности	н	с	в	Σ
f	број ученика	6	10	8	24
f^*	релативна учестаност	0,25	0,42	0,33	1

△

Таблице за дводимензионално обележје

Уколико се посматрају два обележја X и Y истовремено (која су могуће зависна, тј. дводимензионално обележје) табела је облика:

$X \backslash Y$			

Добијена табела се назива и *табела контингенције*. Сам поступак формирања табеле је једноставан. Уколико другачије није наглашено, поступак је следећи: одређују се интервали за свако обележје посебно а затим се реализовани узорак групише по добијеним интервалима.

Наравно, уколико се ради о дискретном обележју као некој од компонената или обема компонентама посматраног дводимензионалног обележја, утврђују се апсолутне (релативне, процентуалне) учестаности одговарајућих парова у реализованом узорку и уносе у табелу.

Пример 8. 38 особа конкурише за једну врсту посла. Послодавца занима њихова стручна и интелектуална способност. Због тога ове особе раде тестове стручности (ТС) и интелигенције (ТИ). Добијени су следећи резултати:

Редни број кандидата	ТС	ТИ	Редни број кандидата	ТС	ТИ
1	70	112	20	55	120
2	75	121	21	60	100
3	80	100	22	58	102
4	85	102	23	60	104
5	75	120	24	74	97
6	48	98	25	48	94
7	52	111	26	53	89
8	50	120	27	58	129
9	51	105	28	79	116
10	55	110	29	82	145
11	46	134	30	84	130
12	87	100	31	81	115
13	72	99	32	52	120
14	70	91	33	55	109
15	63	101	34	68	110
16	56	104	35	73	112
17	60	115	36	46	121
18	72	116	37	52	130
19	78	119	38	80	90

Посматрају се два обележја: стручност и интелигенција, посредством тестова као мерних инструмената за посматрана обележја. Број интервала за оба обележја може бити 6 или 7. Нека се за свако од обележја подаци групишу у по 6 интервала. Сада се одређују дужине интервала за свако обележје посебно. За прво обележје (резултати теста стручности) дужине интервала су $h_1 =$

$(87-46)/6 = 6,83 \approx 7$, тако да се добијају интервали: $[46,53)$, $[53,60)$, $[60,67)$, $[67,74)$, $[74,81)$ и $[81,87]$. За друго обележје (резултати теста интелигенције) дужине интервала су $h_2 = (145-89)/6 = 9,33$ те нека је $h_2 = 10$ ради лакшег рачунања. За други тест се добијају следећи интервали: $[89,99)$, $[99,109)$, $[109,119)$, $[119,129)$, $[129,139)$ и $[139,145]$.

Сада се врши пребројавање података по интервалима из следеће табеле:

ТС\ТИ	$[89,99)$	$[99,109)$	$[109,119)$	$[119,129)$	$[129,139)$	$[139,145]$
$[46,53)$						
$[53,60)$						
$[60,67)$						
$[67,74)$						
$[74,81)$						
$[81,87]$						

Одговарајуће апсолутне учестаности су:

ТС\ТИ	$[89,99)$	$[99,109)$	...	$[129,139)$	$[139,145]$
$[46,53)$	2	1	...	2	0
$[53,60)$	1	2	...	1	0
$[60,67)$	0	3	...	0	0
$[67,74)$	1	1	...	0	0
$[74,81)$	2	1	...	0	0
$[81,87]$	0	2	...	1	1

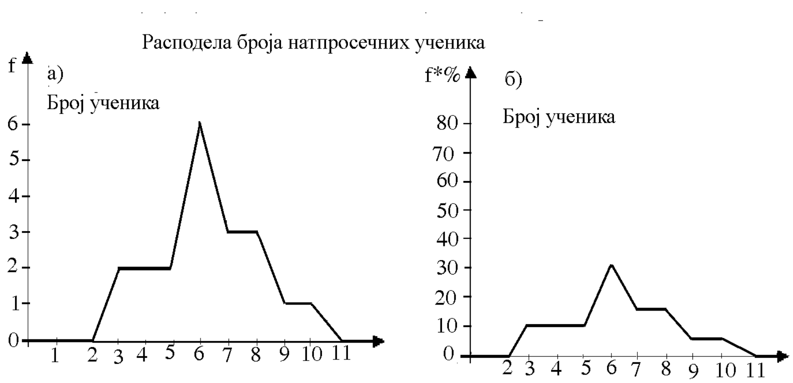
$\triangle$

2.2.2 Графички методи приказа података

Расподела обележја графички се приказује преко (обичних) учестаности или преко збирних учестаности (нарочито збирних релативних учестаности, тј. емпиријске функције расподеле).

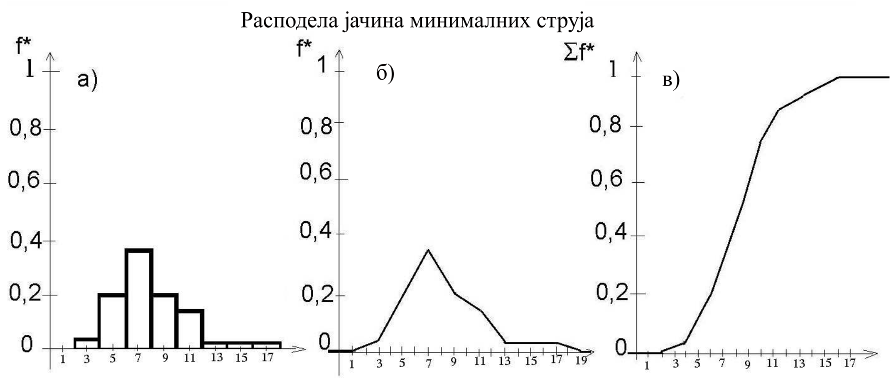
Графички методи приказа података су најчешће: полигон, хистограм, кумулативна крива, разни дијаграми и слично.

На сликама од 2.1 до 2.4 приказани су подаци који се односе на Пример 4. Фигуре на сликама 2.1 а), б) и 2.3 б) су полигони, на сликама 2.2, 2.4 б) су тракасти дијаграми, на слици 2.3 а) је хистограм, а на слици 2.8 је кутија дијаграм. На слици 2.9 је тзв. стереограм који служи за графичко представљање дводимензионог обележја.



Слика 2.1: Полигони: а) апсолутних учестаности; б) релативних процентуалних учестаности из Примера 4.

Слика 2.2: Тракасти дијаграм апсолутних учестаности из Примера 4.



Слика 2.3: а) Хистограм релативних учестаности; б) полигон релативних учестаности; в) кумулативна крива из Примера 5.

1. **Хистограм**

Хистограм се може примењивати само за графичко приказивање реализованог узорка из популације са обележјем X апсолутно непрекидног типа. За узорак који је у том случају интервално сређен, подаци се приказују на следећи начин. Област вредности посматраног обележја је разбијена на интервале дужине h. Ови интервали се приказују на апсцисној оси координатног система припремљеног за графичко представљање реализованог узорка $\mathbf{x}$ обима n. Над сваким од интервала се конструише правоугаоник чија је висина $\nu/(nh)$, односно површина ν/n, где је ν број елемената реализованог узорка који припадају уоченом интервалу. Фигура која представља унију управо конструисаних правоугаоника зове се хистограм релативних учестаности.

За случајни узорак $\mathbf{X}$ обима n, количник ν/n је случајна промењива. Ако се има у виду да је n произвољан природан број, може се говорити о низу случајних промењивих за који важи Бернулијев закон великих бројева,

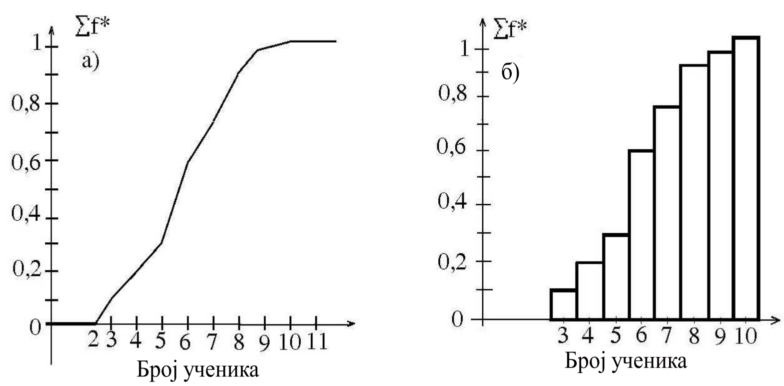
$$P\left\{\left|\frac{\nu}{n}-p\right|>\varepsilon\right\}\longrightarrow 0\,, \quad n\to\infty\,,$$

где је p вероватноћа да обележје X има вредност у одговарајућем интервалу. Ако је дужина интервала h довољно мала, а густина f обележја непрекидна, тада је та вероватноћа приближно једнака $f(z)h$, где је z средина одговарајућег интервала. То значи да је при великом обиму узорка и малој дужини интервала, висине конструисаних правоугаоника могуће посматрати као приближне вредности густине расподеле које одговарају срединама интервала, односно, горња граница овако дефинисаног хистограма релативних учестаности се може посматрати као статистички аналогон густине расподеле посматраног обележја.

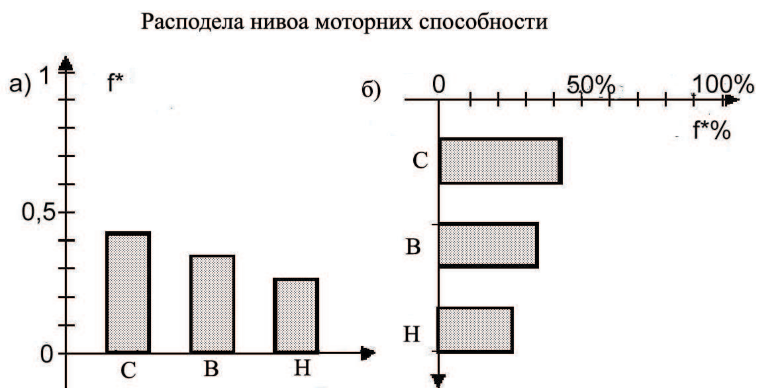
Аналогно описаном поступку може се конструисати и хистограм збирних (кумулативних) релативних учестаности, чија би се горња граница могла посматрати као статистички аналогон функције расподеле посматраног обележја.

Веома често примењује се и хистограм одговарајућих процентуалних учестаности.

Важно је, међутим, нагласити да је хистограм примењив само у почетној фази истраживања. Ово отуда што се не смеју занемарити његови недостаци, а то су неодређеност у начину формирања интервала и губитак информација при груписању података, јер се користи само број који показује колико је елемената реализованог узорка припало одређеном интервалу, а не и сами елементи узорка.



Слика 2.4: а) Кумулативна крива релативних учестаности (огива релативних учестаности); б) тракасти дијаграм збирних релативних учестаности из Примера 4.



Слика 2.5: а) Вертикални и б) хоризонтални тракасти дијаграм за Пример7.

2. **Дијаграм**

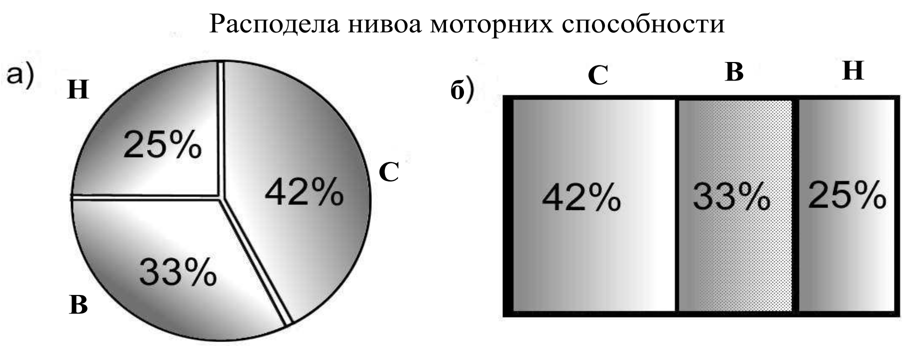
   Дијаграми се користе за графичко представљање реализованих вредности обележја дискретног типа квантитативних или квалитативних. Могу бити линијски и површински. У линијске дијаграме спадају и сви полигони учестаности, али и више од тога. На пр. звездасти дијаграм (видети слику 2.7 б)). Површински дијаграм може бити тракасти (слике 2.2, 2.4 и 2.5) или кружни (слика 2.6 а)) или правоугаони (слика 2.6 б)) и слично. Површински дијаграми се раде по принципу делова површи сразмерних одговарајућим учестаностима.

3. **Кутија дијаграм**

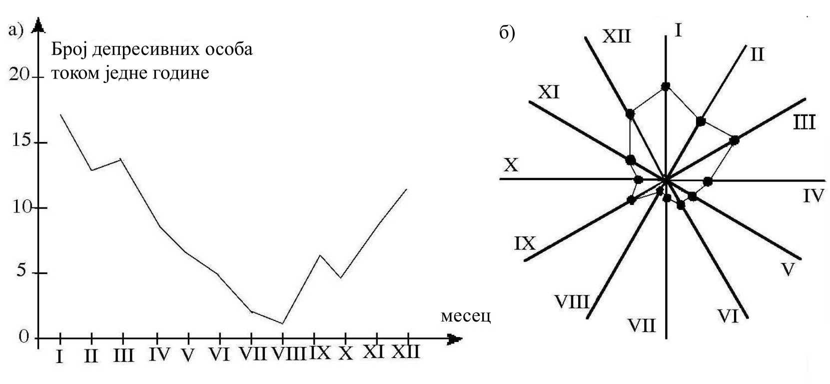
   Кутија дијаграм[1], је брз начин да се испита неки скуп података графички. На овај дијаграм се може да гледа као на примитивнији у односу, на пример, на хистограм. Међутим, он ипак има неке предности. Једна од њих је да је потребно много мање простора да се упореде расподеле више обележја на основу учињених посматрања и која су, у том случају, представљена у облику неколико група података.

   Кутија дијаграм се може да црта вертикално или хоризонтално. Објаснићемо како се конструише хоризонтални кутија дијаграм, док ће се вертикални конструисати на исти начин уз

[1]На енглеском познат као *boxplot*.



Слика 2.6: Подела а) круга ("пита") и б) правоугаоника за приказивање учестаности из Примера 7.



Слика 2.7: Илустрација појаве која има циклични карактер: а) линијски дијаграм; б) звездасти дијаграм.

замену ”лево” са ”доле” и ”десно” са ”горе”, а хоризонтална бројна оса се замењује вертикалном.

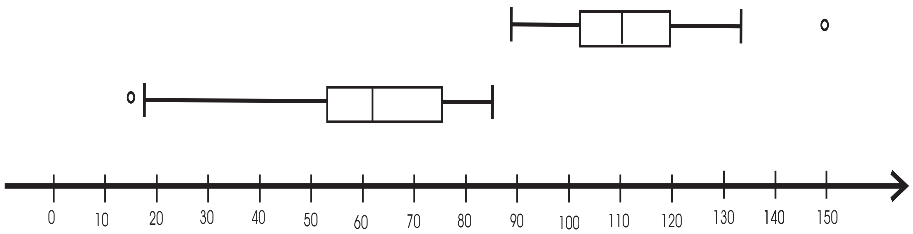
За дати скуп података израчунамо сва три узорачка квартила, означимо их са Q_1, Q_2, Q_3. Уочимо да је Q_2 заправо медијана узорка. Узорачки квартили су статистике које се дефинишу као тачкасте оцене квартила (види Главу 3). Иза тога израчунамо узорачки међуквартилни (интерквартилни) распон који чини разлика између трећег и првог узорачког квартила, $IQR = Q_3 - Q_1$. Конструишемо правоугаоник тако да су му странице паралелне са хоризонталном бројном осом дужине једнаке IQR, док су му друге две странице конструисане тако да је лева на ординати која одговара првом узорачком квартилу, а десна на ординати која одговара трећем узорачком квартилу. У добијеном правоугаонику конструишемо и део ординате која одговара узорачкој медијани. Средина узорка такође може бити означена на дијаграму.

Сваки податак, односно, свака вредност у реализованом узорку која је мања од $1,5 \cdot IQR$ од првог квартила и аналогно већа за $1,5 \cdot IQR$ од трећег квартила ће овде бити сматрана аутлејером[2], односно податком који ”штрчи”, нетипичним податком. (У статистици нема прецизне дефиниције када се неки податак сматра аутлејером, односно, податком који штрчи, међутим, познато је више начина да се при оцењивању непознатих параметара расподеле или при некој другој примени статистичких метода, овакви подаци посебно третирају.) Индикација да подаци нису аутлејери се приказује помоћу хоризонталне дужи која спаја ординату најмањег такавог податка са конструисаним правоугаоником, као и највећег таквог податка са правоугаоником. Опционо се поменуте ординате могу да истакну. Обично се екстремни подаци, најчешће они кији су мањи за $3 \cdot IQR$ од првог, односно за толико већи од трећег узорачког квартила, такође истичу и приказују тачкама на графикону тако да им се графички приказ разликује од претходно поменутих тачака.

Пример 9. Упоредити расподеле два обележја применом кутија дијаграма. Варијациони низ реализованог узорка првог обележја је 15, 17, 48, 48, 50, 51, 52, 52, 52, 53, 55, 55, 55,

[2]На енглеском *outlier.*

56, 58, 58, 60, 60, 60, 63, 68, 70, 70, 72, 72, 73, 74, 75, 75, 78, 79, 80, 80, 81, 82, 84, 85, 87, а варијациони низ реализованог узорка другог обележја је 89, 90, 91, 94, 97, 98, 99, 100, 100, 100, 101, 102, 102, 104, 104, 105, 109, 110, 110, 111, 112, 112, 115, 115, 116, 116, 119, 120, 120, 120, 120, 121, 121, 129, 130, 130, 134, 150. △



Слика 2.8: Кутија дијаграм за Пример 9.

4. **Полигон**

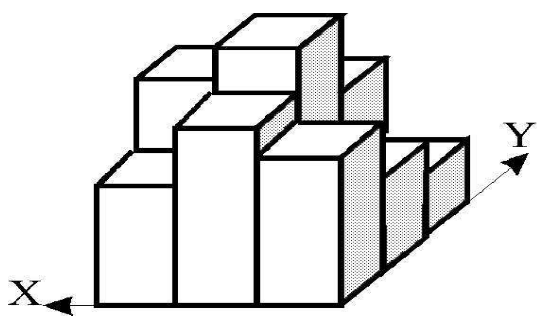
Из хистограма релативних и хистограма збирних релативних учестаности и одговарајућих тракастих дијаграма (код обележја дискретног типа) добијају се полигон релативних и полигон збирних релативних учестаности. Оба се конструишу тако што се средине горњих страница суседних правоугаоника одговарајућег хистограма или тракастог дијаграма споје дужима, чиме уочене средине постају темена полигоналних линија. Полигон релативних учестаности има прво теме на апсцисној оси у средини интервала (кога такође конструишемо) који непосредно претходи првом интервалу у коме је учестаност различита од нуле, а последње теме му је такође на истој оси у средини непосредно суседног интервала последњем интервалу у коме је релативна учестаност различита од нуле. Полигон збирних релативних учестаности има прво теме где и претходни, а последње му је теме над интервалом у коме је збирна учестаност једнака јединици. Оба су полигона отворена и логички се настављају хоризонталним полуправим на оба краја.

Полигон збирних релативних учестаности се зове и **кумулативна крива** или **огива**. Може се вршити изглађивање кумулативне криве при чему се уместо полигоналне линије добија глатка крива која пролази кроз темена ове полигоналне линије.

Полигон се конструише и на основу тракастог дијаграма квантитативног обележја на исти начин као у наведеном случају када "настаје" од хистограма.

5. **Стереограм**

   Постоји могућност графичког приказивања реализованог узорка неког дводимензионог обелжја. Овакав реализовани узорак се приказује **стереограмом**. Тако, ако је у питању дводимензионо обележје (X, Y), стереограм би био на пр.



Слика 2.9: Стереограм

2.3 Појам случајног броја

За избор репрезентативног узорка препоручује се случајни избор, тј. избор елемената популације у узорак на случајан начин. Да би се реализовао случајни избор често се користи таблица случајних бројева.

Размотримо декадни бројни систем. За записивање неког реалног броја у декадном бројном систему користи се дест цифара: 0,1, 2, 3, 4, 5, 6, 7, 8, 9. Ако претпоставимо да вршимо експеримент у коме је једнако вероватан избор било које од наведених десет цифара, свака цифра ће бити изабрана са вероватноћом 0,1. Случајна променљива којом се описује овај експеримент има дискретну униформну расподелу. Понављањем експеримента произвољан број пута (под истим условима, при чему су познати сви могући исходи експеримента, али ни у једном појединoм експерименту није унапред познат исход – статистички експеримент) добио би се низ случајних бројева или, прецизније, случајних цифара. Потреба за оваквим низом и формирањем читаве таблице случајних бројева (Таблица 6) биће јаснија у наредним поглављима. Таблица 6 је само део таблице од 1 000 000 случајних цифара сачињене 1955. године у САД од стране корпорације под називом "Rand Corporation". Техника којом је добијена таблица користи идеју рулета. Наиме, поменута таблица добијена је помоћу рулета са десет поља од којих је свако поље одговарало по једној декадној цифри (при чему је електроника и механика система морала да задовољи посебно високе захтеве). Отуда се статистичка техника која користи случајне бројеве зове метод Монте Карло, према граду познатом по коцкарницама.

Према таблици случајних бројева декадног бројног система могуће је направити и таблице случајних бројева других бројних система, на пр. бинарног бројног система идентификујући, рецимо, све парне цифре са 0, а непарне са 1.

Постоје и неки други физички системи који су се користили као генератори случајних бројева. Један пример је емисија честица радиоактивног извора зрачења, при чему се бележи број честица исте врсте у јединици времена регистрованих на баријери.

Треба нагласити да су неки ирационални бројеви, односно њихове значајне цифре, као што су број π и $\sqrt{7}$, изванредни природни генератори низа случајних цифара. До овог сазнања се дошло тек са применом моћних рачунара.

Како се користи таблица случајних бројева?

Пример 10. Ако би нам из било ког разлога било потребно да имамо 15 двоцифрених бројева не већих од 63, који су униформно расподељени, тј. ако је у питању случајни експеримент са расподелом $P\{X = n\} = \frac{1}{90}$ где је n двоцифрен број, требало би из таблице по некој стратегији (или редом) читати групе од по две цифре изостављајући оне групе које почињу нулом све док не изаберемо 15 двоцифрених бројева не већих од 63. На путу до тог циља игнорисали бисмо све групе цифара које би протумачили као двоцифрен број већи од 63 на које бисмо у таблици наишли. На пример, читајмо групе од по две цифре из првог и другог реда Таблице 6.

Добијамо редом

$$51, 77, 27, 46, 40, 42, 33, 12, 90, 44, 46, 62, 12, 40, 33, 23, 49$$

и одбацујемо бројеве 77 и 90. Ако би за експеримент било прихватљиво да се бројеви понављају, у овом тренутку бисмо завршили читање. Међутим, ако се бројеви не смеју понављати, игнорисали бисмо $46, 12, 40$ и 33 када се други пут јаве у прочитаном низу и прочитали бисмо још наредне бројеве

$$49, 18, 35, 87, 06, 56, 82, 19$$

и одбацили 87, 06 и 82. $\triangle$

О примени таблице случајних бројева биће надаље још речи.

Нагласимо да је са појавом рачунара поменута таблица изгубила на значају, али не и метод. Наиме, таблица није погодна за коришћење при обради података на рачунару, јер пре свега успорава рад паралелним радом, а друго није погодно ни да се таблица унесе у меморију рачунара јер би заузела, односно, блокирала велики део меморије за активно коришћење. С тога се приликом рада на рачунару користе тзв. псеудослучајни бројеви. Псеудослучајни бројеви "доста добро" са статистичке тачке гледишта апроксимирају таблицу случајних бројева, тј. поменуту униформну расподелу, а генеришу се помоћу алгоритама програмираних на рачунару. Један од најчешће коришћених алгоритама је линеарни конгруентни метод код кога се низ бројева $x_0, x_1, x_2, \dots$ добија преко формуле

$$x_{n+1} = (ax_n + c) mod M\,.$$

Читава теорија је усмерена на то да се константе x_0, a, c и M одаберу тако да се добије што дужи низ бројева. Према дефиницији је јасно да је дужина низа различитих бројева највише M. Најчешће је $M = 2^k, \quad k \geq 1$ (k се по правилу узима као врло велики број).

Квалитет који треба да задовољи алгоритам да би добијени низ "довољно добро" апроксимирао низ случајних бројева је предмет из домена тестирања статистичких хипотеза.

2.4 Случајни избори без и са враћањем

Без обзира да ли је популација коначна (обима N, $N < \infty$) или бесконачна, могуће је из ње на различите начине изабрати узорке истог обима n (за коначну популацију $n < N$) за различите природне бројеве n. Дакле, може се говорити о колекцији $\mathcal{S}$ свих узорака из исте популације Ω, $\mathcal{S} = \{s\}$, где је са s означен произвољан узорак посматране популације, $s \subset \Omega$, док ће обим узорка s бити обележен са $n(s)$. Ако у узорку s има истих елемената, са $\nu(s)$ можемо означити број различитих елемената у узорку s.

Дефиниција 5. Број различитих елемената у узорку је *ефективни обим узорка.* $\diamondsuit$

У вези са ефективним обимом узорка за коначну популацију уводи се појам стопе избора:

Дефиниција 6. *Стопа избора узорка* или *фракција узорка* је функција од узорка s дефинисана као количник ефективног обима узорка и обима популације,

$$f(s) = \frac{\nu(s)}{N} \,. \diamondsuit$$

Ако узорак s редукујемо само на различите елементе, добићемо узорак $\tilde{s}$ чији ће обим бити $\nu(s)$, тј. $\nu(\tilde{s}) = \nu(s) = n(\tilde{s})$. Ако међу свим елементима скупа $\mathcal{S}$ извршимо овакву редукцију, добићемо скуп $\tilde{\mathcal{S}} = \{\tilde{s}\}$.

Закључивање на основу узорка по правилу зависи од начина избора елемената популације у узорак. Начин избора узорка зове се *план узорка* или *стратегија избора.* Формална дефиниција плана је следећа:

Дефиниција 7. *План узорка* је закон расподеле случајне променљиве S дефинисане на скупу $\mathcal{S}$, тј. $\{P(S = s), s \in \mathcal{S}\}$. $\diamondsuit$

Надаље ће бити коришћена ознака $P(S = s) = p(s)$.

У том смислу се уопштава појам случајног узорка о коме је већ било речи, подразумевајући да је узорак случајан и када је добијен на основу познатог плана, тј. на основу задате расподеле вероватноћа.

За $\omega \in \Omega$ дефинисаћемо индикатор, у смислу да ли уочени елемент популације припада изабраном узорку s, на следећи начин:

$$I_s(\omega) = \begin{cases} 1, & \omega \in s \\ 0, & \omega \notin s \,. \end{cases}$$

При томе је

$$P(I_s = 1) = \sum_{s \ni \omega} p(s) \,.$$

Краће ћемо означити

$$P(I_s = 1) = \pi\,.$$

За пребројиву, а нарочито за коначну популацију, по потреби се дефинише бијективна функција на скуп природних бројева, тј. на првих N природних бројева, за коначну популацију, чиме се сваки елемент популације идентификује са својим "местом" у популацији. Тада је могуће дефинисати *индикатор укључења i–тог елемента популације у узорак* као случајну промењливу

$$I_s(\omega_i) = \begin{cases} 1, & \omega_i \in s \\ 0, & \omega_i \notin s \end{cases}$$

са расподелом

$$P(I_s(\omega_i) = 1) = \sum_{s \ni \omega_i} p(s)\,.$$

Користићемо ознаку

$$P(I_s(\omega_i) = 1) = \pi_i\,.$$

Поменута бијекција се примењује када је битан редослед избора елемената у узорак.

Нека је из популације Ω на којој посматрамо обележје X узет случајни узорак обима n, $\mathbf{X} = (X_1, X_2, \ldots, X_n)$. Као што је већ речено, X_i је вредност обележја на i–том елементу узорка, односно, $(X_1, X_2, \ldots, X_n)$ је низ случајних променљивих. Са гледишта теорије вероватноће, најједноставнији је *прост* случајни узорак код кога се претпоставља да су случајне променљиве $X_1, X_2, \ldots, X_n$ независне и да свака има исту расподелу као обележје X. У терминима планова узорака то значи да су све вероватноће $p(s)$ при $n(s) = n$ међу собом једнаке.

Најопштија подела планова случајних узорака је на узорке са враћањем (понављањем) и узорке без враћања (понављања). Узорак са враћањем претпоставља стратегију избора код које се један исти елемент популације може више пута јавити у узорку, односно бити изабран. То би се могло догодити уколико се после избора елемнта у узорак и регистровања вредности обележја на њему, он поново враћа у популацију. Отуда и назив ове стратегије. Код узорка без враћања таква могућност не постоји, односно по избору елемента у узорак једном, он се више не враћа у популацију. Код бесконачне популације се ове две стратегије у пракси не разликују, јер је мала вероватноћа поновног избора истог елемента популације у узорак. Уколико је популација коначна, али далеко већег обима него што је обим узорка, вероватноће свих узорака, $p(s)$, константног обима су приближно једнаке, па се узорак изабран по било којој од наведених стратегија може сматрати простим са вероватноћом 1. (Да би ова тврдња опстала, морало би се приступити опширном доказивању, што овде неће бити спроведено.) Ситуација у којој се избор са враћањем и избор без враћања битно разликују је избор из коначне популације из које се узима узорак чији обим није занемарљиво мали у односу на обим популације. Надаље ће бити више речи о томе.

* * *

За избор узорка из уређене популације може се користити таблица случајних бројева.

Уколико нам је потребан узорак обима 20 из популације обима 1000, која је уређена, читали бисмо групе од по три цифре заједно из таблице случајних бројева. Добијене бројеве бисмо тумачили као редне бројеве елемената популације. Уколико би међу прочитаним групама била група 000, то бисмо протумачили као да је реч о последњем елементу популације. Број група које бисмо прочитали би зависио од стратегије избора, а не само од обима узорка. За узорак са враћањем прочитали бисмо тачно 20 група. За узорак без враћања бисмо морали да изоставимо сваку поновљену групу и да наставимо читање док не прочитамо 20 различитих група цифара, тј. редних бројева.

Пример 11. Нека је дата популација од 100 елемената. Користећи таблицу случајних бројева моделирати реализовани узорак без враћања од 20 елемената из ове популације.

Ова популација је очигледно уређена, или се може уредити. Из таблице случајних бројева читаћемо редне бројеве елемената популације које ћемо узети у узорак. Ако се одлучимо за петнаести ред Таблице 6 и читемо по две цифре добијамо:

$$85, 65, 93, 60, 81, 50, 88, 41, 40, 70, 74, 95,$$

сад можемо да наставимо са читањем у шеснаестом реду, при чему можемо да "прочитамо" и 0 на крају дела таблице случајних бројева која се налази на крају петнаестог реда Таблице 6 или да је изоставимо. Рецимо да је "прочитамо", добијамо

$$05, 51, 89, 00, 56, 52, 53, 11.$$

Број 74 се јавља два пута и њега изостављамо на месту када се други пут јави (између група "00" и "56"), јер смо већ претходно узели 74–ти елемент популације у узорак, а узорак је без враћања. Јасно да 05 налаже да узмемо 5–ти елемент популације у узорак, а 00 да узмемо 100–ти. $\triangle$

Пример 12. Планира се сондирање терена у 8 тачака ради испитивања састава тла. Прецизност мерења је до $0,10m$, а површина испитиваног терена је $68a$(ари).

У мапу те локације треба унети Декартов координатни систем тако да уцртане осе буду тангенте испитиване парцеле и утврдити димензије минималног правоугаоника који у потпуности покрива испитивани терен са двема страницама на осама, а затим места за сондирање одредити уз помоћ таблице случајних бројева.

Нека је правоугаоник димензије $100 \times 80m$. Читаћемо из Таблице 6 из првог и другог, а затим из трећег и четвртог реда упоредо групе од по прве 4 цифре из сваке групе колона ради формирања уређених парова координата и множити добијене бројеве са 10^{-2}:

$$x: \ 51,77 \quad 74,64 \quad 42,33 \quad 29,04 \quad 46,62 \quad 45,93 \quad 60,17 \quad 52,07 \quad 25,42\ldots$$

$$y: \ 24,03 \quad 23,49 \quad 83,58 \quad 06,56 \quad 21,96 \quad 30,58 \quad 02,13 \quad 75,79 \quad 45,40\ldots$$

Уређени пар $(42,33; 83,58)$ се одбацује јер излази из подручја дефинисаног правоугаоника. Такође ће бити тачака које треба одбацити јер не припадају дефинисаном подручју које се испитује. $\triangle$

2.4.1 Узорак без враћања из коначне популације

План узорка обима n без враћања дефинисан је на колекцији $\tilde{\mathcal{S}}$ узорака без понављања елемената:

$$p(s) = \begin{cases} \frac{1}{\binom{N}{n}}, & n(s) = \nu(s) = n \\ 0 \quad , & \text{у осталим случајевима} \end{cases} \quad , \quad s \in \tilde{\mathcal{S}}$$

уколико је популација коначна, $N \geq n$. То отуда што је сваки узорак $\tilde{s}$ комбинација без понављања n–те класе од N елемената, а у случају да је избор "фер", сваки такав узорак биће изабран са вероватноћом $\frac{1}{\binom{N}{n}}$. Оваквих узорака који садрже фиксирани елемент ω има тачно $\binom{N-1}{n-1}$ па је вероватноћа

$$\pi = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}$$

за неуређену популацију. Уколико је пак популација уређена, план узорка без враћања биће

$$p(s) = \begin{cases} \frac{1}{n!\binom{N}{n}}\,, & n(s) = \nu(s) = n \\ 0 \quad , & \text{у осталим случајевима} \end{cases} \quad , \quad s \in \tilde{\mathcal{S}}\,,$$

јер се ради о варијацијама без понављања, а вероватноћа π_i је

$$\pi_i = \sum_{s \ni \omega_i} \frac{1}{n!\binom{N}{n}} = \frac{n}{N}\,.$$

Дакле, вероватноћа избора произвољног али фиксираног елемента популације у узорак по стратегији избора без враћања, за задати обим узорка (n) је константна и износи $\frac{n}{N}$ без обзира да ли је популација уређена или не.

Генерално гледано, случајни избор без враћања добија се било извлачењем свих n елемената из популације одједном, било извлачењем једног по једног елемента не враћајући га више у популацију.

Избор без враћања има константну стопу избора $f = n/N$, јер је $\nu(s) = n(s) = n$.

2.4.2 Узорак са враћањем из коначне популације

План избора са враћањем заснива се на чињеници да је у сваком извлацењу вероватноћа избора појединог елемента популације у узорак иста и једнака $1/N$. Према томе

$$p(s) = \begin{cases} \frac{1}{N^n}\,, & n(s) = n \\ 0 \quad , & \text{иначе} \end{cases} \quad , \quad s \in \mathcal{S}\,,$$

без обзира да ли је популација уређена или не, а вероватноћа да i–ти члан популације буде укључен у узорак је

$$\pi_i = 1 - \left(\frac{N-1}{N}\right)^n \,, \quad i = 1, 2, ..., N\,,$$

дакле иста као и када се ради о вероватноћи π за елемент ω неуређене популације.

Овакав узорак нема константну стопу избора, јер $\nu(s)$ варира за исти обим узорка n.

2.5 Неки специјални планови узорака

2.5.1 Стратификовани узорак

У многим реалним ситуацијама природно је поделити популацију на подгрупе које треба проучавати.

Пример 13. Треба извршити анкетно истраживање у предузећима државног и приватног сектора. Предузећа су елементи популације из које треба узети узорак. Међутим, нека предузећа су врло велика и запошљавају више хиљада радника, док су друга мала и запошљавају свега неколико лица. Било која оцена на основу директног случајног узорка изабраног из целине скупа предузећа неће бити реална. Поступак којим се постиже значајно побољшање прецизности закључивања на основу случајног узорка јесте стратификација. Тако се предузећа могу поделити према броју радника на велика, средња и мала. $\triangle$

Дефиниција 8. *Стратификација (раслојавање)* подразумева поделу популације на делове – *стратуме (слојеве)*, дисјунктне подскупове чија унија обухвата целу популацију, са захтевом постизања што веће хомогености унутар стратума (слоја). $\diamondsuit$

Хомогеност се остварује према неком заједничком својству елемената популације, на пр. старосна доб, пол, тип предузећа и сл.

Као што је речено, стратуми су међу собом дисјунктни, а сви заједно обухватају целу популацију. Дакле, чине једно разбијање популације потпуним системом догађаја.

Како је основни циљ математичке статистике оцењивање расподеле обележја посматраног на популацији, циљ стратификације је да се постигне већа тачност оцене, економичност или једноставност испитивања и слично. У неким ситуацијама је испитивање и једино могуће спровести по стратумима.

Техника стратификације подразумева решавање одређених задатака, односно проналажење одговора на следећа питања:

- Како формирати стратуме и колико њих ?

- Како изабрати (расподелити, алоцирати) укупан узорак уочавајући поједине стратуме, тј. алоцирати узорак по стратумима ?

- Како спровести статистичко закључивање на основу добијеног стратификованог узорка ?

На ова питања се може и треба вратити касније, пошто се обраде тачкасте и интервалне оцене параметара, док ћемо се овде још мало задржати на дефинисању стратификованог узорка.

Нека је популација обима N подељена на L дисјунктних стратума обима N_l, где је $l = 1, 2, \ldots, L$, при чему је $N_1+N_2+\ldots+N_L = N$. Претпоставимо да су обими стратума познате величине. Удео l–тог стратума у узорку може се мерити величином $w_l = \frac{N_l}{N}$. Очигледно је $w_1 + w_2 + \ldots + w_L = 1$. Дакле, w_l би се могла протумачити као класична дефиниција вероватноће да се при случајном избору елемената популације, изабере елемент l–тог стратума.

Стратуми се у истој популацији могу одредити на различите начине.

Пример 14. Ради испитивања успеха на студијама на Природно-математичком факултету треба извршити стратификацију свих уписаних студената у прву годину студија.

Већ при самом упису студенти су подељени по одсецима које можемо прихватити као стратуме (слојеве). Дакле, студенти једног одсека би чинили један стратум. У том случају било би онолико стратума колико има одсека на Природно-математичком факултету.

Међутим, с обзиром на циљ истраживања, интуитивно би било прихватљивије дефинисати стратуме према постигнутом успеху у средњој школи. Дакле, ако прихватимо четири уобичајене категорије успеха: одличан, врло добар, добар и довољан, популацију студената уписаних у прву годину студија посматраног факултета поделили би на четири стратума.

Коначно за који начин поделе на стратуме бисмо се определили зависило би и од одговора који се истраживањем тражи, тј. да ли је акценат, рецимо, на професионалној оријентацији средњошколаца (прва подела) или на валидности оцењивања у средњим шкоалама (друга подела). $\triangle$

Статистички критеријум за "бољу" стратификацију спада у домен тестирања статистичких хипотеза.

Из стратификоване популације се узима узорак s чији је обим $n(s)$, а кога чине подскупови (подузорци) $s_1, s_2, \ldots, s_L$ при чему је s_i, $i = 1, 2, \ldots, L$ део узорка s који је узет из i–тог стратума. Дакле, $s_i \cap s_j = \emptyset$, $\sum_{i=1}^{L} s_i = s$, па важи

$$n(s_1) + n(s_2) + \cdots + n(s_L) = n(s).$$

По правилу се уводи претпоставка да су извлачења из различитих стратума независна. Стопа избора i–тог стратума је $f_i = \frac{n(s_i)}{N_i}$. Уколико је $f_i = c$ за свако $i = 1, 2, \ldots, L$, реч је о пропорционалној расподели обима узорка по стратумима. Свакако најједноставнији метод размештаја узорка по стратумима је избор једнаког броја елемената из сваког слоја, тј. ако важи $n(s_i) = \frac{n}{L}$ за свако $i = 1, 2, \ldots, L$. Овакав размештај (алокација) узорка по слојевима се углавном примењује када су слојеви приближно истог обима. У противном се, по правилу, користи алокација са константном фракцијом.

Извлачења из стратума се могу вршити такође са и без враћања при чему добијамо стратификовани случајни узорак са враћањем или без враћања.

Уколико се на i–том стратуму вредност посматраног обележја на популацији X, означи са $X^{(i)}$, случајни узорак ће бити вектор

$$\mathbf{X} = \left(X_1^{(1)}, X_2^{(1)}, \ldots, X_{n(s_1)}^{(1)}, X_1^{(2)}, X_2^{(2)}, \ldots, X_{n(s_2)}^{(2)}, \ldots, X_1^{(L)}, X_2^{(L)}, \ldots, X_{n(s_L)}^{(L)}\right)$$

где се уочавају подвектори који одговарају појединим стратумима.

2.5.2 Групни узорак

Групни узорак такође подразумева претходну поделу целе популације на дисјунктне делове. Код групног узорка се, међутим, не претпоставља подела према заједничком својству у циљу постизања хомогености групе. Напротив, принцип поделе на групе је практичне природе и може бити по територијалном принципу или неком сличном.

Нека је разматрана популација подељена по неком принципу на више дисјунктних група. За стратификовани узорак је потребно из сваке групе изабрати одређени број елемената популације. Насупрот томе, за групни узорак треба изабрати одређени број група на случајан начин и узети све елементе из изабраних група у узорак. Групни узорак се још зове узорак скупина или узорак серија (серија у производњи неког артикла нпр.). Основна јединица избора овог типа узорка је група (за разлику од случајног узорка и стратификованог узорка код којих је основни елемент избора био елемент популације).

Пример 15. На територији Србије треба спровести анкету о коришћењу једног лека за контролу крвног притиска.

Непостојање употребљивих спискова становника Србије у моменту спровођења анкете само је један од разлога који онемогућава избор простог случајног узорка или стратификованог узорка. Други би разлог могао бити економски аспект истраживања, јер би било неекономично да персонал који анкетира становништво обилази све делове територије Србије. Због свега тога је оправдано извршити груписање становништва по територијалним јединицама, рецимо општинама, па анкетирати све становнике случајно изабраних општина. $\triangle$

Групни узорак подразумева да у дефинисаним групама има коначно много елемената популације $N_1, N_2, \ldots, N_i, \ldots$ па се такав узорак може представити случајним вектором

$$\mathbf{X} = \left(X_1^{(k_1)}, X_2^{(k_1)}, \ldots, X_{N_{k_1}}^{(k_1)}, X_1^{(k_2)}, X_2^{(k_2)}, \ldots, X_{N_{k_2}}^{(k_2)}, \ldots, X_1^{(k_l)}, X_2^{(k_l)}, \ldots, X_{N_{k_l}}^{(k_l)} \right),$$

где је k_i ознака групе са укупно N_{k_i} елемената популације у себи.

Избор са и без враћања код групног узорка односио би се на поновни, или не, избор истих група у узорак.

2.5.3 Систематски узорак

Веома погодан метод избора узорка из коначне уређене популације састоји се у следећем:

Нека је $N = nk$, где је n задати обим узорка и k такође природан број. Узима се случајан број између 1 и k (из таблице случајних бројева), претпоставимо да је то број i. Тада се узорак формира од елемената популације чији су редни бројеви

$$i, i+k, i+2k, \ldots, i+(n-1)k\,,$$

тј. узорак садржи први случајно изабрани елемент популације и сваки следећи k–ти по реду бројећи од тог првог.

Погодност овог узорка састоји се у томе што први корак (избор првог елемента узорка) одређује узорак у целини. Међутим, ова стратегија избора је примењива само уколико је редослед елемената популације у уређењу које је на популацији дефинисано **случајан**. О овоме ће још бити речи.

Уочимо да је за дату популацију процедура систематског узорка у ствари избор једне од k група (на које је подељена цела популација) са вероватноћом $\frac{1}{k}$. У овом случају групе одређују скупови индекаса:

$$\{1, k+1, 2k+1, \dots, (n-1)k+1\}, \dots, \{i, k+i, 2k+i, \dots, (n-1)k+i\}, \dots, \{k, 2k, 3k, \dots, nk\}.$$

Вероватноћа да оваквим подскупом i–ти елемент популације буде изабран у узорак је $\pi_i = \frac{1}{k}$.

У случају када је $N = nk + c$, $c < k$, c природан број, неке групе садрже n елемената популације, а друге $n+1$ елемената, тј. величине група нису исте од групе до групе, а вероватноћа избора елемената популације у узорак је такође $\pi_i = \frac{1}{k}$ било да је узорак обима n или $n+1$. (Наиме, ако је случајно изабрани редни број првог елемента који треба узети у узорак из уређене популације i такав да је $i \le c < k$, описаним правилом добиће се узорак обима $n+1$.)

Систематски узорак се још зове *периодични* или *механички* узорак.

Примена таблице случајних бројева није од суштинског значаја за избор систематског узорка. Отуда се често први елемент периодичног узорка бира као средишњи у првом интервалу избора.

Систематски узорак има одређене предности над случајним узорком, јер је правило избора сасвим просто, не захтева таблице случајних бројева, па ни потпуну нумерацију популације, а изводи се знатно брже.

Пример 16. Треба проценити учестаност јављања алергијског бронхитиса међу пацијентима једне здравствене установе.

С обзиром да пацијенте репрезентују њихови здравствени картони који се налазе у картотеци здравствене установе, из картотеке треба узети картоне пацијената који ће чинити узорак уз помоћ лењира дефинишући дужинско растојање између два изабрана картона. При томе су мала одступања од задате дужине без значаја, као и то да ли су картони уредно поређани по бројевима.

Обратимо пажњу на чињеницу да је редослед пристизања пацијената у здравствену установу, чиме је утврђен редослед отворених картона, случајан. Због тога редни број картона чини случајно уређење у популацији. △

Систематски узорак је интуитивно прихватљив – "равномерно" је распоређен по популацији, не допушта случајна груписања или "пропуштања" неких делова популације, што се код случајног избора може десити.

Систематски узорак се може упоредити са стратификованим узорком код кога стратуми представљају елементе на интервалу дужине k, при чему се из сваког од њих бира по један елемент.

Ако је испуњена претпоставка да су елементи популације случајно распоређени у низ, или да је обележје које се посматра независно од распореда елемената популације, систематски узорак постаје само један вид случајног узорка без враћања. Иако се ова логика веома често користи треба бити опрезан. На пример, када се прате сезонске појаве, тј. обележја која имају сезонска колебања (као што је температура ваздуха, број туриста и сл.), може се десити да се сезонска колебања у вредности обележја поклопе са периодом избора и дају погрешну слику о обележју. Због тога се о овоме мора водити рачуна при доношењу одлуке о систематс ком избору.

2.5.4 Вишеетапни узорак

Групни узорак је по структури једноставан, али када је обим група велики може бити непрактичан, или давати мању тачност. Повезивање метода групног и стратификованог или групног и систематског узорка даје нам идеју избора узорка у две или више етапа. Наиме, у првој етапи од свих (дисјунктних) група на које је популација подељена бирамо на случајан начин одређени број група, а затим у другој етапи из сваке групе изабране у првој етапи бирамо такође на случајан начин одређени број елемената. Овакав узорак зове се двоетапни узорак.

Пример 17. Ако у претходном примеру не вршимо анкетирање свих становника одабраних општина, већ одабраног дела становништва из сваке одабране општине (групе) добићемо двоетапни узорак. Прву етапу чини избор група, тј. општина из којих ће се у другој етапи бирати одређени елементи – становници који им припадају. (Уколико би се у првој етапи изабрале све постојеће групе на које је популација подељена, двоетапни узорак би се свео на стратификовани.) $\triangle$

Пример 18. Ако у одабраним општинама у првој етапи претходног примера уочимо месне заједнице, па изаберемо на случајан начин одређен број месних заједница из одабраних група за даљу анализу, а затим у трећој етапи одаберемо на случајан начин по одређени број становника у узорак за анкетирање добићемо такође троетапни узорак.

Уколико би групе прве етапе биле подељене на дисјунктне подгрупе, па из сваке од група изаберемо у другој етапи подгрупе из којих ћемо тек у трећој етапи бирати елементе у узорак, формирали бисмо троетапни узорак. $\triangle$

По истом принципу може се формирати било који вишеетапни узорак са унапред дефинисаним коначним бројем етапа.

Систематски узорак се може користити такође у комбинацији са осталим методима избора узорка. На пример, код стратификованог узорка се елементи унутар стратума могу бирати периодично. Код групног узорка се групе могу бирати периодично. Код вишеетапног се могу комбиновати систематски и случајни избор на више начина – у свакој од етапа избор може бити периодичан или случајан.

* * *

Констатујмо да, свака од стратегија избора има за последицу одређену тачност у оцењивању непознатих параметара обележја, као и тестирању одговарајућих хипотеза.

На тачност статистичог закључка утиче и обим узорка. О одређивању обима узорка биће више речи у оквиру Главе 3.

2.6 Емпиријска функција расподеле

Вратимо се случајном узорку уопште и размотримо још неке важне појмове везане за узорак.

Окосница научне области коју зовемо математичком статистиком или, једноставно, статистиком, је функција од узорка описана следећом дефиницијом:

Дефиниција 9. *Статистика* је функција од узорка чији аналитички израз не зависи од непознатих параметара обележја, тј. функција од узорка и познатих констаната. $\diamondsuit$

Примери неких статистика су:

$$T_n = \sum_{i=1}^{n} X_i \quad - \text{тотал узорка}$$

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \quad - \text{средина узорка}$$

$$\overline{S}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \quad - \text{дисперзија узорка}$$

$$\overline{S}_n = \sqrt{\overline{S}_n^2} \quad - \text{узорачка стандардна девијација}$$

$$\widetilde{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \quad - \text{поправљена дисперзија узорка}$$

$$R = X_{max} - X_{min} \quad - \text{распон узорка}.$$

За два обележја X и Y и узорак $((X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n))$ из популације на којој се посматра дводимензионо обележје (X, Y) може се дефинисати статистика

$$R_{XY} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{\overline{S}_X \overline{S}_Y} - \text{узорачки коефицијент корелације},$$

где су са $\overline{S}_X$ и $\overline{S}_Y$ означене узорачке стандардне девијације за обележја X и Y редом.

Посебно место међу статистикама имају тзв. статистике поретка. Ове се статистике дефинишу посредством варијационог низа:

Дефиниција 10. *Варијациони низ* чине елементи узорка поређани у неопадајућем поретку. $\diamondsuit$

За узорак $(X_1, X_2, \ldots, X_n)$ варијациони низ чини низ случајних променљивих сачињен од елемената овог узорка у ознаци

$$X_{(1)}, X_{(2)}, \ldots, X_{(n)}$$

за који важи

$$X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}\,.$$

За реализоване вредности варијационог низа користи се исти термин *варијациони низ*, без опасности од забуне, а означавају се малим словима:

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}\,.$$

Дефиниција 11. *Статистика поретка реда* k узорка обима n, $1 \leq k \leq n$, је k–ти елемент варијационог низа посматраног узорка, дакле случајна променљива $X_{(k)}$. $\diamondsuit$

У дефиницији функције расподеле, у овој књизи, биће коришћена непрекидност с десна, тј.

$$F(x) = P\{X \leq x\}\,, \quad x \in R\,.$$

Нека је $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ случајни узорак из популације са обележјем X чија је функција расподеле F. За свако $x \in R$ дефинисаћемо случајну величину $\mu_n(x)$ као број елемената узорка $\mathbf{X}$ који су мањи или једнаки x, тј.

Дефиниција 12.

$$\mu_n(x) = card\{j | X_j \leq x, j = 1, 2, \ldots, n\}\,, \quad x \in R\,. \diamondsuit$$

Надаље се може дефинисати случајна променљива $S_n(x)$ која даје вредности случајне променљиве $\mu_n(x)$ у релативном односу према обиму узорка:

Дефиниција 13. *Емпиријска функција расподеле* узорка $\mathbf{X}$ је статистика

$$S_n(x) \stackrel{\text{def}}{=} \frac{\mu_n(x)}{n}\,, \quad x \in R\,. \diamondsuit$$

Случајна променљива $S_n(x)$ је статистика чији је кодомен скуп

$$\{0, 1/n, 2/n, \ldots, (n-1)/n, 1\}$$

или његов прави подскуп са вероватноћама

$$P\{S_n(x) = k/n\} = P\{\mu_n(x) = k\} = \binom{n}{k} (F(x))^k (1 - F(x))^{n-k}, \qquad k = 0, 1, \ldots, n\,.$$

Ово отуда што, према дефиницији, случајна променљива $\mu_n(x)$ има биномну расподелу, $\mathcal{B}(n,p)$ са $p = P\{X \le x\} = F(x)$, $x \in R$. Статистику $S_n(x)$, $x \in R$, можемо посматрати и као аритметичку средину индикатора

$$I_{A_i} = \begin{cases} 1, & \omega \in A_i \\ 0, & \omega \notin A_i \end{cases},$$

$A_i = \{\omega | X_i(\omega) \le x\}$, тј.

$$S_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{A_i}, \quad x \in R .$$

С обзиром да је $E(I_{A_i}) = F(x)$ за фиксирано $x \in R$, важи теорема:

Теорема 1. *За фиксирано $x \in R$, $S_n(x) \longrightarrow F(x)$, $n \to \infty$ скоро извесно, тј.*

$$P\left\{S_n(x) \to F(x)\, ,\, n \to \infty\right\} = 1.$$

Доказ. Тврђење следи на основу Бореловог закона великих бројева.□

За реализовани узорак $(x_1, x_2, \ldots, x_n)$, $S_n(x)$, $x \in R$, је монотоно неопадајућа функција са могућим скоковима у тачкама варијационог низа $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$:

$$S_n(x) = \frac{k}{n}, \quad x \in \left[x_{(k)}, x_{(k+1)}\right), \quad k = 0, 1, \ldots, n .$$

При томе су уведене ознаке $x_{(0)} = -\infty$, и у том случају је и лева граница интервала отворена, и $x_{(n+1)} = +\infty$. Уколико су сви елементи у реализованом узорку различити, скокови су величине $1/n$.

Конвергенција о којој је било речи у претходној теореми, остварује се и униформно по $x \in R$. О томе говори тзв. *централна теорема математичке статистике.* Један од њених облика је следећи.

Теорема 2. (Гливенко-Кантели) *Нека је F функција расподеле обележја X и $S_n(x)$, $x \in R$, емпиријска функција расподеле узорка обима n из популације са обележјем X. Тада важи*

$$P\left\{\sup_{x \in R} |S_n(x) - F(x)| \to 0\, ,\, n \to \infty\right\} = 1. \tag{2.1}$$

Доказ. Нека је F произвољна функција расподеле обележја дискретног или апсолутно непрекидног типа и нека је ε произвољан реалан број за који важи $0 < \varepsilon < 1$. За тако изабрано ε и задату функцију F, могуће је изабрати коначан број тачака

$$z_0, z_1, \ldots, z_N \in \overline{R} \qquad (\overline{R} = R \cup \{-\infty, +\infty\})$$

таквих да је

$$-\infty = z_0 < z_1 < \ldots < z_{N-1} < z_N = +\infty$$

$$F(z_k - 0) - F(z_{k-1}) \leq \varepsilon, \quad k = 1, \ldots, N.$$

На пример, могуће је изабрати скуп $\{z_j\}$ тако да он садржи све тачке прекида (ако има тачака прекида) функције F у којима је скок функције F већи од $\varepsilon/2$. Тада за произвољно $z \in [z_{k-1}, z_k)$ важи

$$S_n(z) - F(z) \leq S_n(z_k - 0) - F(z_{k-1}) \leq S_n(z_k - 0) - F(z_k - 0) + \varepsilon.$$

Слично и

$$S_n(z) - F(z) \geq S_n(z_{k-1}) - F(z_k - 0) \geq S_n(z_{k-1}) - F(z_{k-1}) - \varepsilon.$$

Дефинишимо следеће скупове

$$\overline{B}_k = \{\omega | (S_n(z_k - 0))(\omega) \longrightarrow F(z_k - 0), \;\; n \to \infty\}$$

$$B_k = \{\omega | (S_n(z_k))(\omega) \longrightarrow F(z_k), \;\; n \to \infty\}$$

$$B = \bigcap_{k=0}^{N} B_k \overline{B}_k.$$

Тада, према претходној теореми, догађаји B_k и $\overline{B}_k$ се реализују скоро извесно, тј.

$$P(B_k) = P(\overline{B}_k) = 1.$$

Отуда је

$$P(B) = 1.$$

Ово с тога што се за свако $\omega \in B$ може наћи узорак довољно великог обима $n(\omega)$ такав да кадгод је $n \geq n(\omega)$ тада је $B_0\overline{B}_0 \subset B_1\overline{B}_1 \subset \ldots \subset B_N\overline{B}_N$.

Дакле, за довољно велико $n \geq n(\omega)$, биће

$$|S_n(z_k - 0) - F(z_k - 0)| < \varepsilon, \quad k = 0, 1, \ldots, N$$

и

$$|S_n(z_k) - F(z_k)| < \varepsilon, \quad k = 0, 1, \ldots, N,$$

за свако k, те је

$$\sup_{z \in R} |S_n(z) - F(z)| \leq 2\varepsilon.$$

Тиме је теорема доказана.□

Следеће две теореме говоре о расподели важних статистика базираних на емпиријској функцији расподеле. Овде ћемо их навести без доказа.

Теорема 3. **(Колмогорова)** *Ако је функција F непрекидна, тада за произвољно фиксирано $t > 0$ статистика $D_n = \sup_{x \in R} |S_n(x) - F(x)|$ има расподелу за коју важи*

$$\lim_{n \to \infty} P\{\sqrt{n} D_n \leq t\} = K(t) = \sum_{j=-\infty}^{+\infty} (-1)^j e^{-2j^2 t^2}.$$

Теорема 4. **(Смирнова)** *Нека су S_{1n_1} и S_{2n_2} две емпиријске функције расподеле сачињене на основу два независна узорка обима n_1 и n_2 из исте популације са обележјем X и*

$$D_{n_1 n_2} = \sup_{x \in R} |S_{1n_1}(x) - S_{2n_2}(x)|.$$

Тада, ако је теоријска функција расподеле F непрекидна, за произвољно фиксирано $t > 0$,

$$\lim_{n_1, n_2 \to \infty} P\left\{\sqrt{\frac{n_1 n_2}{n_1 + n_2}} D_{n_1 n_2} \leq t\right\} = K(t).$$

2.7 Моделирање расподела методом Монте Карло

Идеја статистичког моделирања (симулације) се састоји у следећем:

Треба одредити приближну вредност неке реалне величине a. У том циљу бира се случајна величина X са расподелом таквом да је $E(X) = a$. На основу реализованог узорка

$$\mathbf{x} = (x_1, x_2, \dots, x_n)$$

из популације са обележјем X, одређује се приближна вредност величине a као оцена математичког очекивања $E(X)$,

$$\hat{a} = \frac{1}{n} \sum_{i=1}^{n} x_i \, .$$

Више речи о самој оцени биће у наредном поглављу. Овде треба прокоментарисати са којом тачношћу се врши овакво оцењивање. Наиме, према централној граничној теореми,

$$P\{|\overline{X}_n - a| \le \varepsilon\} \approx 2\Phi\left(\varepsilon \frac{\sqrt{n}}{\overline{s_n}}\right),$$

где је

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{t^2}{2}} dt \, , \quad \text{и} \quad \overline{s_n} = \sqrt{\overline{s_n}^2},$$

односно

$$\overline{s_n}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x_n})^2.$$

Тачност ове оцене је реда $1/\sqrt{n}$, што се најчешће не сматра великом тачношћу.

Надаље ћемо се бавити практичним решавањем проблема статистичког моделирања саме расподеле обележја X које је у дефиницији проблема истакнуто.То подразумева моделирање реализованог узорка из популације са овим обележјем.

2.7.1 Моделирање дискретне расподеле са коначно много вредности

Када је обележје X са дискретном расподелом са коначно много вредности, задатак се састоји у томе да је потребно симулирати расподелу случајне промењиве

$$X: \quad \begin{pmatrix} x_1 & x_2 & \dots & x_k \\ p_1 & p_2 & \dots & p_k \end{pmatrix}, \quad \sum_{i=1}^{k} p_i = 1.$$

С обзиром да је $p_i \in [0,1]$ за $i = 1, 2, \dots, k$, поделимо сегмент $[0,1]$ на k подинтервала

$$\Delta_1 = [0, p_1), \quad \Delta_2 = [p_1, p_1 + p_2), \dots, \Delta_k = [p_1 + p_2 + \dots + p_{k-1}, 1].$$

Иза тога се врши избор n случајних бројева између 0 и 1 са жељеним бројем значајних цифара из таблице случајних бројева или на неки други начин. Тиме је дефинисана нова случајна променљива Z као случајно изабрани број. Рецимо да је таквим избором добијен низ бројева $z_1, z_2, \dots, z_n$. Надаље, за сваки од добијених бројева z утврђујемо ком интервалу Δ_j, $j = 1, 2, \dots, k$, припада. Нека је $z \in \Delta_l$, $(l = 1, 2, \dots, k)$. Тада прихватимо да је реализован догађај $\{X = x_l\}$ и тако редом. Дакле,

$$P\{X = x_l\} = P\{Z \in \Delta_l\} = d(\Delta_l) = p_l.$$

Пример 19. Нека случајна променљива X има расподелу следећег облика:

$$\begin{pmatrix} -1 & 0 & 1 \\ 0,1 & 0,3 & 0,6 \end{pmatrix}$$

Моделирати десет вредности ове случајне променљиве, користећи таблицу случајних бројева.

Сегмент $[0,1]$ делимо на подинтервале $[0;0,1), [0,1;0,4), [0,4;1]$. Читамо бројеве из Таблице 6 случајних бројева из, на пр., трећег реда. Сада узимамо сваку тако прочитану цифру из таблице случајних бројева и множимо са 10^{-1}. Тако се од низа цифара

$$4,\ 5,\ 9,\ 3,\ 9,\ 6,\ 0,\ 1,\ 7,\ 3$$

добијају бројеви

$$0,4;\ 0,5;\ 0,9;\ 0,3;\ 0,9;\ 0,6;\ 0;\ 0,1;\ 0,7;\ 0,3\,.$$

Број $0,4$ припада интервалу $[0,4;1]$, те прихватамо да се реализовала вредност 1 случајне променљиве X (једна моделирана вредност). Затим исти принцип закључивања примењујемо и на остале изабране бројеве, и коначно добијамо следећи низ моделираних вредности:

$$1, 1, 1, 0, 1, 1, -1, 0, 1, 0\,.$$

$\triangle$

2.7.2 Моделирање расподела апсолутно непрекидног типа

1. **Моделирање униформне расподеле $\mathcal{U}[a, b]$**

   Како сама таблица случајних бројева одсликава униформну расподелу, означимо са η избор k–тоцифрених природних бројева из Таблице 6. Дакле, прочитане групе од по k цифара сматрамо k–тоцифреним бројевима η и затим извршимо множења са 10^{-k} са циљем да моделирамо вредности случајне променљиве $\xi : \mathcal{U}[0, 1]$, $\xi = \eta \cdot 10^{-k}$. Веза између случајних променљивих $X : \mathcal{U}[a, b]$, $a < b$ и $\xi : \mathcal{U}[0, 1]$ је

   $$X = a + (b - a)\xi.$$

   Последњом трансформацијом се управо изврши моделирање случајне променљиве X, односно њених реализованих вредности.

2. **Општи случај**

   Користићемо случајни избор броја из интервала $[0, 1]$, тј. случајну променљиву $\xi : \mathcal{U}[0, 1]$. Посредством те случајне променљиве моделираћемо вредности сваке друге случајне променљиве апсолутно непрекидног типа.

Теорема 5. *Нека је случајна променљива X апсолутно непрекидног типа са функцијом расподеле F. Решење случајне једначине*

$$F(X) = \xi \tag{2.2}$$

по непознатој X, при чему је $\xi : \mathcal{U}[0, 1]$, је случајна променљива чија је функција расподеле баш F.

Доказ. Како је случајна променљива X по претпоставци апсолутно непрекидног типа, онда постоји интервал $(a, b) \subset R$ на коме је функција F монотоно растућа за $x \in (a, b)$ при чему

a може бити и $-\infty$, а b може бити $+\infty$. Нека је најпре F монотоно растућа за свако $x \in R$. Следи да за свако $y \in (0,1)$ постоји тачно један $x \in R$ такав да је $F(x) = y$, тј. на интервалу (a,b) функција F је бијективна функција, те има инверзну. Дакле, једначина (2.2) има јединствено решење на том интервалу. Отуда

$$P\{X \leq x\} = P\{F^{-1}(\xi) \leq x\} = P\{\xi \leq F(x)\} = F_\xi(F(x)) = F(x)\,,$$

$$x \in R\,.$$

Друга могућност је да постоји интервал $[a,b) \subset R$ на коме је F монотоно растућа и важи

$$F(x) = \begin{cases} 0, & x < a \\ g(x), & a \leq x < b \\ 1, & x \geq b \end{cases}\;.$$

То значи да постоји g^{-1} на $[a,b)$ па је за

$$\begin{array}{ll} x < a & P\{X \leq x\} = F(x) = 0 \\ a \leq x < b & P\{X \leq x\} = P\{g^{-1}(\xi) \leq x\} = P\{\xi \leq g(x)\} = g(x) \\ x \geq b & P\{X \leq x\} = F(x) = 1\,. \end{array}$$

□

Пример 20. **(Експоненцијална расподела)** Нека случајна променљива X има густину расподеле

$$f(x) = \begin{cases} 0, & x < 0 \\ \lambda e^{-\lambda x}, & x \geq 0 \end{cases} \quad , \quad x \in R \quad , \quad \lambda > 0,$$

односно функцију расподеле

$$F(x) = \begin{cases} 0, & x < 0 \\ 1 - e^{-\lambda x}, & x \geq 0 \end{cases} \quad , \quad x \in R \quad , \quad \lambda > 0.$$

Тада се решавањем једначине (2.2) по X за $x \geq 0$ добија

$$X = -\frac{1}{\lambda}\ln(1-\xi)\,. \triangle$$

Пример 21. (Нормална расподела) Нека случајна променљива $X : \mathcal{N}(m\,,\,\sigma^2)$. Њене вредности ћемо симулирати помоћу стандардизоване случајне променљиве

$$X^* = \frac{X - m}{\sigma} : \mathcal{N}(0\,,\,1)$$

Посматраћемо два случаја:

- $(i) x^* < 0$
  Решавамо једначину (2.2) за
  $$F(x^*) = 0,5 - \Phi(-x^*)\,.$$
  Другим речима из таблице за нормалну нормирану расподелу читамо вредност $-x^*$ за коју важи
  $$\Phi(-x^*) = 0,5 - \xi\,,$$
  где је ξ реализована вредност случајне променљиве ξ и добијамо вредност за x као $x = \sigma x^* + m$.
- $(ii) x^* \geq 0$
  Решавамо једначину (2.2) за
  $$F(x^*) = 0,5 + \Phi(x^*)\,.$$
  Другим речима из таблице за нормалну нормирану расподелу читамо вредност x^* за коју важи
  $$\Phi(x^*) = \xi - 0,5$$
  и затим долазимо до решења као у случају (i)

До бројева ξ долазимо из таблице случајних бројева, или неким генератором случајних бројева. Уколико добијемо $\xi \in [0\,,\,1/2)$ применићемо решење (i), а уколико добијемо $\xi \in [1/2\,,\,1]$ применићемо (ii). $\triangle$

Глава 3

Оцењивање параметара

Један од најчешће решаваних проблема у математичкој статистици је оцењивање непознатог параметра, или више њих, расподеле обележја на основу узорка. При томе се примењују два типа оцена: тачкасте и интервалне оцене.

Нека је X обележје које се посматра на популацији. Проблем избора расподеле обележја X из фамилије допустивих расподела $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ се може дефинисати и као избор из фамилије функција расподеле $\{F(x;\theta), \theta \in \Theta\}$ или густина расподеле $\{f(x;\theta), \theta \in \Theta\}$. Дакле, сваким од поменута три скупа се на одређени начин дефинише фамилија допустивих расподела за обележје X.

3.1 Тачкасто оцењивање

Већ смо истакли да је основни задатак математичке статистике да на основу експеримента одреди расподелу посматраног обележја на популацији. Дакле, ако на основу неких претходних истраживања или на неки други начин дођемо до фамилије допустивих расподела за посматрано обележје X, проблем одређивања конкретне расподеле се своди на одређивање тачне вредности параметра θ. Међутим, статистичким поступцима само можемо оцењивати непознати параметар на основу узорка и то са одређеном тачношћу.

Тачкасто оцењивање је један од начина за оцењивање праве вредности непознатог параметра и њиме ћемо се најпре бавити. Тачкасте оцене врло често носе назив *дескриптивне статистике.*

Метод тачкастог оцењивања састоји се у следећем:

Треба дефинисати статистику $Y = u(X_1, X_2, \ldots, X_n)$ тако да за реализовани узорак $(x_1, x_2, \ldots, x_n)$, број $y = u(x_1, x_2, \ldots, x_n)$ буде "добра" оцена за θ.

Обично, али не увек, скуп вредности оцене Y се поклапа са Θ. Ваљаност оцене утврђује се на основу одређених критеријума о којима ће надаље бити речи.

Дефиниција 14. Статистика $Y = u(X_1, X_2, \ldots, X_n)$ на основу узорка $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ из популације са обележјем X, чија расподела припада фамилији допустивих расподела $\{f(x;\theta), \theta \in \Theta\}$, је *непристрасна* или *центрирана оцена параметра* θ, ако је њено математичко очекивање једнако вредности параметра θ, тј.

$$E(Y) = \theta \,. \diamondsuit$$

Пример 22. (Оцена за математичко очекивање) Нека је дат случајни узорак

$$\mathbf{X} = (X_1, X_2, \ldots, X_n)$$

из популације са обележјем X чија расподела припада фамилији допустивих расподела

$$\{f(x;\theta), \theta \in \Theta\} \quad \text{за коју је} \quad E(X) = \theta \,.$$

Наћи непристрасну оцену параметра θ.

Пример такве фамилије је

$$\{\mathcal{N}(m, 1), m \in R\} \,.$$

Посматрајмо статистику

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Докажимо да је статистика $\overline{X}_n$ непристрасна оцена параметра θ.

На основу особина математичког очекивања следи

$$E(\overline{X}_n) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n} E\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} EX_i = \frac{1}{n}\sum_{i=1}^{n} EX = \\ = \frac{1}{n} n\theta = \theta.$$

$\triangle$

Приметимо да је средина узорка непристрасна оцена математичког очекивања произвољне расподеле која има математичко очекивање.

Пристрасне оцене се описују помоћу помераја $b(\theta) = E(Y) - \theta$. Очигледно, непристрасне оцене су оне за које је $b(\theta) = 0$.

Пример 23. (Оцена за дисперзију) Под условом да обележје X има дисперзију и да је $\theta = D(X)$ испитајмо да ли је узорачка дисперзија $\overline{S}_n^2$, добијена на основу простог случајног узорка, непристрасна оцена за θ.

Како је узорак прост,

$E(\overline{S}_n^2) = \frac{1}{n}\sum_{i=1}^n E(X_i - \overline{X}_n)^2 = \frac{1}{n}\sum_{i=1}^n E(X_i^2 - 2X_i\overline{X}_n + \overline{X}_n^2) =$
$= \frac{1}{n}\sum_{i=1}^n (E(X_i^2) - 2\frac{1}{n}\sum_{j=1}^n E(X_iX_j) + E(\overline{X}_n^2)) =$
$= \frac{1}{n}\sum_{i=1}^n (E(X^2) - 2\frac{n-1}{n}(EX)^2 - 2\frac{1}{n}E(X^2) + E(\frac{1}{n^2}\sum_{i,j=1}^n X_iX_j)) =$
$= \frac{1}{n}\sum_{i=1}^n \left(\frac{n-2}{n}E(X^2) - 2\frac{n-1}{n}(EX)^2 + \frac{1}{n^2}nE(X^2) + \frac{n^2-n}{n^2}(EX)^2\right) =$
$= \frac{1}{n}n\frac{n-1}{n}(EX^2 - (EX)^2) = \frac{n-1}{n}D(X)\,.\ \triangle$

Дефиниција 15. Статистика $Y = u(X_1, X_2, \ldots, X_n)$ на основу узорка $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ из популације са обележјем X, чија расподела припада фамилији допустивих расподела $\{f(x;\theta), \theta \in \Theta\}$, је *асимптотски непристрасна* или *асимптотски центрирана оцена* параметра θ, ако

$$E(Y) \to \theta \quad , \quad n \to \infty\,.\ \diamondsuit$$

Пример 24. Како је

$$E(\overline{S}_n^2) = \frac{n-1}{n}D(X)\,,$$

закључујемо да је дисперзија узорка асимптотски непристрасна оцена дисперзије обележја (уколико је узорак прост).

Асимптотски непристрасна оцена се може поправити до непристрасне оцене. Тако, када је реч о дисперзији узорка, на основу ње се може дефинисати поправљена дисперзија узорка:

$$\widetilde{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2 = \frac{n}{n-1}\overline{S}_n^2.$$

У узорцима већег обима **пристрасност** оцене $\overline{S}_n^2$ је занемарљива (на пример, ако се ради са тачношћу 10^{-2} поменута пристрасност се не осећа за узорак чији је обим већи од 200). Међутим, када је обим узорка мали, као оцена за дисперзију обележја X узима се **поправљена дисперзија** узорка. $\triangle$

Једна од мера блискости оцене и праве вредности параметра је средњеквадратно одступање статистике (оцене) Y од праве вредности параметра:

$$E(Y-\theta)^2 = E(Y - E(Y) + E(Y) - \theta)^2 = D(Y) + b^2(\theta)$$

Дакле, у случају да је оцена Y непристрасна, средњеквадратно одступање те оцене од праве вредности параметра је управо дисперзија статистике Y.

Дефиниција 16. Статистика $Y = u(X_1, X_2, \ldots, X_n)$ је *најбоља оцена параметра* θ ако је непристрасна и ако је дисперзија $D(Y)$ мања или једнака од дисперзије било које друге непристрасне оцене за θ. $\diamondsuit$

И поред погодности које пружа класа непристрасних статистика, овакву класу не треба идеализовати. Некада је захтев да је оцена непристрасна престрог, а некада непристрасна оцена и не постоји.

Није увек неопходно оценити сам параметар θ расподеле $f(x;\theta)$, већ треба оценити неку функцију од параметра $\tau(\theta)$, $\theta \in \Theta$. Тада се од оцене $T = T(X_1, X_2, \ldots, X_n)$ за функцију τ тражи да буде непристрасна, тј. да је $E(T) = \tau(\theta)$, што није тривијална последица постојања непристрасне оцене за θ.

Пример 25. Нека обележје X има Пуасонову расподелу $\mathcal{P}(\theta)$. Оценити функцију $\tau(\theta) = 1/\theta$ на основу узорка обима $n = 1$. Лако је уочити да је средина узорка непристрасна оцена за θ. Међутим, ако тражимо непристрасну оцену функције τ, $T = T(X_1)$, она мора да задовољи услов $E(T) = 1/\theta$. Дакле,

$$E(T) = \sum_{x=0}^{\infty} T(x) e^{-\theta} \frac{\theta^x}{x!} = \frac{1}{\theta}\,,$$

односно, важила би једнакост

$$\sum_{x=0}^{\infty} T(x) \frac{\theta^{x+1}}{x!} = e^{\theta}\,,$$

Примењујући Тејлоров развој функције e^θ, добија се

$$\theta \sum_{x=0}^{\infty} T(x)\frac{\theta^x}{x!} = \sum_{t=0}^{\infty} \frac{\theta^t}{t!}$$

што значи да функција T зависи од θ. Међутим, ово противуречи захтеву да T буде статистика. Дакле, у овом случају не постоји непристрасна оцена функције τ на основу задатог узорка. $\triangle$

Друга могућа мера одступања оцене од праве вредности параметра је *апсолутно одступање*, $|Y-\theta|$, које представља једну случајну промењљиву, те се као мера ваљаности оцене може узети вероватноћа постизања горње границе апсолутне грешке,

$$P\{|Y-\theta|<\varepsilon\} = p.$$

Тада се за унапред задато довољно мало $\varepsilon > 0$, одређује p или обрнуто, на основу познавања броја p одређује се ε. Дакле, ако је позната расподела вероватноћа за Y, проблем је, са теоријске тачке гледишта, решив. Међутим, и када нам ова расподела није позната, могуће је под одређеним условима одредити ε за задато p. Ако је, рецимо, ε облика

$$\varepsilon = k\sqrt{D(Y)}, \quad \text{за неко} \quad k \geq 1$$

и Y непристрасна оцена параметра θ, према Чебишевљевој неједнакости је

$$1 - P\left\{|Y-\theta| < k\sqrt{D(Y)}\right\} = P\left\{|Y-\theta| \geq k\sqrt{D(Y)}\right\} \leq \frac{D(Y)}{k^2 D(Y)} = \frac{1}{k^2}\,,$$

тј.

$$p \geq 1 - \frac{1}{k^2}\,.$$

Тако, за $k = 2$ је $p \geq 0,75$. Слично се може одредити k за задато p. Овај се резултат може користити и за одређивање обима узорка уколико се задају вредности за ε и p, о чему ће надаље бити речи.

У вези са апсолутним одступањем може се исказати следећи критеријум ваљаности тачкастих оцена.

Дефиниција 17. Статистика $Y = u(X_1, X_2, \ldots, X_n)$ је *постојана оцена* за θ ако она конвергира у вероватноћи ка θ, тј.

$$Y \xrightarrow{P} \theta\,, \quad n \to \infty\,.$$

Ако је ова конвергенција скоро извесно (скоро сигурно),

$$Y \xrightarrow{s.i.} \theta\,, \quad n \to \infty\,,$$

онда је оцена *строго постојана.* $\diamondsuit$

3.2 Одређивање обима узорка

Повећање обима узорка је поступак на који се у статистици по правилу рачуна. Међутим, практично, кадгод обим узорка није случајан, статистички поступак се спроводи над тачно одређеним, коначним обимом узорка n. Отуда да би примена појединог статистичког поступка дала задовољавајуће резултате, неопходно је приликом планирања експеримента предвидети и обим узорка, n.

Планирање експеримента је по својој суштини поступак којим се планира добијање одређене количине информација. "Производ" који се експериментом ствара јесте информација. Циљ је добити што већу количину информација правећи при томе што мање трошкове, дакле добити што квалитетнији производ по што нижој цени. План експеримента директно утиче на количину добијених информација при сваком мерењу. У том светлу биће говора о избору обима узорка n. Један од првих корака које истраживач мора да начини у свом истраживању је одређивање обима узорка. Обим узорка директно утиче на прецизност оцене, односно на ваљаност закључака добијених применом статистичких метода. Прецизност оцене се може исказати величином грешке добијене оцене. При томе примењени метод за процену ваљаности оцене, односно величину грешке, јесте у вези са обимом узорка.

Поступак за одређивање броја n зависи од параметра који се оцењује, статистике којом се оцењује и од тога да ли су други релевантни параметри обележја познати или се такође оцењују на основу узорка (уколико је реч о оцењивању параметара). Имајући све то у виду, дефинише се максимална величина грешке са којом треба радити у оцењивању.

Поменућемо само неке критеријуме који се користе за одређивање обима узорка у зависности од задате тачности оцењивања:

- средњеквадратно одступање, $E(\widehat{\theta}-\theta)^2$, и стандардно одступање, $\sqrt{E(\widehat{\theta}-\theta)^2}$

- апсолутна грешка, мања од ε, оцене, вероватноће (поверења) $1-\alpha$, тј. $P\left\{|\widehat{\theta}-\theta|<\varepsilon\right\}=1-\alpha$

- коефицијент варијације оцене, $C_v(\widehat{\theta})=\dfrac{\sqrt{E(\widehat{\theta}-\theta)^2}}{E(\widehat{\theta})}$

- релативна грешка оцене, $\delta=\dfrac{|\widehat{\theta}-\theta|}{\theta}$

Обим узорка се планира тако да се постигне задата тачност оцене.

Пример 26. Из узорка $(X_1, X_2, \ldots, X_n)$ треба постићи тачност оцене непознатог параметра θ са апсолутном грешком не већом од ε поверења $1-\alpha$.

Дакле,

$$P\{|\theta-\widehat{\theta}|<\varepsilon\}=1-\alpha$$

$$P\{-\varepsilon<\theta-\widehat{\theta}<\varepsilon\}=1-\alpha$$

$$P\{\widehat{\theta}-\varepsilon<\theta<\widehat{\theta}+\varepsilon\}=1-\alpha\,.$$

Како је

$$\widehat{\theta}=\widehat{\theta}(X_1,\ldots,X_n)=\widehat{\theta}(n)\,,$$

то ће се избором ε одредити најмањи n за који се постиже ниво поверења $1-\alpha$. $\triangle$

Очигледно, да бисмо одредили n, треба да знамо расподелу статистике $\widehat{\theta}$. Овај проблем се решава конкретно, мада најчешће применом централне граничне теореме. При томе може да наступи нови проблем због непознавања математичког очекивања или дисперзије обележја. Тада се ови параметри морају оценити на основу претходног узорка мањег обима, или на основу непотпуне информације о овим карактеристикама обележја, као што је, рецимо информација $D(X)\le\sigma_0^2$.

Пример 27. Нека се оцењује математичко очекивање, m, обележја X бесконачне популације. Непристрасна оцена овог параметра је $\overline{X}_n$(средина узорка). Дакле,

$$P\{|m-\overline{X}_n|<\varepsilon\}=1-\alpha \quad , \quad m=EX=E\overline{X}_n$$

$$P\left\{\frac{|m-\overline{X}_n|}{\frac{\sigma}{\sqrt{n}}}<\frac{\varepsilon\sqrt{n}}{\sigma}\right\}=1-\alpha \quad , \quad \sigma=\sqrt{D(X)}\,.$$

Према централној граничној теореми случајна променљива $\frac{\overline{X}_n-m}{\sigma}\sqrt{n}$ има $\mathcal{N}(0,1)$ расподелу када $n\to\infty$, па се у горњој једнакости вероватноћа може одредити из

$$2\Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right)=1-\alpha\,,$$

одакле следи да је

$$\frac{\varepsilon\sqrt{n}}{\sigma}=z_{\frac{1-\alpha}{2}}\,,\quad \text{односно}\,,\quad n\geq\left(\frac{\sigma z_{\frac{1-\alpha}{2}}}{\varepsilon}\right)^2$$

гарантује тражену тачност.

За задато α вредност $z_{\frac{1-\alpha}{2}}$ се може прочитати из таблице, те се и одређује на приказани начин. Уколико је σ непознато, али се зна да је $\sigma\leq\sigma_0$, ова се информација може употребити тако што ће се на основу

$$n\geq\left(\frac{\sigma z_{\frac{1-\alpha}{2}}}{\varepsilon}\right)^2$$

изабрати природан број n_0 за 1 већи од највећег целог дела израза на десној страни, тј.

$$n_0=\left[\left(\frac{\sigma_0 z_{\frac{1-\alpha}{2}}}{\varepsilon}\right)^2\right]+1\,. \tag{3.1}$$

Исти начин закључивања би се применио и код одређивања обима узорка са враћањем из коначне популације. Разуме се, када је $n\ll N$. $\triangle$

Пример 28. Уколико је у питању узорак без враћања из коначне популације биће

$$D(\overline{X}_n)=\frac{D(X)}{n}\left(1-\frac{n}{N}\right)$$

(видети Главу 6), па за оцењивање математичког очекивања по истом критеријуму као у претходном примеру имамо

$$P\left\{\frac{|m-\overline{X}_n|}{\sqrt{\frac{\sigma^2}{n}\left(1-\frac{n}{N}\right)}}<\frac{\varepsilon}{\sqrt{\frac{\sigma^2}{n}\left(1-\frac{n}{N}\right)}}\right\}=1-\alpha$$

$$\frac{\varepsilon\sqrt{nN}}{\sigma\sqrt{N-n}} = z_{\frac{1-\alpha}{2}}\,,$$

односно,

$$n \geq \frac{1}{\left(\frac{\varepsilon}{\sigma z_{\frac{1-\alpha}{2}}}\right)^2 + \frac{1}{N}}$$

гарантује тражену тачност.

Поново, ако не познајемо σ већ знамо само да је $\sigma \leq \sigma_0$, имаћемо да обим узорка за захтевану тачност оцене треба да задовољи услов

$$n \geq \frac{1}{\left(\frac{\varepsilon}{\sigma_0 z_{\frac{1-\alpha}{2}}}\right)^2 + \frac{1}{N}}\,.$$

Означимо са n_1 најмањи природни број који задовољава последњу неједнакост. Дакле,

$$n_1 = \left[\frac{1}{\left(\frac{\varepsilon}{\sigma_0 z_{\frac{1-\alpha}{2}}}\right)^2 + \frac{1}{N}}\right] + 1\,. \tag{3.2}$$

Уколико упоредимо обим узорка n_1 са обимом n_0 из претходног примера можемо констатовати да се код узорка без враћања захтева мањи обим узорка за постизање исте тачности код тачкастог оцењивања очекиване вредности обележја, него код узорка са враћањем. Заиста, из (3.1)

$$n_0 - 1 + t = \left(\frac{\sigma_0 z_{\frac{1-\alpha}{2}}}{\varepsilon}\right)^2, \quad \text{где је}, \quad 0 \leq t < 1\,,$$

тј.

$$\left(\frac{\varepsilon}{\sigma_0 z_{\frac{1-\alpha}{2}}}\right)^2 = \frac{1}{n_0 - 1 + t}\,,$$

па се заменом у (3.2) добија релација

$$n_1 = \left[\frac{N(n_0 - 1 + t)}{N + n_0 - 1 + t}\right] + 1\,.$$

Како је

$$\frac{1}{n_0 - 1 + t} > \frac{1}{n_0}$$

то је

$$\frac{1}{n_0 - 1 + t} + \frac{1}{N} > \frac{1}{n_0},$$

односно,

$$\left[\frac{1}{\frac{1}{n_0-1+t} + \frac{1}{N}}\right] + 1 \leq n_0 .$$

Другим речима,

$$n_1 \leq n_0 ,$$

што је и требало доказати. $\triangle$

3.3 Довољне статистике

Још један од начина да се говори о ваљаности тачкасте оцене за непознати параметар расподеле обележја је критеријум довољности.

Нека је дата статистика $Y_1 = u_1(X_1, \ldots, X_n)$ за оцену непознатог параметра $\theta \in \Theta \subset R$. Посматрајмо истовремено још $(n-1)$–ну статистику истог узорка:

$$\begin{aligned} Y_2 &= u_2(X_1, \ldots, X_n) \\ &\vdots \\ Y_n &= u_n(X_1, \ldots, X_n) \end{aligned}$$

и то таквих да је трансформација

$$\begin{aligned} y_1 &= u_1(x_1, \ldots, x_n) \\ y_2 &= u_2(x_1, \ldots, x_n) \\ &\vdots \\ y_n &= u_n(x_1, \ldots, x_n) \end{aligned} \tag{3.3}$$

”1−1”. Заједничка густина расподеле вектора статистика $(Y_1, Y_2, \ldots, Y_n)$ је:

$$g(y_1, y_2, \ldots, y_n; \theta) = |J| f(w_1(y_1, \ldots, y_n), \cdots, w_n(y_1, \ldots, y_n); \theta), \tag{3.4}$$

где је са f означена заједничка густина вектора $\mathbf{X}$, за обележје апсолутно непрекидног типа и без фактора $|J|$, односно Јакобијана, за обележје дискретног типа. Са $\mathbf{w} = (w_i, i = 1, 2, \ldots, n)$ је означена инверзна трансформација за трансформацију $\mathbf{u} = (u_i, i = 1, 2, \ldots, n)$.

Уколико трансформација (3.3) није "$1-1$" у целом R^n, онда је десна страна релације (3.4) сума од, рецимо, k израза тог облика. Број сабирака, k, одговара укупном броју области на које је извршено разбијање простора R^n и то таквих да је у свакој од њих трансформација (3.3) "$1-1$".

Уколико је узорак прост, заједничка густина вектора $\mathbf{X}$ би била

$$f(x_1, x_2, \ldots, x_n; \theta) = f(x_1; \theta) f(x_2; \theta) \cdots f(x_n; \theta)$$

што би произвело и адекватне промене у релацији (3.4).

Условна густина за $(Y_2, Y_3, \ldots, Y_n)$ под условом $Y_1 = y_1$ дата је са

$$h(y_2, \ldots, y_n | y_1; \theta) = \frac{g(y_1, y_2, \ldots, y_n; \theta)}{g_1(y_1; \theta)}, \quad \text{за} \quad g_1 > 0\,,$$

где је $g_1(y_1; \theta)$ маргинална густина за Y_1. У општем случају $h(y_2, \ldots, y_n | y_1; \theta)$ зависи од θ. Међутим, посебно важни случајеви су они у којима густина h не зависи од параметра θ.

Дефиниција 18. Нека је n фиксиран природни број и $(X_1, X_2, \ldots, X_n)$ узорак обима n из популације са обележјем чија расподела припада фамилији допустивих расподела $\{f(x; \theta), \theta \in \Theta\}$. Статистика $Y_1 = u_1(X_1, \ldots, X_n)$ је *довољна статистика* за θ ако и само ако за било које друге статистике $Y_2 = u_2(X_1, \ldots, X_n), \ldots, Y_n = u_n(X_1, \ldots, X_n)$, за које Јакобијан припадајуће трансформације није нула, условна густина расподеле $h(y_2, \ldots, y_n | y_1)$ случајних променљивих $Y_2, \ldots, Y_n$ под условом $Y_1 = y_1$, не зависи од параметра θ за било коју фиксирану вредност y_1. $\diamondsuit$

При томе се подразумева не само да аналитички израз условне густине не зависи од θ, већ и њена област дефинисаности такође. Упозоравамо на недозвољену зависност следећим примером.

Пример 29. Функција

$$f(x) = \begin{cases} \frac{1}{2}, & \theta - 1 < x < \theta + 1, \\ 0, & \text{ван} \end{cases}$$

где $\theta \in (-\infty, +\infty)$, зависи од θ. $\triangle$

Пример 30. Нека је (X_1, X_2) прост случајни узорак обима 2 из расподеле $\{\mathcal{B}(2,\theta),\ 0<\theta<1\}$. Испитајмо да ли $Y_1 = X_1 + X_2$ задовољава тражени услов, тј. да ли је довољна статистика за параметар θ ове биномне расподеле.

У ту сврху посматрамо

$$
\begin{aligned}
&f(x_1;\theta)f(x_2;\theta) = \\
&= \begin{cases} \binom{2}{x_1}\theta^{x_1}(1-\theta)^{2-x_1}\binom{2}{x_2}\theta^{x_2}(1-\theta)^{2-x_2} &, \ (x_1,x_2)\in\{(0,0),\ldots,(2,2)\} \\ 0 &, \ \text{иначе} \end{cases} \\
&= \begin{cases} \frac{2!2!\theta^{x_1+x_2}(1-\theta)^{4-(x_1+x_2)}}{x_1!(2-x_1)!x_2!(2-x_2)!} &, \ (x_1,x_2)\in\{(0,0),\ldots,(2,2)\} \\ 0 &, \ \text{иначе} \end{cases} .
\end{aligned}
$$

Одавде добијамо да је:

$$
\begin{aligned}
g_1(y_1) &= \begin{cases} \binom{4}{y_1}\theta^{y_1}(1-\theta)^{4-y_1} &, \ y_1\in\{0,1,2,3,4\} \\ 0 &, \ \text{иначе} \end{cases} \\
&= \begin{cases} \frac{4!\theta^{y_1}(1-\theta)^{4-y_1}}{y_1!(4-y_1)!} &, \ y_1\in\{0,1,2,3,4\} \\ 0 &, \ \text{иначе} \end{cases} .
\end{aligned}
$$

Узмимо другу статистику $Y_2 = X_2$ са расподелом $\mathcal{B}(2,\theta)$, па ће заједничка густина статистика Y_1 и Y_2 бити:

$$
g(y_1,y_2;\theta) = \begin{cases} \frac{4\theta^{y_1}(1-\theta)^{4-y_1}}{(y_1-y_2)!(2-y_1+y_2)!(2-y_2)!y_2!} &, \ (y_1,y_2)\in\{(0,0),\ldots,(4,2)\} \\ 0 &, \ \text{иначе} \end{cases} .
$$

Условна густина ће бити:

$$
\begin{aligned}
h(y_2|y_1;\theta) &= \begin{cases} \frac{\frac{4\theta^{y_1}(1-\theta)^{4-y_1}}{(y_1-y_2)!(2-y_1+y_2)!(2-y_2)!y_2!}}{\frac{4!\theta^{y_1}(1-\theta)^{4-y_1}}{y_1!(4-y_1)!}} &, \ (y_1,y_2)\in\{(0,0),\ldots,(4,2)\} \\ 0 &, \ \text{иначе} \end{cases} \\
&= \begin{cases} \frac{y_1!(4-y_1)!}{3!(y_1-y_2)!(2-y_1+y_2)!(2-y_2)!y_2!}, & (y_1,y_2)\in\{(0,0),\ldots,(4,2)\} \\ 0, & \text{иначе} \end{cases} .
\end{aligned}
$$

Дакле, условна густина за Y_2 под условом $Y_1 = y_1$ не зависи од θ, чиме смо доказали да је статистика Y_1 довољна. $\triangle$

3.3.1 Критеријуми егзистенције довољне статистике

Провера довољности неке статисртике директно по дефиницији се ретко користи због сложеног експлицитног одређивања густина расподеле које у дефиницији учествују. Једноставнији начин за проверу довољности дају наредне теореме.

Теорема 6. (Фишер-Нејманов критеријум) *Нека је* $(X_1, X_2, \ldots, X_n)$ *случајни узорак из популације са обележјем* X *са густином расподеле која припада фамилији допустивих расподела* $\{f(x;\theta),\, \theta \in \Theta\}$*. Нека је* $Y_1 = u_1(X_1, \ldots, X_n)$ *статистика чија је густина расподеле* $g_1(y_1;\theta)$*. Тада је* Y_1 *довољна статистика за* θ *ако и само ако*

$$f(x_1, x_2, \ldots x_n; \theta) = g_1(u_1(x_1, x_2, \ldots, x_n); \theta) H(x_1, x_2, \ldots, x_n) \tag{3.5}$$

где за сваку вредност функције u_1*, функција* H *не зависи од* θ*.*

Доказ. Доказ ћемо извести само за обележје апсолутно непрекидног типа.

Најпре уведимо следећу бијективну трансформацију $\mathbf{u} : R^n \to R^n$ и њој инверзну $\mathbf{w} : R^n \to R^n$, $\mathbf{u} = (u_1, \ldots, u_n)$, $\mathbf{w} = (w_1, \ldots, w_n)$:

$$\begin{array}{rclc} y_1 & = & u_1(x_1, \ldots, x_n) & x_1 = w_1(y_1, \ldots, y_n) \\ y_2 & = & u_2(x_1, \ldots, x_n) & x_2 = w_2(y_1, \ldots, y_n) \\ & \vdots & & \\ y_n & = & u_n(x_1, \ldots, x_n) & x_n = w_n(y_1, \ldots, y_n) \end{array} \tag{3.6}$$

такве да је Јакобијан

$$J = \left| \frac{\partial x_i}{\partial y_j} \right|_{i,j=1,\ldots,n} \neq 0.$$

Претпоставимо да важи услов (3.5) и докажимо да је Y_1 довољна статистика.

Одговарајућа заједничка густина расподеле статистика $Y_1, Y_2, \ldots, Y_n$ биће:

$$\begin{array}{rcl} g(y_1, y_2, \ldots, y_n; \theta) & = & f(w_1(y_1, \ldots, y_n), \ldots, w_n(y_1, \ldots, y_n); \theta)|J| \\ & = & g_1(y_1;\theta) H(w_1(y_1, \ldots, y_n), \ldots, w_n(y_1, \ldots, y_n))|J|\,. \end{array}$$

С друге стране, како је условна густина дата са:

$$\begin{array}{rcl} h(y_2, \ldots, y_n | y_1) & = & \dfrac{g(y_1, \ldots, y_n; \theta)}{g_1(y_1;\theta)} = \\ & = & \dfrac{g_1(y_1;\theta) H(w_1(y_1, \ldots, y_n), \ldots, w_n(y_1, \ldots, y_n))|J|}{g_1(y_1;\theta)} = \\ & = & H(w_1(y_1, \ldots, y_n), \ldots, w_n(y_1, \ldots, y_n))|J| \end{array}$$

и очигледно не зависи од параметра θ, следи да је Y_1 довољна статистика.

Докажимо и други смер исказа теореме.

Претпоставимо сада да је Y_1 довољна статистика и докажимо да важи услов (3.5). У том случају условна густина не зависи од параметра θ:

$$h(y_2,\ldots,y_n|y_1) = \frac{g(y_1,\ldots,y_n;\theta)}{g_1(y_1;\theta)}\,,$$

тј.

$$g(y_1,\ldots,y_n;\theta) = g_1(y_1;\theta)h(y_2,\ldots,y_n|y_1)\,.$$

Користећи инверзну трансформацију $\mathbf{w}$ чији је Јакобијан $J^* = \left|\frac{\partial y_i}{\partial x_j}\right|_{i,j=1,\ldots,n} \neq 0$, добијамо да је заједничка густина вектора $\mathbf{X}$

$$\begin{aligned} &f(x_1,\ldots,x_n;\theta) = g(u_1(x_1,\ldots,x_n),\ldots,u_n(x_1,\ldots,x_n);\theta)|J^*| = \\ &= g_1(u_1(x_1,\ldots,x_n;\theta))h(u_2(x_1,\ldots,x_n),\ldots,u_n(x_1,\ldots,x_n)|u_1(x_1,\ldots,x_n))|J^*|\,. \end{aligned}$$

Очигледно за $H(x_1,\ldots,x_n)$ се може узети $h(u_2(x_1,\ldots,x_n),\ldots,u_n(x_1,\ldots,x_n)|u_1(x_1,\ldots,x_n))|J^*|$, те је и други смер исказа теореме доказан.

За дискретан случај доказ се разликује само у томе што изостаје Јакобијан. $\square$

Дакле, Фишер-Нејманов критеријум нам показује да чињеница да је нека статистика довољна, није условљена избором преосталих $n-1$ статистика (под условом да је Јакобијан коришћене трансформације различит од нуле).

Пример 31. Нека су $Y_1,\ldots,Y_n$ статистике поретка на основу простог случајног узорка $\mathbf{X} = (X_1,\ldots,X_n)$, $Y_i = X_{(i)}\,, i = 1,\ldots,n$ из расподеле чија је густина

$$f(x;\theta) = \begin{cases} e^{-(x-\theta)} & , \quad \theta < x < \infty \quad , -\infty < \theta < \infty \\ \qquad 0 & , \quad \text{иначе}\,. \end{cases}$$

Густина расподеле статистике Y_1 је

$$g_1(y_1;\theta) = \begin{cases} ne^{-n(y_1-\theta)} & , \quad \theta < y_1 < \infty \\ \qquad 0 & , \quad \text{иначе}\,. \end{cases}$$

Заједничка густина расподеле вектора $\mathbf{X}$ је

$$e^{-(x_1-\theta)}e^{-(x_2-\theta)}\cdots e^{-(x_n-\theta)} = g_1(\min_i x_i;\theta)\left\{\frac{exp(-x_1-\cdots-x_n)}{n\cdot exp(-n\cdot\min_i x_i)}\right\}\,.$$

Дакле, користећи се Фишер–Нејмановим критеријумом, закључујемо да је статистика поретка реда један довољна статистика за оцену параметра θ посматране расподеле. $\triangle$

Фишер–Нејманов критеријум захтева познавање густине расподеле за Y_1, што може да отежа или чак онемогући његову практичну примену. Наредни критеријум избегава неопходност познавања ове густине.

Теорема 7. (Теорема факторизације) *Нека је $(X_1, X_2, \ldots, X_n)$ узорак из популације са обележјем чија густина расподеле припада фамилији допустивих расподела $\{f(x;\theta), \theta \in \Theta\}$. Статистика $Y_1 = u_1(X_1, \ldots, X_n)$ је довољна статистика за θ ако и само ако постоје две ненегативне функције k и K, такве да је*

$$f(x_1, x_2, \cdots, x_n; \theta) = k(u_1(x_1, x_2, \ldots, x_n); \theta) K(x_1, x_2, \ldots, x_n)\,, \qquad (3.7)$$

где за сваку конкретну вредност $u_1(x_1, \ldots, x_n)$ функција $K(x_1, \ldots, x_n)$ не зависи од θ.

Доказ. Доказ изводимо само за апсолутно непрекидни случај, док је за дискретни аналоган.

Претпоставимо да важи факторизација (3.7). Дефинишимо исту трансформацију као у доказу Фишер-Нејмановог критеријума. Заједничка густина расподеле статистика $Y_1, \ldots, Y_n$ је

$$\begin{aligned} g(y_1, \ldots, y_n; \theta) &= f(w_1(y_1, \ldots, y_n), \ldots, w_n(y_1, \ldots, y_n); \theta)|J| \\ &= k(y_1; \theta) K(w_1(y_1, \ldots, y_n), \ldots, w_n(y_1, \ldots, y_n))|J|\,. \end{aligned}$$

Сада треба доказати да условна густина не зависи од параметра θ или да важи теорема Фишер–Нејмана. У ту сврху одредимо маргиналну густину за Y_1.

$$g_1(y_1; \theta) = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} g(y_1, y_2, \ldots, y_n; \theta) dy_2 \cdots dy_n =$$
$$= \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} k(y_1; \theta) K(w_1(y_1, \ldots, y_n), \ldots, w_n(y_1, \ldots, y_n))|J| dy_2 \ldots dy_n =$$
$$= k(y_1; \theta) m(y_1)$$

где је

$$m(y_1) = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} K(w_1(y_1, \ldots, y_n), \ldots, w_n(y_1, \ldots, y_n))|J| dy_2 \ldots dy_n$$

За $m(y_1) = 0$ биће $g_1(y_1;\theta) = 0$. Међутим, за $m(y_1) \neq 0$,

$$f(x_1,\ldots,x_n;\theta) = g_1(u_1(x_1,\ldots,x_n);\theta)\frac{K(x_1,\ldots,x_n)}{m(u_1(x_1,\ldots,x_n))}\,.$$

Према Фишер–Нејмановом критеријуму, статистика Y_1 је довољна статистика за параметар θ.

Обрнуто, претпоставимо да је Y_1 довољна статистика. Тада се за k из факторизације (3.7) може узети функција $g_1(u_1(x_1,\ldots,x_n);\theta)$, а за K функција $h(u_2(x_1,\ldots,x_n),\ldots,u_n(x_1,\ldots,x_n)|u_1(x_1,\ldots,x_n))$ из дефиниције довољне статистике, чиме се доказује да факторизација важи. $\square$

Теорема 8. *Ако је $Y_1 = u_1(X_1,\ldots,X_n)$ довољна статистика за параметар θ расподеле обележја X из које је узет узорак, тада произвољна бијективна функција $Z = u(Y_1)$ која не зависи од θ, тј. $Z = u(u_1(X_1,\ldots,X_n)) = v(X_1,\ldots,X_n)$, је такође довољна статистика за θ.*

Доказ. Према теореми факторизације је

$$f(x_1,\ldots,x_n;\theta) = k(u_1(x_1,\ldots,x_n);\theta)K(x_1,\ldots,x_n).$$

С обзиром на бијекцију $z = u(y_1)$, следи да постоји инверзна функција w таква да је $y_1 = w(z)$. С друге стране је $y_1 = u_1(x_1,\ldots,x_n)$, те је

$$u_1(x_1,\ldots,x_n) = w(z) = w(u(u_1(x_1,\ldots,x_n))) = w(v(x_1,\ldots,x_n))$$

која не зависи од θ, те је

$$\begin{aligned} f(x_1,\ldots,x_n;\theta) &= k(w(v(x_1,\ldots,x_n));\theta)K(x_1,\ldots,x_n) = \\ &= k(w(z);\theta)K(x_1,\ldots,x_n) = k_1(z;\theta)K(x_1,\ldots,x_n)\,,\ \text{где је } k_1 = k\circ w. \end{aligned}$$

Како је k_1 функција (од узорка) само од z и од θ, а K не зависи од θ, то је, према теореми факторизације, Z такође довољна статистика за параметар θ.$\square$

3.3.2 Најбоља оцена за параметар

Довољне статистике играју важну улогу у одређивању добре оцене за параметар. Ако је $\widehat{\theta}$ непристрасна оцена за θ, а Y довољна статистика за исти параметар, тада је могуће наћи неку функцију од Y која ће такође бити непристрасна оцена за θ, а неће имати већу дисперзију од $\widehat{\theta}$. Теоријску подлогу за тај закључак даје следећа теорема коју наводимо без доказа.

Теорема 9. (Рао-Блекведова теорема) *Нека су X и Y случајне променљиве такве да Y има очекивање θ и коначну дисперзију $D(Y)$. Нека је $E(Y|X=x)=\varphi(x)$. Тада је $E(\varphi(X))=\theta$ и $D(\varphi(X))\leq D(Y)$.*

Последица 1. *Нека је $Y_1 = u_1(X_1,\ldots,X_n)$ довољна статистика за параметар θ. Нека је $Y_2 = u_2(X_1,\ldots,X_n)$ нека друга статистика (која није функција од узорка само посредством Y_1) која је непристрасна оцена параметра θ. Тада $E(Y_2|Y_1=y_1)=\varphi(y_1)$ дефинише статистику $\varphi(Y_1)$. Ова статистика, која је функција од довољне статистике за θ, је непристрасна оцена за θ и њена дисперзија је мања од дисперзије за Y_2.*

Доказ. Пошто је Y_1 довољна статистика за θ, условна густина расподеле за Y_2 под условом $Y_1 = y_1$ не зависи од θ, тако да је $E(Y_2|Y_1=y_1)=\varphi(y_1)$ функција само од y_1 (а не и од θ). Дакле, $\varphi(Y_1)$ је статистика. Према Рао-Блекведовој теореми,

$$E(\varphi(Y_1)) = E(Y_2) = \theta\,,$$

јер је Y_2 непристрасна оцена параметра θ и

$$D(\varphi(Y_1)) \leq D(Y_2)\,.$$

Но, како Y_2 није функција од узорка само преко Y_1, то је

$$D(\varphi(Y_1)) < D(Y_2)\,.\square$$

Ова последица нам указује на то да у трагању за најбољом оценом параметра θ, пажњу можемо да ограничимо на довољну статистику ако она постоји, јер полазећи од ње долазимо до непристрасне оцене за параметар чија је дисперзија мања од дисперзије било које друге непристрасне оцене.

3.3.3 Комплетност

Дефиниција комплетности припада теорији мера. Ми ћемо је овде користити да искажемо једно важно својство довољних статистика.

Дефиниција 19. Нека је дато обележје X чија расподела припада фамилији допустивих расподела

$$\{f(x;\theta), \theta \in \Theta\}. \tag{3.8}$$

Изаберимо произвољну Борелову функцију u где је $u(X)$ случајна променљива која не зависи од параметра θ и такву да постоји $E(u(X))$. Ако је $E(u(X)) = 0$ за свако $\theta \in \Theta$, само под условом да је $u(x) = 0$ у свакој тачки $x \in R$ у којој је бар један елемент фамилије (3.8) строго позитиван (може бити једнака нули само на скупу мере нула), тада је фамилија (3.8) *комплетна.*$\diamondsuit$

Надаље ћемо овде користити (због једноставности доказивања) само непрекидне функције u, а не Борелове уопште, како је исказано дефиницијом.

Пример 32. Нека је фамилија допустивих расподела $\{\mathcal{U}(0;\theta), \theta > 0\}$, тј.

$$f(x;\theta) = \begin{cases} \dfrac{1}{\theta}, & 0 < x < \theta \\ 0, & \text{иначе} \end{cases} .$$

Докажимо да је фамилија $\{f(x;\theta), \theta > 0\}$ комплетна.

Пођимо од претпоставке да је $E(u(X)) = 0$, где је u непрекидна функција. Утврдимо када је ова претпоставка могућа. Дакле,

$$E(u(X)) = \int_0^\theta u(x)\frac{1}{\theta}dx = \frac{1}{\theta}\int_0^\theta u(x)dx = 0\,,$$

а с друге стране последња једнакост је тачна ако и само ако је

$$\int_0^\theta u(x)dx = 0\,.$$

Ако је $\theta > 0$ и нађемо извод по горњој граници, тј. по θ, (што је читаоцу познато као извод параметарског интеграла) добићемо:

$$J'_\theta = \int_0^\theta u'_\theta(x)dx + \theta' \cdot u(\theta) - 0 = 0 \quad .$$

Из последње једнакости добијамо да је $u(\theta) = 0$ за свако $\theta > 0$, а одавде следи да је и $u(x) = 0$, за свако $x > 0$, па је фамилија комплетна. $\triangle$

Пример 33. Нека је у питању дискретна расподела и нека је дата фамилија $\{\mathcal{B}(1;\theta), 0 \leq \theta < 1\}$. Докажимо да је она комплетна. Дакле,

$$f(x;\theta) = \begin{cases} \theta^x(1-\theta)^{1-x}, & x = 0,1 \\ 0, & \text{иначе} \end{cases} .$$

Пођимо, као и у предходном примеру, од претпоставке да је $E(u(X)) = 0$ и да је функција u непрекидна. Тада је:

$$0 = u(0)\theta^0(1-\theta)^1 + u(1)\theta^1(1-\theta)^0 = u(0) - \theta u(0) + u(1)\theta = (u(1)-u(0))\theta + u(0).$$

Линеарна функција је идентички једнака нули ако и само ако су јој коефицијенти једнаки нули, тј.

$$u(0) = 0$$

$$u(1) - u(0) = 0\,,$$

а одавде добијамо да је

$$u(0) = u(1) = 0\,,$$

чиме смо доказали комплетност фамилије. $\triangle$

3.3.4 Јединственост најбоље оцене за параметар

Видели смо да, ако је $Y_1 = u_1(X_1,\ldots,X_n)$ била довољна статистика за параметар θ и ако је постојала било која непристрасна оцена Y_2 за параметар θ, која није била функција од узорка само преко Y_1, тада је постојала бар још једна функција од Y_1 различита од Y_1, која је била непристрасна оцена за θ. Тако се наше трагање за бољом статистиком (по критеријуму средње квадратног одступања) за θ може да сузи само на функције од Y_1.

Теорема 10. *Нека је $Y_1 = u_1(X_1,\ldots,X_n)$ довољна статистика за параметар θ расподеле обележја X посматраног на популацији из које је узет узорак, и нека је фамилија густина расподеле статистике Y_1*

$$\{g_1(y_1;\theta), \theta \in \Theta\} \tag{3.9}$$

комплетна. Ако постоји непрекидна функција φ од Y_1 (дефинисана у Последици 1) која је непристрасна оцена параметра θ, (тј. постоји статистика $\varphi(Y_1)$ таква да је $E(\varphi(Y_1)) = \theta$) тада је функција φ скоро сигурно јединствена најбоља оцена за параметар θ.

Доказ. Претпоставимо супротно тј. да сем функције $\varphi(Y_1)$ постоји још нека функција $\psi(Y_1)$, и нека су обе непрекидне функције које не зависе од параметра θ, а непристрасне су оцене параметра θ. Тада је:

$$E(\varphi(Y_1) - \psi(Y_1)) = E(\varphi(Y_1)) - E(\psi(Y_1)) = \theta - \theta = 0.$$

Како је (3.9) комплетна фамилија, то значи да је функција $\varphi - \psi = 0$ за све вредности аргумента за које је бар један елемент фамилије (3.9) строго позитиван. Дакле,

$$\varphi(Y_1) - \psi(Y_1) = 0$$

скоро сигурно, тј.

$$\varphi(Y_1) = \psi(Y_1)$$

скоро сигурно. Да је функција φ и најбоља статистика закључује се према теореми Рао–Блеквела из чињенице да је $D(\varphi(Y_1)) \leq D(Y_2)$ за било коју другу непристрасну оцену Y_2 на основу истог узорка. $\square$

Ова теорема је само специјалан случај општије теореме Лемана и Шефеа.

Као што је у дефиницији комплетне фамилије наглашено да се она може исказати и за функције које нису непрекидне (у том случају се доказ изводи апаратом теорије мера), тако се и теорема о јединствености најбоље оцене за параметар која се односи на комплетну фамилију може исказати и без претпоставке о непрекидности, чиме се у оквиру овог курса нећемо бавити.

Исказ да је Y_1 довољна статистика за параметар θ, $\theta \in \Theta$, за коју је фамилија

$$\{g_1(y_1;\theta), \theta \in \Theta\}$$

функција густина расподеле комплетна, замењује се, ради једноставнијег изражавања, исказом *комплетна довољна статистика за* θ.

Пример 34. Нека је дата Пуасонова расподела $\{\mathcal{P}(\theta),\, 0 < \theta < \infty\}$ са густином

$$f(x;\theta) = \begin{cases} \dfrac{\theta^x e^{-\theta}}{x!}, & x \in \{0,1,2,\ldots\} \\ 0, & \text{иначе} \end{cases}$$

и прост случајни узорак $\mathbf{X}$ обима n из те расподеле. Доказати да је $Y_1 = \sum_{i=1}^{n} X_i$ комплетна довољна статистика за параметар θ.

Довољност статистике следи на основу

$$\begin{aligned}\prod_{i=1}^{n} f(x_i;\theta) &= \prod_{i=1}^{n} \frac{\theta^{x_i} e^{-\theta}}{x_i!} = \\ &= \exp\left(\ln\theta \cdot \sum_{i=1}^{n} x_i - n\theta\right) \cdot \frac{1}{\prod_{i=1}^{n} x_i!}\end{aligned}$$

и теореме факторизације.

Комплетност фамилије $\{g_1(y_1;\theta)\,,\, 0 < \theta < \infty\}$ следи из чињенице да је фамилија Пуасонових расподела комплетна, тј. чињенице да Y_1 има $\mathcal{P}(n\theta)$ расподелу (што је познато из Теорије вероватноће). Према томе Y_1 има густину

$$g_1(y_1;\theta) = \begin{cases} \dfrac{(n\theta)^{y_1} e^{-n\theta}}{y_1!}\,, & y_1 \in \{0,1,2,\ldots\} \\ 0\,, & \text{иначе} \end{cases}$$

Узмимо произвољну непрекидну функцију u за коју је $E(u(Y_1)) = 0$ и докажимо да је то могуће само ако је $u(y_1) = 0$ у тачкама $y_1 = 0, 1, 2, \ldots$.

Дакле, за свако $\theta > 0$

$$0 = E(u(Y_1)) = \sum_{y_1=0}^{\infty} u(y_1) \frac{(n\theta)^{y_1} e^{-n\theta}}{y_1!} =$$

$$= e^{-n\theta}\left(u(0) + u(1)\frac{n\theta}{1!} + u(2)\frac{(n\theta)^2}{2!} + \cdots\right).$$

Пошто $e^{-n\theta}$ није једнако нули, биће

$$0 = u(0) + nu(1)\theta + \left(\frac{n^2 u(2)}{2}\right)\theta^2 + \cdots.$$

Међутим, ако такав бесконачни (степени) ред конвергира ка нули за свако $\theta > 0$, тада сваки коефицијенат уз одговарајући степен променљиве мора бити 0. То јест

$$u(0) = 0\,, \quad nu(1) = 0\,, \quad \frac{n^2 u(2)}{2} = 0\,, \quad \ldots$$

и отуда

$$0 = u(0) = u(1) = u(2) = \cdots$$

што је и требало доказати.

Дакле, $E(Y_1) = n\theta$, па је тражена функција φ која даје јединствену најбољу оцену за параметар θ

$$\varphi(Y_1) = \frac{Y_1}{n} = \overline{X}_n \quad . \triangle$$

Наведимо још једну, из оперативних разлога важну, теорему:

Теорема 11. *Нека је $(X_1, \dots, X_n)$ узорак из расподеле $\{f(x;\theta), \theta \in \Theta\}$ где је Θ интервал. Нека је $Y_1 = u_1(X_1, \dots, X_n)$ довољна статистика за θ са комплетном фамилијом густина расподеле $\{g_1(y_1;\theta), \theta \in \Theta\}$. Нека је $Z = u(X_1, \dots, X_n)$ било која друга статистика (која није само функција од Y_1). Ако расподела за Z не зависи од θ, тада је Z независна случајна променљива у односу на случајну променљиву Y_1.*

Доказ. Доказ ћемо извести само у специјалном случају и то када је обележје X апсолутно непрекидног типа и када је условна густина расподеле статистике Z под условом $Y_1 = y_1$ непрекидна функција од y_1.

Нека су редом:

$$\begin{array}{ll} g_2(z) & \text{–маргинална густина за } Z, \\ h(z|Y_1 = y_1) & \text{–условна густина за } Z \text{ при услову } Y_1 = y_1, \\ g(y_1, z; \theta) & \text{–заједничка густина за } (Y_1, Z). \end{array}$$

Тада је интеграл:

$$\int_{-\infty}^{+\infty} h(z|y_1) g_1(y_1;\theta) dy_1 = \int_{-\infty}^{+\infty} \frac{g(y_1, z;\theta)}{g_1(y_1;\theta)} g_1(y_1;\theta) dy_1 = g_2(z)\,.$$

Одатле следи да је

$$\int_{-\infty}^{+\infty} (g_2(z) - h(z|y_1)) g_1(y_1;\theta) dy_1 = 0\,.$$

Како је Y_1 довољна статистика за параметар θ, то $h(z|y_1)$ не зависи од θ и, како је фамилија комплетна и функција $g_2(z)$ константа по y_1 и не зависи од θ, то је и разлика $g_2(z) - h(z|y_1)$ непрекидна функција од y_1 и не зависи од θ. Дакле, биће $g_2(z) - h(z|y_1) = 0$. Следи $g_2(z) = h(z|y_1)$. Односно, добили смо да је:

$$h(z|y_1) = \frac{g(y_1, z; \theta)}{g_1(y_1; \theta)}$$
$$g(y_1, z; \theta) = g_1(y_1; \theta) h(z|y_1)$$
$$g(y_1, z; \theta) = g_1(y_1; \theta) g_2(z)\,.$$

Дакле, заједничка густина је производ маргиналних густина, а самим тим су Y_1 и Z независне случајне променљиве. □

Пример 35. Сада смо у могућности да додокажемо да су средина и дисперзија простог случајног узорка из нормалне расподеле $\{\mathcal{N}(m, \sigma^2), m \in R, \sigma^2 \in R^+\}$ независне случајне променљиве. У Примеру 36 ћемо се уверити да је средина узорка $\overline{X}_n$ за свако дато $\sigma^2 > 0$ комплетна довољна статистика за m где је $-\infty < m < \infty$, а сада ћемо само користити ту чињеницу. Да бисмо доказали поменуту независност довољно је показати (према Теореми 11) да расподела за $\overline{S}_n^2$ не зависи од m. У ту сврху користићемо функцију генератрисе момената (види Главу 6) случајне променљиве $\overline{S}_n^2$:

$$M(t) = E(e^{t\overline{S}_n^2}) =$$
$$= \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \exp\left(\frac{t}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2\right)\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\sum_{i=1}^{n}\frac{(x_i - m)^2}{2\sigma^2}\right) dx_1 \ldots dx_n.$$

Функција $M(t)$ је дефинисана за $t < \frac{n}{2\sigma^2}$.

Написани интеграл је компликован за израчунавање и нећемо га израчунавати, а чињеницу да $M(t)$ не зависи од m добићемо на следећи начин. Уводимо $1-1$ трансформацију

$$\omega_1 = x_1 - m \quad , \quad \omega_2 = x_2 - m \quad , \ldots , \quad \omega_n = x_n - m\,.$$

Очигледно, њен Јакобијан је једнак 1. Користићемо ову трансформацију за увођење смене у интеграл којим је дефинисана посматрана функција генератриса момената, па су суме

$$\sum_{i=1}^{n}(x_i - m)^2 = \sum_{i=1}^{n}\omega_i^2\,, \quad \sum_{i=1}^{n}(x_i - \overline{x}_n)^2 = \sum_{i=1}^{n}(\omega_i - \overline{\omega})^2\,, \quad \overline{\omega} = \frac{1}{n}\sum_{j=1}^{n}\omega_j\,.$$

Дакле,

$$M(t) =$$
$$= \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \exp\left(\frac{t}{n}\sum_{i=1}^{n}(\omega_i - \overline{\omega})^2\right)\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\omega_i^2\right) d\omega_1 \ldots d\omega_n$$

што, очигледно, не зависи од m. $\triangle$

У случају регуларне експоненцијалне класе, о чему ће надаље бити речи, теорема је типа *ако и само ако.* Теорема важи и у случају вишедимензионог параметра.

Уопштење је могуће вршити и у правцу вишедимензионог обележја са вишедимензионим параметром.

3.3.5 Довољна статистика за вишедимензиони параметар

У многим случајевима је параметар који треба оценити вишедимензиони. Дакле, у општем случају ће бити $\boldsymbol{\Theta} \subset R^r$ где је r неки природан број. Међутим, врло често је $r = 2$. Отуда и појам довољности треба проширити тако да обухвати и вишедимензионе параметре. Проширују се и теореме о којима смо говорили у једнодимензионом случају.

Дефиниција 20. Нека је $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ случајни узорак из популације са обележјем X чија расподела припада фамилији допустивих расподела $\{f(x;\boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$. Нека је $\mathbf{Y}$ m–димензиони случајни вектор статистика $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_m)$, где је $Y_i = u_i(X_1, X_2, \ldots, X_n)$, $i = 1, \ldots, m$. Нека је $f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})$, $\mathbf{y} \in R^m$ густина расподеле вектора $\mathbf{Y}$. Тада је $\mathbf{Y}$ *довољна статистика* за $\boldsymbol{\theta}$ ако и само ако количник

$$\frac{f(x_1, x_2, \ldots, x_n;\boldsymbol{\theta})}{f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})} = H(x_1, x_2, \ldots, x_n) \tag{3.10}$$

не зависи од $\boldsymbol{\theta}$ за све вредности x_i обележја X.$\diamondsuit$

У случају обележја X апсолутно непрекидног типа, количник у (3.10) је баш условна густина.

Треба уочити да је у општем случају $m \neq r$, мада у већини случајева важи једнакост.

Проширење, на пример, теореме факторизације на вишедимензиони параметар, било би:

Вектор статистика Y је довољна статистика за параметар $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ ако и само ако се могу да нађу две ненегативне функције k и K такве да је

$$f(x_1, x_2, \ldots, x_n;\boldsymbol{\theta}) = k(\mathbf{y};\boldsymbol{\theta})K(x_1, x_2, \ldots, x_n)$$

где функција K не зависи од $\boldsymbol{\theta}$.

Концепт комплетности на пр., као и други појмови везани за довољне статистике за једнодимензиони параметар се такође проширује на довољне статистике за вишедимензиони параметар на одговарајући начин, али о томе овде неће бити речи.

Пример 36. Нека је дат прост узорак $\mathbf{X} = (X_1, \ldots, X_n)$ из популације са нормалном расподелом обележја, тј. са фамилијом допустивих расподела

$$\{\mathcal{N}(\theta_1, \theta_2)\,,\, -\infty < \theta_1 < +\infty\,,\, 0 < \theta_2 < \infty\}\,.$$

Густина расподеле обележја X је

$$f(x; \theta_1, \theta_2) = \exp\left(-\frac{1}{2\theta_2}x^2 + \frac{\theta_1}{\theta_2}x - \frac{\theta_1^2}{2\theta_2} - \ln\sqrt{2\pi\theta_2}\right)\,.$$

Посматрајмо статистике

$$Y_1 = \sum_{i=1}^{n} X_i \quad \text{и} \quad Y_2 = \sum_{i=1}^{n} X_i^2\,.$$

Лако је доказати да је (Y_1, Y_2) дводимензиона довољна комплетна статистика за параметар (θ_1, θ_2). Штавише, ако дефинишемо пресликавање $1-1$ на следећи начин:

$$Z_1 = \frac{Y_1}{n} = \overline{X}_n \quad , \quad Z_2 = \frac{Y_2 - \frac{Y_1^2}{n}}{n-1} = \frac{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}{n-1}\,,$$

онда ће (Z_1, Z_2) бити дводимензиона комплетна довољна статистика за (θ_1, θ_2) и једина таква да је непристрасна, тј.

$$E(Z_1) = \theta_1 \quad , \quad E(Z_2) = \theta_2\,.$$

Детаљније ћемо се овде позабавити само довољношћу статистике (Y_1, Y_2). Заиста, заједничка густина овог простог узорка је

$$\prod_{i=1}^{n} f(x_i; \theta_1, \theta_2) = \prod_{i=1}^{n} \frac{1}{(2\pi\theta_2)^{\frac{1}{2}}} \exp\left(-\frac{(x_i - \theta_1)^2}{2\theta_2}\right) =$$

$$= (2\pi)^{-\frac{n}{2}} \theta_2^{-\frac{n}{2}} \exp\left(-\frac{1}{2\theta_2}\left(\sum_{i=1}^{n} x_i^2 - 2\theta_1 \sum_{i=1}^{n} x_i + n\theta_1^2\right)\right)$$

где се може узети

$$k(y_1, y_2; \theta_1, \theta_2) = \exp\left(-\frac{1}{2\theta_2}\left(y_2 - 2\theta_1 y_1 + n\theta_1^2\right) - \frac{n}{2}\ln(2\pi\theta_2)\right)$$

и

$$K = 1$$

па према теореми факторизације следи тврђење. $\triangle$

3.4 Регуларна фамилија густина расподеле

У даљем излагању у вези са оцењивањем параметара биће од интереса да заједничку густину расподеле случајног узорка посматрамо као функцију од непозатог параметра, једнодимензионог или вишедимензионог, зависно од дефиниције проблема који ћемо решавати.

Дефиниција 21. Нека је дат узорак $\mathbf{X} = (X_1, \ldots, X_n)$ из популације са обележјем X чија расподела припада фамилији допустивих расподела $\{f(x;\theta),\, \theta \in \mathbf{\Theta}\}$. *Функција веродостојности* за параметар θ, у ознаци $L(\theta)$, је заједничка густина расподеле вектора $\mathbf{X}$ посматрана као функција од θ

$$L(\theta) = L(\theta; x_1, \ldots, x_n) = f(x_1, \ldots, x_n; \theta) \quad , \quad (x_1, \ldots, x_n) \in R^n \,.\, \diamondsuit$$

Уведимо сада дефиницију регуларне фамилије густина расподеле за једнодимензиони параметар:

Дефиниција 22. Нека је дата функција $u : R^n \longrightarrow R$ и фамилија допустивих расподела $\{f(x;\theta), \theta \in \Theta\}$ за обележје X неке популације из које се узима узорак обима n. Казаћемо да је *фамилија допустивих расподела регуларна*[1] ако за функцију u и фамилију допустивих расподела важе следећи *услови регуларности*:

- (R_1) Θ је интервал (или цео скуп R)
- (R_2) Скуп $K = \{\mathbf{x} = (x_1, \ldots, x_n) : L(\theta, \mathbf{x}) > 0\} \subset R^n$ не зависи од θ

[1] Услови регуларности се могу и другачије исказати или поставити.

- (R_3) Функција $L(\theta)$ је диференцијабилна по θ и важи[2]

$$\frac{\partial}{\partial\theta}\int_{\mathbf{R}^n} L(\theta,\mathbf{x})d\mathbf{x} = \int_{\mathbf{R}^n}\frac{\partial L(\theta,\mathbf{x})}{\partial\theta}d\mathbf{x}$$

$$\frac{\partial}{\partial\theta}\int_{\mathbf{R}^n} u(\mathbf{x})L(\theta,\mathbf{x})d\mathbf{x} = \int_{\mathbf{R}^n}u(\mathbf{x})\frac{\partial L(\theta,\mathbf{x})}{\partial\theta}d\mathbf{x}$$

и аналогно за дискретан случај

$$\frac{\partial}{\partial\theta}\sum_{\mathbf{x}\in\mathbf{K}} L(\theta,\mathbf{x}) = \sum_{\mathbf{x}\in\mathbf{K}}\frac{\partial L(\theta,\mathbf{x})}{\partial\theta}$$

$$\frac{\partial}{\partial\theta}\sum_{\mathbf{x}\in\mathbf{K}} u(\mathbf{x})L(\theta,\mathbf{x}) = \sum_{\mathbf{x}\in\mathbf{K}}u(\mathbf{x})\frac{\partial L(\theta,\mathbf{x})}{\partial\theta}\,.\diamondsuit$$

Теорема 12. (Рао-Крамерова доња граница) *Нека је дат узорак* $\mathbf{X}=(X_1,\ldots,X_n)$ *из популације са обележјем* X *чија расподела припада фамилији допустивих расподела* $\{f(x;\theta),\theta\in\Theta\},\Theta\subset R$, *која је регуларна. Нека је* Y *статистика посматраног узорка са коначним другим моментом и таква да је непристрасна оцена диференцијабилне функције* $\varphi(\theta)$ *параметра* $\theta\in\Theta$, $\varphi:\Theta\longrightarrow R$, *посматране фамилије. Тада важи*

$$D(Y)\geq\frac{(\varphi'(\theta))^2}{E\left(\frac{\partial\ln L(\theta)}{\partial\theta}\right)^2}$$

за свако $\theta\in\Theta$, *где је са* $L(\theta)$ *означена случајна променљива* $L(\theta)=L(\theta,X_1,\ldots,X_n)$.

Доказ. Доказ ћемо навести за апсолутно непрекидан случај, а дискретан случај се разматра аналогно.

Како је функција веродостојности функција густине расподеле вектора $\mathbf{X}$, то је

$$\int_{\mathbf{R}^n} L(\theta;\mathbf{x})d\mathbf{x}=1\,. \tag{3.11}$$

Нека је

$$Y=u(\mathbf{X})\,.$$

[2] У даљем тексту биће коришћено $\int_{\mathbf{R}^n} f(\mathbf{x})d\mathbf{x}=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}f(x_1,\ldots,x_n)dx_1\ldots dx_n\,.$

Тада, због апсолутне непрекидности и непристрасности статистике Y имамо

$$E(Y) = \int_{\mathbf{R}^n} u(\mathbf{x}) L(\theta; \mathbf{x}) d\mathbf{x} = \varphi(\theta) \quad . \tag{3.12}$$

Диференцирањем по θ једнакости (3.11) и (3.12) и на основу особине регуларности, добијамо:

$$\int_{\mathbf{R}^n} \frac{\partial L(\theta; \mathbf{x})}{\partial \theta} d\mathbf{x} = \frac{\partial}{\partial \theta} \int_{\mathbf{R}^n} L(\theta; \mathbf{x}) d\mathbf{x} = 0\,, \tag{3.13}$$

као и

$$\frac{\partial}{\partial \theta} \int_{\mathbf{R}^n} u(\mathbf{x}) L(\theta, \mathbf{x}) d\mathbf{x} = \frac{\partial \varphi(\theta)}{\partial \theta}\,,$$

тј.

$$\int_{\mathbf{R}^n} u(\mathbf{x}) \frac{\partial L(\theta; \mathbf{x})}{\partial \theta} d\mathbf{x} = \varphi'(\theta)\,. \tag{3.14}$$

Из (3.13) добијамо

$$\begin{aligned} 0 &= \int_{\mathbf{R}^n} \frac{\partial L(\theta; \mathbf{x})}{\partial \theta} d\mathbf{x} = \int_{\mathbf{R}^n} \left(\frac{1}{L(\theta; \mathbf{x})} \frac{\partial L(\theta; \mathbf{x})}{\partial \theta} \right) L(\theta; \mathbf{x}) d\mathbf{x} = \\ &= \int_{\mathbf{R}^n} \frac{\partial \ln L(\theta; \mathbf{x})}{\partial \theta} L(\theta; \mathbf{x}) d\mathbf{x} = E\left(\frac{\partial \ln L(\theta; \mathbf{X})}{\partial \theta} \right)\,, \end{aligned}$$

одакле

$$0 = 0 \cdot \varphi(\theta) = \int_{\mathbf{R}^n} \varphi(\theta) \frac{\partial \ln L(\theta; \mathbf{x})}{\partial \theta} L(\theta; \mathbf{x}) d\mathbf{x}\,. \tag{3.15}$$

Из (3.14) добијамо

$$\varphi'(\theta) = \int_{\mathbf{R}^n} u(\mathbf{x}) \frac{\partial L(\theta; \mathbf{x})}{\partial \theta} d\mathbf{x} = \int_{\mathbf{R}^n} u(\mathbf{x}) \frac{\partial \ln L(\theta; \mathbf{x})}{\partial \theta} L(\theta; \mathbf{x}) d\mathbf{x}\,. \tag{3.16}$$

Одузимањем (3.15) од (3.16) добијамо:

$$\varphi'(\theta) - 0 = \int_{\mathbf{R}^n} (u(\mathbf{x}) - \varphi(\theta)) \frac{\partial \ln L(\theta; \mathbf{x})}{\partial \theta} L(\theta; \mathbf{x}) d\mathbf{x}\,,$$

тј.

$$\varphi'(\theta) = \int_{\mathbf{R}^n} (u(\mathbf{x}) - \varphi(\theta)) \frac{\partial \ln L(\theta; \mathbf{x})}{\partial \theta} L(\theta; \mathbf{x}) d\mathbf{x}\,.$$

Користећи неједнакост Коши–Шварц–Буњаковског добијамо:

$$\begin{aligned}
|\varphi'(\theta)|^2 &= \left|\int_{\mathbf{R}^n}(u(\mathbf{x})-\varphi(\theta))\frac{\partial \ln L(\theta;\mathbf{x})}{\partial\theta}L(\theta;\mathbf{x})d\mathbf{x}\right|^2 \\
&\leq \int_{\mathbf{R}^n}(u(\mathbf{x})-\varphi(\theta))^2L(\theta;\mathbf{x})d\mathbf{x}\int_{\mathbf{R}^n}\left(\frac{\partial \ln L(\theta;\mathbf{x})}{\partial\theta}\right)^2L(\theta;\mathbf{x})d\mathbf{x} = \\
&= D(Y)\cdot D\left(\frac{\partial \ln L(\theta;\mathbf{X})}{\partial\theta}\right) = D(Y)E\left[\left(\frac{\partial \ln L(\theta;\mathbf{X})}{\partial\theta}\right)^2\right],
\end{aligned}$$

односно

$$D(Y) \geq \frac{|\varphi'(\theta)|^2}{E\left(\frac{\partial \ln L(\theta;X_1,\ldots,X_n)}{\partial\theta}\right)^2} \quad ,$$

што је и требало доказати. $\square$

Приметимо израз који игра важну улогу у дефинисању Рао–Крамерове доње границе:

$$I(\theta) = E\left(\frac{\partial \ln L(\theta;X_1,\ldots,X_n)}{\partial\theta}\right)^2$$

и који је у статистичкој литератури познат под називом *Фишерова количина информација.*

Рао–Крамерова неједнакост нам омогућава да код регуларних фамилија допустивих расподела одредимо доњу границу дисперзија свих непристрасних статистика за оцену непознатог параметра расподеле. Детаљније ћемо се бавити само једнодимензионим случајем.

Имајући у виду дефиницију најбоље статистике за параметар, закључујемо да је у регуларном случају то она статистика чија дисперзија достиже Рао-Крамерову доњу границу. У том случају дефинишемо најефикаснију статистику за параметар.

Дефиниција 23. *Ефикасност непристрасне статистике* у регуларном случају тачкастог оцењивања параметра расподеле је количник Рао–Крамерове доње границе и дисперзије статистике која се користи као оцена за посматрани параметар.$\diamondsuit$

Дефиниција 24. *Најефикаснија статистика* за параметар регуларне фамилије је непристрасна статистика чија је ефикасност једнака јединици.$\diamondsuit$

Пример 37. Статистика $\overline{X}_n$ је најефикаснија статистика за параметар θ Пуасонове расподеле $\mathcal{P}(\theta)$ на основу простог случајног узорка. Доказати! $\triangle$

Пример 38. Нека је дат прост случајни узорак обима n, $n > 1$, из популације са обележјем X чија расподела припада фамилији нормалних расподела $\{\mathcal{N}(m, \theta); \theta > 0\}$. Раније смо доказали да је статистика $\widetilde{S}_n^2$ непристрасна оцена параметра θ. Лако се може проверити да је Рао–Крамерова доња граница за параметар θ нормалне расподеле (m је познати параметар) једнака $\frac{2\theta^2}{n}$, а њена дисперзија $\frac{2\theta^2}{n-1}$, јер је

$$E(X-m)^4 = 3\theta^2\,,$$

па је њена ефикасност $\frac{n-1}{n}$. Закључујемо да је испитивана статистика само асимптотски најефикаснија. $\triangle$

Навешћемо без доказа још неке чињенице у вези са Рао–Крамеровом доњом границом.

Ако за неку непристрасну оцену утврдимо да јој дисперзија достиже Рао–Крамерову доњу границу, тада је могуће доказати да је та статистика такође и довољна статистика за посматрани параметар. Другим речима, класа најефикаснијих статистика је поткласа класе довољних статистика.

Најзад, констатујмо да постоји одговарајућа неједнакост и за нерегуларне густине расподела, али се тиме овде нећемо бавити.

3.4.1 Експоненцијална класа функција густина расподеле

Дефиниција 25. Нека је $\{f(x;\theta), \theta \in \Theta\}$ фамилија допустивих густина расподеле обележја X. Нека је A носач за X и $\Theta = \{\theta | \gamma < \theta < \delta\}$, $\gamma, \delta \in R \bigcup \{-\infty, +\infty\}$, $\gamma < \delta$ тј., интервал. Када је X апсолутно непрекидног типа нека је $A = (a, b)$ где су a и b такође из $R \bigcup \{-\infty, +\infty\}$. Уколико је X дискретног типа, нека је $A = \{a_1, a_2, \ldots\}$. Ако је тада

$$f(x;\theta) = \tag{3.17}$$

$$= \begin{cases} \exp\left(p(\theta)K(x) + S(x) + q(\theta)\right), & x \in A \\ 0, & \text{иначе} \end{cases},$$

при чему је $p(\theta)$ нетривијална непрекидна функција параметра θ, а функције $S(x)$ и $K'(x)$ су непрекидне по x, а $K'(x)$ није идентички једнака 0. Тада се фамилија расподела дефинисана густином (3.17) зове *експоненцијална класа густина расподеле.*$\diamondsuit$

Када $A = (a, b)$, у апсолутно непрекидном случају, односно његове границе, и скуп $A = \{a_1, a_2, \ldots\}$ у дискретном случају, не зависе од θ, може се доказати да је експоненцијална класа регуларна, тј. да имамо *регуларни случај експоненцијалне класе.*

Пример 39. Нека је дата фамилија $\{\mathcal{N}(0,\theta), \theta > 0\}$. Дакле,

$$f(x;\theta) = \frac{1}{\sqrt{2\pi\theta}} e^{-\frac{x^2}{2\theta}},$$

односно,

$$f(x;\theta) = e^{-\ln\sqrt{2\pi\theta}} e^{-\frac{x^2}{2\theta}} = \exp\left(-\frac{1}{2\theta}x^2 - \ln\sqrt{2\pi\theta}\right)$$

је густина расподеле обележја X са нормалном расподелом. Овде су очигледно функције:

$$p(\theta) = -\frac{1}{2\theta}, \quad K(x) = x^2, \quad S(x) = 0, \quad q(\theta) = -\ln\sqrt{2\pi\theta}, \quad x \in R.$$

Видимо да је p нетривијална, S и K' су непрекидне, па су услови експоненцијалности задовољени. $\triangle$

Пример 40. Нека је дата расподела дискретног типа $\{\mathcal{P}(\theta), 0 < \theta < \infty\}$ где је

$$f(x;\theta) = \begin{cases} \frac{\theta^x e^{-\theta}}{x!}, & x = 0, 1, 2, \ldots \\ 0, & \text{иначе} \end{cases}.$$

Одавде се добија

$$\frac{\theta^x e^{-\theta}}{x!} = \exp(x\ln\theta - \ln(x!) - \theta)$$

при чему је:

$$q(\theta) = -\theta, \quad p(\theta) = \ln\theta, \quad K(x) = x, \quad S(x) = -\ln(x!),$$

па би Пуасонова расподела била пример експоненцијалне класе у дискретном случају. $\triangle$

Ако узмемо прост случајни узорак $\mathbf{X} = (X_1, \ldots, X_n)$ из популације чија расподела припада класи експоненцијалних функција густина расподеле, онда би овај вектор имао густину расподеле дату са

$$f(x_1;\theta)\cdots f(x_n;\theta) =$$
$$= \begin{cases} \exp(p(\theta)\sum_{i=1}^n K(x_i) + nq(\theta))\exp(\sum_{i=1}^n S(x_i)), \ x_i \in A, \ i = 1, \ldots, n \\ \hfill 0, \quad \text{иначе} \quad . \end{cases}$$

У вези са простим случајним узорком наводимо без доказа следећу теорему.

Теорема 13. *Нека је за обележје X $\{f(x;\theta), \gamma < \theta < \delta\}$, $\gamma, \delta \in R$, фамилија допустивих густина расподеле из експоненцијалне класе и нека имамо прост случајни узорак $\mathbf{X} = (X_1, \ldots, X_n)$ константног обима n из популације са обележјем X. Статистика*

$$Y_1 = \sum_{i=1}^n K(X_i)$$

је довољна статистика за θ, а фамилија $\{g_1(y_1;\theta), \gamma < \theta < \delta\}$ густина расподеле за Y_1 је комплетна, тј. Y_1 је комплетна довољна статистика за θ.

Још једно важно својство регуларног случаја експоненцијалне класе је да Теорема 11 даје потребне и довољне услове уколико фамилија допустивих расподела припада овој класи.

Експоненцијална класа густина расподеле се дефинише и у случају вишедимензионог параметра.

Дефиниција 26. Нека је $\mathbf{\Theta} = \{\boldsymbol{\theta} | \boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)'\}$ параметарски простор. Фамилија допустивих расподела $\{f(x;\boldsymbol{\theta}); \boldsymbol{\theta} \in \mathbf{\Theta}\}$ обележја X посматраног на популацији је *експоненцијална*, ако има густину

$$f(x;\boldsymbol{\theta}) = C(\boldsymbol{\theta}) \exp\left(\sum_{j=1}^k Q_j(\boldsymbol{\theta}) T_j(x)\right) U(x) \, ,$$

где су C и Q_j мерљиве функције на параметарском простору $\mathbf{\Theta}$, а T_j и U су мерљиве функције по x и још важи да је $C(\boldsymbol{\theta}) > 0$ и $U(x) \geq 0$ и то за свако $\boldsymbol{\theta} \in \mathbf{\Theta}$ и свако x из скупа вредности за обележје X.$\diamondsuit$

3.5 Методи тачкастог оцењивања параметара

3.5.1 Метод максималне веродостојности

Метод максималне веродостојности је општи метод за оцењивање непознатих параметара расподеле и може се примењивати (са више или мање успеха) код регуларних и нерегуларних фамилија, затим код произвољних узорака, простих или не итд.

Метод максималне веродостојности се примењује, како за једнодимензиони, тако и за вишедимензиони параметар.

Нека је дата фамилија допустивих густина расподеле $\{f(x;\boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ обележја X и узорак $(X_1,\ldots,X_n)$ са заједничком густином расподеле $f(x_1,\ldots,x_n;\boldsymbol{\theta})$ из популације са обележјем X. Како смо већ истакли, заједничку густину можемо да посматрамо као функцију параметра $\boldsymbol{\theta}$ и у том случају се ова функција зове функција веродостојности и најчешће се означава са

$$L(\boldsymbol{\theta}) = L(\boldsymbol{\theta}; x_1,\ldots,x_n) = f(x_1,\ldots,x_n;\boldsymbol{\theta}).$$

Дефиниција 27. *Оценом максималне веродостојности* параметра $\theta \in \Theta \subset R$ зваћемо ону статистику $Y = u(X_1,\ldots,X_n)$ која даје максимум функцује L по $\theta \in \Theta$ за свако $\mathbf{x} = (x_1,\ldots,x_n)$ из узорачког простора $\mathcal{X} \subset R^n$ реализованих узорака фиксног обима n. Ознака за оцену максималне веродостојности биће $\widehat{\theta} = u(X_1,\ldots,X_n)$, тј. за произвољни $\mathbf{x} \in \mathcal{X}$, $\widehat{\theta} = u(x_1,\ldots,x_n)$ је она функција за коју је

$$L(\widehat{\theta}; x_1,\ldots,x_n) = \max_{\theta\in\Theta} L(\theta; x_1,\ldots,x_n)\,.\diamondsuit$$

Ако је $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_r)$ вишедимензиони параметар, оцена максималне веродостојности ће, јасно, бити $r-$димензиона статистика $\widehat{\boldsymbol{\theta}} = (\widehat{\theta_1},\ldots,\widehat{\theta_r})$ која максимализује функцију веродостојности $L(\boldsymbol{\theta})$.

У случају да је узорак прост, заједничка густина је једнака производу маргиналних густина, па је:

$$L(\boldsymbol{\theta}; x_1,\ldots,x_n) = f(x_1;\boldsymbol{\theta})f(x_2;\boldsymbol{\theta})\ldots f(x_n;\boldsymbol{\theta})\,.$$

Пример 41. Нека је дат прост случајни узорак $(X_1,\ldots,X_n)$ из популације са обележјем X чија расподела припада фамилији допустивих расподела

$$\{\mathcal{N}(\theta,1), \theta \in R\}.$$

Оцена максималне веродостојности ће се добити на следећи начин:

$$\begin{aligned} L(\theta; x_1,\ldots,x_n) &= f(x_1;\theta)f(x_2;\theta)\ldots f(x_n;\theta) = \\ &= \frac{1}{\sqrt{2\pi}}e^{-\frac{(x_1-\theta)^2}{2}}\frac{1}{\sqrt{2\pi}}e^{-\frac{(x_2-\theta)^2}{2}}\cdots\frac{1}{\sqrt{2\pi}}e^{-\frac{(x_n-\theta)^2}{2}} = \\ &= (2\pi)^{-\frac{n}{2}}e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2}. \end{aligned}$$

Логаритмовањем леве и десне стране, добићемо:

$$\ln L(\theta) = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2.$$

Да бисмо нашли оцену максималне веродостојности параметра θ треба наћи максимум функције L, и то ћемо учинити посредно одређујући максимум функције $\ln L$. Дакле, из

$$\frac{\partial \ln L(\theta)}{\partial\theta} = -\frac{1}{2}\sum_{i=1}^{n}2(x_i-\theta)(-1) = 0$$

следи

$$\sum_{i=1}^{n}x_i - n\theta = 0\,,$$

односно,

$$\theta = \frac{1}{n}\sum_{i=1}^{n}x_i\,,$$

па је тражена статистика

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n}X_i\,,$$

тј.

$$\widehat{\theta} = \overline{X_n}\,.$$

Уочимо да је ова оцена непристрасна. $\triangle$

Пример 42. Нека је дат прост случајни узорак $(X_1,\ldots,X_n)$ из популације са обележјем X чија расподела припада фамилији допустивих расподела

$$\{\mathcal{U}(0,\theta]; 0<\theta<1\}.$$

Густина је

$$f(x;\theta) = \begin{cases} \frac{1}{\theta}, & 0<x\leq\theta \\ 0, & \text{иначе} \end{cases}.$$

Функција веродостојности је

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta) = \frac{1}{\theta^n}, \quad 0 < x_i \le \theta, \quad i = 1,\dots,n.$$

Ова функција ће имати максимум када је θ најмање могуће, па ћемо за оцену максималне веродостојности узети статистику

$$\widehat{\theta} = \max_{i\in\{1,\dots,n\}} X_i \;\; = \;\; X_{(n)}\,.$$

Дакле, оцена максималне веродостојности је статистика поретка реда n. Ова оцена није непристрасна, али јесте асимптотски непристрасна. $\triangle$

Теорема 14. *Ако постоји довољна статистика $Y = u(X_1,\dots,X_n)$ за параметар θ на основу узорка $\mathbf{X} = (X_1,\dots,X_n)$ из популације са обележјем X чија расподела припада фамилији допустивих расподела која зависи од θ и ако, такође, постоји оцена максималне веродостојности $\widehat{\theta}$ за параметар θ и јединствена је, тада је $\widehat{\theta}$ функција од довољне статистике Y, тј. $\widehat{\theta} = v(Y)$.*

Доказ. Нека је $g(y;\theta)$ густина расподеле статистике Y. Тада је функција веродостојности (према Фишер–Нејмановом критеријуму):

$$L(\theta;x_1,\dots,x_n) \;\; = \;\; g(u(x_1,\dots,x_n);\theta)H(x_1,\dots,x_n),$$

па ће максимум функције веродостојности бити функција од узорка само преко функције u, што значи да је $\widehat{\theta}$ функција од довољне статистике. $\square$

Под одређеним условима оцена максималне веродостојности је строго постојана оцена за параметар θ, а у регуларном случају је асимптотски нормална и асимптотски најефикаснија.

3.5.2 Метод момената

Још један метод тачкастог оцењивања параметара увео је Пирсон. То је тзв. метод момената. Да бисмо изложили овај метод увешћемо још неке дефиниције теорије узорака, као и претпоставку да посматрано обележје X има моменте до неког реда r.

Дефиниција 28. За узорак $(X_1, \ldots, X_n)$ обима n из популације са обележјем X, *узорачки момент реда* k је статистика

$$A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k \quad , \quad k = 1, 2, \ldots, r\,,$$

а *узорачки централни момент реда* k је статистика

$$M_k = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X}_n)^k \quad , \quad k = 1, 2, \ldots, r\,.\diamondsuit$$

Специјални случајеви овако дефинисаних момената су средина и дисперзија узорка, помињани већ раније.

Да би се истакла зависност узорачких момената од обима узорка n, користе се и ознаке A_{nk} и M_{nk} за обичан и централни узорачки момент реда k.

Теорема 15. *Узорачки момент реда k простог случајног узорка је непристрасна и постојана оцена теоријског момента реда k датог обележја.*

Доказ. Означимо са $\alpha_k = EX^k$ $(k \leq r)$ теоријски момент реда k обележја X популације из које је узет узорак $(X_1, \ldots, X_n)$. Тада је:

$$E(A_{nk}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i^k\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i^k) = \frac{1}{n} n\alpha_k = \alpha_k,$$

што доказује да је узорачки момент непристрасна оцена. Докажимо постојаност ове оцене користећи неједнакост Чебишева:

$$P\{|A_{nk} - \alpha_k| \geq \varepsilon\} \leq \frac{E(A_{nk} - \alpha_k)^2}{\varepsilon^2} = \frac{D(A_{nk})}{\varepsilon^2}\,,$$

где је

$$D(A_{nk}) = E(A_{nk}^2) - (EA_{nk})^2 = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i^k\right)^2 - \alpha_k^2\,.$$

Како је

$$E\left(\frac{1}{n}\sum_{i=1}^{n} X_i^k\right)^2 \quad = \quad E\left(\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} X_i^k X_j^k\right) =$$

$$\begin{aligned} &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}E(X_i^kX_j^k) = \\ &= \frac{1}{n^2}nE\left(X^{2k}\right)+\frac{1}{n^2}(n^2-n)E\left(X^k\right)E\left(X^k\right) = \\ &= \frac{1}{n}\alpha_{2k}+\frac{n-1}{n}\alpha_k^2 \quad , \end{aligned}$$

то је

$$D(A_{nk}) = \frac{1}{n}\alpha_{2k}+\frac{n-1}{n}\alpha_k^2-\alpha_k^2 = \frac{\alpha_{2k}-\alpha_k^2}{n}\,,$$

односно,

$$P\{|A_{nk}-\alpha_k|\geq\varepsilon\}\leq\frac{\alpha_{2k}-\alpha_k^2}{n\varepsilon^2} \quad \longrightarrow 0 \quad , \quad n\to\infty\,.\square$$

Аналогно тврђење овоме важи и за централне моменте M_{nk}, као и за било које друге узорачке карактеристике које су дефинисане као непрекидне функције од коначног броја статистика A_{nk}. Ово је тачно као последица следеће теореме теорије вероватноће коју наводимо без доказа:

Теорема 16. *Нека низ r–димензионих случајних променљивих чији је општи члан $(\xi_1(n),\ldots,\xi_r(n))$, конвергира у вероватноћи ка константи $(c_1,\ldots,c_r)$. Тада за непрекидну функцију $\varphi: R^r \longrightarrow R$, важи*

$$\zeta(n) = \varphi(\xi_1(n),\ldots,\xi_r(n)) \xrightarrow{P} \varphi(c_1,\ldots,c_r)\,, \quad n\longrightarrow\infty\,.$$

Теорема 17. *Узорачки момент A_{nk} простог случајног узорка је асимптотски нормалан, тј. има приближно $\mathcal{N}(\alpha_k,\frac{\alpha_{2k}-\alpha_k^2}{n})$ расподелу када обим узорка неограничено расте.*

Доказ. Полазимо од простог случајног узорка $(X_1,\ldots,X_n)$ из популације са обележјем X произвољне расподеле која има моменте $E(X^k)=\alpha_k$ и $D(X^k)=\alpha_{2k}-\alpha_k^2$, $k\leq r$. Тада је, као у Теореми 15,

$$A_{nk}=\frac{1}{n}\sum_{i=1}^{n}X_i^k\,, \quad E(A_{nk})=\alpha_k\,, \quad D(A_{nk})=\frac{\alpha_{2k}-\alpha_k^2}{n}\,.$$

Користећи централну граничну теорему закључујемо да низ случајних променљивих дефинисаних општим чланом

$$\eta(n)=\frac{\frac{1}{n}\sum_{i=1}^{n}X_i^k-\alpha_k}{\sqrt{\frac{\alpha_{2k}-\alpha_k^2}{n}}}=\frac{A_{nk}-\alpha_k}{\sqrt{\alpha_{2k}-\alpha_k^2}}\sqrt{n}$$

конвергира у расподели ка случајној промењливој са $\mathcal{N}(0,1)$ расподелом. Изражавајући A_{nk} као функцију од $\eta(n)$, добијамо

$$A_{nk} = \eta(n)\sqrt{\frac{\alpha_{2k} - \alpha_k^2}{n}} + \alpha_k\,.$$

Како је A_{nk} линеарна функција од $\eta(n)$ и $\eta(n)$ има, као граничну, када $n \to \infty$, нормалну расподелу, то и A_{nk} има, за довољно велики обим узорка, приближно нормалну расподелу $\mathcal{N}(\alpha_k, \frac{\alpha_{2k}-\alpha_k^2}{n})$. $\square$

Ова теорема омогућава да се за велики обим узорка оцени грешка оцењивања при оцени теоријског момента одговарајућим узорачким моментом:

$$P\{|A_{nk} - \alpha_k| < t\} = 2\Phi\left(t\sqrt{\frac{n}{\alpha_{2k} - \alpha_k^2}}\right)\,.$$

Поступак метода момената

Нека је дата фамилија допустивих расподела $\{f(x;\boldsymbol{\theta}), \boldsymbol{\theta} = (\theta_1, \ldots, \theta_r) \in \boldsymbol{\Theta}\}$ обележја X и узорак $(X_1, \ldots, X_n)$ из популације са овим обележјем. Треба оценити θ_i, $i = 1, \ldots, r$.

Постојање момената α_k, $k = 1, \ldots, r$, где је $\alpha_k = \alpha_k(\boldsymbol{\theta})$, је само потребан услов за примену метода момената. Сам поступак се састоји у следећем.

Нека су a_{nk} реализоване вредности узорачких момената A_{nk}. Изједначаваћемо теоријске моменте са реализованим вредностима узорачких момената, тј.

$$\alpha_k(\theta_1, \ldots, \theta_r) \equiv \alpha_k(\boldsymbol{\theta}) = a_{nk}\,, \quad k = 1, \ldots, r\,.$$

На тај начин се добија систем од r једначина по непознатим параметрима $\theta_1, \ldots, \theta_r$. Ако је овај систем решив, добићемо тачкасте оцене параметара. Довољно је да кореспонденција буде обострано једнозначна између параметарских промењливих $\theta_1, \ldots, \theta_r$ и реализованих вредности статистика, $a_{n1}, \ldots, a_{nr}$, тј. да постоје такве функције φ_i помоћу којих можемо да нађемо јединствено решење система у облику:

$$\theta_i = \varphi_i(a_{n1}, \ldots, a_{nr})\,, \quad i = 1, \ldots, r\,.$$

Заменом реализованих вредности статистика на месту аргумената функција φ_i самим статистикама, добићемо статистике, које ћемо звати статистикама (оценама) метода момената:

$$\widetilde{\theta}_i = \varphi_i(A_{n1}, \ldots, A_{nr}), \quad i = 1, \ldots, r.$$

Ако су φ_i непрекидне функције тада ће оцене параметара θ_i добијене методом момената бити постојане.

Пример 43. (Гама расподела) Нека обележје X има густину расподеле

$$\{f(x;\boldsymbol{\theta}), \boldsymbol{\theta} = (\theta_1, \theta_2) \in \boldsymbol{\Theta} = (0, +\infty) \times (0, +\infty)\},$$

где је

$$f(x;\boldsymbol{\theta}) = \begin{cases} \dfrac{x^{\theta_2 - 1} e^{-\frac{x}{\theta_1}}}{\theta_1^{\theta_2}\Gamma(\theta_2)}, & x > 0 \\ 0, & x \leq 0 \end{cases}.$$

Моменти гама расподеле су

$$\alpha_k = \int_0^{+\infty} \frac{x^{\theta_2 + k - 1} e^{-\frac{x}{\theta_1}}}{\theta_1^{\theta_2}\Gamma(\theta_2)} dx = \theta_1^k \theta_2(\theta_2 + 1) \cdots (\theta_2 + k - 1), \quad k = 1, 2, \ldots.$$

За $k = 1, 2$ добија се систем који треба решити по θ_1 и θ_2:

$$\alpha_1 = \theta_1\theta_2 \quad , \quad \alpha_2 = \theta_1^2\theta_2(\theta_2 + 1).$$

Решење система је:

$$\theta_1 = \frac{\alpha_2 - \alpha_1^2}{\alpha_1} \quad , \quad \theta_2 = \frac{\alpha_1^2}{\alpha_2 - \alpha_1^2},$$

одакле следе оцене методом момената

$$\widetilde{\theta}_1(X) = \frac{A_{n2} - A_{n1}^2}{A_{n1}}, \quad \text{тј.} \quad \widetilde{\theta}_1(X) = \frac{\overline{S}_n^2}{\overline{X}_n}$$

и

$$\widetilde{\theta}_2(X) = \frac{A_{n1}^2}{A_{n2} - A_{n1}^2}, \quad \text{однодсно}, \quad \widetilde{\theta}_2(X) = \frac{\overline{X}_n^2}{\overline{S}_n^2}. \triangle$$

3.6 Статистике поретка

С обзиром на то да су статистике поретка основ за добијање оцена извесног броја параметара расподеле обележја, задржаћемо се мало детаљније на проблему одређивања расподела статистика поретка, као и на неким конкретним применама.

Нека је $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ прост случајни узорак из популације са обележјем X чија је функција расподеле F. Као што смо рекли, низ $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$, где је $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ је низ статистика поретка датог узорка.

Означимо са F_k функцију расподеле статистике поретка реда k, $k = 1, 2, \ldots, n$. Тада је за свако $x \in R$

$$\begin{aligned} F_n(x) &= P\left\{X_{(n)} \leq x\right\} = P\left(\cap_{i=1}^n \{X_{(i)} \leq x\}\right) = P\left(\cap_{i=1}^n \{X_i \leq x\}\right) = \\ &= \prod_{i=1}^n P\{X_i \leq x\} = \prod_{i=1}^n F(x) = (F(x))^n . \end{aligned}$$

Такође,

$$F_1(x) = P\{X_{(1)} \leq x\} = 1 - P\{X_{(1)} > x\} = 1 - \prod_{i=1}^n (1 - F(x)) = 1 - (1 - F(x))^n$$

или уопште:

$$F_k(x) = P\{X_{(k)} \leq x\} = P\{\text{бар } k \text{ елемената у узорку је мање или једнако } x\} = \sum_{i=k}^n \binom{n}{i} (F(x))^i (1 - F(x))^{n-i}.$$

На основу овако добијених функција расподеле релативно је једноставно одредити густине расподела. Ми ћемо се задржати само на густинама за обележје апсолутно непрекидног типа.

Због једноставности, прећи ћемо на ознаке $Y_k \equiv X_{(k)}$.

Теорема 18. *Нека је $Y_1 \leq Y_2 \leq \ldots \leq Y_n$ варијациони низ простог узорка $(X_1, \ldots, X_n)$ обележја X чија је густина расподеле $f(x)$, непрекидна и строго позитивна за $x \in (a, b) \subset R$. Тада је заједничка густина расподеле вектора $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$:*

$$g(y_1, \ldots, y_n) = \begin{cases} n! f(y_1) f(y_2) \ldots f(y_n), & a < y_1 < y_2 < \ldots < y_n < b \\ 0, & \text{иначе} \end{cases} .$$

Доказ нећемо изводити у општем случају, већ само илустрацију доказа на примеру обима узорка $n = 3$.

Дакле, нека је дат прост случајни узорак (X_1, X_2, X_3). Уредивши га добијамо $\left(X_{(1)}, X_{(2)}, X_{(3)}\right)$, односно статистике $Y_1 \leq Y_2 \leq Y_3$ где је:

$$Y_1 = X_{(1)}\,, \quad Y_2 = X_{(2)}\,, \quad Y_3 = X_{(3)}\,.$$

Како је

$$P\{a < X_1 = X_2 < b, a < X_3 < b\} = \int_a^b \int_a^b \int_{x_2}^{x_2} f(x_1)f(x_2)f(x_3)dx_1dx_2dx_3 = \\ = 0\,,$$

посматраћемо само случај строге неједнакости.

Заједничка густина расподеле за вектор (X_1, X_2, X_3) је:

$$f(x_1, x_2, x_3) = \begin{cases} f(x_1)f(x_2)f(x_3)\,, & a < x_i < b\,, \quad i = 1, 2, 3 \\ 0\,, & \text{иначе}\,. \end{cases}$$

Посматрајмо следеће скупове:

$$A_1 = \{(x_1, x_2, x_3) : a < x_1 < x_2 < x_3 < b\}$$

$$A_2 = \{(x_1, x_2, x_3) : a < x_1 < x_3 < x_2 < b\}$$

$$A_3 = \{(x_1, x_2, x_3) : a < x_3 < x_1 < x_2 < b\}$$

$$A_4 = \{(x_1, x_2, x_3) : a < x_3 < x_2 < x_1 < b\}$$

$$A_5 = \{(x_1, x_2, x_3) : a < x_2 < x_1 < x_3 < b\}$$

$$A_6 = \{(x_1, x_2, x_3) : a < x_2 < x_3 < x_1 < b\}.$$

Нека је $y_1 = \min\{x_1, x_2, x_3\}$, $y_3 = \max\{x_1, x_2, x_3\}$. Свака од области $A_i, i = 1, \ldots, 6$ се може обострано једнозначно да преслика у област $B = \{(y_1, y_2, y_3) : a < y_1 < y_2 < y_3 < b\}$ и за свако такво пресликавање одредимо Јакобијан:

$$J_k = \left|\frac{\partial x_i}{\partial y_j}\right|_{i,j=1,2,3}\,, \quad k = 1, 2, \ldots, 6\,.$$

Апсолутна вредност сваког од ових Јакобијана износи 1 тј.

$$J_1 = \begin{vmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{vmatrix} = 1, \quad \ldots\,, \qquad J_5 = \begin{vmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{vmatrix} = -1, \quad \ldots\,.$$

Тако је заједничка густина расподеле статистика Y_1, Y_2, Y_3 :

$$g(y_1,y_2,y_3) = \begin{cases} |J_1|f(y_1)f(y_2)f(y_3) + \cdots + |J_6|f(y_1)f(y_2)f(y_3) & , \\ \qquad a < y_1 < y_2 < y_3 < b & \\ 0 & , \\ \qquad \text{у осталим случајевима}, & \end{cases}$$

$$= \begin{cases} (3!)f(y_1)f(y_2)f(y_3), & a < y_1 < y_2 < y_3 < b \\ 0, & \text{у осталим случајевима} \end{cases}$$

што је и требало доказати.

Приметимо да су статистике поретка, за разлику од случајних променљивих из простог случајног узорка из кога су настале, стохастички зависне.

Статистике поретка су основ за оцењивање неких важних параметара обележја, као што је, рецимо, медијана или, уопште, квантил реда p.

3.6.1 Примена статистика поретка у одређивању тачкастих оцена за квантиле

Наведимо, најпре, дефиницију квантила. У литератури се, углавном, могу наћи следеће дефиниције овог појма:

Дефиниција 29. *Квантил реда* p, у ознаци M_p, $0 < p < 1$, случајне променљиве X са функцијом расподеле F, је било које решење једначине $F(x) = p$, $0 < p < 1$, по $x \in R$, односно, решење система неједначина

$$P\{X < x\} \leq p \leq P\{X \leq x\}, \quad 0 < p < 1,$$

по $x \in R$. $\diamondsuit$

Дефиниција 30. *Квантил реда* p, у ознаци M_p, $0 < p < 1$, случајне променљиве X са функцијом расподеле F, је

$$M_p = \inf_{x \in R}\{F(x) \geq p\}. \diamondsuit$$

Овако дефинисана функција од p носи назив *квантилна функција.*

Систем неједначина се користи поготову уколико решење поменуте једначине није јединствено.

Ако је обележје X апсолутно непрекидног типа, са строго растућом функцијом расподеле F за свако $x \in R$ за које је $0 < F(x) < 1$, обе дефиниције квантила реда p дају јединствену вредност која се, практично, добија решавањем једначине:

$$F(x) = p\,, \quad 0 < p < 1\,,$$

по $x \in R$.

Код обележја дискретног типа ситуација је сложенија утолико што се другом од наведених дефиниција долази до јединствене бројне вредности која се назива квантилом, док се првом дефиницијом у неким случајевима добија јединствена вредност као квантил, а у неким се долази до полуотвореног интервала чије све тачке задовољавају дефиницију. Дакле, у том случају, квантил није јединствена бројчана вредност, него цео интервал. У практичним применама се, у таквим случајевима, може користити произвољна тачка поменутог интервала. Најчешће се користи средина интервала или доња граница.

Посебно често коришћени квантил је квантил реда $p = 0,5$, који се због своје важности и посебно именује. Његово име је *медијана.*

Пример 44. Нека је дата случајна променљива X са расподелом

$$X\,:\,\begin{pmatrix} 1 & 2 & 3 \\ 1/4 & 1/4 & 1/2 \end{pmatrix}.$$

Квантил реда $1/3$ је $M_{1/3} = 2$, јер је $P\{X < 2\} = 1/4 < 1/3$, а $P\{X \le 2\} = 1/2 > 1/3$. Медијана је, међутим по првој дефиницији, цео интервал $[2, 3)$. $\triangle$

Када се ред квантила изражава у процентима, $100p\%$, уместо термина квантил користи се термин *перцентил.* За неке квантиле се користе посебни називи везано за ред квантила. Тако, за $p \in \{0,25; 0,50; 0,75\}$ користи се назив *квартил*, за $p \in \{0,1; 0,2; 0,3; 0,4; 0,5; 0,6; 0,7; 0,8; 0,9\}$ се користи назив *децил* и слично.

Приступимо сада оцењивању квантила на основу узорка. Статистике које се користе као оцене квантила једним именом се зову *узорачки квантили.*

У случају дискретног обележја или узорка малог обима n, квантил реда p се оцењује одговарајућим линеарним комбинацијама статистика поретка и то, ако је

- $(n-1)p$ природан број, квантил M_p се оцењује статистиком
$$\widehat{M_p} = X_{(1+(n-1)p)},$$
- $(n-1)p$ није природан број, квантил M_p се оцењује статистиком
$$\begin{aligned}\widehat{M_p} &= (k-(n-1)p)X_{(k)} + (1+(n-1)p-k)X_{(k+1)} = \\ &= X_{(k)} + (1+(n-1)p-k)(X_{(k+1)} - X_{(k)}),\end{aligned}$$
где је k природан број такав да је $k < 1+(n-1)p < k+1$.

Изложени поступак, заправо, идентификује члан варијационог низа (ако такав постоји) који дели варијациони низ у односу $100p\%$ према $100(1-p)\%$ по аналогији са теоријским значењем квантила случајне променљиве.

О оцени квантила обележја апсолутно непрекидног типа биће речи у оквиру наредног потпоглавља.

3.7 О још неким примерима тачкастих оцена

Не улазећи у то којим методом су оцене добијене, навешћемо неке конкретне примере тачкастих оцена параметара обележја о којима до сада није било речи.

Реализовани узорци из популације са обележјем апсолутно непрекидног типа се, по правилу, сређују интервално. Већ је речено да се код интервалног сређивања узорка губи извесна количина информација коју узорак у себи носи. У оценама које треба дефинисати за интервално сређен узорак се не користе сами елементи узорка, већ репрезентанти интервала, најчешће средине интервала, као објекти над којима се дефинишу статистике (док је узорак случајан и интервали су случајни, па и њихови репрезентанти). Тако, на пример, средина узорка се одређује као аритметичка средина средина интервала по којима се узорак сређује. Но, није увек оправдано користити средине интервала, као што ће се видети из неких ниже наведених примера.

Оцена квантила обележја апсолутно непрекидног типа

Узорачки квантил узорка обима n за обележје апсолутно непрекидног типа је статистика која се базира на границама интрвала и апсолутним учестаностима елемената узорка у интервалима по којима се реализовани узорак сређује.

Узорачки квантил реда p је статистика

$$M_p = a_p + h_p \frac{np - \Sigma_p}{N_p},$$

где је a_p – доња граница квантилног интервала, h_p – дужина квантилног интервала, N_p – апсолутна учестаност у квантилном интервалу, а Σ_p – збирна апсолутна учестаност до границе a_p, тј. до квантилног интервала.

Квантилни интервал је интервал у коме је збирна апсолутна учестаност први пут већа или једнака np.

Оцене мода

Још један од центара груписања вредности случајне променљиве је мод случајне променљиве. Из практичних разлога наводимо овде, не само његову оцену, односно статистику којом се оцењује, већ и саму дефиницију.

Дефиниција 31. *Мод* случајне променљиве је она вредност случајне променљиве за коју густина расподеле достиже локални максимум. ♢

Из дефиниције следи да случајна променљива може да има више модова, па се према броју модова користе називи: *унимодална, бимодална, тримодална* и *мултимодална* случајна променљива.

За мод случајне променљиве X се најчешће користи ознака $M_o(X)$.

Што се тиче оцена мода, разликују се поступци оцењивања за дискретна и обележја апсолутно непрекидног типа.

Код дискретног обележја, оцена мода (модова) је елемент варијационог низа који има највећу апсолутну учестаност у својој околини.

Код обележја апсолутно непрекидног типа, тј. код узорка који се сређује интервално, најчешће су у употреби следеће статистике:

- $M_o = a_\mu + h_\mu \dfrac{N_0 - N_1}{(N_0 - N_1) + (N_0 - N_2)}$
- $M_o = a_\mu + h_\mu \dfrac{N_2}{N_1 + N_2}$

где је a_μ – лева граница модалног интервала, h_μ – дужина модалног интервала, N_0 – апсолутна учестаност у модалном интервалу, N_1 – апсолутна учестаност у интервалу испред модалног и N_2 – апсолутна учестаност у интервалу иза модалног. Модални интервал је интервал у коме је апсолутна учестаност елемената реализованог узорка већа у односу на суседне интервале – претходни и наредни. Уколико је први интервал модални, онда је $N_1 = 0$, а уколико је последњи интервал модални, тада је $N_2 = 0$.

* * *

Приметимо да је у претходна два случаја параметара обележја коришћена иста ознака за параметар расподеле и његову оцену. То у статистичким разматрањима није изузетак и може се толерисати кадгод нема опасности од забуне.

Оцене коефицијента асиметрије

Један од важнијих показатеља облика расподеле обележја је коефицијент асиметрије.

Уколико је расподела случајне променљиве X симетрична у односу на праву $x = E(X)$, сви централни моменти непарног реда случајне променљиве X, уколико постоје, једнаки су 0 (обрнуто не мора да важи). Због тога је погодно грубо процењивање облика расподеле неке случајне променљиве X у погледу симетричности у односу на праву $x = E(X)$ коришћењем трећег централног момента. У ту сврху уводи се тзв. коефицијент асиметрије.

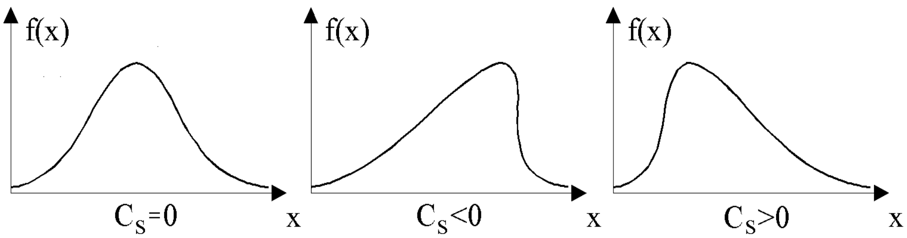
Дефиниција 32. *Коефицијент асиметрије* је

$$C_s = \frac{\mu_3}{\mu_2^{3/2}} = \frac{\mu_3}{\sigma^3},$$

где је $\mu_3 = E(X - E(X))^3$, а $\mu_2 = D(X)$.$\diamondsuit$

Очигледно је да је и ова нумеричка карактеристика неименован број.

За симетричну расподелу је $C_s = 0$, док се за расподелу асиметричну на десно добија позитивна вредност коефицијента асиметрије, а за расподелу асиметричну на лево добија се негативна вредност (слика 3.1).



Слика 3.1: Зависност облика функције густине расподеле и знака коефицијента асиметрије.

Пример 45. Добро позната случајна променљива $X : \mathcal{N}(m, \sigma^2)$ има функцију густине расподеле симетричну у односу на праву $x = m$. Рачунањем је једноставно проверити да је код ове расподеле $\mu_3 = 0$, тј. $C_s = 0$. $\triangle$

Не сме се, међутим, изгубити из вида да има расподела чији је непарни централни момент једнак 0, а да при томе саме расподеле нису симетричне.

Пример 46. Случајна променљива

$$X : \begin{pmatrix} -4 & 1 & 5 \\ 1/3 & 1/2 & 1/6 \end{pmatrix}$$

нема симетричну расподелу, а $E(X) = 0$, $D(X) = 10$ и $E(X - E(X))^3 = 0$, те је $C_S = 0$. $\triangle$

Као тачкасте оцене овог параметра се користи више статистика. Неке од њих су

- Пирсонов коефицијент асиметрије: $\dfrac{\overline{X}_n - M_o}{\overline{S}_n}$,
- Фишеров коефицијент асиметрије: $\dfrac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^3}{(\overline{S}_n)^3}$,
- Јулов коефицијент асиметрије: $\dfrac{Q_1 + Q_3}{2M_e} - 1$, где су Q_1 и Q_3 први и трећи узорачки квартил.

Набројани узорачки коефицијенти асиметрије дефинисани су тако да имају вредност једнаку нули код симетричног обележја, позитивну вредност код искошености на десно и негативну код искошености на лево. Међутим, у случају слабо изражене асиметрије нису поуздани.

Оцене коефицијента спљоштености

Још један показатељ облика расподеле обележја је коефицијент спљоштености (конвексности, конкавности). Он указује на положај максимума густине расподеле, као и на такозване *тешке* или *дебеле репове* густине расподеле.

Дефиниција 33. *Коефицијент спљоштености (конвексности, конкавности)* је

$$C_k = \frac{\mu_4}{\mu_2^2} = \frac{\mu_4}{\sigma^4},$$

где је $\mu_4 = E(X - E(X))^4$. $\diamondsuit$

Овај број мери удаљавање по y-правцу функције $y = f(x)$ густине расподеле случајне промењиве X од апсцисне осе, тј. удаљеност дискретних тачака графика густине расподеле у случају да је X случајна промењива дискретног типа. У литератури је још познат као *картозис*[3].

Како је за нормалну расподелу $C_k = 3$, то се узимајући за "нулти" ниво спљоштености криву густине расподеле случајне промењиве са нормалном расподелом, дефинише коефицијент ексцеса.

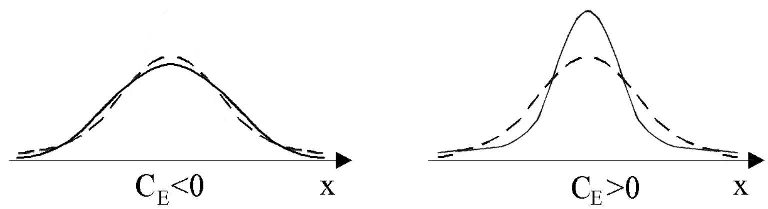
Дефиниција 34. *Коефицијент ексцеса* или *ексцесни коефицијент спљоштености* је

$$C_E = C_k - 3. \diamondsuit$$

Према дефиницији је $C_E = 0$ за случајну промењиву са нормалном расподелом, док је за случајну промењиву чија густина расподеле има максимум изнад максимума нормалне криве, коефицијент екцеса позитиван а у супротном случају негативан (слика 3.2). Такође треба напоменути да је коефицијент ексцеса код постојања тешких репова позитиван.

Статистике којима се оцењује коефицијент ексцеса, тј. *узорачки коефицијенти ексцеса* су рецимо

[3] На енглеском *kurtosis*.



Слика 3.2: Знак коефицијента ексцеса одређује се у односу на нормалну расподелу која је приказана испрекиданом линијом.

- Пирсонов коефицијенти ексцеса: $\dfrac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^4}{(\overline{S}_n^2)^2} - 3$, и
- модални коефицијенти ексцеса: $2,5\dfrac{N_0\overline{S}_n}{nh_0} - 1$ где је h_0–дужина модалног интервала, N_0–учестаност у модалном интервалу.

Оба ова коефицијента имају својство да мере конкавност у односу на функцију густине расподеле случајне променљиве чија је расподела $\mathcal{N}(0,1)$, па имају негативну вредност у случају да више приањају уз апсцисну осу, тј. да им је конкавност мање изражена од поменуте криве и позитивну вредност у супротном.

Оцена коефицијента корелације

Бројне су статистике којима се оцењује коефицијент корелације обележја. Овде ћемо се, међутим, задржати само на једној од њих, док се више података о оцењивању коефицијента корелације може наћи у књизи: Б. Поповић, М. Ристић: "Статистика у психологији", Мрљеш, Београд, 2000.

Статистика коју наводимо је у литератури позната под називом *узорачки* или *Пирсонов коефицијент корелације.*

За испитивање линеарне повезаности два обележја X и Y на елементима исте популације узима се узорак облика

$$((X_1,Y_1),\ldots,(X_n,Y_n))\,, \tag{3.18}$$

при чему је обим узорка n константан. Пар случајних променљивих (X_i,Y_i) представља вредност дводимензионог обележја на i–том елементу узорка.

Сама статистика је већ помињана у претходној глави ове књиге и дефинисана је са

$$R_{X,Y} = \frac{\frac{1}{n}\Sigma_{i=1}^{n}(X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{\overline{S}_X \overline{S}_Y},$$

где су $\overline{S}_X$ и $\overline{S}_Y$ узорачке стандардне девијације за обележја X и Y редом на основу узорка (3.18).

* * *

Следећи пример илуструје све горе наведене оцене.

Пример 47. На бази анализе плувиографских трака, за једну кишомерну станицу током 40 година, добијени су интензитети киша различитих трајања. У доњој табели приказани су максимални годишњи интензитети (mm/h) киша трајања $2h$ и $1/2h$ у посматраном периоду.

Година	$i_{2h,max}$	$i_{1/2h,max}$	Година	$i_{2h,max}$	$i_{1/2h,max}$
1	3,1	10,0	21	12,9	32,2
2	7,5	30,8	22	4,6	19,2
3	12,6	42,2	23	5,2	13,3
4	10,4	36,2	24	3,1	12,2
5	4,9	15,4	25	4,9	12,2
6	7,1	20,0	26	5,2	25,2
7	5,3	14,2	27	4,6	18,5
8	9,7	24,5	28	11,8	32,3
9	6,2	19,3	29	12,4	33,8
10	10,0	18,8	30	10,1	29,8
11	6,0	16,2	31	6,0	19,7
12	5,8	18,2	32	8,2	22,5
13	4,5	15,8	33	9,8	28,7
14	5,3	17,3	34	11,7	36,9
15	9,3	21,3	35	10,0	27,8
16	8,5	26,8	36	6,7	23,1
17	11,6	29,0	37	13,4	44,6
18	10,5	28,8	38	8,4	23,8
19	11,6	31,1	39	7,9	27,2
20	9,9	27,9	40	8,0	19,2

У овом моменту ће се надаље посматрати само једно обележје: вредност максималних интензитета 2–часовних падавина у mm/h ($i_{2h,max}$) током 40 година.

Дакле, варијациони низ ($x_{(i)}$) је: 3,1; 3,1; 4,5; 4,6; 4,6; 4,9; 4,9; 5,2; 5,2; 5,3; 5,3; 5,8; 6,0; 6,0; 6,2; 6,7; 7,1; 7,5; 7,9; 8,0; 8,2; 8,4; 8,5; 9,3; 9,7; 9,8; 9,9; 10,0; 10,0; 10,1; 10,4; 10,5; 11,6; 11,6; 11,7; 11,8; 12,4; 12,6; 12,9; 13,4.

Интервално сређивање узорка:

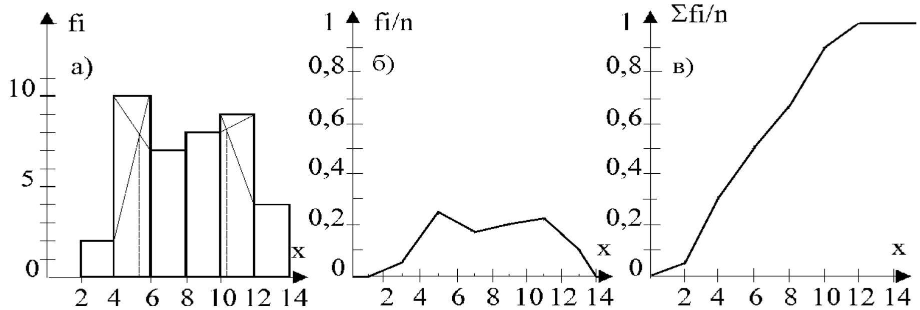
- број класа (интервала): $k = 1 + 3,322 \log_{10} 40 = 6,32 \Rightarrow$ усвојено 6 класа;
- дужина интервала: $h = (13,4 - 3,1)/6 = 1,71 \Rightarrow$ усвојена дужина 2,0.

Одлука о броју и дужини интервала прилично је препуштена искуству истраживача, а груписањем података делом се губи информација о узорку у потпуности.

При распоређивању елемената узорка у интервале добија се следећа табела

Интервал	$[2;4)$	$[4;6)$	$[6;8)$	$[8;10)$	$[10;12)$	$[12;14]$
x_i	3,0	5,0	7,0	9,0	11,0	13,0
f_i	2	10	7	8	9	4
Σf_i	2	12	19	27	36	40
$f_i \cdot x_i$	6,0	50,0	49,0	72,0	99,0	52,0
f_i/n	0,05	0,25	0,175	0,20	0,225	0,10
$\Sigma f_i/n$	0,05	0,30	0,475	0,675	0,90	1,00
f_i/n [%]	5	25	17,5	20	22,5	10
$\Sigma(f_i/n)$ [%]	5	30	47,5	67,5	90	100
$x_i - \overline{x}_n$	-5,2	-3,2	-1,2	0,8	2,8	4,8
$(x_i - \overline{x}_n)^2$	27,04	10,24	1,44	0,64	7,84	23,04
$(x_i - \overline{x}_n)^3$	-140,61	-32,77	-1,73	0,51	21,95	110,59
$(x_i - \overline{x}_n)^4$	731,16	104,86	2,07	0,41	61,47	530,84

$\Sigma f_i = 40$; $\Sigma f_i x_i = 328,0$; $\overline{x}_n = 1/n \Sigma f_i x_i = 8,2$; $\Sigma f_i/n = 1,00$; $\Sigma f_i/n$ [%] $= 100$; $\Sigma f_i |x_i - \overline{x}_n| = 101,6$; $\Sigma f_i (x_i - \overline{x}_n)^2 = 334,4$; $\Sigma f_i (x_i - \overline{x}_n)^3 = 23,04$; $\Sigma f_i (x_i - \overline{x}_n)^4 = 5205,25$.



Слика 3.3: Расподела максималних интензитета киша $[mm/h]$ трајања 2 часа на једном плувиографу током 40 година осматрања: а) Хистограм апсолутних учестаности; б) полигон релативних учестаности ; в) кумулативна крива учестаности (збирне релативне учестаности)

1. **Мере централне тенденције**:
   - Средина узорка:
     $T_n = \sum f_i \cdot x_i = 328\,, \quad \overline{x}_n = \frac{T_n}{n} = \frac{328}{40} = 8,2$
   - Медијана узорка: медијански интервал је $[8;10) \Rightarrow$
     $M_e = 8,0 + 2,0\frac{20-19}{8} = 8,25$
   - Мод узорка: јављају се два модална интервала $[4;6)$ и $[10;12)$ (види се и на слици 3.3, где је илустрована и прва по реду формула за прорачун вредности мода, узета као меродавна за остале прорачуне у којима фигурише вредност мода):
     $Mo_1 = 4,0 + 2,0\frac{10-2}{(10-2)+(10-7)} = 5,45$
     $(Mo_1 = 4,0 + 2,0\frac{7}{2+7} = 5,56)$
     $Mo_2 = 10,0 + 2,0\frac{9-8}{(9-8)+(9-4)} = 10,33$
     $(Mo_2 = 10,0 + 2,0\frac{4}{8+4} = 10,67)$
   - Квантили:
     (*i*) Квартили:
     $Q_1 = 4,0 + 2,0\frac{40\cdot 0,25-2}{10} = 5,6$
     $Q_2 = 8,0 + 2,0\frac{40\cdot 0,50-19}{8} = 8,25$
     $Q_3 = 10,0 + 2,0\frac{40\cdot 0,75-27}{9} = 10,67$

(ii) Квинтили:

$K_1 = 4,0 + 2,0\frac{40\cdot 0,2-2}{10} = 5,2$
$K_2 = 6,0 + 2,0\frac{40\cdot 0,4-12}{7} = 7,14$
$K_3 = 8,0 + 2,0\frac{40\cdot 0,6-19}{8} = 9,25$
$K_4 = 10,0 + 2,0\frac{40\cdot 0,8-27}{9} = 11,11$

(iii) Секстили:

$S_1 = 4,0 + 2,0\frac{40\cdot 1/6-2}{10} = 4,93$
$S_2 = 6,0 + 2,0\frac{40\cdot 2/6-12}{7} = 6,38$
$S_3 = 8,0 + 2,0\frac{40\cdot 3/6-19}{8} = 8,25$
$S_4 = 8,0 + 2,0\frac{40\cdot 4/6-19}{8} = 9,92$
$S_5 = 10,0 + 2,0\frac{40\cdot 5/6-27}{9} = 11,41$

(iv) Децили:

$D_1 = 4,0 + 2,0\frac{40\cdot 0,1-2}{10} = 4,4$
$D_2 = 4,0 + 2,0\frac{40\cdot 0,2-2}{10} = 5,2$
$D_3 = 4,0 + 2,0\frac{40\cdot 0,3-2}{10} = 6,0$
$D_4 = 6,0 + 2,0\frac{40\cdot 0,4-12}{7} = 7,14$
$D_5 = 8,0 + 2,0\frac{40\cdot 0,5-19}{8} = 8,25$
$D_6 = 8,0 + 2,0\frac{40\cdot 0,6-19}{8} = 9,25$
$D_7 = 10,0 + 2,0\frac{40\cdot 0,7-27}{9} = 10,22$
$D_8 = 10,0 + 2,0\frac{40\cdot 0,8-27}{9} = 11,11$
$D_9 = 10,0 + 2,0\frac{40\cdot 0,9-27}{9} = 12,0$

2. **Мере растурања**:

- Оцене мера груписања око фиксиране вредности:

(а) Средње апсолутно одступање од медијане:
$Md(M_e) = \frac{1}{40}(2\cdot|3-8,25| + 10\cdot|5-8,25| + 7\cdot|7-8,25| +$
$+8\cdot|9-8,25| + 9\cdot|11-8,25| + 4\cdot|13-8,25|) = 2,5375$

(б) Средње апсолутно одступање од средине узорка:
$Md(\overline{X}) = \frac{1}{40}(2\cdot|3-8,2| + 10\cdot|5-8,2| + 7\cdot|7-8,2| +$
$+\,8\cdot|9-8,2| + 9\cdot|11-8,2| + 4\cdot|13-8,2|) = \frac{101,6}{40} = 2,54$

(в) Средње квадратно одступање:
$\overline{s}_n^2 = \frac{1}{40}334,4 = 8,36$

(г) Стандардна девијација:
$\overline{s}_n = \sqrt{8,36} = 2,89$

- Оцене мера груписања дела података:

(а) Распон узорка: $r_x = 13,4 - 3,1 = 10,3$

Уколико би се користили таблично сређени подаци за распон узорка би се добило $r_x = 13 - 3 = 10$.

(б) Квантилни распони:

(i) Квартилни: $Q_3 - Q_1 = 10,67 - 5,6 = 5,07$

(ii) Квинтилни: $K_4 - K_1 = 11,1 - 5,2 = 5,91$

(iii) Секстилни: $S_5 - S_1 = 11,41 - 4,93 = 6,47$

(iv) Децилни: $D_9 - D_1 = 12,0 - 4,4 = 7,6$

3. **Оцене показатеља облика расподеле**:

(а) Узорачки коефицијент асиметрије:

- Пирсонов:
  1)$\frac{8{,}2-5{,}45}{2{,}89} = 0,95$ 2)$\frac{8{,}2-10{,}33}{2{,}89} = -0,74$
  Узорак је бимодалан и добијени резултати управо указују на то да је већина елемената у узорку десно од првог мода, али лево од другог мода.
- Фишеров: $\frac{\frac{1}{40}22{,}96}{2{,}89^3} = 0,024$
- Јулов: $\frac{5{,}6+10{,}67}{2\cdot 8{,}25} - 1 = -0,01$

Сва три типа узорачких коефицијената асиметрије указују на релативно симетричну расподелу.

(б) Узорачки коефицијент ексцеса:

- Пирсонов: $\frac{\frac{1}{40}5205{,}25}{8{,}36^2} - 3 = -1,14$
- Модални:
  1) $2,5\frac{10\cdot 2{,}89}{40\cdot 2} - 1 = -0,10$ 2) $2,5\frac{9\cdot 2{,}89}{40\cdot 2} - 1 = -0,19$
  На основу испитиваног узорка може се закључити да је расподела посматраног обележја са мање израженом конкавношћу у односу на нормалну расподелу.

4. **Оцена мере повезаности два обележја**

- Узорачки коефицијент корелације

Посматрајмо сада оба обележја, при чему ће бити обележје X максимални годишњи интензитет падавина трајања 2 часа, а обележје Y максимални годишњи интензитет падавина трајања 1/2 часа. Користићемо некласирани узорак, тј. узорак који није сређен интервално.

Добијају се следеће вредности:

$\overline{x}_{40} = 8,12$; $\sum_i (x_i - \overline{x}_{40})^2 = 320,68$; $\overline{s}_x = 2,83$;

$\overline{y}_{40} = 24,15$; $\sum_i (y_i - \overline{y}_{40})^2 = 2698,15$; $\overline{s}_y = 8,21$;

и $\sum_i (x_i - \overline{x}_{40})(y_i - \overline{y}_{40}) = 836,29$.

Те је оцена коефицијента линеарне корелације X и Y:

$$r_{X,Y} = \frac{\frac{1}{40}836,29}{2,83 \cdot 8,21} = 0,90.$$

Може се уочити да се овде добијена узорачка средња вредност за обележје X нешто мало разликује од претходно израчунате иако је био коришћен исти реализовани узорак. Разлика је настала, јер је у последњем израчунавању коришћен узорак који није сређен интервално. Већ смо раније нагласили да се интервалним сређивањем узорка губи део информација које узорак носи о расподели обележја!

△

3.8 Области поверења

За разлику од тачкастог оцењивања параметара, код којег се оценом сматра статистика, тј. њена реализована вредност, области поверења имају ту улогу да се процени скуп, односно подскуп, одговарајућег реалног простора унутар кога се може сматрати да ће се "наћи" права вредност параметра. У најједноставнијем случају, тј. код једнодимензионог параметра, област поверења биће интервал као део реалне праве.

3.8.1 Интервали поверења

Разјаснићемо питање интервала поверења на једном примеру.

Пример 48. Нека је $(X_1, X_2, \dots, X_n)$ прост случајни узорак из популације са обележјем X чија расподела припада фамилији допуствих расподела $\{\mathcal{N}(m,4), -\infty < m < \infty\}$. Као што је познато, ако је m права вредност параметра, случајна промењива

$$Z_0 = \frac{\overline{X}_n - m}{2}\sqrt{n} \quad \text{има} \quad \mathcal{N}(0,1) \quad \text{расподелу.}$$

У том случају је за задато α, $0 < \alpha < 1$, односно $\gamma = 1 - \alpha$, могуће увек одредити број $z_{\frac{1-\alpha}{2}}$, тако да је

$$P\{|Z_0| \leq z_{\frac{1-\alpha}{2}}\} = 1 - \alpha.$$

Дакле, биће:

$$P\left\{\left|\frac{\overline{X}_n - m}{2}\sqrt{n}\right| \leq z_{\frac{1-\alpha}{2}}\right\} = 1 - \alpha$$

$$P\left\{\left|\frac{\overline{X}_n - m}{2}\sqrt{n}\right| \leq z_{\frac{1-\alpha}{2}}\right\} = P\{\overline{X}_n - z_{\frac{1-\alpha}{2}}\frac{2}{\sqrt{n}} \leq m \leq \overline{X}_n + z_{\frac{1-\alpha}{2}}\frac{2}{\sqrt{n}}\} = 1-\alpha.$$

Одавде се добија интервал као скуп могућих вредности за параметар m:

$$\left[\overline{X}_n - z_{\frac{1-\alpha}{2}}\frac{2}{\sqrt{n}} \quad , \quad \overline{X}_n + z_{\frac{1-\alpha}{2}}\frac{2}{\sqrt{n}}\right].$$

Овај интервал ћемо звати интервал поверења за непознато математичко очекивање нормалне расподеле нивоа поверења $\gamma = 1 - \alpha$. $\triangle$

Посматрајмо интервал добијен у претходном примеру:

$$\left[\overline{X}_n - z_{\frac{1-\alpha}{2}}\frac{2}{\sqrt{n}} \quad , \quad \overline{X}_n + z_{\frac{1-\alpha}{2}}\frac{2}{\sqrt{n}}\right].$$

Његове границе су случајне величине.

Дефиниција 35. Ако је бар једна граница интервала случајна променљива, интервал се назива *случајни интервал*. $\diamondsuit$

Из дефиниције случајног интервала је јасно да је и његова дужина d случајна величина, па се може говорити о очекиваној дужини, $E(d)$, случајног интервала.

Дакле, у горњем примеру добијен је један случајни интервал, а његова очекивана дужина је: $E(d) = \frac{4z_{\frac{1-\alpha}{2}}}{\sqrt{n}}$.

У овом примеру је дужина интервала неслучајна. Међутим, у следећем примеру је она случајна.

Пример 49. Нека је $X : \mathcal{N}(0, \sigma^2), \sigma > 0$. Колика је вероватноћа да случајни интервал $(|X|, |10X|)$ садржи тачку σ? Која је очекивана дужина овог интервала?

Одговор је једноставан с обзиром на чињеницу да $\frac{X}{\sigma}$ има $\mathcal{N}(0,1)$ расподелу и низ догађаја једнаких вероватноћа (еквивалентни догађаји):

$$P\{|X| < \sigma < 10|X|\} = P\left\{\frac{\sigma}{10} < |X| < \sigma\right\} =$$
$$= P\left\{\frac{\sigma}{10} < X < \sigma \vee -\sigma < X < -\frac{\sigma}{10}\right\}.$$

Дакле, вероватноћа да случајни интервал садржи тачку σ једнака је:

$$P\left\{\frac{\sigma}{10} < |X| < \sigma\right\} = P\left\{\frac{\sigma}{10} < X < \sigma\right\} + P\left\{-\sigma < X < -\frac{\sigma}{10}\right\} =$$

$$= 2P\left\{\frac{\sigma}{10} < X < \sigma\right\} = 2P\left\{\frac{1}{10} < \frac{X}{\sigma} < 1\right\} = 2\Phi(1) - 2\Phi(0,1) = 0,60\,.$$

Дужина посматраног случајног интервала је $10|X| - |X| = 9|X|$. Како је

$$E|X| = 2\int_0^{\infty} \frac{x}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} dx = \sigma\sqrt{\frac{2}{\pi}}\,,$$

Очекивана дужина овог интервала је $9E|X| \approx 7,2\sigma$. $\triangle$

Пример 50. На основу узорка обима n, $\mathbf{X} = (X_1, \ldots, X_n)$, из популације са обележјем X чија расподела припада фамилији допустивих расподела $\{\mathcal{N}(m,\sigma^2),\ -\infty < m < +\infty,\ \sigma^2 > 0\}$ одредити колика је вероватноћа да случајни интервал

$$\left(\overline{X}_n - 2\frac{\sigma}{\sqrt{n}} \quad , \quad \overline{X}_n + 2\frac{\sigma}{\sqrt{n}}\right)$$

садржи тачку m? Колика је очекивана дужина овог случајног интервала?

Као и у претходном примеру, користићемо еквивалентне догађаје:

$$P\left\{\overline{X}_n - 2\frac{\sigma}{\sqrt{n}} < m < \overline{X}_n + 2\frac{\sigma}{\sqrt{n}}\right\} = P\left\{-2\frac{\sigma}{\sqrt{n}} < \overline{X}_n - m < 2\frac{\sigma}{\sqrt{n}}\right\} =$$
$$= P\left\{-2 < \frac{\overline{X}_n - m}{\frac{\sigma}{\sqrt{n}}} < 2\right\}.$$

Како је расподела случајне промењиве

$$\frac{\overline{X}_n - m}{\frac{\sigma}{\sqrt{n}}} : \mathcal{N}(0,1)\,,$$

то је тражена вероватноћа $2\Phi(2) = 0,954$, а очекивана дужина интервала

$$E(d) = 4\frac{\sigma}{\sqrt{n}}\,.\ \triangle$$

Пример 51. Нека је (X_1, X_2, X_3, X_4) прост случајни узорак из популације са обележјем X, чија је густина расподеле

$$f(x) = \begin{cases} \frac{1}{\theta}, & 0 < x < \theta, \theta > 0 \\ 0, & \text{иначе} \end{cases}\,,$$

тј. обележје X има расподелу која припада фамилији допустивих расподела $\{\mathcal{U}(0,\theta), \theta > 0\}$.

Одредити 95%–тни интервал поверења за θ користећи статистику поретка Y_4 овог узорка.

С обзиром да се ради о униформној расподели,

$$F(x) = \begin{cases} 0, & x < 0 \\ \frac{x}{\theta}, & 0 \le x < \theta \\ 1, & x \ge \theta \end{cases}\,,$$

односно, густина расподеле статистике поретка максималног реда је

$$\begin{aligned} g_4(y_4) &= \begin{cases} 4(F(y_4))^3 f(y_4), & 0 < y_4 < \theta \\ 0, & \text{иначе} \end{cases} \\ &= \begin{cases} \frac{4}{\theta^4} y_4^3, & 0 < y_4 < \theta \\ 0, & \text{иначе} \end{cases}\,. \end{aligned}$$

Ако изаберемо реалне бројеве $0 < c_1 < c_2 \le 1$ такве да је:

$$\begin{aligned} 0,95 = P\{c_1\theta < Y_4 < c_2\theta\} = \frac{4}{\theta^4}\int_{c_1\theta}^{c_2\theta} y_4^3 dy_4 = \qquad (3.19) \\ = (c_2 - c_1)(c_2 + c_1)(c_2^2 + c_1^2) \end{aligned}$$

добићемо

$$0,95 = P\left\{\frac{Y_4}{c_2} < \theta < \frac{Y_4}{c_1}\right\}.$$

Дакле, за реализовану вредност за Y_4, y_4, интервал

$$\left(\frac{y_4}{c_2}, \frac{y_4}{c_1}\right)$$

је један 95%–тни интервал поверења за θ. Константе c_1 и c_2 које задовољавају услов (3.19) су, на пример,

$$c_1 = \sqrt[4]{0,05} \quad , \quad c_2 = 1,$$

где смо изабрали $c_2 = 1$, а онда израчунали c_1. $\triangle$

Пример 52. Нека је $(X_1, \ldots, X_{10})$ прост случајни узорак из популације са обележјем X које има расподелу $\mathcal{N}(m, \sigma^2)$, где је m позната карактеристика. Нека је

$$Y = \sum_{i=1}^{10} (X_i - m)^2.$$

Колика је вероватноћа да случајни интервал $(\frac{Y}{20{,}5}, \frac{Y}{3{,}25})$ садржи праву вредност параметра σ^2?

С обзиром на чињеницу о расподели случајне променљиве:

$$\frac{Y}{\sigma^2} : \chi^2_{10}$$

и чињенице да је

$$P\left\{\frac{Y}{20,5} < \sigma^2 < \frac{Y}{3,25}\right\} = P\left\{3,25 < \frac{Y}{\sigma^2} < 20,5\right\}$$

може се одредити тражена вероватноћа и она износи 0,95. Дужина посматраног случајног интервала је $d = Y\left(\frac{1}{3{,}25} - \frac{1}{20{,}5}\right) = 0,26Y$. С друге стране, како је

$$E\left(\frac{Y}{\sigma^2}\right) = 10 \quad \text{следи да је} \quad E(Y) = 10\sigma^2\,,$$

па је очекивана дужина интервала $2,6\sigma^2$. $\triangle$

Општи проблем који је решаван кроз све претходне примере је следећи. Нека је $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ узорак из популације са обележјем X чија расподела припада фамилији допустивих расподела $\{f(x;\theta), \theta \in \Theta\}$. Треба одредити интервал поверења за непознати параметар θ на основу задатог узорка $\mathbf{X}$. Проблем одређивања интервалне оцене параметра θ се своди на то да се одреде две статистике $Y_1 = u_1(X_1, \ldots, X_n)$ и $Y_2 = u_2(X_1, \ldots, X_n)$ такве да је

$$P\{Y_1 \leq Y_2\} = 1$$

и

$$P\{Y_1 \leq \theta \leq Y_2\} = \gamma\,, \quad 0 \leq \gamma \leq 1\,,$$

где је γ задата вероватноћа коју зовемо ниво поверења. Прецизно,

Дефиниција 36. Нека су $Y_1 = u_1(X_1, \ldots, X_n)$ и $Y_2 = u_2(X_1, \ldots, X_n)$ две статистике на основу истог узорка $\mathbf{X} = (X_1, \ldots, X_n)$ из популације са обележјем X чија је фамилија допустивих расподела $\{f(x;\theta), \theta \in \mathit{\Theta}\}$, такве да под условом да је θ права вредност параметра (што је надаље наглашено у индексу мере P), важи да је:

$$P_\theta\{Y_1 \leq Y_2\} = 1$$

и

$$P_\theta\{Y_1 \leq \theta \leq Y_2\} \geq \gamma\,, \quad 0 \leq \gamma \leq 1\,.$$

Тада се интервал $[Y_1, Y_2]$ зове *двострани интервал поверења* за непознати параметар θ са *нивоом поверења* γ. Уколико је једна од граница неслучајна величина, интервал ће бити *једнострани интервал поверења.* Оба једним именом зовемо *интервал поверења.* $\diamondsuit$

Природно је тражити да интервал поверења буде што ужи у том смислу да очекивање дужине интервала поверења, $E(Y_2 - Y_1)$, буде што мање. С друге стране, ниво поверења треба да буде што већи, и обично се узима да је $\gamma = 0,95$ или $\gamma = 0,99$. (Уобичајено је да се ниво поверења изражава у процентима као $100\gamma\%$, што је у претходним примерима већ коришћено.) Ова два захтева су углавном опречна и не може им се истовремено удовољити, а излаз се донекле налази у повећању обима узорка, мада се то не може узети као правило.

Интервал поверења је по дефиницији случајни интервал. Међутим, када се експеримент обави и на основу реализованог узорка добију реализоване вредности статистика које су границе интервала, добија се реализовани интервал поверења који је интервал на реалној правој. И за овако добијени интервал користи се назив *интервал поверења (нивоа поверења γ)* без опасности од забуне. Важно је, међутим, правилно разумевање нивоа поверења.

Осврнимо се на тумачење вероватноће γ.

Погрешно је тумачити да са вероватноћом γ реализовани интервал поверења садржи параметар θ, већ је вероватноћа γ само вероватноћа да случајни интервал $[Y_1\,, Y_2]$ прекрије, односно садржи, непознату праву вредност параметра θ. Вероватноћу γ можемо интерпретирати и овако: замислимо да смо "узели" 100 реализованих узорака истог обима n и добили низове бројева

$$(x_{11},\dots,x_{1n}),(x_{21},\dots,x_{2n}),\dots,(x_{100;1},\dots,x_{100;n}),$$

а затим на основу њих израчунали интервале поверења

$$[y_{11}\,, y_{12}]\,,\quad [y_{21}\,, y_{22}]\,,\quad \dots\quad ,\quad [y_{100;1}\,, y_{100;n}]\,.$$

Тада на те интервале можемо гледати као на реализације случајног интервала $[Y_1, Y_2]$. Како је $P\{Y_1 \le \theta \le Y_2\} = \gamma$ и тумачећи вероватноћу као граничну вредност релативних учестаности, можемо рећи да приближно $100\gamma\%$ реализованих интервала прекрива непознати параметар θ, а осталих $100(1-\gamma)\%$ реализованих интервала га не прекрива.

Пређимо сада на општи поступак добијања интервала поверења за произвољни непознати једнодимензиони параметар расподеле.

Дакле, нека је обележје X са расподелом која припада фамилији допустивих расподела $\{f(x;\theta), \theta \in \Theta\}$. Претпоставићемо још да је Θ неки интервал у R (иначе интервално оцењивање не би имало никаквог смисла). Уочимо, такође, да је узорачки простор реализованих узорака $\mathcal{X} \subset R^n$. Изаберимо реалан број γ, $0 \le \gamma \le 1$. Полазећи од узорка $\mathbf{X} = (X_1, X_2, \dots, X_n)$, тражићемо функције $u_1(\mathbf{x})$ и $u_2(\mathbf{x})$, такве да је $u_1 \le u_2$ за свако $\mathbf{x} \in \mathcal{X}$ и да је за свако $\theta \in \Theta$:

1) скуп $\{\mathbf{x} : \theta \in [u_1(\mathbf{x}), u_2(\mathbf{x})]\}$ мерљив,

2) $P_\theta\{\omega : \theta \in [u_1(\mathbf{X}), u_2(\mathbf{X})]\} \ge \gamma$.

Дефиниција 37. Претпоставимо да је $\{f(x;\theta), \theta \in \Theta\}$ фамилија допустивих расподела обележја X, $\mathbf{X} = (X_1, \dots, X_n)$ узорак обима n из популације са обележјем X и $T : \mathcal{X} \times \Theta \longrightarrow R$ функција која задовољава следећа два услова:

- Расподела вероватноћа случајне величине $T(X_1, \dots, X_n; \theta)$ **не зависи** од параметра θ.

- За свако $\mathbf{x} = (x_1, \ldots, x_n) \in \mathcal{X}$ функција $T(x_1, \ldots, x_n; \theta)$ је **непрекидна** и **строго монотона** функција аргумента θ. При томе карактер монотоности не сме да зависи од $\mathbf{x}$.

Тада се случајна величина $T(X_1, \ldots, X_n; \theta)$ зове *централна статистика* за параметар θ. $\diamondsuit$

Доказаћемо следећу теорему:

Теорема 19. *Нека постоји централна статистика T таква да $T : \mathcal{X} \times \Theta \longrightarrow R$ која је за свако фиксирано $\theta \in \Theta$ мерљива функција на скупу $\mathcal{X}$. Означимо са A кодомен функције T тј. $A = T(\mathcal{X} \times \Theta) \subset R$ и још претпоставимо да је за свако $a \in A$ и за свако $\mathbf{x} \in \mathcal{X}$ решива по θ једначина $a = T(\mathbf{x}, \theta)$. Када је θ фиксирано, користићемо ознаку $T(\mathbf{x}, \theta) \equiv T_\theta(\mathbf{x})$. Тада је функција $T_\theta(\mathbf{X})$ случајна променљива чија расподела не зависи од θ и на основу ње је увек могуће наћи интервал поверења за параметар θ.*

Доказ. Под условима наведеним у теореми, за свако фиксирано $\gamma \in [0, 1]$ могуће је одредити два реална броја $t_1(\gamma)$ и $t_2(\gamma)$ таква да за свако $\theta \in \Theta$ важи да је:

$$P_\theta\{t_1(\gamma) \leq T_\theta(\mathbf{X}) \leq t_2(\gamma)\} \geq \gamma\,,$$

јер је у питању строго монотона функција по θ па сигурно можемо наћи $t_1(\gamma)$ и $t_2(\gamma)$, с тим што то не морају да буду јединствени реални бројеви.

Пошто је функција T монотона по θ за фиксирано $\mathbf{x}$, решићемо једначине по параметру θ и добићемо јединствена решења која су границе интервала.

Значи, имамо следећи систем једначина и ограничења

$$t_1(\gamma) = T(\mathbf{x}, \theta) \tag{3.20}$$

$$t_2(\gamma) = T(\mathbf{x}, \theta) \tag{3.21}$$

$$t_1(\gamma) \leq t_2(\gamma)\,.$$

Означимо решења овог система са $T_{1\theta}(\mathbf{x}) = u_1(\mathbf{x})$ и $T_{2\theta}(\mathbf{x}) = u_2(\mathbf{x})$. Дакле, решавањем овог система по θ добијамо границе интервала, и ако уместо фиксираног узмемо случајни вектор $\mathbf{X}$, добићемо случајни интервал и то:

$$[T_{1\theta}(\mathbf{X}), T_{2\theta}(\mathbf{X})]\,.\square$$

Показаћемо како се применом централне статистике у случају расподеле апсолутно непрекидног типа може ефективно одредити интервал поверења за непознати параметар θ.

Нека је g густина расподеле случајне променљиве $T = T(X_1, \ldots, X_n; \theta)$, која не зависи од параметра θ. Нека је $\gamma \in (0,1)$ задати ниво поверења. Када је T апсолутно непрекидног типа, из услова:

$$\gamma = P_\theta\{t_1 \le T(X_1, \ldots, X_n; \theta) \le t_2\} = \int_{t_1}^{t_2} g(t)dt$$

одређујемо константе t_1 и t_2. Те константе нису једнозначно одређене. Фиксирајмо две константе за које је

$$\gamma = \int_{t_1}^{t_2} g(t)dt.$$

За тако фиксиране константе t_1 и t_2 и фиксирано $(x_1, \ldots, x_n)$ одредимо решења једначина $T(x_1, \ldots, x_n; \theta) = t_1$ и $T(x_1, \ldots, x_n; \theta) = t_2$ по θ, која су јединствена због монотоности функције T. Означимо решења са $y_1 = u_1(x_1, \ldots, x_n)$, $y_2 = u_2(x_1, \ldots, x_n)$ и одговарајуће статистике су $Y_1 = u_1(X_1, \ldots, X_n)$, $Y_2 = u_2(X_1, \ldots, X_n)$. Тада ће за њих важити

$$P_\theta\{Y_1 < \theta < Y_2\} = \gamma.$$

Приметимо да укључивање граница интервала у сам интервал нема значаја код обележја апсолутно непрекидног типа.

За дискретан случај поступак је аналоган, само што се уместо интеграла јавља сума и мора се узети у обзир да тражена вероватноћа мора да буде већа или једнака γ.

Ако претпоставимо (без смањења општости) да је T монотоно опадајућа функција по θ и означимо са $T_1(\mathbf{x}, \theta)$ решење једначине (3.21) , а са $T_2(\mathbf{x}, \theta)$ решење једначине (3.20), очигледно $T_2(\mathbf{x}, \theta) \ge T_1(\mathbf{x}, \theta)$.

Ако за неко θ при фиксираном $\mathbf{x}$ важи неједнакост:

$$t_1(\gamma) \le T(\mathbf{x}, \theta) \le t_2(\gamma) \tag{3.22}$$

тада је и

$$T_1(\mathbf{x}, \theta) \le \theta \le T_2(\mathbf{x}, \theta)\,. \tag{3.23}$$

Обрнуто, свако $\mathbf{x}$ које задовољава релацију (3.23), задовољава такође и релацију (3.22), па је

$$\{\mathbf{x} : t_1(\gamma) \leq T(\mathbf{x}, \theta) \leq t_2(\gamma)\} = \{\mathbf{x} : \theta \in [T_1(\mathbf{x}, \theta), T_2(\mathbf{x}, \theta)]\}.$$

Тада је $[T_1(\mathbf{x}, \theta), T_2(\mathbf{x}, \theta)]$ тј. $[T_1(\mathbf{X}, \theta), T_2(\mathbf{X}, \theta)]$ интервал поверења са нивоом поверења γ.

Задржаћемо се сада на специјалним интервалима поверења од којих смо неке већ разматрали кроз претходне примере.

Оцена за математичко очекивање m код нормалне расподеле $\mathcal{N}(m, \sigma^2)$ када је σ^2 **познато**:

Интервал поверења за параметар m имали смо у Примеру 48 и он износи:

$$I_m = \left[\overline{X}_n - z_{\frac{1-\alpha}{2}} \frac{\sigma}{\sqrt{n}} \quad , \quad \overline{X}_n + z_{\frac{1-\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right].$$

Оцена за математичко очекивање m код нормалне расподеле $\mathcal{N}(m, \sigma^2)$ када σ^2 **није познато**:

Најпре треба оценити σ^2 па користимо следећу централну статистику која има χ^2 расподелу са $(n-1)$ степени слободе:

$$\frac{n\overline{S}_n^{\,2}}{\sigma^2} = \chi^2_{n-1} \quad , \quad \overline{S}_n^{\,2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \,.$$

Дефинишимо случајну променљиву:

$$\frac{Z^*}{\sqrt{\frac{\chi^2_{n-1}}{n-1}}} = t_{n-1} \,,$$

где је

$$Z^* = \frac{\overline{X}_n - m}{\sqrt{\frac{\sigma^2}{n}}}$$

са нормалном нормираном расподелом, па статистика t_{n-1} има студентову расподелу са $(n-1)$ степени слободе, јер су Z^* и $\overline{S}_n^{\,2}$ независне случајне променљиве. Отуда

$$\frac{\overline{X}_n - m}{\sqrt{\overline{S}_n^2}}\sqrt{n-1} = t_{n-1} \,.$$

Ово ће бити тражена централна статистика за параметар m на основу које ћемо одредити интервал поверења:

$$I_m = \left[\overline{X}_n - t_{n-1;\frac{1-\alpha}{2}} \sqrt{\frac{\overline{S}_n^2}{n-1}} \quad , \quad \overline{X}_n + t_{n-1;\frac{1-\alpha}{2}} \sqrt{\frac{\overline{S}_n^2}{n-1}} \right]$$

са нивоом поверења $1-\alpha$, где је број $t_{n-1;\frac{1-\alpha}{2}}$ одређен из услова

$$P\left\{ |t_{n-1}| \leq t_{n-1;\frac{1-\alpha}{2}} \right\} = 1-\alpha \,.$$

Интервал поверења за разлику математичких очекивања $m_1 - m_2$ два независна обележја са нормалним расподелама и једнаким дисперзијама када је σ^2 **непознато**[4]:

Нека су дата два независна обележја са нормалном расподелом и једнаким дисперзијама $X : \mathcal{N}(m_1, \sigma^2)$ и $Y : \mathcal{N}(m_2, \sigma^2)$. Њихова разлика, као што је познато, има такође нормалну расподелу. Користићемо ово својство и на основу два независна узорка, по један из сваке од ових расподела, оценићемо разлику њихових очекивања.

Нека су поменути узорци редом $(X_1, X_2, \ldots, X_{n_1})$ и $(Y_1, Y_2, \ldots, Y_{n_2})$. Уочимо статистику

$$\overline{X}_{n_1} - \overline{Y}_{n_2} : \mathcal{N}\left(m_1 - m_2 \,,\, \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} \right) \,.$$

Посматрајмо следеће случајне променљиве:

$$\frac{\overline{X}_{n_1} - \overline{Y}_{n_2} - (m_1 - m_2)}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}} : \mathcal{N}(0,1) \,,$$

$$\frac{n_1 \overline{S}_{n_1}{}^2}{\sigma^2} = \chi^2_{n_1 - 1} \,,$$

$$\frac{n_2 \overline{S}_{n_2}{}^2}{\sigma^2} = \chi^2_{n_2 - 1}$$

и

$$\frac{n_1 \overline{S}_{n_1}{}^2}{\sigma^2} + \frac{n_2 \overline{S}_{n_2}{}^2}{\sigma^2} = \chi^2_{n_1 + n_2 - 2} \,.$$

[4] За случај познате дисперзије процедура ће бити аналогна, али са одговарајућим статистикама и њено спровођење се оставља читаоцу.

Одавде ћемо имати статистику која има студентову расподелу са $n_1 + n_2 - 2$ степена слободе:

$$\frac{\frac{\overline{X}_{n_1} - \overline{Y}_{n_2} - (m_1 - m_2)}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}}}{\sqrt{\frac{\frac{n_1 \overline{S}_{n_1}{}^2}{\sigma^2} + \frac{n_2 \overline{S}_{n_2}{}^2}{\sigma^2}}{n_1 + n_2 - 2}}} = t_{n_1+n_2-2}\,.$$

На тај начин смо изабрали централну статистику за параметар $m_1 - m_2$ на основу које ћемо одредити тражени интервал поверења

$$I_{m_1-m_2} = \left[\overline{X}_{n_1} - \overline{Y}_{n_2} - t_{n_1+n_2-2;\frac{1-\alpha}{2}} \sqrt{\frac{n_1 \overline{S}_{n_1}^2 + n_2 \overline{S}_{n_2}^2}{n_1 n_2 (n_1 + n_2 - 2)} (n_1 + n_2)} \quad , \right.$$

$$\left. \overline{X}_{n_1} - \overline{Y}_{n_2} + t_{n_1+n_2-2;\frac{1-\alpha}{2}} \sqrt{\frac{n_1 \overline{S}_{n_1}^2 + n_2 \overline{S}_{n_2}^2}{n_1 n_2 (n_1 + n_2 - 2)} (n_1 + n_2)} \right].$$

Оцена дисперзије σ^2 код нормалне расподеле када је m **познато**:

Нека је дата фамилија допустивих расподела $\{\mathcal{N}(m, \sigma^2), \sigma^2 > 0\}$, где је очекивање m познато. Уочимо случајну промењљиву са χ^2 расподелом и n степени слободе, која ће имати улогу централне статистике

$$Y = \sum_{i=1}^{n} \frac{(X_i - m)^2}{\sigma^2}\,.$$

Даље, ако је $\gamma = 1 - \alpha$ ниво поверења, треба одредити реалне бројеве a и b, $a < b$, такве да је:

$$P\{a \leq Y \leq b\} = \gamma = 1 - \alpha\,.$$

У претходним процедурама смо користили симетричност расподеле централне статистике за одређивање граница интервала и бројеве a и b смо одређивали на јединствен начин. С обзиром да χ^2–расподела има густину која је асиметрична, питање одређивања ових бројева је интересантно утолико што се поштује исти принцип у размишљању као и претходно, те се они одређују тако да важи

$$P\{\chi_n^2 < a\} = P\{\chi_n^2 > b\} = \frac{\alpha}{2}$$

(мада не неопходно). Отуда ће надаље бити коришћене ознаке за одговарајуће квантиле χ^2–расподеле.

$$P\left\{\chi^2_{n;\alpha/2} < \sum_{i=1}^{n} \frac{(X_i - m)^2}{\sigma^2} < \chi^2_{n;1-\alpha/2}\right\} = 1 - \alpha\,,$$

тј.

$$P\left\{\frac{\sum_{i=1}^{n}(X_i - m)^2}{\chi^2_{n;1-\alpha/2}} < \sigma^2 < \frac{\sum_{i=1}^{n}(X_i - m)^2}{\chi^2_{n;\alpha/2}}\right\} = 1 - \alpha\,.$$

Одавде је, очигледно, интервал поверења за дисперзију σ^2 када је очекивање познато :

$$I_{\sigma^2} = \left[\frac{\sum_{i=1}^{n}(X_i - m)^2}{\chi^2_{n;1-\alpha/2}} \quad , \quad \frac{\sum_{i=1}^{n}(X_i - m)^2}{\chi^2_{n;\alpha/2}}\right].$$

Помоћу овог интервала се може одредити и одговарајући интервал поверења за стандардну девијацију користећи дефиницију стандардне девијације:

$$I_{\sigma} = \left[\sqrt{\frac{\sum_{i=1}^{n}(X_i - m)^2}{\chi^2_{n;1-\alpha/2}}} \quad , \quad \sqrt{\frac{\sum_{i=1}^{n}(X_i - m)^2}{\chi^2_{n;\alpha/2}}}\right].$$

Оцена дисперзије σ^2 код нормалне расподеле када је m **непознато**:

Полази се од исте фамилије $\{\mathcal{N}(m,\sigma^2), \sigma^2 > 0\}$, и користи статистика која има χ^2 расподелу са $(n-1)$ степени слободе:

$$\frac{n\overline{S}_n^{\,2}}{\sigma^2} = \chi^2_{n-1}.$$

Идеја је иста као и у предходном случају,

$$P\left\{a \leq \frac{n\overline{S}_n^{\,2}}{\sigma^2} \leq b\right\} = \gamma = 1 - \alpha,$$

па кад се обави неопходно израчунавање, добијамо следеће границе случајног интервала:

$$I_{\sigma^2} = \left[\frac{n\overline{S}_n^{\,2}}{b}\,,\,\frac{n\overline{S}_n^{\,2}}{a}\right].$$

При томе се бројеви a и b одређују као и код претходне оцене, тј. из услова (који није обавезан, али је уобичајен)

$$P\{\chi^2_{n-1} \leq a\} = \frac{\alpha}{2} \quad , \quad P\{\chi^2_{n-1} \leq b\} = 1 - \frac{\alpha}{2}\,.$$

Интервал поверења за количник дисперзија два независна обележја са нормалним расподелама:

Нека су дата два независна узорка $(X_1, \ldots, X_{n_1})$ и $(Y_1, \ldots, Y_{n_2})$ за обележја редом $X : \{\mathcal{N}(m_X, \sigma_X^2)\,,\, \sigma_X^2 > 0\}$ и $Y : \{\mathcal{N}(m_Y, \sigma_Y^2), \sigma_Y^2 > 0\}$ при чему очекивања нису позната. Уколико се дисперзије ових обележја упоређују међу собом кроз њихов количник, има смисла одређивање интервала поверења за количник

$$\frac{\sigma_Y^2}{\sigma_X^2}\,.$$

Оцена се добија на основу сазнања да случајне промењиве имају расподеле како следи

$$\frac{n_1\overline{S}_X^2}{\sigma_X^2} \,=\, \chi^2_{n_1-1}\,, \qquad \frac{n_2\overline{S}_Y^2}{\sigma_Y^2} \,=\, \chi^2_{n_2-1}\,,$$

где су $\overline{S}_X^2$ и $\overline{S}_Y^2$ дисперзије узорака за обележја X и Y редом,

$$\frac{\frac{n_1\overline{S}_X^2}{\sigma_X^2(n_1-1)}}{\frac{n_2\overline{S}_Y^2}{\sigma_Y^2(n_2-1)}} = F_{n_1-1;n_2-1}\,.$$

Сам интервал поверења ће бити:

$$I_{\frac{\sigma_Y^2}{\sigma_X^2}} = \left[a\frac{\frac{n_2\overline{S}_Y^2}{n_2-1}}{\frac{n_1\overline{S}_X^2}{n_1-1}} \quad , \quad b\frac{\frac{n_2\overline{S}_Y^2}{n_2-1}}{\frac{n_1\overline{S}_X^2}{n_1-1}} \right],$$

за a и b који задовољавају услов

$$P\{a \leq F_{n_1-1;n_2-1} \leq b\} = 1 - \alpha$$

и, по правилу, услов

$$P\{F_{n_1-1;n_2-1} < a\} = P\{F_{n_1-1;n_2-1} > b\} = \frac{\alpha}{2}\,.$$

* * *

Сви наведени интервали поверења су тзв. двострани, а задржимо се још мало на једностраним интервалима. Поготову код дисперзије и стандардне девијације су у употреби, тзв. једнострани интервали поверења код којих је само једна граница случајна. Овакав интервал се користи у прилици када је једна граница од већег интереса за истраживање у коме се примењује интервално оцењивање као статистичка процедура. Изложићемо овај начин размишљања на примеру дисперзије.

Параметарски простор за дисперзију је по правилу $\Theta = [0\,, +\infty)$. Једнострани горњи интервал поверења (горња граница му је случајна) би се добио на основу размишљања

$$P\{0 \leq \sigma^2 \leq \varphi(X_1, X_2, \ldots, X_n)\} = 1 - \alpha\,,$$

а једнострани доњи (доња граница му је случајна)

$$P\{\varphi(X_1, X_2, \ldots, X_n) \leq \sigma^2 < +\infty\} = 1 - \alpha\,.$$

3.8.2 Непараметарски интервали поверења за квантиле

Постоје интервали поверења чије границе нису у вези са расподелом обележја чији се параметри оцењују, односно не зависе од те расподеле. У том случају се за интервал поверења каже да је непараметарски. Пример непараметарских интервала поверења су интервали поверења за квантиле.

Интервална оцена квантила се, као што је био случај и са тачкастом, базира на статистикама поретка. Она се одређује из услова

$$P\{X_{(r)} \leq M_p \leq X_{(s)}\} \geq \gamma \quad , \quad 1 \leq r < s \leq n\,.$$

Показаћемо, да ако X има расподелу апсолутно непрекидног типа, онда случајни интервал $\left[X_{(r)}, X_{(s)}\right]$, $1 \leq r < s \leq n$ прекрива праву вредност квантила M_p са вероватноћом која зависи од r, s и p, али не и од расподеле обележја X. Отуда ће он представљати непараметарски интервал (за разлику од претходних примера интервала поверења који су зависили од расподеле обележја X која је била нормална, и све коришћене централне статистике су биле засноване на тој претпоставци).

Заиста,

$$\{\omega : X_{(r)}(\omega) \leq M_p\} =$$
$$= \{\omega : X_{(r)}(\omega) \leq M_p \wedge X_{(s)}(\omega) \geq M_p\} \cup \{\omega : X_{(r)}(\omega) \leq M_p \wedge X_{(s)} < M_p\}\,,$$

те како је

$$\{\omega : X_{(s)}(\omega) < M_p\} \subset \{\omega : X_{(r)}(\omega) \leq M_p\}\,,$$

следи да је

$$\begin{aligned} P\{X_{(r)} \leq M_p\} &= \\ = P\{X_{(r)} \leq M_p \wedge X_{(s)} \geq M_p\} &+ P\{X_{(r)} \leq M_p \wedge X_{(s)} < M_p\} = \\ = P\{X_{(r)} \leq M_p \leq X_{(s)}\} &+ P\{X_{(s)} < M_p\}\,, \end{aligned}$$

а одавде је (у апсолутно непрекидном случају)

$$\begin{aligned} P\{X_{(r)} \leq M_p \leq X_{(s)}\} &= P\{X_{(r)} \leq M_p\} - P\{X_{(s)} < M_p\} = \\ &= \sum_{i=r}^{s-1} \binom{n}{i} p^i (1-p)^{n-i}, \end{aligned}$$

што је и требало доказати.

Овде је још важно констатовати да s и r нису на јединствен начин одређени, али увођењем неког додатног критеријума као на пример код интервала поверења чија је централна статистика имала χ^2–расподелу, може се превазићи неодређеност.

Пример 53. Нека је 3,1; 3,1; 4,5; 4,6; 4,6; 4,9; 4,9; 5,2; 5,2; 5,3; 5,3; 5,8; 6,0; 6,0; 6,2; 6,7; 7,1; 7,5; 7,9; 8,0; 8,2; 8,4; 8,5; 9,3; 9,7; 9,8; 9,9; 10,0; 10,0; 10,1; 10,4; 10,5; 11,6; 11,6; 11,7; 11,8; 12,4; 12,6; 12,9; 13,4 варијациони низ реализованог узорка обима 40. Одредити интервал поверења нивоа поверења 0,90 за квантил реда $p = 0,25$ (први квартил) на основу овог узорка.

У варијационом низу реализованог узорка траже се бројеви $x_{(r)}$ и $x_{(s)}$ који би били границе интервала поверења за $M_{0,25}$ са нивоом поверења $0,9$. Дакле, $n = 40$, $p = 0,25$ дају

$$\sum_{k=r}^{s-1} \binom{40}{k} \cdot 0,25^k \cdot 0,75^{40-k} = P\{x_{(r)} \leq X \leq x_{(s)}\} = P\{r \leq Y \leq s\} =$$

$$= P\left\{\frac{r - 40 \cdot 0,25}{\sqrt{40 \cdot 0,25 \cdot 0,75}} \leq \frac{Y - 10}{\sqrt{7,5}} \leq \frac{s - 10}{\sqrt{7,5}}\right\} = P\{c_1 \leq Y^* \leq c_2\} = 0,90\ ,$$

где је случајна променљива $Y : \mathcal{B}(n, p)$ и њена апроксимација нормалном расподелом $Y^* \equiv Z \sim \mathcal{N}(0, 1)$. Тражимо r и s тако да је

$$\frac{1}{\sqrt{2\pi}} \int_{c_1}^{c_2} e^{-\frac{z^2}{2}} dz = 0,90.$$

Ако се изабере $r = 6$ и $s = 15$ добија се

$$P\{-1,46 \leq Y^* \leq 1,83\} = 0,42786 + 0,46562 = 0,89348 \approx 0,9.$$

Дакле, тражени интервал има границе $x_{(6)} = 4,9$ и $x_{(15)} = 6,2$, тј. 90%-тни интервал поверења за непознати квартил реда $0,25$ је:

$$I_{M_{0,25}} = [4,9\,;\, 6,2].$$

Јасно да овај избор није јединствен, већ је само један од могућих избора. $\triangle$

3.8.3 Вишедимензионе области поверења

У многим практичним примерима једнодимензионог параметра, могуће је конструисати интервале поверења, али је то увек повезано са претпоставком о монотоности за централну статистику, као и претпоставком да је Θ интервал. Услов монотоности, међутим, није увек могуће остварити, а ако Θ не би био интервал, интервал поверења не би имао никаквог смисла.

Области поверења решавају општији проблем.

Поново посматрамо мерљив узорачки простор $\mathcal{X}$ за популацију са обележјем X чија расподела припада фамилији допустивих расподела $\mathcal{F}_{\boldsymbol{\theta}} = \{f(x; \boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$, при чему је $\boldsymbol{\Theta}$ вишедимензиони мерљив скуп, тј. подскуп одговарајућег вишедимензионог реалног простора, а $\mathbf{X} = (X_1, \dots, X_n)$ случајни узорак.

Нека је $K \subset \mathcal{X} \times \boldsymbol{\Theta}$, $K_{\boldsymbol{\theta}} = \{\mathbf{x} : (\mathbf{x}, \boldsymbol{\theta}) \in K\}$, $K(\mathbf{x}) = \{\boldsymbol{\theta} : (\mathbf{x}, \boldsymbol{\theta}) \in K\}$. Нека је $K_{\boldsymbol{\theta}}$ мерљив подскуп од $\mathcal{X}$ за свако $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Такође претпоставимо да је за свако $\mathbf{x} \in \mathcal{X}$, $K(\mathbf{x}) \neq \emptyset$, тачније, да је $\{\mathbf{X} : K(\mathbf{X}) = \emptyset\} \subset L$ за који важи да за свако $\boldsymbol{\theta} \in \boldsymbol{\Theta} : P_{\boldsymbol{\theta}}(L) = 0$. Нека је β реалан број, $0 < \beta < 1$, такав да је

$$\beta \leq \inf_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} P_{\boldsymbol{\theta}}(K_{\boldsymbol{\theta}}(\mathbf{X}))$$

где је

$$K_{\boldsymbol{\theta}}(\mathbf{X}) = \{\omega : \mathbf{X}(\omega) \in K_{\boldsymbol{\theta}}\}.$$

Тада је

$$P_{\boldsymbol{\theta}}\{\mathbf{X} \in K_{\boldsymbol{\theta}}\} \geq \beta\,,$$

а област $K_{\boldsymbol{\theta}}(\mathbf{X})$ ћемо звати област поверења за параметар $\boldsymbol{\theta}$ са нивоом поверења β.

Пример 54. Наћи област поверења за дводимензиони параметар нормалне расподеле $\mathcal{N}(m, \sigma^2)$ нивоа поверења β.

За област поверења дводимензионог параметра (m, σ^2) нормалне расподеле није могуће узети правоугаоник

$$\left[\overline{X}_n - \frac{\overline{S}_n}{\sqrt{n-1}} t_{n-1,\frac{\gamma_1}{2}}, \overline{X}_n + \frac{\overline{S}_n}{\sqrt{n-1}} t_{n-1,\frac{\gamma_1}{2}}\right] \times \left[\frac{n\overline{S}_n^{\,2}}{\chi^2_{n-1,\frac{1+\gamma_2}{2}}}, \frac{n\overline{S}_n^{\,2}}{\chi^2_{n-1,\frac{1-\gamma_2}{2}}}\right],$$

$\gamma_1 = 1-\alpha_1$, $\gamma_2 = 1-\alpha_2$, јер случајне величине којима је правоугаоник дефинисан $\frac{\overline{X}_n - m}{\overline{S}_n}$ и $\overline{S}_n^{\,2}$ нису независне.

Област поверења могуће је базирати на дводимензионој статистици $(\overline{X}_n, \overline{S}_n^{\,2})$ чије су компоненте независне.

Уочимо скуп:

$$K_{\boldsymbol{\theta}} = \left\{\mathbf{x} : \sqrt{\frac{n}{\sigma^2}}|\overline{x}_n - m| \leq t = z_{\frac{1-\alpha_1}{2}} \wedge t_1(\beta_2) = \chi^2_{n-1,\frac{\alpha_2}{2}} \leq \frac{n\overline{s}_n^{\,2}}{\sigma^2} \leq t_2(\beta_2) = \right.$$
$$\left. = \chi^2_{n-1,1-\frac{\alpha_2}{2}}\right\}.$$

Нека је $\beta = \beta_1\beta_2$ и $\Phi(z_{\frac{1-\alpha_1}{2}}) = \frac{1-\alpha_1}{2}$, где је $\beta_1 = 1-\alpha_1$ и $\beta_2 = 1-\alpha_2$. На основу независности $\overline{X}_n$ и $\overline{S}_n^{\,2}$ следи да је:

$$P_\theta\{\mathbf{X} \in K_\theta\} = P_\theta\left\{\frac{|\overline{X}_n - m|}{\sqrt{\frac{\sigma^2}{n}}} \leq z_{\frac{1-\alpha_1}{2}}\right\} P_\theta\left\{t_1(\beta_2) \leq \frac{n\overline{S}_n^{\,2}}{\sigma^2} \leq t_2(\beta_2)\right\} =$$
$$= \beta_1\beta_2 = \beta\,.$$

Област поверења ће бити облика

$$K_{\boldsymbol{\theta}} = \{\mathbf{x} : \mathbf{x} = (x_1, x_2)\}\,,$$

ограничена параболом $x_2 = \dfrac{(\overline{x}_n - x_1)^2 n}{z_{\frac{1-\alpha_1}{2}}}$, и двема паралелним правама

$$x_2 = \frac{n\overline{s}_n^{\,2}}{t_2(\beta_2)}\,, \quad x_2 = \frac{n\overline{s}_n^{\,2}}{t_1(\beta_2)}\,.$$

Прецизније

$$K_{\boldsymbol{\theta}} = \left\{ \mathbf{x} : \mathbf{x} = (x_1, x_2) \,\wedge\, x_2 \geq \frac{(\overline{x}_n - x_1)^2 n}{z_{\frac{1-\alpha_1}{2}}} \wedge x_2 \geq \frac{n\overline{s}_n^2}{t_2(\beta_2)} \wedge x_2 \leq \frac{n\overline{s}_n^2}{t_1(\beta_2)} \right\}.$$

$\triangle$

Глава 4

Тестирање статистичких хипотеза

Тестирање статистичких хипотеза представља вид статистичког закључивања који се примењује у ситуацијама у којима се унапред претпоставља постојање одређене везе међу изучаваним појавама или када се разматра расподела обележја којом се посматрана појава карактерише. У психологији се рецимо повезују: емоције и израз лица, утицај прве импресије и тумачење каснијих података, у контроли квалитета: повезује се производна смена са бројем дефектних производа и још многи други примери се могу дати у том смислу. Са друге стране, може да се изнесе било каква претпоставка о обележју које карактерише изучавану појаву, као, на пример: висина становништва једне регије (или у целини) прати нормалну расподелу и слично. Претпоставке могу да се односе и само на поједине карактеристике обележја као што је очекивање, медијана и слично. Свака од претпоставки може да буде тачна или погрешна.

4.1 Основни појмови

Дефиниција 38. Тврђење о посматраним појавама и процесима на једној или више популација, које може да се искаже као тврђење о расподели једног или више обележја је *статистичка хипотеза.* $\diamondsuit$

Другим речима, свака претпоставка да обележје $\mathbf{X}$ има расподелу која припада неком скупу допустивих расподела назива се статистичка хипотеза.

Дефиниција 39. *Тестирање статистичке хипотезе* је поступак провере хипотезе. ♢

Циљ тестирања хипотезе је њено прихватање или одбацивање при чему се верификација обавља неким унапред утврђеним статистичким методом.

Хипотеза се тестира на основу узорка и доноси се одлука о њеном прихватању или одбацивању, при чему се говори и о вероватноћи погрешног закључка.

Пример 55. Испитују се моторне способности радника запослених у два погона једне фабрике. При томе треба утврдити да ли постоји разлика у моторној снази шаке радника који припадају двема дефинисаним групама. С тим у вези природно је поставити две хипотезе:

- Постоје разлике у моторној снази шака радника у првом и другом погону, и
- Не постоје разлике у моторној снази шака радника у првом и другом погону. △

У овом случају поступак статистичког закључивања подразумева да се једна од постављених хипотеза узима за полазну или тзв. **нулту хипотезу** и означава се са H_0. Друга хипотеза назива се, у том случају, **алтернативна хипотеза** и означава са H_1, или, ређе, H_a. Која хипотеза од постављених се узима као нулта, зависи од самог проблема. Општи је принцип да се за нулту хипотезу по правилу узима она која се лакше верификује, тј. за коју је лакше утврдити вероватноће извођења погрешних закључака.

Хипотеза може бити проста или сложена.

Дефиниција 40. Статистичка хипотеза је *проста* ако у потпуности одређује расподелу обележја којом се бави, у противном је *сложена*. ♢

Наводимо неколико примера најраспрострањенијих математичких формулација статистичких хипотеза.

- Хипотеза о облику расподеле посматраног обележја X

$$H_0 : F_X(x) = F_0(x)\,, \quad x \in R\,,$$

где је F_0 потпуно одређена задата расподела или

$$H_0 : F_X \in \mathcal{F}\,,$$

где је $\mathcal{F}$ задата фамилија функција расподеле.

- Хипотеза о хомогености (истоврсности) којом се проверава једнакост расподела, рецимо k, случајних вектора истих димензија

$$H_0 : F_1(\mathbf{x}) = \ldots = F_k(\mathbf{x})\,, \quad \mathbf{x} = (x_1, \ldots, x_n) \in R^n\,,$$

где је F_i функција расподеле вектора $(X_{i1}, \ldots, X_{in}), i = 1, \ldots, k$. Оваква хипотеза се примењује, на пример, у ситуацијама када се за више узорака истог обима проверава да ли су узети из исте популације.

- Хипотеза о независности два обележја X и Y

$$H_0 : F(x, y) = F_X(x) F_Y(y)\,, \quad (x, y) \in R^2\,,$$

где су $F(x, y)\,, (x, y) \in R^2$ функција расподеле случајног вектора (X, Y), $F_X(x)\,, x \in R$, функција расподеле случајне променљиве X, а $F_Y(y)\,, y \in R$, функција расподеле случајне променљиве Y. Оваква хипотеза може бити исказана и за више од два обележја.

- Хипотеза о случајности се односи на вишедимензионо обележје или било који вектор случајних променљивих

$$\mathbf{X} = (X_1, \ldots, X_n)$$

и њоме се тестира да ли су компоненте X_i независне и јендако расподељене, тј. да ли је могуће разматрати $\mathbf{X}$ као прост случајни узорак из популације са обележјем ξ

$$H_0 : F_{\mathbf{X}}(\mathbf{x}) = F_\xi(x_1) \ldots F_\xi(x_n)\,, \quad \mathbf{x} = (x_1, \ldots, x_n) \in R^n\,,$$

где је F_ξ функција расподеле случајне променљиве ξ.

Правило за тестирање хипотезе H_0 је **статистички тест**. Правило тестирања најчешће користи неку статистику. Статистика чијим се посредством врши тестирање зове се **тест статистика**.

Статистички тестови код којих расподела тест статистике зависи од расподеле посматраног обележја су **параметарски тестови**, а тестови код којих расподела тест статистике не зависи од расподеле посматраног обележја су **непараметарски тестови**.

Параметарски тестови најчешће служе за проверу хипотезе о параметрима посматраних расподела, а непараметарски за проверу облика расподеле, зависности обележја (два или више), случајности низа догађаја и слично. Надаље ћемо се бавити неким конкретним тестовима из сваке од ових група.

Сваки реализовани узорак обима n, $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ дефинише једну тачку n–димензионог реалног еуклидског простора, $(x_1, x_2, \ldots, x_n) \in R^n$. При тестирању статистичких хипотеза по правилу се дефинише скуп $C \subset R^n$ који служи као критеријум за одбацивање, односно прихватање нулте хипотезе и то: ако $(x_1, x_2, \ldots, x_n) \in C$, тада се хипотеза H_0 одбацује, а ако $(x_1, x_2, \ldots, x_n) \in C^c$, тада нема разлога да се H_0 одбаци.

Скуп свих тачака $C \subset R^n$ за које се H_0 одбацује је **критична област теста**. Критична област се најчешће исказује преко критичне вредности тест статистике. Тако, ако је тест статистика једнодимензиона функција n–димензионог аргумента, посредством тест статистике се n–димензиона област простора R^n преводи у једнодимензиону, тј. $C \subset R$. При томе се, без опасности од забуне, користи иста ознака C.

Одлуком о прихватању или одбацивању нулте хипотезе могуће је начинити две врсте грешака. Могуће је да је нулта хипотеза одбачена, а да је она фактички тачна. Таквим закључивањем чини се **грешка прве врсте**. Вероватноћа да се учини грешка прве врсте се најчешће означава са α. Ситуација при којој се чини ова грешка, у поступку у коме се примењује критична област, настаје када реализовани узорак припадне критичној области, иако је H_0 тачна, па се H_0 одбаци. Према томе, α може да се изрази као:

$$\alpha = P_{H_0}\{(X_1, X_2, \ldots, X_n) \in C\},$$

како се најчешће означава чињеница да је α условна вероватноћа

$$\alpha = P\{(X_1, X_2, \ldots, X_n) \in C | H_0\}.$$

Тада се још каже да је C **критична област величине** α или да је α **величина критичне области** C. У пракси се величина критичне области исказује у процентима, а најчешће коришћене вредности за α су 1% и 5%, тј. $\alpha = 0,01$ и $\alpha = 0,05$.

Вероватноћа α се зове **праг значајности** или **ниво значајности** теста.

Грешка друге врсте чини се када се нулта хипотеза прихвати, а заправо није тачна. То се дешава ако реализовани узорак не припадне критичној области, а нулта хипотеза фактички није тачна. Вероватноћа да се начини грешка друге врсте најчешће се означава са β и може да се изрази као:

$$\beta = P_{H_1}\{(X_1, X_2, \ldots, X_n) \in C^c\},$$

односно

$$\beta = P\{(X_1, X_2, \ldots, X_n) \in C^c | H_1\}.$$

Дакле, грешка друге врсте чини се када је фактички тачна алтернативна хипотеза H_1, а прихвати се хипотеза H_0.

Шематски се вероватноће правилних и погрешних одлука приказују на следећи начин:

стварна ситуација $\rightarrow$ одлука $\downarrow$	H_0	H_1
H_0	$1-\alpha$	β
H_1	α	$1-\beta$

Природно је да тежимо да нађемо тест, тј. критичну област C, такву да вероватноће грешака, α и β, буду што мање. Такав захтев у општем случају је противуречан, јер најчешће смањивање α доводи до повећања β, и обратно. Отуда се у статистици поступа тако што се једна од двеју вероватноћа α или β фиксира (најчешће α), а онда се друга одреди да буде најмања могућа у задатим условима тестирања. Каже се да се одређује **најбоља критична област** величине α, када се за фиксирано α одређује критична област за коју је β најмање међу свим критичним областима величине α. У том смислу скуп C мора да задовољи одређене критеријуме оптималности о којима ће надаље бити још речи.

Поменимо овде и термин **значајност теста** који је, по правилу, у вези са једнодимензионом критичном облашћу везаном за одређену тест статистику. Наиме, под значајношћу теста подразумева се величина критичне области чија је граница реализована вредност тест статистике у конкретно решаваном проблему тестирања поједине статистичке хипотезе. У том смислу се у комуникацијама, односно размени информација о неком експерименту, користи израз "значајност већа (рецимо) од 5%", када се нулта хипотеза прихвата, или "значајност мања од 5%", па се нулта хипотеза одбацује.

Уместо термина значајност, данас је све више у употреби термин p–**вредност**.

Посматрајмо надаље обележје X чија расподела припада фамилији допустивих расподела $\{F(x;\theta), \theta \in \Theta\}$. Статистичком хипотезом се често дефинише подскуп овог скупа расподела као скуп допустивих расподела. Формално, нека је $\Lambda \subset \Theta$. Тада се нулта и алтернативна хипотеза могу исказати на следећи начин

$$H_0 : \theta \in \Lambda\,, \quad H_1 : \theta \in \Delta \qquad \text{за неки скуп} \quad \Delta \subseteq \Lambda^c = \Theta \setminus \Lambda\,. \tag{4.1}$$

Ако је Λ једночлан скуп, нулта хипотеза је проста, у противном она је сложена. На исти начин се може говорити и о алтернативној хипотези.

О вероватноћи одбацивања нулте хипотезе може да се говори и у терминима функције моћи теста.

Дефиниција 41. *Функција моћи* статистичког теста за тестирање нулте хипотезе H_0 против алтернативне H_1 је вероватноћа да узорак припадне критичној области теста ако је H_1 тачна, тј. вероватноћа одбацивања нулте хипотезе ако је она фактички погрешна, $P_{H_1}\{(X_1, \ldots, X_n) \in C\} = 1 - \beta$. $\diamondsuit$

Уколико се хипотезе исказују у терминима избора расподеле из фамилије допустивих расподела избором вредности параметра као што је наведено у (4.1), функција моћи се може посматрати као функција од θ

$$M(\theta) = P\{(X_1, \ldots, X_n) \in C\}\,, \tag{4.2}$$

где је θ вредност параметра.

Вредност функције моћи за поједину вредност параметра θ (или у општем случају за просту нулту хипотезу) зове се **моћ теста**.

Насупрот функцији моћи је функција коју зовемо оперативном карактеристиком теста, која даје вероватноћу прихватања нулте хипотезе, тј. вероватноћу да узорак не припадне критичној области.

Дефиниција 42. *Оперативна карактеристика теста* је функција која даје вероватноћу прихватања нулте хипотезе ако, $P\{(X_1,\ldots,X_n)\in C^c\}$. $\diamondsuit$

Дакле, ако се хипотезе исказују вредношћу параметра,

$$N(\theta)=1-M(\theta) \quad , \quad \theta\in\Theta,$$

где је $M(\theta)$ дефинисано у (4.2), односно

$$N(\theta)=P\{(X_1,\ldots,X_n)\in C^c\}.$$

Пример 56. Нека је $\Theta=\{\theta_0,\theta_1\}$ двочлани скуп. Нека су нулта и алтернативна хипотеза редом $H_0:\theta=\theta_0$ и $H_1:\theta=\theta_1$. Тада је моћ теста за вредност аргумента $\theta=\theta_0$ праг значајности теста,

$$M(\theta_0)=P_{\theta_0}\{(X_1,\ldots,X_n)\in C\}=\alpha,$$

а за вредност аргумента $\theta=\theta_1$ оперативна карактеристика теста ће бити вероватноћа грешке друге врсте

$$N(\theta_1)=1-M(\theta_1)=P_{\theta_1}\{(X_1,\ldots,X_n)\in C^c\}=\beta\,.\,\triangle$$

Ефективни поступак за одређивање најбоље критичне области задате величине α за тестирање нулте просте против алтернативне просте хипотезе даје следећа теорема позната као теорема Нејман-Пирсона.

Теорема 20. (Нејман–Пирсон) *Нека је* $(X_1,\ldots,X_n)$ *узорак из популације са обележјем* X *чија расподела припада фамилији допустивих расподела* $\{f(x;\theta),\theta\in\Theta\}$, $\Theta=\{\theta_0,\theta_1\}$, $\theta_0\neq\theta_1$ *и нека је* $L(\theta;x_1,\ldots,x_n)$ *функција веродостојности посматраног узорка. Нека је изабран реалан број* $k>0$ *и нека је скуп* $C\subset\mathbf{R}^n$ *такав да важи:*

- *(1)*
$$\frac{L(\theta_0;x_1,\ldots,x_n)}{L(\theta_1;x_1,\ldots,x_n)}\le k \quad \text{за свако} \quad (x_1,\ldots,x_n)\in C$$
- *(2)*
$$\frac{L(\theta_0;x_1,\ldots,x_n)}{L(\theta_1;x_1,\ldots,x_n)}\ge k \quad \text{за свако} \quad (x_1,\ldots,x_n)\in C^c$$
- *(3)*
$$\alpha=P_{\theta_0}\{(X_1,\ldots,X_n)\in C\}.$$

Тада је C најбоља критична област величине α за тестирање нулте просте хипотезе $H_0(\theta = \theta_0)$ против алтернативне, такође просте хипотезе, $H_1(\theta = \theta_1)$.

Доказ. Доказ наводимо само за апсолутно непрекидно обележје, а у дискретном случају доказ је аналоган.

Ако је C, дефинисана условима теореме, једина критична област величине α, доказ је завршен. Ако постоји још нека критична област величине α означимо је са A,

$$A \subset \mathbf{R}^n \quad \text{таква да је} \quad P_{\theta_0}\{(X_1, \ldots, X_n) \in A\} = \alpha\,,$$

треба показати да је вероватноћа грешке друге врсте мања за област C него за област A, тј. да важи:

$$P_{\theta_1}\{(X_1, \ldots, X_n) \in C^c\} \leq P_{\theta_1}\{(X_1, \ldots, X_n) \in A^c\}$$

што је еквивалентно са

$$P_{\theta_1}\{(X_1, \ldots, X_n) \in C\} \geq P_{\theta_1}\{(X_1, \ldots, X_n) \in A\}\,. \tag{4.3}$$

Означимо са:

$$\int\int \cdots \int_B L(\theta_i; x_1, \ldots, x_n) dx_1 \ldots dx_n = \int_B L(\theta_i) \quad , \quad i = 0, 1$$

за произвољну обаласт $B \subseteq R^n$. Треба, дакле, показати да је:

$$\int_C L(\theta_1) - \int_A L(\theta_1) \geq 0.$$

Како је $C = (C \cap A) \cup (C \cap A^c)$ и $A = (A \cap C) \cup (A \cap C^c)$, добијамо

$$\int_C L(\theta_1) - \int_A L(\theta_1) = \int_{C \cap A^c} L(\theta_1) - \int_{A \cap C^c} L(\theta_1).$$

Како је с друге стране $L(\theta_1; x_1, \ldots, x_n) \geq \frac{1}{k} L(\theta_0; x_1, \ldots, x_n)$ за свако $(x_1, \ldots, x_n) \in C$, иста неједнакост важи и за свако $(x_1, \ldots, x_n) \in C \cap A^c$, па је

$$\int_{C \cap A^c} L(\theta_1) \geq \frac{1}{k} \int_{C \cap A^c} L(\theta_0).$$

Слично је $L(\theta_1; x_1, \ldots, x_n) \leq \frac{1}{k} L(\theta_0; x_1, \ldots, x_n)$ за свако $(x_1, \ldots, x_n) \in A \cap C^c$ па је

$$\int_{A \cap C^c} L(\theta_1) \leq \frac{1}{k} \int_{A \cap C^c} L(\theta_0).$$

Тако добијамо да је

$$\int_{C\cap A^c} L(\theta_1) - \int_{A\cap C^c} L(\theta_1) \geq \frac{1}{k}\left[\int_{C\cap A^c} L(\theta_0) - \int_{A\cap C^c} L(\theta_0)\right].$$

Међутим, с друге стране је

$$\int_{C\cap A^c} L(\theta_0) - \int_{A\cap C^c} L(\theta_0) = \int_{C\cap A^c} L(\theta_0) + \int_{C\cap A} L(\theta_0) - \int_{C\cap A} L(\theta_0) - \\ - \int_{A\cap C^c} L(\theta_0) = \int_C L(\theta_0) - \int_A L(\theta_0) = 0\,.$$

Дакле, следи да је

$$\int_C L(\theta_1) - \int_A L(\theta_1) \geq \frac{1}{k}\left[\int_{C\cap A^c} L(\theta_0) - \int_{A\cap C^c} L(\theta_0)\right] = 0\,,$$

што је и требало доказати. □

Приметимо следеће. Како су θ_0 и θ_1 познати бројеви, то је $\frac{L(\theta_0;X_1,\ldots,X_n)}{L(\theta_1;X_1,\ldots,X_n)}$ једна статистика, а како је C скуп свих могућих вредности $(x_1,\ldots,x_n)$ узорка $(X_1,\ldots,X_n)$ за које је

$$\frac{L(\theta_0;x_1,\ldots,x_n)}{L(\theta_1;x_1,\ldots,x_n)} \leq k\,,$$

величину критичне области α можемо одредити из

$$\alpha = P_{\theta_0}\{(X_1,\ldots,X_n)\in C\} = P_{\theta_0}\left\{\frac{L(\theta_0;X_1,\ldots,X_n)}{L(\theta_1;X_1,\ldots,X_n)} \leq k\right\}\,,$$

као што из исте једнакости за задато α можемо одредити k. У апсолутно непрекидном случају је то могуће одредити на јединствен начин, док ће код дискретног обележја, могуће, бити потребна додатна информација.

Приметимо, такође, да у доказу теореме није коришћена чињеница да су хипотезе биле дефинисане избором параметра расподеле. Осим тога, истакнимо да теорема важи и за узорак који није прост, што такође следи из доказа. Наведимо један пример одређивања најбоље критичне области Нејман–Пирсоновом теоремом код кога су две просте хипотезе другачије дефинисане.

Пример 57. Нека је дат прост случајни узорак $\mathbf{X} = (X_1,\ldots,X_n)$ из популације са обележјем X чија расподела припада скупу допустивих расподела

$$\{f_1(x), f_2(x)\}\,,$$

$$f_1(x) = \begin{cases} \frac{e^{-1}}{x!}, & x = 0, 1, 2, \ldots \\ 0, & \text{иначе} \end{cases}$$

$$f_2(x) = \begin{cases} (\frac{1}{2})^{x+1}, & x = 0, 1, 2, \ldots \\ 0, & \text{иначе} \end{cases} .$$

Поставимо две просте хипотезе

$$H_0 : f = f_1 \,, \quad H_1 : f = f_2 \,.$$

Количник одговарајућих функција веродостојности ће тада бити

$$\frac{g(x_1, \ldots, x_n)}{h(x_1, \ldots, x_n)} = \frac{\frac{e^{-n}}{x_1! \ldots x_n!}}{(\frac{1}{2})^n (\frac{1}{2})^{x_1 + \ldots + x_n}} = \frac{(2e^{-1})^n 2^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n (x_i!)} \,.$$

За реалан број $k > 0$ размотримо неједнакост

$$\frac{g(x_1, \ldots, x_n)}{h(x_1, \ldots, x_n)} \leq k,$$

тј.

$$\frac{(2e^{-1})^n 2^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n (x_i!)} \leq k \,.$$

Из ње следи

$$n \ln 2 - n + \left(\sum_{i=1}^n x_i \right) \ln 2 - \sum_{i=1}^n \ln(x_i!) \leq \ln k$$

$$\left(\sum_{i=1}^n x_i \right) \ln 2 - \sum_{i=1}^n \ln(x_i!) \leq \ln k + n - n \ln 2$$

$$\sum_{i=1}^n \ln \left(\frac{2^{x_i}}{x_i!} \right) \leq k_1 \,.$$

Дакле, најбоља критична област величине α за дефинисано тестирање је облика

$$C = \left\{ (x_1, x_2, \ldots, x_n) : \sum_{i=1}^n \ln \left(\frac{2^{x_i}}{x_i!} \right) \leq c \right\} . \triangle$$

Још један аспект ове теореме завређује посебну пажњу. Ради се о броју параметара који се јавља у расподели посматраног обележја, тј. о димензији параметра. Пажљивом анализом доказа види се да димензија параметра није од значаја за доказ, нити се исказ теореме везује за димензију параметра. Дакле, густина расподеле обележја које је предмет тестирања може да зависи од параметра произвољне димензије, односно од произвољног коначног броја параметара. Оно што је битно, то је да су обе, и нулта и алтернативна хипотеза, просте, тј. да у потпуности одређују расподелу.

4.1.1 Униформно најмоћнији тестови

Надаље ћемо разматрати могућности за тестирање нулте просте против алтернативне сложене хипотезе. Размотримо пажљиво следећи пример.

Пример 58. Нека обележје X има густину расподеле која припада фамилији $\{\mathcal{N}(\theta,1), \theta \in \{\theta_0, \theta_1\}, \theta_0 < \theta_1\}$. Применом Нејман–Пирсонове теореме одредити најбољу критичну област за тестирање просте нулте хипотезе $H_0 : \theta = \theta_0$ против просте алтернативне $H_1 : \theta = \theta_1$ на основу простог случајног узорка $\mathbf{X} = (X_1, \ldots, X_n)$.

Функција веродостојности је дата са:

$$L(\theta_j; x_1, \ldots, x_n) = \frac{1}{(\sqrt{2\pi})^n} \exp\left\{-\sum_{i=1}^{n} \frac{(x_i - \theta_j)^2}{2}\right\}, \quad j = 0, 1\,.$$

Да бисмо нашли критичну област C по теореми Нејман–Пирсона, посматрамо следећу неједнакост:

$$\frac{L(\theta_0)}{L(\theta_1)} = \frac{\frac{1}{(\sqrt{2\pi})^n} \exp\left\{-\sum_{i=1}^{n} \frac{(x_i-\theta_0)^2}{2}\right\}}{\frac{1}{(\sqrt{2\pi})^n} \exp\left\{-\sum_{i=1}^{n} \frac{(x_i-\theta_1)^2}{2}\right\}} =$$

$$= \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left[(x_i - \theta_0)^2 - (x_i - \theta_1)^2\right]\right\} \le k.$$

Из ове неједнакости следи да је

$$-\frac{1}{2}\sum_{i=1}^{n}\left[(x_i - \theta_0)^2 - (x_i - \theta_1)^2\right] \le \ln k\,.$$

После краћег рачунања имаћемо низ неједнакости

$$(\theta_0 - \theta_1) \sum_{i=1}^{n} (\theta_0 + \theta_1 - 2x_i) \geq -2 \ln k \,,$$

$$\sum_{i=1}^{n} (\theta_0 + \theta_1) - 2 \sum_{i=1}^{n} x_i \leq \frac{2 \ln k}{\theta_1 - \theta_0} \,,$$

$$\frac{1}{n} \sum_{i=1}^{n} x_i \geq \frac{\ln k}{(\theta_0 - \theta_1) n} + \frac{\theta_0 + \theta_1}{2} \,,$$

$$\frac{1}{n} \sum_{i=1}^{n} x_i \geq c \,, \quad c = const \,. \, \triangle$$

Обратимо пажњу да је за облик критичне области у наведеном примеру од значаја била чињеница да је $\theta_0 < \theta_1$. Међутим, још важније је уочити да би критична област задржала овај облик (не и вредност константе c) и за сваки други број $\theta_1 \in (\theta_0, +\infty)$.

Ограничење $\theta_0 < \theta_1$ има за последицу да смо били у могућности да на јединствен начин одредимо најбољу критичну област. С тим у вези можемо проблем тестирања да поставимо на следећи начин.

Пример 59. Нека обележје X има густину расподеле која припада фамилији $\{\mathcal{N}(\theta, 1), \theta \in R\}$. Тестирати хипотезу $H_0 : \theta = \theta_0$ против алтернативне $H_1 : \theta > \theta_0$ са прагом значајности α на основу простог случајног узорка $\mathbf{X} = (X_1, \ldots, X_n)$.

Из претходног примера следи да ће за свако фиксирано $\theta_1 \in (\theta_0, +\infty)$ критична област величине α бити одређена као

$$C = [c, +\infty) \,,$$

где је c одређено по критеријуму

$$\alpha = P \left\{ \sum_{i=1}^{n} X_i \geq c \right\} \,.$$

Прецизније,

$$C = \left\{ (x_1, \ldots, x_n) : \sum_{i=1}^{n} x_i \geq \frac{n}{2} (\theta_1 + \theta_0) - \frac{\ln k}{\theta_1 - \theta_0} \right\} .$$

Констатујемо да ће према Нејман–Пирсоновој теореми област C бити најбоља критична област за тестирање нулте просте против сваке алтернативне просте хипотезе садржане у алтернативној сложеној хипотези. $\triangle$

Дефиниција 43. Критична област C је *униформно најмоћнија област* величине α за тестирање просте хипотезе H_0 против алтернативне сложене хипотезе H_1 ако је скуп C најбоља критична област величине α за тестирање H_0 против сваке просте хипотезе садржане у H_1. Тест дефинисан овом критичном облашћу, зове се *униформно најмоћнији тест* са прагом значајности α за тестирање просте хипотезе H_0 против алтернативне сложене H_1. $\diamondsuit$

Униформно најмоћнији тест не мора увек да постоји, међутим, када постоји, Нејман–Пирсонова теорема даје технику за његово налажење.

Пример 60. Посматрајмо обележје X и фамилију његових допустивих расподела дефинисану у претходном примеру. Може ли се одредити униформно најмоћнији тест за тестирање

$$H_0 : \theta = \theta_0 \quad \text{против} \quad H_1 : \theta \neq \theta_0 \ ?$$

Можемо констатовати да ако дефинишемо нулту и алтернативну хипотезу на следећи начин

$$H_0 : \theta = \theta_0 \quad , \quad H_1 : \theta \neq \theta_0$$

неће постојати униформно најмоћнија област, па ни униформно најмоћнији тест. Заиста, за $\theta_1 < \theta_0$ критична област ће бити одређена са

$$\sum_{i=1}^{n} x_i \leq \frac{n}{2}(\theta_1 + \theta_0) - \frac{\ln k}{\theta_1 - \theta_0} \, . \ \triangle$$

4.1.2 Тест количника веродостојности

Поставља се питање можемо ли тестирати нулту сложену хипотезу против алтернативне такође сложене. У ту сврху користимо интуитивни тест, тест количника веродостојности, који користи идеју теореме Нејман–Пирсона.

Пример 61. Нека је дато обележје $X : \{\mathcal{N}(\theta_1, \theta_2), -\infty < \theta_1 < +\infty, \theta_2 > 0\}$. Тестирати нулту сложену хипотезу $H_0 : (\theta_1 = 0, \theta_2 > 0)$ против алтернативне сложене хипотезе $H_1 : (\theta_1 \neq 0, \theta_2 > 0)$ на основу простог случајног узорка $\mathbf{X} = (X_1, \ldots, X_n)$.

Посматрајмо параметарске просторе дефинисане на следећи начин:

$$A_0 = \{(\theta_1, \theta_2), \theta_1 = 0, \theta_2 > 0\}$$

$$\Theta = \{(\theta_1, \theta_2), -\infty < \theta_1 < +\infty, \theta_2 > 0\}\,.$$

Сада се постављене хипотезе могу исказати у терминима ових простора

$$H_0 : (\theta_1, \theta_2) \in A_0 \quad \text{и} \quad H_1 : (\theta_1, \theta_2) \in A_0^c\,,$$

где је $A_0^c = \Theta \setminus A_0$.

Пођимо од функције веродостојности за постављене хипотезе, која је сада функција две променљиве, односно од дводимензионог параметра, разликујући јој вредност на скупу Θ и скупу A_0:

$$L(\theta_1, \theta_2; x_1, \ldots, x_n) = \frac{1}{(2\pi\theta_2)^{\frac{n}{2}}} \exp\left\{-\frac{1}{2\theta_2}\sum_{i=1}^{n}(x_i - \theta_1)^2\right\} = L(\Theta)$$

$$L(\theta_1, \theta_2; x_1, \ldots, x_n) = \frac{1}{(2\pi\theta_2)^{\frac{n}{2}}} \exp\left\{-\frac{1}{2\theta_2}\sum_{i=1}^{n}x_i^2\right\} = L(A_0)\,.$$

Уочимо количник $\frac{L(A_0)}{L(\Theta)}$. Како аналитички изрази за $L(A_0)$ и $L(\Theta)$ зависе од непознатих параметара, то се једноставном заменом реализованог узорка случајним неће добити статистике. Међутим, баш због поменуте зависности има смисла одређивати максимум функција $L(A_0)$ и $L(\Theta)$ по $\theta = (\theta_1, \theta_2)$ на A_0 и на Θ, редом, за сваки од реализованих узорака $\mathbf{x} = (x_1, \ldots, x_n) \in \mathcal{X}$, па се затим посматра следећи количник

$$\frac{\max_{(\theta_1,\theta_2)\in A_0} L(A_0)}{\max_{(\theta_1,\theta_2)\in \Theta} L(\Theta)} = \frac{L(\hat{A}_0)}{L(\hat{\Theta})} = \lambda = \lambda(x_1, \ldots, x_n)\,,$$

чији аналитички израз неће зависити од непознатих параметара.

Спроведени поступак, заправо, значи замену непознатих параметара њиховим оценама максималне веродостојности.

Добијени количник се зове **количник веродостојности**. Како је $A_0 \subset \Theta$, онда је количник мањи од јединице, па је $0 \le \lambda \le 1$. За нулту хипотезу су проблематичне мале вредности овог количника, односно, функције $\lambda(x_1, \ldots, x_n)$, $(x_1, \ldots, x_n) \in \mathcal{X}$, јер указују на то да је веродостојност H_0 мала у поређењу са H_1, тј. да подаци фаворизују H_1 у односу на H_0.

Нека је λ_0 позитиван прави разломак. Принцип тестирања количником веродостојности налаже да се хипотеза $H_0 : (\theta_1, \theta_2) \in A_0$ одбаци ако и само ако је

$$\lambda(x_1, \ldots, x_n) = \lambda \le \lambda_0$$

за неко $0 < \lambda_0 < 1$. Функција $\lambda(X_1, \ldots, X_n)$ је случајна променљива, тј. статистика, па је праг значајности овог теста дат са

$$\alpha = P_{H_0}\{\lambda(X_1, \ldots, X_n) \leq \lambda_0\}.$$

Остаје још да пример довршимо ефективним одређивањем максимума функција

$$\ln L(A_0) = -\frac{n}{2}\ln(2\pi\theta_2) - \frac{1}{2\theta_2}\sum_{i=1}^{n} x_i^2$$

и

$$\ln L(\Theta) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\theta_2 - \frac{1}{2\theta_2}\sum_{i=1}^{n}(x_i - \theta_1)^2.$$

Уобичајеним поступком одређивања парцијалних извода по θ_1 и θ_2 следи:

$$\frac{\partial \ln L(A_0)}{\partial \theta_2} = -\frac{n}{2}\frac{1}{\theta_2} + \frac{1}{2\theta_2^2}\sum_{i=1}^{n} x_i^2 = 0\,,$$

$$\theta_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2\,.$$

Ову вредност уврстимо у израз за $L(A_0)$ да бисмо добили $L(\hat{A}_0)$:

$$L(\hat{A}_0) = \left(\frac{2\pi}{n}\sum_{i=1}^{n} x_i^2\right)^{-\frac{n}{2}} e^{-\frac{n}{2}} = \left(\frac{ne^{-1}}{2\pi\sum_{i=1}^{n} x_i^2}\right)^{\frac{n}{2}}.$$

Ово исто урадимо и за $L(\hat{\Theta})$, дакле,

$$\frac{\partial \ln L(\Theta)}{\partial \theta_1} = \frac{\sum_{i=1}^{n}(x_i - \theta_1)}{\theta_2}$$

$$\frac{\partial \ln L(\Theta)}{\partial \theta_2} = \frac{\sum_{i=1}^{n}(x_i - \theta_1)^2}{2\theta_2^2} - \frac{n}{2\theta_2}\,.$$

Изједначавањем ових извода са нулом, добићемо вредности за θ_1 и θ_2 које износе редом:

$$\theta_1 = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}_n$$

$$\theta_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2 = \overline{s}_n{}^2.$$

Коначно се добија да је:

$$\lambda = \frac{1}{\left(1 + \frac{n\overline{x}_n{}^2}{\sum_{i=1}^{n}(x_i - \overline{x}_n)^2}\right)^{\frac{n}{2}}}, \quad \lambda \leq \lambda_0$$

односно, даље

$$\frac{n\overline{x}_n{}^2}{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2} \geq (n-1)(\lambda_0^{-\frac{2}{n}} - 1),$$

тј.

$$\frac{\sqrt{n}|\overline{x}_n|}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2}} \geq \sqrt{(n-1)(\lambda_0^{-\frac{2}{n}} - 1)} = c\,.$$

У овом конкретном примеру добили смо статистику

$$\frac{|\overline{X}_n|}{\sqrt{\widetilde{S}_n^2}}\sqrt{n}$$

која има студентову расподелу са $n-1$ степени слободе и која омогућава да n–димензиону критичну област теста преведемо на једнодимензиону, а која ће бити унија интервала, $(-\infty, -c] \cup [c, +\infty)$. Дакле, кад израчунамо c тражимо λ_0 и враћамо се на критичну област $C = [0, \lambda_0]$. $\triangle$

Тест количника веродостојности је интуитивни тест и не постоји строги разлог за његову примену у смислу да је униформно најмоћнији. Међутим, за већину практичних проблема описаним поступком количника веродостојности се добија најмоћнији могући тест за задате услове. Нажалост, за статистику $\lambda(X_1, \ldots, X_n)$ овог теста се увек не налази расподела међу познатим расподелама. Два најчешћа услова која су потребна да би се одредила расподела тест статистике до које се долази методом количника веродостојности су:

- регуларност (можда и у некој ослабљеној форми, али се углавном ови услови базирају на диференцијабилности) и

- да област у којој је функција веродостојности строго позитивна **не зависи** од непознатих вредности параметара.

Размотримо чињеницу да смо до резултата добијеног у претходном примеру могли доћи и другачијим размишљањем.

Пример 62. Нека расподела обележја X припада фамилији допустивих расподела

$$\{\mathcal{N}(m,\sigma^2),\, m \in (-\infty,+\infty),\, \sigma^2 \in (0,\infty)\}.$$

На основу простог узорка обима n тестирати нулту хипотезу $H_0 : m = m_0$ против алтернативне $H_1 : m \neq m_0$.

У вези са тачкастим оценама параметара, навели смо да је статистика (Z_1, Z_2) дводимензиона комплетна довољна статистика за дводимензиони параметар (m, σ^2) где је $Z_1 = \overline{X}_n$ и $Z_2 = \frac{n}{n-1}\overline{S}_n^2$. Дакле, наш тест се мора да базира на овим статистикама, па ће количник функција веродостојности бити:

$$\frac{L(m_0; x_1, \ldots, x_n)}{L(m_1; x_1, \ldots, x_n)} = \varphi(z_1, z_2) = \varphi(z_1(x_1, \ldots, x_n), z_2(x_1, \ldots, x_n))$$

за неко $m_1 \neq m_0$.

Користећи се овом чињеницом, прибећићемо поступку који је једноставнији од директног израчунавања.

Доказали смо раније да су $\overline{X}_n$ и $\overline{S}_n^2$ независне случајне променљиве па су такве и

$$\frac{\overline{X}_n - m_0}{\frac{\sigma}{\sqrt{n}}} \quad , \quad \frac{n\overline{S}_n^2}{\sigma^2}\,.$$

За последње знамо да имају следеће расподеле:

$$\frac{\overline{X}_n - m_0}{\frac{\sigma}{\sqrt{n}}} : \mathcal{N}(0,1) \quad , \quad \frac{n\overline{S}_n^2}{\sigma^2} : \chi^2_{n-1}.$$

Њихов количник има Студентову расподелу са $n-1$ степени слободе:

$$t_{n-1} = \frac{\frac{\overline{X}_n - m_0}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{\frac{n\overline{S}_n^2}{\sigma^2}}{n-1}}} = \frac{\overline{X}_n - m_0}{\overline{S}_n}\sqrt{n-1} \quad , \quad \overline{S}_n = \sqrt{\overline{S}_n^2}\,.$$

С обзиром на симетричност студентове расподеле око своје очекиване вредности, дакле око нуле, и чињенице да ако је m_0 тачна вредност параметра m,

$$E_{m_0}\left(\frac{\overline{X}_n - m_0}{\overline{S}_n}\sqrt{n-1}\right) = E(t_{n-1}) = 0 \quad \text{и} \quad H_1(m \neq m_0)\,,$$

критичну област C изабраћемо на следећи начин:

$$C = \left\{ (x_1, \ldots, x_n) : \frac{|\overline{x}_n - m_0|}{\overline{s}_n} \sqrt{n-1} \geq k \right\}, \quad k > 0\,.$$

Дакле, за дату вредност α број k одређујемо из услова:

$$P_{m_0} \left\{ \frac{|\overline{X}_n - m_0|}{\overline{S}_n} \sqrt{n-1} \geq k \right\} = P\{|t_{n-1}| \geq k\} = \alpha\,,$$

где k одређујемо из таблице за Студентову расподелу. $\triangle$

Начинимо сада генерализацију.

Нека је $(X_1, \ldots, X_n)$ узорак обима n из популације са обележјем X чија расподела припада фамилији допустивих расподела $\{f(x;\theta_1, \theta_2, \ldots, \theta_m), (\theta_1, \theta_2, \ldots, \theta_m) \in \Theta\}$. Задржаћемо се на параметарском простору $\Theta = \{(\theta_1, \theta_2, \ldots, \theta_m)\}$. Нека је $A_0 \subset \Theta$. Желимо да тестирамо (просту или сложену) хипотезу $H_0 : (\theta_1, \theta_2, \ldots, \theta_m) \in A_0$, $dim A_0 = k < m$, против свих алтернативних. Дефинишимо функције веродостојности посматраног узорка

$$L(A_0) = \varphi(x_1, \ldots, x_n; \theta_1, \theta_2, \ldots, \theta_m)\,, \quad (\theta_1, \theta_2, \ldots, \theta_m) \in A_0$$

и

$$L(\Theta) = \varphi(x_1, \ldots, x_n; \theta_1, \theta_2, \ldots, \theta_m)\,, \quad (\theta_1, \theta_2, \ldots, \theta_m) \in \Theta\,.$$

Нека су, као и раније, $L(\widehat{A}_0)$ и $L(\widehat{\Theta})$ максимуми горњих функција по $(\theta_1, \theta_2, \ldots, \theta_m) \in A_0$ и $(\theta_1, \theta_2, \ldots, \theta_m) \in \Theta$ редом (за које претпостављамо да постоје). Посматрамо као и раније количник

$$\frac{L(\widehat{A}_0)}{L(\widehat{\Theta})} = \lambda = \lambda(x_1, \ldots, x_n)$$

који се зове **количник веродостојности**. Нека је λ_0 позитиван реалан број мањи од јединице. Принцип тестирања количником веродостојности налаже да се хипотеза $H_0 : (\theta_1, \theta_2, \ldots, \theta_m) \in A_0$ одбаци ако и само ако

$$\lambda(x_1, \ldots, x_n) = \lambda \leq \lambda_0\,.$$

Функција $\lambda(X_1, \ldots, X_n)$ је случајна променљива, па је праг значајности овог теста дат са

$$\alpha = P_{H_0}\{\lambda(X_1, \ldots, X_n) \leq \lambda_0\}\,.$$

Следећи пример илуструје генерализацију.

Пример 63. Нека су обележја X и Y независна са расподелама $\mathcal{N}(\theta_1, \theta_3)$ и $\mathcal{N}(\theta_2, \theta_3)$ редом, где су θ_1, θ_2 и θ_3 непознати параметри дефинисани параметарским простором

$$\Theta = \{(\theta_1, \theta_2, \theta_3); -\infty < \theta_1 < \infty, -\infty < \theta_2 < \infty, 0 < \theta_3 < \infty\}.$$

Нека су $(X_1, \ldots, X_n)$ и $(Y_1, \ldots, Y_m)$ независни прости узорци из ових расподела. Нека је $A_0 = \{(\theta_1, \theta_2, \theta_3); -\infty < \theta_1 = \theta_2 < \infty, 0 < \theta_3 < \infty\}$. Тестирати хипотезу $H_0 : (\theta_1, \theta_2, \theta_3) \in A_0$ против свих алтернативних.

Функција веродостојности се формира из простог узорка обима $n + m > 2$, $(X_1, \ldots, X_n, Y_1, \ldots, Y_m)$ па је,

$$L(\Theta) = \left(\frac{1}{2\pi\theta_3}\right)^{\frac{n+m}{2}} \exp\left\{-\frac{\sum_{i=1}^n (x_i - \theta_1)^2 + \sum_{j=1}^m (y_j - \theta_2)^2}{2\theta_3}\right\}$$

и

$$L(A_0) = L(\Theta)\Big|_{\theta_1 = \theta_2}.$$

Решење се добија уобичајеним одређивањем максимума из једначина добијених помоћу парцијалних извода

$$\frac{\partial \ln L(A_0)}{\partial \theta_1}, \frac{\partial \ln L(A_0)}{\partial \theta_3}, \frac{\partial \ln L(\Theta)}{\partial \theta_1}, \frac{\partial \ln L(\Theta)}{\partial \theta_2}, \frac{\partial \ln L(\Theta)}{\partial \theta_3}$$

и њиховим изједначавањем са нулом. Дакле,

$$L(\widehat{A_0}) = (2\pi e u_2)^{-\frac{n+m}{2}},$$

$$u_2 = \frac{\sum_{i=1}^n (x_i - u_1)^2 + \sum_{j=1}^m (y_j - u_1)^2}{n+m},$$

$$u_1 = \frac{\sum_{i=1}^n x_i + \sum_{j=1}^m y_j}{n+m}$$

и

$$L(\widehat{\Theta}) = (2\pi e v_3)^{-\frac{n+m}{2}},$$

$$v_3 = \frac{\sum_{i=1}^n (x_i - v_1)^2 + \sum_{j=1}^m (y_j - v_2)^2}{n+m},$$

$$v_2 = \frac{\sum_{j=1}^m y_j}{m},$$

$$v_1 = \frac{\sum_{i=1}^m x_i}{n}.$$

Отуда,

$$\lambda(x_1, \ldots, x_n, y_1, \ldots, y_m) = \left(\frac{v_3}{u_2}\right)^{\frac{n+m}{2}} \leq \lambda_0$$

води до тест статистике

$$T = \frac{\overline{X}_n - \overline{Y}_m}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 + \sum_{j=1}^{m}(Y_j - \overline{Y}_m)^2}} \sqrt{\frac{nm(n+m-2)}{n+m}},$$

која има студентову расподелу са $n+m-2$ степена слободе. $\triangle$

У вези са генерализацијом овог поступка наведимо без доказа следећу теорему.

Теорема 21. *Нека узорак $(X_1, \ldots, X_n)$ има заједничку густину расподеле, односно функцију веродостојности $L(\theta)$, где је $\theta \in \Theta$ вишедимензиони параметар. Нека је r димензија параметарског простора Θ, а k димензија параметра дефинисаног хипотезом $H_0 : \theta \in A_0$, тј. димензија простора A_0. Тада за велико n, статистика*

$$-2 \ln \lambda(X_1, \ldots, X_n) = -2 \ln \frac{L(\widehat{A}_0)}{L(\widehat{\Theta})}$$

има приближно χ^2 расподелу са $r-k$ степени слободе.

Ова теорема омогућава да одредимо границу критичне области за велики обим узорка, без обзира на расподелу посматраног обележја. Међутим, који је обим узорка, n, довољно велики није могуће одредити у општем случају, већ ће брзина конвергенције зависити од расподеле обележја које се посматра.

4.2 Параметарски тестови

У овом одељку навешћемо само неколико важнијих тестова за тестирање параметарских хипотеза. Сви тестови овог поглавља су тестови количника веродостојности.

Већ смо истакли да не постоји "универзални" обим узорка који ће гарантовати ваљаност статистичких закључака са задатом тачношћу. Када је реч о тестирању параметарским тестовима, у том смислу је посебно занимљив тест за непознато математичко очекивање обележја, који за довољно велики обим узорка, може да се третира као непараметарски у смислу горе наведене дефиниције.

Наиме, ако се располаже простим узорком довољно великог обима, тест статистика ће имати асимптотски нормалну нормирану расподелу, без обзира на расподелу обележја о чијем се очекивању ради.

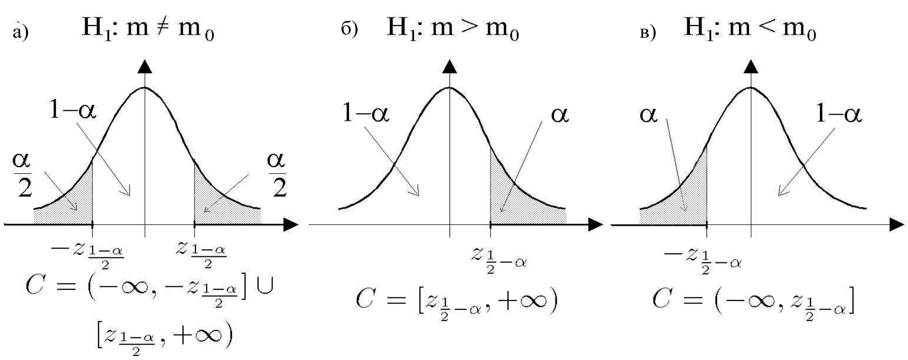
Формулација "велики узорак" у смислу тестова овог одељка је узорак чији је обим $n \geq 30$.

4.2.1 Тест за средњу вредност обележја за велике узорке

Код теста за средњу вредност, m, тестира се нулта хипотеза $H_0(m = m_0)$, против алтернативне хипотезе која може да буде тројака: $H_1(m \neq m_0)$, $H_1(m > m_0)$ или $H_1(m < m_0)$, на основу простог узорка $\mathbf{X} = (X_1, \ldots, X_n)$. Тестирање средње вредности се базира на средини узорка, $\overline{X}_n$. У случају да дисперзија обележја X чија се средња вредност оцењује, није позната, користи се статистика

$$Z_0 = \frac{\overline{X}_n - m_0}{\widetilde{S}_n} \cdot \sqrt{n} \tag{4.4}$$

која има приближно нормалну расподелу $\mathcal{N}(0,1)$ за **велики обим узорка**, без обзира на расподелу обележја X. Јасно да је $\widetilde{S}_n$ оцена непознате стандардне девијације обележја X, те је $\widetilde{S}_n/\sqrt{n}$ оцена параметра $\sqrt{D(\overline{X}_n)}$.



Слика 4.1: Тестирање средње вредности (m) обележја за велике узорке: области прихватања нулте хипотезе $H_0(m = m_0)$ и критичне области за различите алтернативне хипотезе H_1. Критична област је део апсцисне осе "испод" шрафираног дела.

Критична област величине α за тестирање H_0 против $H_1(m \neq m_0)$, одређује се из услова

$$P_{H_0}\{|Z_0| \geq c\} = \alpha.$$

С обзиром на расподелу статистике Z_0, вредност c се најчешће записује као $c = z_{\frac{1-\alpha}{2}}$ (слика 4.1). Или уопште, у односу на све алтернативне хипотезе критична област величине α одређује се према табели:

H_0	H_1	H_0 се одбацује ако се за реализовани узорак добије
$m = m_0$	$m \neq m_0$	$\lvert \frac{\overline{x}_n - m_0}{\widetilde{s}_n/\sqrt{n}} \rvert \geq z_{0,5-\alpha/2}$
$m = m_0$	$m > m_0$	$\frac{\overline{x}_n - m_0}{\widetilde{s}_n/\sqrt{n}} \geq z_{0,5-\alpha}$
$m = m_0$	$m < m_0$	$\frac{\overline{x}_n - m_0}{\widetilde{s}_n/\sqrt{n}} \leq -z_{0,5-\alpha}$

Међутим, ако је обим узорка мали, тест статистика (4.4) нема нормалну расподелу чак ни код нормалне расподеле обележја X. Када узорак има n–димензиону нормалну расподелу, а обим узорка је мали, (4.4) има Студентову расподелу, о чему ће надаље бити речи.

Уколико је дисперзија посматраног обележја (однекуд) позната, користиће се $\sigma/\sqrt{n}$, $\sigma = \sqrt{D(X)}$, за стандардизацију статистике $\overline{X}_n$, тј. тест статистика ће бити

$$Z_0 = \frac{\overline{X}_n - m_0}{\sigma}\sqrt{n}$$

и њена расподела ће бити такође нормална нормирана за велики обим простог узорка, без обзира на расподелу посматраног обележја.

4.2.2 Параметарска тестирања код нормалне расподеле

Наредних шест група тестова односе се на нормалну расподелу, тј. на обележје X чија расподела припада фамилији допустивих расподела $\{\mathcal{N}(m, \sigma^2),\ m \in R,\ \sigma^2 \in R^+\}$.

1. Тестира се хипотеза о непознатом математичком очекивању обележја X, $H_0(m = m_0)$. При томе се разликују два случаја:

(а) σ^2 познато, или σ^2 непознато а обим узорка велики. Тест статистика за случај познате дисперзије је

$$Z_0 = \frac{\overline{X}_n - m_0}{\sigma}\sqrt{n}\,.$$

У овом случају Z_0 има тачну расподелу $\mathcal{N}(0,1)$, док за непознато σ^2

$$Z_0 = \frac{\overline{X}_n - m_0}{\overline{S}_n}\sqrt{n-1}\,,$$

има само асимптотски расподелу $\mathcal{N}(0,1)$. Критичне области које одговарају појединим алтернативним хипотезама за реализовану вредност z_0 статистике Z_0 приказане су у табели:

H_0	H_1	C
$m = m_0$	$m \neq m_0$	$\lvert z_0\rvert \geq z_{0,5-\alpha/2}$
$m = m_0$	$m > m_0$	$z_0 \geq z_{0,5-\alpha}$
$m = m_0$	$m < m_0$	$z_0 \leq -z_{0,5-\alpha}$

Критичне области су приказане на слици 4.1.

Пример 64. За следеће резултате:

Бр. поена	Бр. студената
[50,60)	4
[60,70)	17
[70,80)	24
[80,90)	10
[90,100]	5

који су добијени тестирањем прага осетљивости, тестирати хипотезу да је средња вредност једнака 75 за праг значајности $\alpha = 0,01$, ако је дисперзија позната и износи 100.

Тестира се хипотеза $H_0(m = 75)$ против алтернативне $H_1(m \neq 75)$. Дисперзија је позната и износи 100, праг значајности је $\alpha = 0,01$, док је узорачка средина једнака

$$\overline{x}_{60} = \frac{1}{60}(4 \cdot 55 + 17 \cdot 65 + 24 \cdot 75 + 10 \cdot 85 + 5 \cdot 95) = 74,17.$$

Тако се добија да је реализована вредност тест статистике једнака

$$\frac{74,17-75}{10}\sqrt{60}=-0,64.$$

Како је $z_{0,495}=2,575$, то је критична област

$$C=(-\infty;-2,575]\cup[2,575;+\infty).$$

Како $-0,64\notin C$, то се хипотеза H_0 прихвата. (Значајност, односно p–вредност, овог теста је 0,5222, дакле већа од 0,01.) $\triangle$

(б) <u>σ^2 непознато и обим узорка мали.</u> Тест статистика

$$t_0=\frac{\overline{X}_n-m_0}{\widetilde{S}_n}\sqrt{n}=\frac{\overline{X}_n-m_0}{\overline{S_n}}\sqrt{n-1}$$

има Студентову расподелу са $n-1$ степени слободе. Табела одговарајућих критичних области је:

H_0	H_1	C
$m=m_0$	$m\neq m_0$	$\|t_0\|\geq t_{n-1;\frac{1-\alpha}{2}}$
$m=m_0$	$m>m_0$	$t_0\geq t_{n-1;0,5-\alpha}$
$m=m_0$	$m<m_0$	$t_0\leq -t_{n-1;0,5-\alpha}$

,

где се константе $t_{n-1,\frac{1-\alpha}{2}}$ и $t_{n-1,\frac{1}{2}-\alpha}$, тј. границе критичне области, читају из таблице Студентове расподеле. Графички приказ би био аналоган ономе са слике 4.1

2. Тестира се хипотеза о једнакости средњих вредности двају независних обележја X и Y са претпостављеним расподелама:

$$X\ :\ \mathcal{N}(m_X,\sigma_X^2)\,,\quad Y\ :\ \mathcal{N}(m_Y,\sigma_Y^2)$$

на основу простих независних узорака $\mathbf{X}=(X_1,\ldots,X_{n_X})$ и $\mathbf{Y}=(Y_1,\ldots,Y_{n_Y})$ обима n_X и n_Y редом,

$$H_0(m_X=m_Y),\quad \text{односно},\quad H_0(m_X-m_Y=0).$$

Разматрају се два случаја:

(а) σ_X^2 и σ_Y^2 познате, или σ_X^2 и σ_Y^2 непознате и обими узорака велики.

За тестирање наведене нулте хипотезе користи се тест статистика

$$Z_0 = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}}, \text{ односно }, Z_0 = \frac{\overline{X} - \overline{Y}}{\sqrt{\widehat{D}(\overline{X} - \overline{Y})}},$$

где је

$$\widehat{D}(\overline{X} - \overline{Y}) = \frac{\overline{S}_{n_X}^2}{n_X} + \frac{\overline{S}_{n_Y}^2}{n_Y} \quad \text{и} \quad \overline{X} \equiv \overline{X}_{n_X}, \quad \overline{Y} \equiv \overline{Y}_{n_Y}$$

од којих се прва користи за познате дисперзије и има тачно нормалну нормирану расподелу, а друга има приближно $\mathcal{N}(0,1)$ расподелу ако је хипотеза H_0 тачна.

Критичне области величине α за одговарајуће алтернативне хипотезе H_1 дате су табелом:

H_0	H_1	C
$m_X = m_Y$	$m_X \neq m_Y$	$\lvert z_0 \rvert \geq z_{0,5-\alpha/2}$
$m_X = m_Y$	$m_X > m_Y$	$z_0 \geq z_{0,5-\alpha}$
$m_X = m_Y$	$m_X < m_Y$	$z_0 \leq -z_{0,5-\alpha}$

.

(б) σ_X^2 и σ_Y^2 непознате, обим узорка мали и $\sigma_X^2 = \sigma_Y^2$. Користи се тест статистика

$$t_0 = \frac{(\overline{X} - \overline{Y})\sqrt{\frac{n_X \cdot n_Y}{n_X + n_Y}(n_X + n_Y - 2)}}{\sqrt{(n_X - 1)\widetilde{S}_X^2 + (n_Y - 1)\widetilde{S}_Y^2}}, \tag{4.5}$$

која, под наведеним условима, има приближно $t_{n_X+n_Y-2}$ расподелу. Табела одговарајућих критичних области изгледа овако:

H_0	H_1	C
$m_X = m_Y$	$m_X \neq m_Y$	$\lvert t_0 \rvert \geq t_{n_X+n_Y-2;0,5-\alpha/2}$
$m_X = m_Y$	$m_X > m_Y$	$t_0 \geq t_{n_X+n_Y-2;0,5-\alpha}$
$m_X = m_Y$	$m_X < m_Y$	$t_0 \leq -t_{n_X+n_Y-2;0,5-\alpha}$

.

Тестирање у оквиру ове тачке се може вршити и у случају непознатих, а различитих дисперзија σ_X^2 и σ_Y^2. Тест статистика је поново облика (4.5), али се границе критичних области одређују тзв. апроксимацијом Кохрена о којој овде неће бити речи.

Пример 65. Посматране су две групе радника једне фабрике и мерен је њихов коефицијент интелигенције. За прву групу од 16 радника добијено је да је $\overline{x}_{16} = 114$ и $\overline{s}_X = 82$. За другу групу од 14 радника добијено је да је $\overline{y}_{14} = 121$ и $\overline{s}_Y = 60$. Да ли постоје битне разлике између средњих вредности коефицијената интелигенције ових двеју група радника ако је праг значајности $\alpha = 0,05$?

Тестира се хипотеза $H_0(m_1 = m_2)$ против хипотезе $H_1(m_1 \neq m_2)$, при чему су дисперзије непознате. Тест статистика има реализовану вредност $-2,258$. Како је $t_{28;0,475} = 2,048$, то је критична област $C = (-\infty; -2,048] \cup [2,048; +\infty)$. Како $-2,258 \in C$, то се хипотеза H_0 одбацује, тј. закључује се да постоје битне разлике између коефицијената интелигенције посматраних двеју група радника. (Значајност, односно p–вредност, овог теста је 0,032 што је мање од 0,05.) $\triangle$

3. Тестира се хипотеза о једнакости средњих вредности двају обележја X и Y посматраних истовремено на истој популацији са претпостављеним расподелама:

$$X \ : \ \mathcal{N}(m_X, \sigma_X^2)\,, \quad Y \ : \ \mathcal{N}(m_Y, \sigma_Y^2)$$

на основу дводимензионог простог узорка обима n. Најчешће се ради о томе да се, заправо, на истим јединкама (људима, животињама) утврђују вредности једног испитиваног обележја при постојању различитих услова извођења експеримента, па се резултати једног мерења означе са $X^{(1)}$, а другог са $X^{(2)}$. "Резултат" оба извршена мерења на узорку обима n је случајни вектор:

$$\left(\left(X_1^{(1)}, X_1^{(2)}\right), \left(X_2^{(1)}, X_2^{(2)}\right), \ldots, \left(X_n^{(1)}, X_n^{(2)}\right)\right)$$

Тестира се хипотеза:

$$H_0(m_1 = m_2), \text{ односно, } \quad H_0(m_1 - m_2 = 0)$$

и каже се да се ради о **тесту за математичко очекивање код спарених узорака**. Тест статистика која се при томе користи

$$t_0 = \frac{\overline{D}_n}{\sqrt{\frac{\sum_{i=1}^{n}(D_i - \overline{D}_n)^2}{n(n-1)}}}\,,$$

где је

$$\overline{D}_n = \frac{1}{n}\sum_{i=1}^{n} D_i\,, \quad D_i = X_i^{(1)} - X_i^{(2)}\,,\ \ i = 1, \ldots, n\,,$$

има Студентову расподелу са $n-1$ степени слободе. Критичне области се одређују као код теста за математичко очекивање нормалне расподеле са непознатом дисперзијом и малим обимом узорка.

Пример 66. Нека је на групи од 10 људи мерен број позитивних реакција под дејством два стресора и то најпре физичке природе, при чему је стрес изазиван електричним шоком, а затим психолошке природе – гласна музика. У табели је дат број позитивних реакција:

Стресор \ Особа	1	2	3	4	5	6	7	8	9	10
Електрошок	6	8	4	8	6	4	5	5	6	7
Гласна музика	8	9	9	12	9	7	9	9	8	11
d_i	-2	-1	-5	-4	-3	-3	-4	-4	-2	-4

Тестирати хипотезу о једнакости математичких очекивања за $\alpha = 0,05$.

Добија се да је $\overline{d}_{10} = -3,2$ и $\sum(d_i - \overline{d}_{10})^2 = 13,6$, тако да тест статистика има реализовану вредност

$$t_0 = \frac{-3,2}{\sqrt{13,6}} \cdot \sqrt{9} = -2,603.$$

Критична област је

$$C = (-\infty; -2,262] \cup [2,262; +\infty)$$

и како $-2,603 \in C$, то се хипотеза H_0 одбацује, тј. закључује се да има разлике у очекиваном броју позитивних реакција под дејством физичког и психолошког стресора код посматране групе испитаника. (Значајност или p–вредност овог теста је 0,029, дакле мања од 0,05.) $\triangle$

4. Тест који се односи на тестирање дисперзије обележја са нормалном расподелом има нулту хипотезу:

$$H_0(\sigma^2 = \sigma_0^2),$$

где је σ_0^2 фиксиран позитиван реалан број. У том случају, тест статистика је

$$\chi_0^2 = \frac{(n-1)\widetilde{S}_n^2}{\sigma_0^2}$$

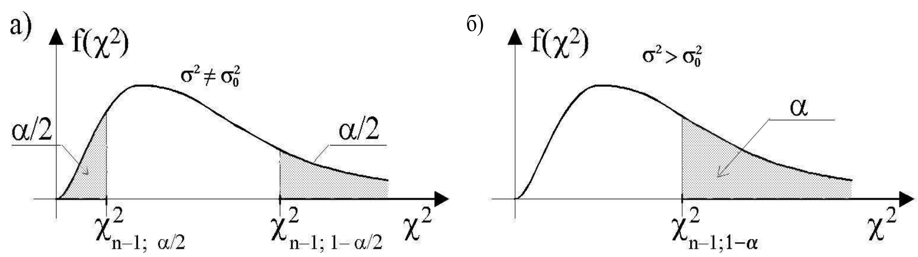
која има приближно χ^2 расподелу са $n-1$ степени слободе: χ_{n-1}^2.

Критичне области величине α, за различите алтернативне хипотезе, дате су у табели:

H_0	H_1	C
$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	$\chi_0^2 \leq \chi^2_{n-1;\frac{\alpha}{2}} \vee \chi_0^2 \geq \chi^2_{n-1;1-\frac{\alpha}{2}}$
$\sigma^2 = \sigma_0^2$	$\sigma^2 > \sigma_0^2$	$\chi_0^2 \geq \chi^2_{n-1;1-\alpha}$
$\sigma^2 = \sigma_0^2$	$\sigma^2 < \sigma_0^2$	$\chi_0^2 \leq \chi^2_{n-1;\alpha}$

.

Део апсцисне осе испод шрафиране површине је критична област величине α и на слици 4.2 а) одговара области одбацивања разматране нулте хипотезе против алтернативне $H_1(\sigma^2 \neq \sigma_0^2)$ и на одговарајући начин 4.2 б).



Слика 4.2: Тестирање дисперзије (σ^2) обележја за велике узорке: критичне области за $H_0(\sigma^2 = \sigma_0^2)$ против алтернативних хипотеза а) $H_1(\sigma^2 \neq \sigma_0^2)$, и б) $H_1(\sigma^2 > \sigma_0^2)$.

Наведена статистика се користи за случај непознатог математичког очекивања. За случај познатог m, користи се тест статистика

$$\chi_0^2 = \frac{\sum_{i=1}^n (X_i - m)^2}{\sigma_0^2}$$

која има χ^2 расподелу са n степени слободе, па се у том смислу и критичне области разликују од горе наведених, тј. разликују се само као последица промене у броју степени слободе.

Пример 67. Мерењем коефицијента интелигенције 50 ученика добијено је да је дисперзија једнака $\overline{s}_{50}^2 = 2,45$. Тестирати хипотезу да је стандардно одступање веће од 2 за праг значајности $\alpha = 0,05$.

Тестира се хипотеза $H_0(\sigma^2 = 4)$ против хипотезе $H_1(\sigma^2 > 4)$. Како је $\chi^2_{49;0,95} = 67,5$, то је критична област $C = [67,5; +\infty)$. Тест статистика има реализовану вредност $(49 \cdot 2,45)/4 = 30,01$ и она не припада области C, тако да се хипотеза H_0 прихвата. (Значајност или p–вредност овог теста је 0,99, дакле, већа од 0,05.) $\triangle$

5. Често се указује потреба за упоређивањем два обележја по њиховим дисперзијама. Користе се два независна проста узорка $(X_1, \ldots, X_{n_X})$ и $(Y_1, \ldots, Y_{n_Y})$ обележја X и Y чије су расподеле редом $\mathcal{N}(m_X, \sigma_X^2)$ и $\mathcal{N}(m_Y, \sigma_Y^2)$. Тестира се нулта хипотеза

$$H_0(\sigma_X^2 = \sigma_Y^2)$$

против одговарајуће сложене хипотезе, као и у претходним тестовима. У случају да су m_X и m_Y познате величине, користи се тест статистика

$$F_0 = \frac{n_X \sum\limits_{i=1}^{n_Y} (Y_i - m_Y)^2}{n_Y \sum\limits_{i=1}^{n_X} (X_i - m_X)^2}$$

која има Фишерову расподелу F_{n_Y, n_X}. Када m_X и m_Y нису познате, користи се

$$F_0 = \frac{\widetilde{S}_{n_Y}^2}{\widetilde{S}_{n_X}^2}$$

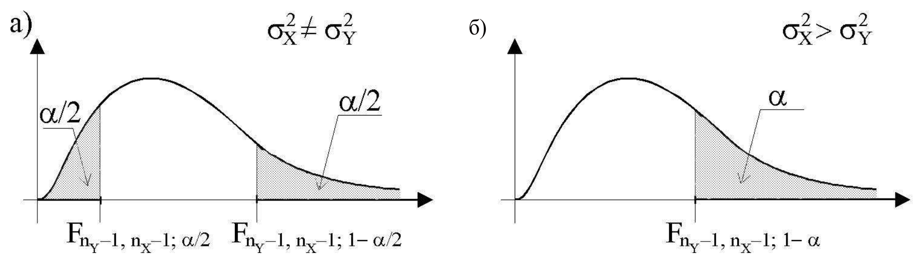
која има приближно Фишерову расподелу $F_{n_Y - 1, n_X - 1}$. Табела одговарајућих критичних области у последњем случају је:

H_1	C
$\sigma_X^2 \neq \sigma_Y^2$	$F_0 \leq F_{n_Y-1,n_X-1;\frac{\alpha}{2}} \vee F_0 \geq F_{n_Y-1,n_X-1;1-\frac{\alpha}{2}}$
$\sigma_X^2 > \sigma_Y^2$	$F_0 \geq F_{n_Y-1,n_X-1;1-\alpha}$
$\sigma_X^2 < \sigma_Y^2$	$F_0 \leq F_{n_Y-1,n_X-1;\alpha}$

.

док за случај познатих m_X и m_Y треба на одговарајући начин прилагодити број степени слободе статистика у табели.

Критичне области за одговарајуће сложене алтернативне хипотезе приказане су на слици 4.3.



Слика 4.3: Критичне области код примене Фишерове расподеле за тестирање нулте против алтернативних хипотеза а) $H_1(\sigma_X^2 \neq \sigma_Y^2)$ и б) $H_1(\sigma_X^2 > \sigma_Y^2)$, када очекивања нису позната.

6. Тестирање коефицијента корелације

   За случајни вектор (X, Y) чија расподела припада фамилији дводимензионих нормалних расподела $\{\mathcal{N}(m_X, m_Y,\, \sigma_X^2,\, \sigma_Y^2,\, \rho)\,, m_X \in R\,, m_Y \in R\,, \sigma_X^2 \in R^+\,, \sigma_Y^2 \in R^+\,, |\rho| \in [0,1]\}$ тестира се нулта хипотеза

$$H_0 : \rho = \rho_0, \quad -1 < \rho_0 < 1$$

   против одговарајућих алтернативних. Разликују се два случаја: $\rho_0 = 0$ и $\rho_0 \neq 0$. Ово је последица различитих тест статистика имплицираних поменутим вредностима коефицијента корелације. Дакле, за два обележја тестира се постојање линеарне везе међу њима на следећи начин:

   (а) Тестирање нулте хипотезе

$$H_0 : \rho = 0$$

на основу узорачког коефицијента корелације и узорка обима n.

Под претпоставком да је нулта хипотеза тачна, тест статистика

$$t_0 = \frac{R_{XY}\sqrt{n-2}}{\sqrt{1-R_{XY}^2}}$$

има Студентову расподелу са $n-2$ степена слободе. R_{XY} је, као и до сада, узорачки коефицијент корелације. Отуда је табела одговарајућих критичних области за праг значајности α:

H_0	H_1	C
$\rho = 0$	$\rho \neq 0$	$\lvert t_0\rvert \geq t_{n-2;0,5-\alpha/2}$
$\rho = 0$	$0 < \rho < 1$	$t_0 \geq t_{n-2;0,5-\alpha}$
$\rho = 0$	$-1 < \rho < 0$	$t_0 \leq -t_{n-2;0,5-\alpha}$

.

Треба се подсетити раније изнете чињенице да уколико вектор (X, Y) има дводимензионалну нормалну расподелу, сазнање о томе да је $H_0 : \rho = 0$ тачна, значи не само некорелисаност, већ и независност обележја X и Y.

Пример 68. Група од 16 студената показала је на испиту из математике следећи успех:

Писмени	90	90	80	90	92	88	90	63
Усмени	84	84	82	94	90	85	89	62

Писмени	70	54	78	86	99	84	56	85
Усмени	65	52	72	90	98	89	58	85

.

Да ли је на 5% прагу значајности коефицијент корелације близак нули?

Тестира се хипотеза $H_0(\rho = 0)$ против хипотезе $H_1(\rho \neq 0)$. Одговарајуће узорачке средине су $\overline{x}_{16} = 80,94$ и $\overline{y}_{16} = 79,94$, док су узорачке дисперзије $\overline{s}_X = 166,9$ и $\overline{s}_Y = 177,66$, респективно. Узорачки коефицијент корелације има вредност $r_{XY} = 0,964$. Према томе, вредност тест статистике је

$$t_0 = \frac{0,964\sqrt{14}}{\sqrt{1-0,964^2}} = 13,565\,.$$

Критична област је $C = (-\infty; -2,145] \cup [2,145; +\infty)$ и како 13,565 припада области C, то се хипотеза H_0 одбацује. △

(б) Тестирање нулте хипотезе

$$H_0 : \rho = \rho_0 \,,$$

где је $\rho_0 \neq 0$, тј. $\rho_0 \in (-1,0) \cup (0,1)$, такође на бази узорка обима n.

Користи се тест статистика

$$Z = \frac{1}{2} \ln \left(\frac{1 + R_{XY}}{1 - R_{XY}} \right)$$

која, под претпоставком да је нулта хипотеза тачна, има приближно нормалну расподелу

$$Z : \mathcal{N} \left(\frac{1}{2} \ln \left(\frac{1 + \rho_0}{1 - \rho_0} \right) + \frac{\rho_0}{2(n-1)} \,,\ \frac{1}{n-3} \right).$$

Стандардизовањем овакве случајне променљиве омогућено је коришћење таблице за нормалну нормирану расподелу и табела одговарајућих критичних области величине α као у претходном случају, где је z_0 реализована вредност статистике

$$Z_0 = \frac{Z - \frac{1}{2} \ln(\frac{1+\rho_0}{1-\rho_0}) - \frac{\rho_0}{2(n-1)}}{1/\sqrt{n-3}} \,.$$

4.2.3 Тестирање параметра биномне расподеле

Ако треба тестирати вероватноћу p реализације неког догађаја A преко узорка обима n, заправо се врши тестирање хипотезе о параметру биномне расподеле, тј. тестирање нулте хипотезе $H_0(p = p_0)$ против свих алтернативних, под претпоставком да је узорак узет из популације са обележјем $S_n : \mathcal{B}(n,p)$. За мали обим узорка, критичне области се одређују директно из дефиниције биномне расподеле, о чему се читалац може више информисати у тексту посвећеном тесту знакова. Међутим, за велики обим узорка, што овде подразумева $n > 50$ и $np_0 > 10$, користи се нормална апроксимација биномне расподеле и статистика

$$Z_0 = \frac{S_n - np_0}{\sqrt{np_0(1-p_0)}}$$

која има приближно $\mathcal{N}(0,1)$ расподелу. Дакле:

H_0	H_1	C
$p = p_0$	$p \neq p_0$	$\lvert z_0 \rvert \geq z_{0{,}5-\alpha/2}$
$p = p_0$	$p > p_0$	$z_0 \geq z_{0{,}5-\alpha}$
$p = p_0$	$p < p_0$	$z_0 \leq -z_{0{,}5-\alpha}$

.

Пример 69. Анкетом се испитују шансе једног кандидата на изборима. Међу 100 случајно изабраних гласача 55 њих се изјаснило да би гласало за тог кандидата. Нека је p вероватноћа да је случајно изабрани анкетирани симпатизер посматраног кандидата. Тестирати хипотезу да ће посматрани кандидат добити 50% гласова целокупног бирачког тела против свих алтернативних. Дакле, тестира се нулта хипотеза $H_0(p = 0,5)$ и нека је $\alpha = 0,01$, против: а) $H_1(p \neq 0,5)$, б) $H_1(p > 0,5)$ и в) $H_1(p < 0,5)$.

Реализована вредност тест статистике је

$$z_0 = \frac{55 - 100 \cdot 0,5}{\sqrt{100 \cdot 0,5 \cdot 0,5}} = 1.$$

а) Добија се да је $z_{0{,}495} = 2,575$, тако да је критична област $C = (-\infty; -2,575] \cup [2,575; +\infty)$. Како $1 \notin C$, то се хипотеза H_0 прихвата. (Значајност или p–вредност овог теста је 0,317, дакле, већа од 0,01.) То истовремено значи да нема смисла даље вршити тестирања против преостале две алтернативне хипотезе. $\triangle$

4.3 Непараметарски тестови

Две су врсте проблема који се најчешће решавају непараметарским тестовима:

1. проблем једног узорка – испитују се:
   - (а) параметри расподеле (пре свега квантили),
   - (б) случајност узорка, и
   - (ц) сагласност узорка са претпостављеном расподелом.
2. проблем два узорка – испитују се:
   - (а) зависност два обележја,
   - (б) упоређују се расподеле два обележја, и сл.

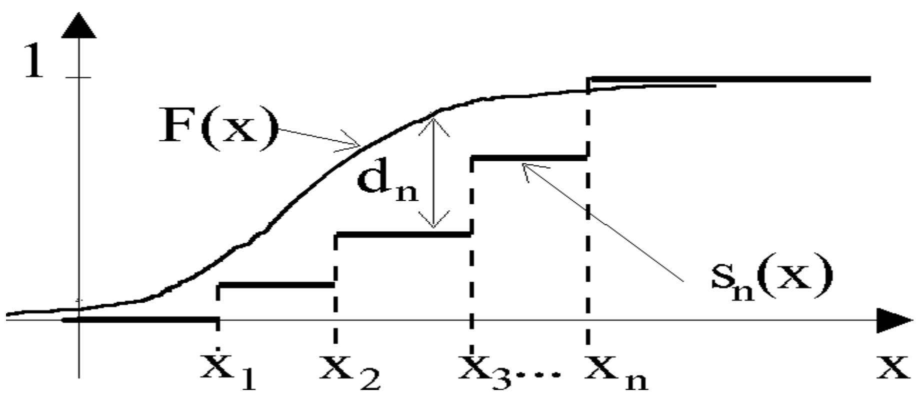
4.3.1 Тест Колмогоров – Смирнова

Тест Колмогоров–Смирнова се користи само код обележја апсолутно непрекидног типа, тј. код таквих обележја код којих је функција расподеле F непрекидна. Тестира се нулта хипотеза

$$H_0 : F(x) = F_0(x)\,, \quad \forall x \in R\,,$$

где је F_0 нека одређена, такође непрекидна, функција расподеле. Тест Колмогоров–Смирнова директно примењује централну теорему математичке статистике, те на основу узорка обима n, $(X_1, \ldots, X_n)$, одређује емпиријску функцију расподеле $S_n(x)$, $x \in R$, и дефинише статистику

$$D_n = \sup_{-\infty<x<+\infty} |S_n(x) - F_0(x)|.$$



Слика 4.4: Одређивање вредности d_n код теста Колмогоров–Смирнова.

Под претпоставком да је нулта хипотеза тачна, статистика D_n има расподелу Колмогорова. На основу реализованог узорка $(x_1, \ldots, x_n)$ треба одредити реализовану вредност статистике D_n:

$$d_n = \sup_{-\infty<x<+\infty} |s_n(x) - F_0(x)|.$$

Реализована вредност емпиријске функције расподеле је степенаста функција $s_n(x)$, $x \in R$, и има коначан број "степеника" (највише $n+1$, у случају да су сви елементи реализованог узорка различити). Уочимо још једном да су s_n и F_0 монотоно неопадајуће функције. Одређивање супремума, у пракси, се, по правилу, своди на одређивање максимума апсолутних разлика $|s_n(x-0) - F_0(x-0)|$ (због непрекидности с десна) на сегментима дефинисаним узорком (слика 4.4).

Граница критичне области за задати праг значајности α: $d_{n,1-\alpha}$ чита се из таблице Колмогорова. Критична област одређује се на следећи начин:

H_0	H_1	C
$F = F_0$	$F \neq F_0$	$d_n \geq d_{n,1-\alpha}$

Пример 70. Следећа табела приказује резултате теста интелигенције 53 дечака:

IQ	Бр. деце
[65,75)	1
[75,85)	2
[85,95)	10
[95,105)	12
[105,115)	14
[115,125)	11
[125,135]	3

.

Испитати сагласност ових података са нормалним законом расподеле користећи тест Колмогоров–Смирнова са 1% прагом значајности.

Нека је обележје X коефицијент интелигенције детета. Тестира се хипотеза

$$H_0(\text{подаци су сагласни са } \mathcal{N}(m, \sigma^2) \text{ расподелом}).$$

Параметри расподеле m и σ^2 су непознати. Параметар m се оцењује узорачком средином а σ^2 узорачком дисперзијом. Тако се добија $\widehat{m} = 105,28$ и $\widehat{\sigma}^2 = 184,047$. Значи, тестира се хипотеза $H_0($ подаци су сагласни са $\mathcal{N}(105,28\,;\,184,047)$ расподелом). Даљи поступак је садржан у следећој табели:

$x_i + 0$	n_i	$\sum_i$	$s_{53}(x_i - 0)$	t_i	$F_0(x_i) = 0,5 \pm \Phi(\lvert t_i \rvert)$	$\lvert s_{53}(x_i-0) - F_0(x_i) \rvert$
75	1	1	0,019	-2,232	0,0129	0,0061
85	2	3	0,057	-1,495	0,0681	0,0111
95	10	13	0,245	-0,758	0,2236	0,0214
105	12	25	0,472	-0,021	0,4920	0,0200
115	14	39	0,736	0,716	0,7642	0,0282
125	11	50	0,943	1,454	0,9265	0,0165
135	3	53	1	2,191	0,9857	0,0143

.

Овде је n_i апсолутна учестаност i-тог интервала, $\sum_i$ збирна учестаност до тог интервала (укључујући и тај интервал), а $t_i = \frac{(x_i-0)-\widehat{m}}{\sqrt{\widehat{\sigma}^2}}$. Примећује се да је максимална разлика $|s_{53}(x-0)-F_0(x)| \equiv |s_{53}(x-0)-F_0(x-0)|$ једнака $0,0282$, тако да је $d_{53} = 0,0282$. Како је $1-\alpha = 0,99$, то из таблице Колмогорова следи да је $d_{53;0,99} = 0,23$, тако да је критична област $C = [0,23; +\infty)$. Како $0,0280 \notin C$, то се хипотеза H_0 прихвата, односно нема разлога да се одбаци. Дакле, на основу резултата теста, може се тврдити да је IQ нормално расподељен на популацији дечака испитиваног узраста. $\triangle$

Тест Колмогоров–Смирнова користи се и за тестирање једнакости расподела двају обележја X и Y апсолутно непрекидног типа на основу независних узорака $(X_1, X_2, \ldots, X_{n_1})$ и $(Y_1, Y_2, \ldots, Y_{n_2})$. Користи се статистика

$$D_{\frac{n_1 n_2}{n_1+n_2}} = \sup_{-\infty<x<+\infty} |S_X(x) - S_Y(x)|$$

за тестирање нулте хипотезе

$$H_0 : F_X = F_Y$$

против алтернативне

$$H_1 : F_X \neq F_Y,$$

где су $S_X(x)$ и $S_Y(x)$, $x \in R$, одговарајуће емпиријске функције расподеле обележја X и Y на основу посматраних узорака. И овде се у пракси супремум замењује максимумом. Одговарајућа критична област величине α је:

H_0	H_1	C
$F_X = F_Y$	$F_X \neq F_Y$	$d_{\frac{n_1 n_2}{n_1+n_2}} \geq d_{\frac{n_1 n_2}{n_1+n_2},1-\alpha}$

.

Пример 71. Случајно изабрани дечаци из две школе подвргнути су тесту агресивности. Добијени су следећи резултати:

Број поена на тесту	[75,85)	[85,95)	[95,105)	[105,115)	[115,125)	[125,135]
Бр. дечака I школе	3	10	12	14	11	3
Бр. дечака II школе	0	2	13	30	5	1

.

Тестирати хипотезу да су узорци из популације са истом расподелом обележја са прагом значајности 0,05.

Тестира се хипотеза $H_0(F_X = F_Y)$ против хипотезе $H_1(F_X \neq F_Y)$. Поступак рачунања дат је у табели:

$x_i + 0$	n_i	$\sum_i$	$s_X(x_i - 0)$	m_i	$\sum_i$	$s_Y(x_i - 0)$	$\lvert s_X(x_i-0) - s_Y(x_i-0) \rvert$
85	3	3	0,0566	0	0	0	0,0566
95	10	13	0,2453	2	2	0,0392	0,2061
105	12	25	0,4717	13	15	0,2941	0,1776
115	14	39	0,7358	30	45	0,8824	0,1466
125	11	50	0,9434	5	50	0,9804	0,0370
135	3	53	1	1	51	1	0

Како је $(53 \cdot 51)/(53 + 51) = 25,99$ а то приближно једнако 26, то је $d_{26} = 0,2061$. С друге стране, критична област је облика $C = [d_{26;0,95}; +\infty) = [0,264; +\infty)$. Вредност 0,2061 не припада критичној области C, што значи да се хипотеза H_0 прихвата, тј. узорци су са истом расподелом обележја (за дати праг значајности). Односно, може се сматрати да у степену агресивности код дечака двеју испитиваних школа нема разлике. $\triangle$

4.3.2 Пирсонов χ^2 тест

Један од најчешће примењиваних непараметарских тестова је Пирсонов или χ^2 тест.

Тест носи назив по свом аутору Карлу Пирсону, који га је дефинисао и увео у статистичку праксу 1900. године. Његов алтернативни назив, χ^2 тест, потиче од расподеле тест статистике којом се користи. Овде ћемо само приближити идеју о расподели тест статистике, а нећемо се бавити доказом управо наведене тврдње.

Пођимо од случајне променљиве у ознаци X_1 са биномном расподелом $\mathcal{B}(n, p_1)$. Стандардизована случајна променљива $X_1^* = \frac{X_1 - np_1}{\sqrt{np_1(1-p_1)}}$ има асимптотски (према Муавр–Лапласовој теореми) нормалну нормирану расподелу. Отуда, када $n \to \infty$, случајна променљива $Q_1 = (X_1^*)^2$ има асимптотски χ^2–расподелу са 1 степеном слободе, χ_1^2. Уводећи нову случајну променљиву $X_2 = n - X_1$ и параметар $p_2 = 1 - p_1$, случајна променљива Q_1 се може да представи као

$$Q_1 = \frac{(X_1 - np_1)^2}{np_1(1-p_1)} = \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_2 - np_2)^2}{np_2}\,.$$

Дакле, збир овако дефинисаних случајних променљивих има асимптотски χ_1^2–расподелу када се n увећава ($n \geq 50$).

Посматрајмо сада случајни вектор $(X_1, X_2, \ldots, X_{k-1})$ димензије $k-1$ са мултиномном расподелом $\mathcal{M}(n, p_1, p_2, \ldots, p_{k-1})$. Дефинишимо случајну промењиву $X_k = n - (X_1 + \cdots + X_{k-1})$ и параметар $p_k = 1 - (p_1 + \cdots + p_{k-1})$. Може се показати да случајна промењива

$$Q_{k-1} = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i}, \quad np_i \geq 5 \text{ за свако } i = 1, \ldots, k$$

конвергира у расподели ка случајној промењивој са χ^2_{k-1} расподелом. У литератури се може да нађе упозорење да је ова апроксимација добра тек када је n довољно велико тако да свако np_i, $i = 1, \ldots, k$ буде најмање 5, чега ћемо се и ми овде придржавати.

Случајна промењива Q_{k-1} служи као основ за дефиницију Пирсоновог теста.

Претпоставимо да се узорачки простор неког експеримента разбија на коначан број међусобно дисјунктних скупова $A_1, \ldots, A_k$. Нека је $P(A_i) = p_i$, $i = 1, \ldots, k$, где је $p_k = 1 - (p_1 + \cdots + p_{k-1})$, што значи да је p_i вероватноћа да је исход овог случајног експеримента у скупу A_i. Претпостављамо да се случајни експеримент понавља n независних пута под истим условима, па ћемо случајном промењивом X_i да означимо колико је пута исход експеримента припао скупу A_i. Другим речима, $X_1, \ldots, X_k$, $X_k = n - (X_1 + \cdots + X_{k-1})$ су апсолутне учестаности са којима исход експеримента припада респективно скуповима $A_1, \ldots, A_k$. Тада је заједничка густина расподеле за $X_1, \ldots, X_{k-1}$ мултиномна са параметрима $n, p_1, \ldots, p_{k-1}$. Надаље се тестирање односи на нулту хипотезу о поменутој мултиномној расподели:

$$H_0 : p_1 = p_{01},\, p_2 = p_{02}, \ldots,\, p_{k-1} = p_{0,k-1}\,,$$

против свих алтернативних, где су $p_{01}, \ldots, p_{0,k-1}$ бројеви, $0 < p_{0i} < 1$, $i = 1, 2, \ldots, k-1$ и $p_{01} + p_{02} + \cdots + p_{0,k-1} < 1$.

Јасно је сада у каквој је вези овај тест са случајном промењивом Q_{k-1}. Ако је H_0 тачна хипотеза, случајна промењива

$$Q_{k-1} = \sum_{i=1}^{k} \frac{(X_i - np_{0i})^2}{np_{0i}}$$

има приближно χ^2_{k-1}–расподелу.

Уочимо следеће. Када је H_0 тачна, онда је np_{0i} очекивање од X_i. Отуда је и интуитивно јасно да експериментална вредност случајне промењливе Q_{k-1} не треба да је велика ако је H_0 тачна. На основу овога, за унапред задати праг значајности α, одређујемо границу критичне области c као

$$P\{Q_{k-1} \geq c\} = \alpha\,.$$

С обзиром на расподелу којој тежи, уместо ознаке Q_{k-1}, или неке друге, за ову случајну промењливу се користи ознака баш χ^2_{k-1}.

Неки најкарактеристичнији примери примене χ^2 теста изнети су у наредним одељцима.

А. Испитивање сагласности узорка са претпостављеном расподелом

Као и код теста Колмогоров–Смирнова, тестира се хипотеза да је непозната расподела F посматраног обележја X једнака задатој – познатој расподели F_0, тј.

$$H_0 : F = F_0,$$

против алтернативне, да су расподеле различите. Битна разлика у односу на тест Колмогоров–Смирнова је у томе што се χ^2 тест не ограничава само на расподеле апсолутно непрекидног типа, већ се може применити на било коју расподелу F_0.

Поступак тестирања спроводи се тако што се област вредности за X дели на одређен број (k) дисјунктних скупова, тј. реална права се подели на k дисјунктних интервала S_1, S_2, …, S_k, чија је унија скуп R:

$$S_1, S_2, \ldots, S_k \subset R\,, \quad \cup_{i=1}^{k} S_i = R\,, \quad S_i \cap S_j = \emptyset \text{ за } i \neq j.$$

Следећи корак је израчунавање теоријских вероватноћа $P\{X \in S_i\} = p_{0i}$, $i = 1, \ldots, k-1$ и $p_{0k} = 1-(p_{01}+\cdots+p_{0,k-1})$, уз претпоставку да је нулта хипотеза тачна. Затим се срачунавају теоријске апсолутне учестаности у интервалима S_i задатог узорка обима n:

$$\hat{n}_i = np_{0i} \equiv e_i\,.$$

Другим речима, $\hat{n}_i$ је очекивани број елемената узорка у сваком интервалу S_i (отуда[1] и ознака e_i која се користи у литератури) под

[1] Од енглеске речи *expected*.

условом да је нулта хипотеза тачна. Он се упоређује са реализованим у експерименту бројем елемената узорка у том интервалу, тј. са n_i (у литератури се користи и ознака[2] o_i). Ако је H_0 тачна, одступања $\hat{n}_i$ од n_i не би смела да буду велика. За меру одступања у свакој групи узима се релативно одступање

$$\frac{(n_i - \hat{n}_i)^2}{\hat{n}_i},$$

па се као тест статистика користи

$$\chi_0^2 = \sum_{i=1}^{k} \frac{(n_i - \hat{n}_i)^2}{\hat{n}_i} = \sum_{i=1}^{k} \frac{n_i^2}{\hat{n}_i} - n\,.$$

Она за велики обим узорка n, под претпоставком да је нулта хипотеза тачна и сви параметри расподеле F_0 у нултој хипотези познати, има приближно χ^2 расподелу са $k-1$ степени слободе.

Критична област величине α добија се из услова

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

јер реализација догађаја $\{\chi_0^2 \geq c\}$ сигнализира велико одступање F од F_0.

Дакле, наредна табела приказују критичну област величине α за разматрани тест.

H_0	H_1	C
$F = F_0$	$F \neq F_0$	$\chi_0^2 \geq \chi^2_{k-1,1-\alpha}$

.

Ова критична област се одређује аналогно са оном приказаном на слици 4.2 *b*).

При разбијању скупа R на интервале не постоји строго правило о броју интервала. У пракси се руководи логиком сређивања реализованог узорка, међутим, не треба бирати ниједан интервал S_i у коме би се добило $\hat{n}_i < 5$, односно такав интервал треба прикључити претходном или наредном интервалу.

Уколико неки од параметара расподеле F_0 треба оценити на основу узорка (параметар није унапред познат) да би се одредиле вероватноће p_{0i}, онда за сваки процењени параметар треба смањити број степени слободе за један. Дакле, ако се оцењује укупно l параметара, број степени слободе статистике χ_0^2 је: $k-l-1$.

[2] Од енглеске речи *observed*.

Пример 72. Тестирањем 200 испитаника добијен је реализовани узорак који је, после интервалног сређивања, како следи:

Број бодова	[10,12)	[12,14)	[14,16)	[16,18)	[18,20)	[20,22]	n
Број испитаника	10	26	56	64	30	14	200

.

Са прагом значајности $\alpha = 0,01$ тестирати хипотезу да је расподела резултата тестова нормална.

Параметре нормалне расподеле m и σ^2 треба оценити, на основу узорка, средином узорка и дисперзијом узорка редом:

$$\bar{x}_{200} = 16,2 \quad \text{и} \quad \bar{s}^2_{200} = 6,08 \quad \Rightarrow \bar{s}_{200} = 2,47.$$

Скуп реалних бројева се дели на интервале: $S_1 = (-\infty, 12)$, $S_2 = [12, 14)$, $S_3 = [14, 16)$, $S_4 = [16, 18)$, $S_5 = [18, 20)$ и $S_6 = [20, +\infty)$. Уколико је нулта хипотеза тачна, тада је

$$\begin{aligned} p_{01} &= P_{H_0}\{X \in (-\infty, 12)\} = \\ &= P_{H_0}\{-\infty < X^* < \frac{12-16,2}{2,47}\} = 0,0446, \\ p_{02} &= P_{H_0}\{X \in [12, 14)\} = 0,1421\,, \\ p_{03} &= 0,2814, \quad p_{04} = 0,2992, \quad p_{05} = 0,1709\,, \\ p_{06} &= 1 - p_{01} - p_{02} - p_{03} - p_{04} - p_{05} = 0,0618. \end{aligned}$$

Дакле, $\hat{n}_1 = 200 \cdot 0,0446 = 8,92$, $\hat{n}_2 = 200 \cdot 0,1421 = 28,42$, $\hat{n}_3 = 56,28$, $\hat{n}_4 = 59,84$, $\hat{n}_5 = 34,18$ и $\hat{n}_6 = 12,36$, те је $\chi^2_{6-2-1} = 1,3562$.

Како је критична област дефинисаног теста $[11,3; +\infty)$, нема разлога да се нулта хипотеза одбаци; дакле, са прагом значајности 0,01 нема значајне разлике расподеле посматраног обележја од нормалне расподеле. (Значајност или p–вредност овог теста је 0,716). $\triangle$

Пример 73. Тестира се хипотеза о нормалној расподели зарада радника једног предузећа према случајном узорку од 70 радника из два погона ($1-50$ први погон, $51-70$ други погон): 970, 650, 890, 1230, 680, 1010, 740, 480, 690, 820, 990, 860, 1040, 820, 1100, 540, 730, 670, 880, 530, 680, 790, 780, 850, 900, 700, 770, 890, 930, 1000, 1180, 1010, 850, 830, 940, 980, 740, 1110, 810, 840, 620, 790, 480, 990, 1060, 800, 700, 590, 920, 810, 1040, 710, 1100, 1070, 1010, 830, 1020, 760, 780, 1140, 700, 750, 900, 780, 970, 960, 710, 660, 410, 560.

интервал	n_i	*)	$\hat{p}_{0i}$	$\hat{n}_i = 70 \cdot \hat{p}_{0i}$
$(-\infty;$ 600)	7	$(-\infty;$ -1,42)	0,0778	5,4
[600; 700)	7	[-1,42; -0,84)	0,1226	8,6
[700; 800)	16	[-0,84; -0,26)	0,1970	13,8
[800; 900)	14	[-0,26; 0,32)	0,2281	16,0
[900; 1000)	11	[0,32; 0,90)	0,1904	13,3
[1000; 1100)	9	[0,90; 1,48)	0,1147	8,0
[1100; $+\infty$)	6	[1,48; $+\infty$)	0,0694	4,9
Σ	70	/	1	70

*)–центрирани и нормирани интервали

Оцењују се параметри m и σ^2 нормалне расподеле

$$\hat{m} = \overline{x}_{70} = \frac{1}{70}(550 \cdot 7 + 650 \cdot 7 + \ldots + 1150 \cdot 6) = 844,29$$

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{\overline{s}_n^2} =$$

$$\sqrt{\frac{1}{70}(550^2 \cdot 7 + 650^2 \cdot 7 + \ldots + 1150^2 \cdot 6) - 844,29^2} = 172,28$$

$$\Rightarrow \hat{p}_{0i} : z_d = \frac{x_d - \hat{m}}{\hat{\sigma}};\ z_g = \frac{x_g - \hat{m}}{\hat{\sigma}}, \quad \hat{p}_{0i} = \Phi(z_g) - \Phi(z_d)$$

$$\chi_0^2 = \frac{(7-5,4)^2}{5,4} + \frac{(7-8,6)^2}{8,6} + \ldots + \frac{(6-4,9)^2}{4,9} = 2,14.$$

Има $k = 7$ интервала и $l = 2$ оцењена параметра, значи да је број степени слободе $k - l - 1 = 7 - 2 - 1 = 4$.

Осим уобичајеног начина закључивања о прихватању или одбацивању нулте хипотезе који је коришћен и у претходном примеру, овде ће бити изложен још један, карактеристичан за χ^2 тест:

С обзиром да је очекивање случајне променљиве χ_ν^2 једнако ν,

$$E(\chi_\nu^2) = \nu,$$

то је, без обзира на праг значајности α, могућ закључак да одступања у узорку од теоријски претпостављене расподеле нису значајна ако је $\chi_0^2 < k - l - 1$, тј. мање од очекиване вредности тест статистике и да у том случају хипотезу H_0 треба прихватити.

Како је у овом примеру $\chi_0^2 = 2,14 < 4$, хипотеза H_0 се прихвата. $\triangle$

Пример 74. Издвојена је група талентованих ученика и бележен је њихов коефицијент интелигенције. Добијени су следећи резултати:

IQ	$[120,124)$	$[124,130)$	$[130,136)$	$[136,140]$
Бр. ученика	10	50	35	5

.

Испитати сагласност ових података са χ^2 расподелом за $\alpha = 0,01$.

Број степени слободе претпостављене χ^2 расподеле треба оценити, рецимо средином узорка јер је $E(\chi^2_\nu) = \nu$, па се добија $\hat{\nu} = \overline{x}_{100} = 129,15 \approx 129$. Тестира се хипотеза $H_0(X$ има χ^2_{129} расподелу). Користимо 4 интервала: $S_1 = (-\infty, 124)$, $S_2 = [124, 130)$, $S_3 = [130, 136)$ и $S_4 = [136, +\infty)$. Број непознатих, тј. оцењених параметара је $l = 1$. Вероватноће су:

$$p_{01} = P_{H_0}\{-\infty < X < 124\} = 0,392,$$

$$p_{02} = P_{H_0}\{124 \leq X < 130\} = 0,149,$$

$$p_{03} = P_{H_0}\{130 \leq X < 136\} = 0,139,$$

$$p_{04} = 1 - (p_{01} + p_{02} + p_{03}) = 0,32.$$

Број степени слободе χ^2_0 статистике је $4 - 1 - 1 = 2$, тако да је $\chi^2_2 = 159,25$. Критична област је $C = \left[\chi^2_{2;0,99}, +\infty\right) = [9,21; +\infty)$. Како $159,25 \in C$, то се хипотеза H_0 одбацује, тј. закључује се да ови подаци нису сагласни са χ^2_{129} расподелом. $\triangle$

У пракси се за тестирање сагласности са задатом расподелом често користи формулација испитивање сагласности очекиваних, e_i и опсервираних вредности, o_i, поготову код расподела дискретног типа. У случају обележја дискретног типа примењује се општи принцип који је горе изложен на веома једноставан начин. Овај случај се посебно истиче и због тога што образлаже како се тестира сагласност расподеле квалитативног обележја са унапред претпостављеном дискретном расподелом.

Пример 75. Произвођач износи на тржиште шест различитих жвакаћих гума у типизираном паковању. На основу узорка обима 60 у коме је регистровано 13, 18, 11, 8, 5, 5 продатих комада по типовима жвакаћих гума редом, тестирати хипотезу да је вероватноћа продаје за сваки тип жвакаће гуме иста. Тестирање извршити са прагом значајности $\alpha = 0,05$.

Нека је A_i, $i = 1, 2, \dots, 6$ тип жвакаће гуме. Тада је

$$H_0 : P(A_i) = p_{0i} = \frac{1}{6}, \quad i = 1, \dots, 6\,,$$

па је

$$e_i = np_{0i} = 60\frac{1}{6} = 10, \quad i = 1, \dots, 6\,,$$

јер треба, заправо, тестирати сагласност узорка са дискретном униформном расподелом.

Ако је X_i учестаност са којом је догађај A_i исход експеримента, тада је

$$\chi^2_5 = \sum_{i=1}^{5} \frac{(o_i - e_i)^2}{e_i} = \frac{(13-10)^2}{10} + \frac{(18-10)^2}{10} + \frac{(11-10)^2}{10} + \frac{(8-10)^2}{10} + {} + \frac{(5-10)^2}{10} + \frac{(5-10)^2}{10} = 12,8$$

реализована вредност тест статистике која има χ^2–расподелу са 5 степени слободе. Граница критичне области задовољава услов

$$P(\chi^2_5 \geq 11,1) = 0,05\,,$$

односно критична област је $C = [11,1; +\infty)$. Како је $12,8 > 11,1$ хипотезу H_0 одбацујемо са 5%–ним прагом значајности. $\triangle$

Б. Тестирање једнакости две мултиномне расподеле

Посматрајмо две независне мултиномне расподеле са параметрима $n_j, p_{1j}, p_{2j}, \dots, p_{kj}$, $j = 1, 2$ респективно. Нека X_{ij}, $i = 1, 2, \dots, k$, $j = 1, 2$ представљају одговарајуће учестаности. Ако су n_1 и n_2 велики, случајна променљива

$$\sum_{j=1}^{2}\sum_{i=1}^{k} \frac{(X_{ij} - n_j p_{ij})^2}{n_j p_{ij}} = \sum_{i=1}^{k} \frac{(X_{i1} - n_1 p_{i1})^2}{n_1 p_{i1}} + \sum_{i=1}^{k} \frac{(X_{i2} - n_2 p_{i2})^2}{n_2 p_{i2}}$$

је збир две стохастички независне случајне променљиве од којих свака има расподелу χ^2_{k-1}. То значи да је наведена случајна променљива са расподелом χ^2_{2k-2}. Тестирамо хипотезу

$$H_0 : p_{11} = p_{12},\, p_{21} = p_{22}, \dots,\, p_{k1} = p_{k2}$$

где су сви $p_{i1} = p_{i2}$, $i = 1, 2, \ldots, k$ непознати. Стога су нам потребне тачкасте оцене ових параметара. Статистика максималне веродостојности за ове параметре у случају када је $p_{i1} = p_{i2}$ је

$$\widehat{p}_{ij} = \frac{X_{i1} + X_{i2}}{n_1 + n_2}\,, \quad i = 1, 2, \ldots, k\,, \quad j = 1, 2\,.$$

Приметимо да нам је потребна само $k-1$ оцена, јер ћемо оцену за $p_{k1} = p_{k2}$ имати самим тим што смо нашли оцене за првих $k-1$ вероватноћа. Дакле, случајна променљива

$$\sum_{j=1}^{2}\sum_{i=1}^{k} \frac{\left(X_{ij} - n_j\frac{X_{i1}+X_{i2}}{n_1+n_2}\right)^2}{n_j\frac{X_{i1}+X_{i2}}{n_1+n_2}}$$

има приближно χ^2 расподелу са $2k - 2 - (k-1) = k - 1$ степени слободе.

Пример 76. Тестирати хипотезу о једнакој заступљености четири типа личности у популацији становништва два града на основу независних узорака обима 100 са 5%–ним прагом значајности:

Тип личности	колерик	сангвиник	меланхолик	флегматик
Остварене учестаности у граду I	30	25	23	22
Остварене учестаности у граду II	25	27	23	25

Оценимо најпре, на основу датог узорка, непознате вероватноће:

$$\widehat{p}_{11} = \widehat{p}_{12} = \frac{30+25}{100+100} = 0,275\,, \quad \widehat{p}_{21} = \widehat{p}_{22} = \frac{25+27}{200} = 0,26\,,$$

$$\widehat{p}_{31} = \widehat{p}_{32} = \frac{23+23}{200} = 0,23 \quad \text{и} \quad \widehat{p}_{41} = \widehat{p}_{42} = \frac{22+25}{200} = 0,235\,.$$

С обзиром да је $n_1 = n_2 = 100$, то је

$$n_j\widehat{p}_{1j} = 27,5\ , \quad n_j\widehat{p}_{2j} = 26\,, \quad n_j\widehat{p}_{3j} = 23\,, \quad n_j\widehat{p}_{4j} = 23,5\ , \quad j = 1, 2\,.$$

Тест статистика за проверу постављене нулте хипотезе

$$\chi_0^2 = \sum_{j=1}^{2}\sum_{i=1}^{k} \frac{(X_{ij} - n_j\widehat{p}_{ij})^2}{n_j\widehat{p}_{ij}}$$

ће имати χ^2–расподелу са $4-1=3$ степена слободе, те је њена реализована вредност

$$\chi_3^2 = \frac{(30-27,5)^2}{27,5} + \frac{(25-26)^2}{26} + \frac{(23-23)^2}{23} + \frac{(22-23,5)^2}{23,5} +$$

$$+ \frac{(25-27,5)^2}{27,5} + \frac{(27-26)^2}{26} + \frac{(23-23)^2}{23} + \frac{(25-23,5)^2}{23,5} = 0,72\,,$$

а критична област је $[7,81;+\infty)$. Закључак је да се са 5%–ним прагом значајности може сматрати да је подједнак број сваког од типова личности у оба града, јер $0,72 \notin C$. $\triangle$

В. Испитивање независности χ^2 тестом (табеле контингенције)

χ^2 тестом често се испитује и независност два обележја исте популације. Дакле, за два обележја X и Y једне популације, тестира се хипотеза

$$H_0 : X \text{ и } Y \text{ су независна обележја}$$

против алтернативне да нису независна.

Тестирање се обавља тако што се на узорку обима n из посматране популације региструју "вредности" оба обележја (обележја не морају бити нумеричка, квантитативна, већ могу бити и квалитативна код каквих се ово тестирање најчешће и спроводи). При томе се формира табела (табела контингенције, случајности) која у пољу (i,j) има податак n_{ij} о броју елемената у узорку, код којих обележје X има вредност x_i, а обележје Y вредност y_j:

$X\downarrow/Y\rightarrow$	y_1	y_2	$\dots$	y_r	$n_{i\bullet}$
x_1	n_{11}	n_{12}	$\dots$	n_{1r}	$n_{1\bullet}=\sum_j n_{1j}$
x_2	n_{21}	n_{22}	$\dots$	n_{2r}	$n_{2\bullet}=\sum_j n_{2j}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_{k1}	n_{k2}	$\dots$	n_{kr}	$n_{k\bullet}=\sum_j n_{kj}$
$n_{\bullet j}$	$n_{\bullet 1}= \sum_i n_{i1}$	$n_{\bullet 2}= \sum_i n_{i2}$	$\dots$	$n_{\bullet r}= \sum_i n_{ir}$	n

.

У табели су коришћене следеће ознаке:

- n_{ij} је број парова за које је $X=x_i$ и $Y=y_j$;

- $n_{i\bullet}$ је број парова за које је $X = x_i$;
- $n_{\bullet j}$ је број парова за које је $Y = y_j$;

Према томе, нулта хипотеза: X и Y су независна обележја, може да се искаже као

$$H_0 : \forall(i,j)\ p_{ij} = p_{i\bullet}p_{\bullet j},$$

а алтернативна

$$H_1 : \exists(i,j)\ p_{ij} \neq p_{i\bullet}p_{\bullet j},$$

где је $p_{ij} = P\{X = x_i \wedge Y = y_j\}$, $p_{i\bullet} = P\{X = x_i\}$, $p_{\bullet j} = P\{Y = y_j\}$.

Вероватноће $p_{i\bullet}$ и $p_{\bullet j}$ се оцењују на основу релативних учестаности

$$\widehat{p}_{i\bullet} = \frac{n_{i\bullet}}{n},\ \widehat{p}_{\bullet j} = \frac{n_{\bullet j}}{n}.$$

Ако је хипотеза H_0 тачна, непозната вероватноћа p_{ij} оцењује се на следећи начин

$$\widehat{p}_{ij} = \widehat{p}_{i\bullet}\widehat{p}_{\bullet j} = \left(\frac{n_{i\bullet}}{n}\right)\cdot\left(\frac{n_{\bullet j}}{n}\right) = \frac{n_{i\bullet}n_{\bullet j}}{n^2}.$$

Очекивани број парова за које је $X = x_i$ и $Y = y_j$ оцењује се са

$$\widehat{n}_{ij} = n\widehat{p}_{ij} = \frac{n_{i\bullet}n_{\bullet j}}{n}.$$

Релативно одступање од хипотезе о независности за "ћелију" (i,j) је

$$\frac{(n_{ij} - \widehat{n}_{ij})^2}{\widehat{n}_{ij}}\,.$$

Отуда је укупно одступање за све ћелије

$$\chi_0^2 = \sum_{i=1}^{k}\sum_{j=1}^{r}\frac{(n_{ij} - \widehat{n}_{ij})^2}{\widehat{n}_{ij}} = \sum_{i=1}^{k}\sum_{j=1}^{r}\frac{n_{ij}^2}{\widehat{n}_{ij}} - n = \sum_{i=1}^{k}\sum_{j=1}^{r}\frac{o_{ij}^2}{e_{ij}} - n\,.$$

Као и раније, "o" потиче од речи "obsrved", а "e" од речи "expected".

Ако је хипотеза H_0 тачна, χ_0^2 има χ^2 расподелу са $(k-1)(r-1)$ степени слободе. Критична област теста је:

H_0	H_1	C
X и Y су независна обележја	X и Y нису независна обележја	$\chi_0^2 \geq \chi^2_{(k-1)\cdot(r-1);1-\alpha}$

Дакле, ако је за реализовани узорак $\chi_0^2 < \chi^2_{(k-1)\cdot(r-1);1-\alpha}, H_0$ се прихвата. Међутим, и овде се може упоређивати реализована вредност статистике и очекивање случајне променљиве са одговарајућом χ^2 расподелом (H_0 се прихвата ако је $\chi_0^2 < (k-1)(r-1)$).

Пример 77. Испитује се независност висине и тежине становника једне регије. Сви резултати мерења су сврстани у следеће категорије: високи и гојазни, високи и негојазни, ниски и гојазни и ниски и негојазни и приказани табелом

Y (висина) $\rightarrow$ X (тежина) $\downarrow$	високи	ниски	$n_{i\bullet}$	$\widehat{p}_{i\bullet}$
гојазни	14	36	50	0,161
негојазни	59	201	260	0,839
$n_{\bullet j}$	73	237	310	/
$\widehat{p}_{\bullet j}$	0,235	0,765	/	1

То даје:

$$\widehat{p}_{11} = \frac{50\cdot 73}{310^2} \;\Rightarrow\; \widehat{n}_{11} = 310\cdot\widehat{p}_{11} = 11,8$$

$$\widehat{p}_{12} = \frac{50\cdot 237}{310^2} \;\Rightarrow\; \widehat{n}_{12} = 310\cdot\widehat{p}_{12} = 38,2$$

$$\widehat{p}_{21} = \frac{260\cdot 73}{310^2} \;\Rightarrow\; \widehat{n}_{21} = 310\cdot\widehat{p}_{21} = 61,2$$

$$\widehat{p}_{22} = \frac{260\cdot 237}{310^2} \;\Rightarrow\; \widehat{n}_{22} = 310\cdot\widehat{p}_{22} = 198,8\,,$$

па је

$$\chi_0^2 = \frac{(14-11,8)^2}{11,8} + \ldots + \frac{(201-198,8)^2}{198,8} = 0,64.$$

Број степени слободе је $(2-1)(2-1) = 1$ па је критична област $[3,84;+\infty)$. Дакле, H_0 се прихвата, тј. може се сматрати да су за ту популацију висина и тежина независна обележја. $\triangle$

Уочимо да је у претходном примеру број степени слободе примењене тест статистике са χ^2–расподелом био 1. По правилу се у таквим случајевима примене Пирсоновог теста, из разлога у које у оквиру овог курса нећемо улазити, врши тзв. Јејцова (Yates' correction) корекција:

$$\chi^2_{\text{кориговано}} = \sum_{i=1}^{k} \frac{(|o_i - e_i| - 0,5)^2}{e_i},$$

где је o_i опсервирана апсолутна учестаност, а e_i очекивана апсолутна учестаност.

Дакле, у последњем примеру би требало извршити корекцију.

Коригована вредност из примера је: $\chi^2_{\text{кориговано}} = 0,38$, а за праг значајности $\alpha = 0,05$ одговарајућа критична област је $C = [3,84; +\infty)$. Како вредност 0,38 не припада скупу C прихвата се хипотеза да су обележја независна, тј. закључак се не мења у односу на онај који је начињен без корекције.

Уочимо да смо кориговањем дошли до мање вредности за тест статистику на основу истог узорка. Ово је закључак који важи увек при примени Јејцове корекције, $\chi^2_{\text{кориговано}} < \chi^2_0$. Отуда, ако је реализована вредност тест статистике мања од леве границе критичне области, корекцију не треба ни вршити, јер неће утицати на закључак о прихватању нулте хипотезе.

Пример 78. Испитано је 500 особа којима је постављено питање да ли би на предстојећим изборима своје поверење поклонили кандидату који би био мушкарац или радије кандидату који би био женског пола и добијени су следећи резултати:

Бирачи \ Кандидат	Мушкарац	Жена
Мушкарци	180	90
Жене	130	100

Испитати да ли избор кандидата зависи од пола бирача за $\alpha = 0,05$.

Тестира се хипотеза H_0(избор кандидата не зависи од пола бирача). Реализована вредност тест статистике је $\chi^2_0 = 5,43$, а број степени слободе јој је 1. Критична област је, као и у претходном примеру, $C = [3,84\,; +\infty)$. Дакле, требало би одбацити нулту хипотезу, односно, закључити да ће ишод избора зависити од пола бирача који буду изашли на изборе. Међутим, применом Јејцове корекције на исти узорак долазимо до податка $\chi^2_{\text{кориговано}} = 3,47$, што значи да се хипотеза H_0 прихвата, односно да избор кандидата (мушкарца или жене) неће зависити од пола гласача који буду изашли на изборе.

Осврнимо се овде још на чињеницу да је кориговања вредност 3,47 релативно близу границе критичне области, што по правилу налаже додатну анализу. У овој ситуацији примерено би било тражити већи обим узорка, или ако то није могуће, закључивати са неким другим прагом значајности. Рецимо, за праг значајности $\alpha = 0,1$, критична област ће бити $C = [2,71\,;\,+\infty)$, па би се на овом нивоу значајности нулта хипотеза одбацила. $\triangle$

$\star \quad \star \quad \star$

Предност непараметарских тестова над параметарским је у томе што нису неопходна никаква претходна знања о облику расподеле обележја у вези са којим се врши тестирање (код параметарских тестова су неопходне претпоставке о расподели обележја које се тестира). Међутим, ако су испуњени критеријуми за примену параметарских тестова, онда њих треба и применити јер су ефикаснији од одговарајућих непараметарских.

Дефиниција 44. Од два теста за тестирање исте нулте хипотезе против исте одговарајуће алтернативне, *ефикаснији* је онај за који је потребан мањи обим узорка да би се постигла једнака вероватноћа одбацивања нулте хипотезе ако је она заиста погрешна, односно једнака вредност функције моћи:

$$P_{H_1}\{(X_1, \ldots, X_n) \in C\} = 1 - P_{H_1}\{(X_1, \ldots, X_n) \in C^c\} = 1 - \beta\,,$$

дакле, онај који има већу моћ. $\diamondsuit$

У наставку су изложени још неки од непараметарских тестова који су често у употреби.

4.3.3 Биномни тест (тест знакова)

Тест знакова се заснива на посматрању реализације неког догађаја A у низу независних опита, за које је вероватноћа реализације догађаја A у сваком појединим опиту $P(A) = p$. Остварење догађаја A се означава са "+", а неостварење са "–".

Навешћемо три примене овог теста.

1. Биномни тест се може користити **за тестирање једнакости расподела два обележја** X и Y апсолутно непрекидног типа:

$$H_0 : F_X = F_Y \quad \text{и} \quad H_1 : F_X \neq F_Y$$

на основу два независна узорка $(X_1, \ldots, X_n)$ и $(Y_1, \ldots, Y_n)$ **истог обима**. Узорци се не смеју сређивати у варијациони низ нити интервално, а упоређивање регистрованих (реализованих) вредности из два узорка има суштинског смисла.

Пример 79. У фабрици постоје две независне линије за производњу једног производа. Тестира се хипотеза да је квалитет производа произведених на овим линијама исти кроз регистровање броја дефектних производа у току 10 дана на обе линије. Регистровани подаци о броју дефектних производа су приказани табелом:

Дан	1	2	3	4	5	6	7	8	9	10
Линија I	172	165	206	184	174	142	190	169	161	200
Линија II	201	179	159	192	177	170	182	179	169	210
	$-$	$-$	$+$	$-$	$-$	$-$	$+$	$-$	$-$	$-$

.

Тестирамо хипотезу о једнакој расподели броја дефектних производа на две независне производне линије (остварен у истом периоду), против алтернативне да су расподеле различите.

За сваки од посматраних 10 дана бележићемо знак "+" уколико је на линији I регистровано више дефектних производа него на линији II, а знак "−" у супротном случају. Означимо са T број знакова "+" у посматраном узорку. За наш реализовани узорак је $t = 2$. (Уколико би се у неком дану констатовао једнак број дефектних призвода на обе производне линије, такав дан би се у узорку игнорисао, односно посматрао би се узорак из кога би били елиминисани овакви дани.)

Нека остварење знака "+" у низу значи реализацију догађаја A. У општем случају, T би била случајна променљива са биномном расподелом. Уколико је нулта хипотеза тачна, у низу регистрованих знакова треба да је приближно једнак број знакова "+" и "−", односно, $P(+) = P(A) = 0,5$. На тај начин се нулта хипотеза $H_0 : F_X = F_Y$ преводи у хипотезу

$$H_0 : \;\; T : \mathcal{B}(10\,;\, 0,5)$$

а алтернативна H_1 би била да T нема назначену биномну расподелу. Баш из овог разлога сам тест носи назив биномни тест.

Из претходне анализе следи да су за нулту хипотезу проблематичне како мале, тако и велике вредности тест статистике T.

Претпоставимо да вршимо тестирање са прагом значајности $\alpha = 0,05$ и $\alpha = 0,1$. Одредимо критичну област.

Ако критичну област чине вредности $T = 0$ и $T = 10$, под условом да је нулта хипотеза тачна, вероватноћа грешке прве врсте би била

$$\alpha = P\{T = 0 \vee T = 10\} = \binom{10}{0} \cdot 0,5^{10} + \binom{10}{10} \cdot 0,5^{10} \;=\; 0,002\,.$$

Међутим, ова вредност за α је сувише мала, па значи да треба проширити критичну област. Ако у критичну област укључимо и $T = 1$ и $T = 9$, добијамо да је $\alpha = 0,022$, што је и даље мала вредност у односу на задате вредности вероватноће грешке прве врсте за ово тестирање. Дакле, још једном проширимо критичну област и дефинишимо је са $T = 0, 1, 2, 8, 9, 10$. У том случају је $\alpha = 0,11$. Уколико смо задовољни овом прецизношћу, можемо да спроведемо жељено тестирање. Према томе, с обзиром да је реализована вредност тест статистике $t = 2$ и да она припада критичној области, нулту хипотезу одбацујемо.$\triangle$

За мали обим узорка, као што је приказано у претходном примеру, директно се користи биномна расподела за одређивање критичне области величине α. Међутим, за грубу процену или за велике узорке, користи се нормална апроксимација биномне расподеле, тј. статистика

$$Z_0 = \frac{T - np_0}{\sqrt{np_0(1 - p_0)}}$$

где је T–број знакова " + " у низу знакова добијеном на основу узорка и која за $n > 50$ има приближно $\mathcal{N}(0, 1)$ расподелу. (Стварна расподела статистике T је $\mathcal{B}(n, p_0)$.)

Критичне области се одређују у зависности од алтернативне хипотезе, као и код параметарског теста за тестирање параметра p биномне расподеле (в. потпоглавље 4.2.3).

Познато је да за мали обим узорка постоји и одговарајућа модификована тест статистика која има приближно Фишерову расподелу, о чему овде неће бити речи.

2. Описани тест се често назива и **тест квантила**, јер се њиме за обележје X може да испитује хипотеза

$$H_0 : P\{X \leq x_0\} = p_0,$$

где су x_0 и p_0 задате вредности и реализација догађаја $\{X \leq x_0\}$ у узорку се означи са $" + "$, а њему супротног са $" - "$. Могуће алтернативне хипотезе су:

$$H_1 : P\{X \leq x_0\} \neq p_0\,,\ \ H_1 : P\{X \leq x_0\} > p_0 \ \text{ и } \ H_1 : P\{X \leq x_0\} < p_0.$$

С обзиром на дефиницију квантила, нулта хипотеза се може да искаже и као:

$$H_0 : M_{p_0} = x_0,$$

а алтернативне редом као

$$H_1 :\ M_{p_0} \neq x_0\,, \quad H_1 :\ M_{p_0} < x_0 \quad \text{ и } \quad H_1 :\ M_{p_0} > x_0\,.$$

Дакле, третира се проблем једног узорка и тестира параметар расподеле.

Пример 80. Посматра се група од 100 двадесетогодишњака и региструје број оних који су конформисти. За дату групу добијено је да 68 њих има дату особину. Тестирати хипотезу да ће конформизам бити особина 75% двадесетогодишњака за $\alpha = 0,05$.

Обим узорка једнак је $n = 100$, а $p_0 = 0,75$ је вероватноћа конформисте међу двадесетогодишњацима. Од укупно посматраних $t = 68$ је број оних који имају особину конформизма. Реализована вредност тест статистике је $z_0 = -1,617$ а критична област је $C = (-\infty; -1,96] \cup [1,96; +\infty)$. Како реализована вредност не припада критичној области, то се нулта хипотеза прихвата, тј. може се сматрати да је 75% конформиста међу двадесетогодишњацима. $\triangle$

Када је $p_0 = 1/2$ за овај тест се користи назив **тест медијане**.

Тест медијане се може применити и за два независна узорка различитих обима када се тестира хипотеза о томе да узорци потичу од два обележја са истом медијаном, или још боље из истог обележја (на једној те истој популацији).

3. Тест квантила се користи и за "спарене узорке", односно за **тестирање хипотезе да није дошло до промене расподеле обележја** X апсолутно непрекидног типа под дејством различитих фактора утицаја на елементе једне популације. За ово тестирање се региструју вредности обележја X на истим елементима популације два пута, тј. под различитим околностима и добија се тзв. спарени узорак

$$\left((X_1^{(1)}, X_1^{(2)}) \ldots, (X_n^{(1)}, X_n^{(2)})\right) .$$

Пример 81. На 12 клијената је примењивана одговарајућа групна психотерапија са циљем да се ублажи или уклони депресивно стање у коме су се налазили. Дати су кодирани нивои психичког стања сваког пацијента пре ($X^{(1)}$) и после ($X^{(2)}$) спроведене терапије.

Кл.	1	2	3	4	5	6	7	8	9	10	11	12
$x^{(1)}$	5,6	7,1	6,4	5,8	4,9	4,7	5,0	4,9	3,6	5,4	4,7	3,1
$x^{(2)}$	5,6	6,3	6,7	5,3	4,0	5,2	4,9	5,2	3,3	4,8	3,2	2,4
Δ	0	+	−	+	+	−	+	−	+	+	+	+

У циљу спровођења статистичког теста дефинише се статистика

$$\Delta = sgn\left(X^{(1)} - X^{(2)}\right),$$

чије су вредности: 0 – уколико је $X^{(1)} = X^{(2)}$, "+" – уколико је $X^{(1)} > X^{(2)}$ и "−" – уколико је $X^{(1)} < X^{(2)}$.

Ако се код теста знакова у низу "+" и "−" јави и 0, онда се, као што смо већ навели, овакав елемент узорка надаље игнорише и ради се са узорком смањеног обима. Ово стога што реализација таквог догађаја има вероватноћу 0, с обзиром на претпоставку о расподели апсолутно непрекидног типа.

Ако нема разлике у расподелама, сматра се да је

$$P\{\Delta = "+"\} = P\{\Delta = "-"\} = 0,5\ ,$$

тј. да је једнако вероватно да је $X^{(1)} > X^{(2)}$ ($P\{X^{(1)} > X^{(2)}\} = 0,5$) као и $X^{(1)} < X^{(2)}$ ($P\{X^{(1)} < X^{(2)}\} = 0,5$).

Тестира се хипотеза да **није** дошло до смањења депресије

$$H_0 : P\left\{X^{(1)} > X^{(2)}\right\} = 0,5$$

против алтернативне

$$H_1 : P\left\{X^{(1)} > X^{(2)}\right\} > 0,5$$

да **је** дошло до смањења депресије.

Први клијент се изоставља из разматрања (јер је код њега регистровано $x^{(1)} = x^{(2)}$), па се ради са $n = 11$ и $t = 8$. С обзиром да обим узорка није "довољно велики", само ће се грубо проценити одговор на постављено питање применом нормалне апроксимација тест статистике, пре свега ради демонстрације поступка, а не за извођење озбиљног и одговорног закључка (где јој, заправо, није место), која даје

$$z_0 = \frac{8 - 11 \cdot 0,5}{\sqrt{11 \cdot 0,5 \cdot 0,5}} = 1,51.$$

За једнострану алтернативу H_1 критичне су велике вредности за z_0, тако да је за $\alpha = 0,05$ граница критичне области $z_{0,5-0,05} = z_{0,495} = 1,645$, односно критична област је $C = [1,645; +\infty)$. Како $z_0 \notin C$ прихвата се H_0, што значи да се не примећује значајна разлика у психичком стању клијената пре и после спроведене терапије. $\triangle$

Уместо понуђене тест статистике, може се користити и статистика

$$Z = \frac{|D| - 1}{\sqrt{n}},$$

где је $D =$број$(+)-$број$(-)$, која такође има апроксимативно $\mathcal{N}(0,1)$ расподелу.

4.3.4 Тест серија (тест корака)

Тест серија се заснива на испитивању случајности међусобног распореда две врсте објеката, нпр. 0 и 1.

Нека се 0 јавља n_1 пута, а 1 јавља n_2 пута, у заједничком низу од $n = n_1 + n_2$ елемената. Серију чини подниз истих елемената, било 0 било 1. Нека је U укупан број серија у добијеном низу. Ако је хипотеза о случајном распореду 0 и 1 тачна, важе формуле:

$$E(U) = \frac{2n_1n_2}{n} + 1\,, \quad D(U) = \frac{(E(U)-1)(E(U)-2)}{n-1}\,, \quad n = n_1 + n_2\,.$$

За расподелу статистике U постоје посебне таблице, али се показало да се већ за $n_1n_2 \geq 9$ може користити нормална апроксимација за стандардизовану вредност

$$Z_0 = \frac{U - E(U)}{\sqrt{D(U)}}\,.$$

Тест корака се користи најчешће за: 1) тестирање случајности низа података, као и за 2) испитивање једнакости расподела двају обележја апсолутно непрекидног типа.

1. У наредном примеру приказано је **тестирање случајности** низа података.

   Пример 82. Разматра се пример о зарадама радника (пример 73). Упоређују се зараде радника из другог погона (последњих 20 радника у узорку) у односу на медијану зарада свих радника у узорку (односно у односу на претпостављену медијану зарада свих запослених у том предузећу), $M_{0,5} = 825$ на основу узорка који није сређен интервално.

   Величина зараде мања од медијане означава се са 0, а већа са 1. Добија се низ: 1, 0, 1, 1, 1, 1, 1, 0, 0, 1, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0. Тестира се хипотеза да је добијени низ случајан са прагом значајности $\alpha = 0,05$. С обзиром да је проблематичан мали број серија у низу података, ради се о једностраној алтернативној хипотези, па је $C = (-\infty; -1,645]$.

   $$n_1 = n_2 = 10, \quad E(U) = 2\frac{10\cdot 10}{10+10} + 1 = 11,$$

   $$D(U) = \frac{(11-1)(11-2)}{20-1} = (2,18)^2,$$

$$z_0 = \frac{10 - 11}{2,18} = -0,46\,.$$

Дакле, $z_0 \notin C$ па се H_0 прихвата, односно, радници другог погона би чинили репрезентативан узорак при испитивању обележја "висина зарада запослених" (у посматраном предузећу). Другим речима, о заради запослених у предузећу које чине два поменута погона може се закључивати само на основу зарада радника другог погона, јер је расподела зарада иста у другом погону и у целом предузећу. $\triangle$

За нормалну апроксимацију, критична област је:

H_0	H_1	C
Низ је случајан	У низу је мали број серија	$z_0 \leq -z_{0,5-\alpha}$

.

Испитивање случајности се врши и у односу на друге нивое осим медијане.

Важно је да се за тестирање случајности тестом корака не сме сређивати узорак ни по ком критеријуму. Дакле, реализовани узорак се користи онакав какав је и настао у серији мерења (посматрања).

2. Код **тестирања једнакости расподела два обележја**, ситуација је следећа.

   Разматрају се два обележја апсолутно непрекидног типа, X и Y, на независним узорцима обима n_1 и n_2 редом. Тестира се хипотеза да оба обележја имају исту расподелу, тј.

$$H_0 : F_X = F_Y$$

   против алтернативне $H_1 : F_X \neq F_Y$.

   Тест серија примењује се тако што се елементи оба реализована узорка поређају у јединствен неопадајући низ, па се елементи првог узорка означавају са 0, другог са 1, чиме се добије низ од $n_1 + n_2$ симбола 0 и 1. Ако је хипотеза H_0 тачна, распоред 0 и 1 је случајан. При томе се као сумњива за хипотезу H_0 сматра само појава малог броја серија (већа груписања 0 и 1) у низу симбола. Према томе, примењује се иста тест статистика Z_0 као и за претходни тест серија, важе исте формуле за $E(U), D(U)$ и n. Критична област дата је у табели

H_0	H_1	C
$F_X = F_Y$	$F_X \neq F_Y$	$z_0 \leq -z_{0,5;-\alpha}$

,

а пре сазнања о границама критичне области, добијене мале вредности за z_0 знак су да са статистичким закључивањем на основу такве серије треба бити обазрив.

4.3.5 Тест рангова (тест Вилкоксон – Ман – Витнија)

Тест рангова се, као и тест серија, заснива на испитивању случајности појављивања 0 и 1 у низу изведеном из реализованог узорка, а осетљивији је од теста серија.

И за њега ћемо навести две могуће примене и то: 1) тестирање хипотезе о случајности и 2) тестирање једнакости двеју расподела апсолутно непрекидног типа.

1. **Пример 83.** Посматрајмо низ нула и јединица из примера 82. Тестирајмо **хипотезу о случајности** применом теста рангова за $\alpha = 0,05$.

   Утврђује се колико укупно јединица има испред сваке појединe нуле. За разматрани низ: 1, 0, 1, 1, 1, 1, 1, 0, 0, 1, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, добија се следећи број инверзија (инверзијом у низу се сматра појава 1 испред 0) 1, 6, 6, 7, 7, 8, 10, 10, 10, 10.

   Као тест статистика користи се $V-$укупан број јединица лево од сваке нуле у низу посебно, односно за разматрани пример је
   $$V = 1+6+6+7+7+8+10+10+10+10 = 75\,.$$

   С обзиром на укупан број нула и јединица у низу:
   $$0 \leq V \leq 100\,.$$

   За хипотезу о случајности критичан је како мали тако и велики број инверзија у низу нула и јединица, јер то указује на груписање нула или јединица на почетку низа. У општем случају, за n_1 нула и n_2 јединица у низу је $0 \leq V \leq n_1 n_2$. Ако је хипотеза о случајности тачна и $n_1 \geq 10$ и $n_2 \geq 10$, статистика V има расподелу веома блиску нормалној са

$$E(V) = \frac{n_1 n_2}{2}, \quad D(V) = E(V) \cdot \frac{n+1}{6}, \quad n = n_1 + n_2 .$$

У разматраном примеру је $n_1 = 10$, $n_2 = 10$ и

$$E(V) = \frac{10 \cdot 10}{2} = 50 , \quad D(V) = 50 \frac{20+1}{6} = 175 , \quad \sqrt{D(V)} = 13,23 .$$

Стога се уместо тачне расподеле користи нормална апроксимација, тј. статистика

$$Z_0 = \frac{V - E(V)}{\sqrt{D(V)}}$$

која има приближно $\mathcal{N}(0;1)$ расподелу. Овде је, значи,

$$z_0 = \frac{75-50}{13,23} = 1,89.$$

Како је проблематичан и велики и мали број инверзија, критична област је:

H_0	H_1	C
Низ је случајан	У низу је премало или превише инверзија	$\lvert z_0 \rvert \geq z_{0,5-\frac{\alpha}{2}}$

.

За разматрани пример је

$$z_{0,475} = 1,96 , \quad \text{те је} \quad C = (-\infty; -1,96] \cup [1,96; +\infty) ,$$

па се, с обзиром да $z_0 = 1,89$ не припада критичној области, прихвата хипотеза о случајности разматране серије. Дакле, дошли смо до истог закључка као и после примене теста серија. Овде је, међутим, важно уочити да је реализована вредност $z_0 = 1,89$ блиска граници критичне области (1,96) чиме је пољуљана "поузданост" закључка. У практичним применама није пожељно ослањати се на такве закључке. Излаз се може тражити у повећању обима узорка или у закључивању на неком другом нивоу значајности. $\triangle$

2. Тест рангова такође се користи и за **тестирање једнакости расподела**. За апсолутно непрекидна обележја X и Y преко независних узорака обима n_1 и n_2 тестира се хипотеза о једнакости расподела полазећи од низа 0 и 1 дужине $n = n_1 + n_2$, који се формира на исти начин као код теста серија. И нулта хипотеза је иста:
$$H_0 : F_X = F_Y \, .$$
Користи се тест статистика Z_0, иста и под истим условима као код испитивања случајности, а критична област дата је у табели

H_0	H_1	C
$F_X = F_Y$	$F_X \neq F_Y$	$\lvert z_0 \rvert \geq z_{0,5-\alpha/2}$

.

Осетљивост овог теста, међутим, састоји се у томе што је погодан за тестирање нулте хипотезе $H_0 : F_X = F_Y$, против алтернатива H_1 : ”X је чешће веће од Y”, и H_1 : ”X је чешће мање од Y”. Одговарајуће критичне области су у том случају:

H_0	H_1	C
$F_X = F_Y$	X је чешће веће од Y	$z_0 \geq z_{0,5-\alpha}$
$F_X = F_Y$	X је чешће мање од Y	$z_0 \leq -z_{0,5-\alpha}$

.

Пример 84. Две групе, свака од по 12 експерименталних мишева оболелих од рака, изложене су хемотерапији да би се проверило њено дејство на ћелије рака. Осим тога, само другој групи мишева дат је истовремено и антитоксин који је имао за циљ да спречи уништавање здравих ћелија приликом примене хемотерапије. Мерено је време (у сатима) преживљавања експерименталних животиња у односу на почетак примене терапије. Експеримент је окончан након 480 сати, тј. након 20 дана, те је за животиње које су преживеле овај период регистровано време живота 480 сати. Времена преживљавања експерименталних животиња дата су табелом

I група	84	128	168	92	184	92	76	104	72	180	144	120
II група	140	184	368	96	480	188	480	244	440	380	480	196

.

Да ли се на основу овог експеримента на нивоу значајности $\alpha = 0,05$ може тврдити да ће применом антитоксин терапије бити продужен живот пацијената који се излажу хемотерапији у односу на пацијенте којима се не укључује антитоксин?

Сређивањем реализованих узорака у јединствен неопадајући низ и додељивањем 0 сваком елементу I експерименталне групе, а 1 сваком елементу II експерименталне групе, добијамо низ

$$0,0,0,0,0,1,0,0,0,1,0,0,0,0,1,1,1,1,1,1,1,1,1,1\,.$$

На основу овог низа тестирамо хипотезу

$$H_0: \text{Нема разлике у дужини живота I и II групе мишева}\;,$$

против алтернативне

$$H_1: \text{Дужина живота мишева II групе је већа у односу на I групу}\,.$$

Дакле,

$$V = 1+1+1+2+2+2+2 = 11\,,\quad E(V) = 72\,,\quad D(V) = 300\,,$$

док је реализована вредност статистике Z_0

$$z_0 = \frac{11-72}{\sqrt{300}} = -3,523\,.$$

Критична област овог теста је $C = (-\infty\,;\,-1,645]$. Како реализована вредност тест статистике припада критичној области, то нулту хипотезу одбацујемо у корист алтернативне. Другим речима, на нивоу значајности $\alpha = 0,05$ се може закључити да је применом антитоксина у комбинацији са хемотерапијом значајно продужен живот експерименталних животиња.$\triangle$

Глава 5

Теорија одлучивања

У многим случајевима коначан циљ статистичке анализе се може интерпретирати у облику одлучивања о одређеном понашању или деловању. Ево неколико примера. При узорачкој контроли производње, треба донети једну од две одлуке: прихватити понуђену партију производа или је одбацити. Затим, лекар на основу анализе симптома болести код одређеног болесника мора да се понесе са болешћу на један од коначно много познатих начина, тј. мора да донесе једну од коначно много одлука како да третира болесника. Приликом анализе случајног процеса са коначним, али непознатим очекивањем, на основу резултата посматрања тог процеса, треба донети одлуку о величини дејства на процес (његову корекцију) за "померање" очекивања, на пример, у нулу. У последњем примеру то дејство може бити изражено неким реалним бројем t, па је отуда у овом случају број могућих одлука бесконачан.

У свим наведеним случајевима одлука се доноси на основу анализе посматрања, узорка $\mathbf{X}$, односно реализованог узорка $\mathbf{x}$ одговарајућег обележја X и као последица тога, **одлука** d представља вредност функције $w(\mathbf{x})$ дефинисане на узорачком простору $\mathcal{X}$ чији је кодомен **скуп могућих одлука** $\mathrm{D} = \{d\}$ у датој ситуацији. На тај начин, статистика $w(\mathbf{X})$,

$$w : \mathcal{X} \longrightarrow \mathrm{D}$$

је правило које сваки резултат посматрања $\mathbf{x} \in \mathcal{X}$ доводи у везу са одлуком $d = w(\mathbf{x}) \in \mathrm{D}$. Функција w се зове **функција одлуке (процедура)** и она се бира на основу неког критеријума оптималности. Принцип решавања тог задатка зове се теорија одлучивања, дефинисао ју је Уалд (Wald) 1950. године.

Нека је $(\mathcal{X}, \mathcal{B}, \mathrm{P})$ простор вероватноћа који одговара статистичком моделу са узорком фиксираног обима. То значи да је $\mathcal{X}$ – коначнодимензиони еуклидски простор, $\mathcal{B}$ – Борелово σ–поље на $\mathcal{X}$, $\mathrm{P} \in \mathcal{P}$, где је $\mathcal{P}$ – фамилија вероватноћа.

С обзиром на циљ овог поглавља, само да представи концепт статиситчке теорије одлучивања, надаље ћемо претпостављати да је фамилија $\mathcal{P}$ описана тако да $\mathrm{P} \in \mathcal{P}$ зависи од параметра θ (једно или вишедимензионог), тј. $\mathcal{P} = \{\mathrm{P}_\theta; \theta \in \mathbf{\Theta}\}$ на $(\mathcal{X}, \mathcal{B})$, у замену за општи модел ове теорије.

Са $\mathbf{\Theta}$ ћемо, као и до сада, означавати параметарски простор, при чему ћемо још сматрати да је $\mathbf{\Theta}$ – отворен скуп k–димензионог еуклидског простора E_k, $k \geq 1$.

Нека су задати фамилија допустивих расподела $\{F(x;\theta), \theta \in \mathbf{\Theta}\}$, којој по претпоставци припада расподела посматранаог обележја X, и скуп одлука $\mathrm{D} = \{d\}$. На основу узорка треба донети одлуку о избору једне од функција расподеле из дате фамилије. С обзиром да ће се избор функције расподеле извршити избором вредности параметра θ, то је у овом случају кодомен функције одлуке скуп $\mathbf{\Theta}$. Према томе,

$$w : \mathcal{X} \longrightarrow \mathbf{\Theta} \subseteq \mathrm{D}.$$

Дакле, на основу узорка $\mathbf{x} \in \mathcal{X}$ треба донети одлуку $d \in \mathrm{D}$, где је скуп D сада скуп расположивих вредности параметра θ.

Да бисмо поставили критеријуме избора функције одлуке, неопходно је упоредити резултате коришћења различитих правила w. У ту сврху се дефинише **функција губитка**:

$$\mathcal{L} : \mathbf{\Theta} \times D \longrightarrow [0, +\infty).$$

Функција губитка мери грешку коју бисмо начинили при доношењу одлуке d, а да је при томе права вредност параметра баш θ. Према томе је $\mathcal{L}(\theta, \theta) = 0$. Функцију $\mathcal{L}$ треба разумевати као губитак (на тачности) услед прихватања одлуке d под условом да је расподела обележја X тачно $F(x;\theta)$.

По правилу се функција губитка тражи у неком одређеном скупу функција, тј. скупу функција са одређеним својством. Тако, на пример, функција губитка се често дефинише као: $\mathcal{L}(\theta, d) = \lambda(\theta) W(|d - \theta|)$, где је W монотоно растућа реална функција реалне променљиве $t \geq 0$, таква да је $W(0) = 0$. Што се тиче функције λ, претпоставља се да је позитивна, коначна и $\mathcal{S}$– измерљива, где је $\mathcal{S}$ Борелово σ– поље на $\mathbf{\Theta}$.

Да би се избегле тешкоће, често се уводи претпоставка да је $\mathcal{L}$ ограничена функција параметра θ за свако $d \in \mathrm{D}$. Ову ћемо претпоставку и ми усвојити. Некада се за функцију $\mathcal{L}$ узимају само непрекидне функције по оба аргумента, а некада равномерно ограничене по (θ, d).

Пример 85. Претпоставимо да параметарски простор $\boldsymbol{\Theta}$ садржи тачно две тачке, тј. $\boldsymbol{\Theta} = \{0, 1\}$, а скуп одлука нека је $\mathrm{D} = \{\mathrm{d} | 0 \leq \mathrm{d} \leq 1\}$. Функцију губитка можемо изабрати на следећи начин

$$\mathcal{L}(\theta, w) = |w(\mathbf{x}) - \theta|^{\alpha} \quad , \quad \alpha \in N \, .$$

Овде је $\lambda \equiv 1$, $W(t) = t^{\alpha}$, $t \geq 0$. $\triangle$

За сваку процедуру w дефинише се даље **функција ризика** у ознаци $R(\theta,w)$, таква да за сваку одлуку $d = w(\mathbf{x}) \in \mathrm{D}$ она представља условно очекивање функције губитка под условом да је θ права вредност параметра:

$$R(\theta, w) = E(\mathcal{L}(\theta, w(\mathbf{X}))|\theta) = E_{\theta}(\mathcal{L}(\theta, d)).$$

Функција ризика даје критеријум упоређивања различитих правила одлучивања на следећи начин: ако имамо две функције одлуке (два правила одлучивања) w_1 и w_2, такве да је

$$R(\theta, w_1) \leq R(\theta, w_2), \quad \forall \theta \in \boldsymbol{\Theta} \tag{5.1}$$

са строгом неједнакошћу за бар једно θ, тада је правило w_1 боље од правила w_2, јер w_1 доводи до мањег очекиваног губитка.

С друге стране, могуће је да функције одлуке w_1 и w_2 буду неупоредиве по наведеном критеријуму, тј. да за неке вредности θ буде $R(\theta, w_1) < R(\theta, w_2)$, док је за остале вредности θ знак неједнакости супротно усмерен. У таквим ситуацијама је неопходно прибавити додатне информације о проблему, да би се могао извршити избор процедуре.

Пример 86. Нека је дат прост узорак обима 25, $\mathbf{X} = (X_1, \ldots, X_{25})$, из популације са обележјем X чија расподела припада фамилији допустивих расподела $\{\mathcal{N}(\theta, 1), -\infty < \theta < \infty\}$, и нека је $Y = \overline{X}_{25}$. Затим, нека је:

$$\mathcal{L}(\theta, w(\mathbf{x})) = (w(\mathbf{x}) - \theta)^2$$

Размотрићемо две могуће функције одлуке: $w_1(\mathbf{x}) = y$ и $w_2(\mathbf{x}) = 0$. Оне редом дају ризике

$$R(\theta, w_1) = E(Y-\theta)^2 = D(Y) = \frac{1}{25} \quad \text{и}$$

$$R(\theta, w_2) = E(0-\theta)^2 = \theta^2 \,.$$

Видимо да, ако је права вредност параметра $\theta = 0$, онда је w_2 одлична функција одлуке јер је функција ризика једнака нули. Међутим, ако се θ разликује од нуле чак и врло мало, рецимо да је $\theta = 2$, онда је:

$$R(2, w_1) = \frac{1}{25} \quad , \quad R(2, w_2) = 4 \,, \quad \text{па је} \quad R(2, w_2) > R(2, w_1) \,.$$

Такође видимо да је

$$R(\theta, w_2) \leq R(\theta, w_1) \quad \text{ако је} \quad -\frac{1}{5} \leq \theta \leq \frac{1}{5} \,. \triangle$$

Функција одлуке w се назива **недопустивом** ако постоји w' која је боља од w у смислу наведеног критеријума (5.1). У супротном, одлука (функција одлуке) w је **допустива**.

Пример 87. Ако бисмо се у примеру 86 ограничили на такве функције одлуке (статистике) за које је $E(w(\mathbf{X})) = \theta$, тада w_2 не би била међу допустивим функцијама одлука. $\triangle$

Ако је класа допустивих функција одлуке једночлана, тада се може говорити о оптималној (најбољој) одлуци .

У теорији одлучивања се традиционално примењују два прилаза (расуђивања): *Бајесов* и *минимаксни.*

5.1 Минимакс одлучивање

Минимакс принцип се састоји у томе да се донесе одлука која минимализује максимални ризик. Ова, иначе једноставна идеја, поставља пред оцењивача често неостварив захтев, јер функција одлуке $w(\mathbf{x})$ која минимализује ризик за једну вредност параметра θ, за другу то не мора да чини, као што смо видели у Примеру 86.

Пример 88. Ако се још једном вратимо на пример 86 и одредимо нови критеријум за функцију одлуке, то може да буде ограничење типа да захтевамо функцију одлуке која минимализује максимум функције ризика. У том случају w_2 не би била добра зато што је $R(\theta, w_2) = \theta^2$ неограничена функција на параметарском простору $\mathbf{\Theta} = (-\infty, +\infty)$. С друге стране:

$$\max_{\theta} R(\theta, w_1) = \max_{\theta} \left(\frac{1}{25}\right) = \frac{1}{25} \quad , \quad -\infty < \theta < \infty.$$

Критеријум који смо овде применили зове се минимаксни критеријум. Према овом критеријуму се може показати да је $w_1(\mathbf{x}) = y = \overline{x}_{25}$ најбоља функција одлуке ако је функција губитка $\mathcal{L}(\theta, w(\mathbf{x})) = (\theta - w(\mathbf{x}))^2$.$\triangle$

Овим примером смо илустровали следеће:

а) Без некаквог ограничења за функцију одлуке веома је тешко наћи функцију одлуке која има функцију ризика униформно мању од функције ризика друге функције одлуке.

б) Принцип избора најбоље функције одлуке који се зове **минимакс принцип**.

Дефинишимо прецизно минимаксну функцију одлуке :

Дефиниција 45. Ако је функција одлуке $w_0(\mathbf{x})$ таква да за сваки $\theta \in \mathbf{\Theta}$

$$\max_{\theta \in \mathbf{\Theta}} R(\theta, w_0(\mathbf{x})) \leq \max_{\theta \in \mathbf{\Theta}} R(\theta, w(\mathbf{x}))$$

за било коју другу функцију одлуке $w(\mathbf{x})$, тада се $w_0(\mathbf{x})$ зове *минимаксна функција одлуке.* $\diamondsuit$

Посветимо сада мало пажње другом основном принципу статистичког закључивања, тестирању стаистичких хипотеза, на бази минимакс принципа. Тестираћемо хипотезу о непознатом параметру θ код двочланог параметарског простора $\mathbf{\Theta} = \{\theta_0, \theta_1\}$.

Раније смо већ истакли да се тестирање хипотезе $H_0 : \theta = \theta_0$ против $H_1 : \theta = \theta_1$ може да изрази у терминима критичне области у узорачком простору. Исто се може учинити и код минимакс поступка за тестирање хипотеза. Наиме можемо да изаберемо подскуп C узорачког простора $\mathcal{X}$ и ако $\mathbf{x} \in C$ прихватамо хипотезу H_1, односно доносимо одлуку $w(\mathbf{x}) = \theta_1$, а уколико $\mathbf{x} \in C^c$ одлучићемо

$w(\mathbf{x}) = \theta_0$. На тај начин критична област C одређује функцију одлуке. У том смислу функцију ризика можемо означити са $R(\theta, C)$ уместо са $R(\theta, w)$. Дакле,

$$R(\theta, C) = R(\theta, w) = \int_{C \cup C^c} \mathcal{L}(\theta, w) dF_\theta(\mathbf{x}).$$

Отуда

$$R(\theta, C) = \int_C \mathcal{L}(\theta, \theta_1) dF_\theta(\mathbf{x}) + \int_{C^c} \mathcal{L}(\theta, \theta_0) dF_\theta(\mathbf{x}).$$

За $\theta = \theta_0$ добија се

$$R(\theta_0, C) = \mathcal{L}(\theta_0, \theta_1) \int_C dF_{\theta_0}(\mathbf{x}),$$

а за $\theta = \theta_1$ се добија

$$R(\theta_1, C) = \mathcal{L}(\theta_1, \theta_0) \int_{C^c} dF_{\theta_1}(\mathbf{x}).$$

Ако је $M(\theta)$ функција моћи теста која одговара критичној области C, тада је

$$R(\theta_0, C) = \mathcal{L}(\theta_0, \theta_1) M(\theta_0) = \mathcal{L}(\theta_0, \theta_1)\alpha$$

и

$$R(\theta_1, C) = \mathcal{L}(\theta_1, \theta_0) \left(1 - \int_C dF_{\theta_1}(\mathbf{x})\right) = \mathcal{L}(\theta_1, \theta_0)(1 - M(\theta_1)) = \mathcal{L}(\theta_1, \theta_0)\beta,$$

где су α и β вероватноћа грешке прве и друге врсте редом.

Минимаксно решење нашег проблема била би критична област C за коју би

$$max\{R(\theta_0, C), R(\theta_1, C)\}$$

био минималан. Доказаћемо да, ако уведемо додатни услов да је

$$R(\theta_0, C) = R(\theta_1, C)\,, \tag{5.2}$$

у том случају минимаксно решење за тестирање просте нулте против такође просте алтернативне хипотезе, Нејман–Пирсонова најбоља критична област величине α за тестирање истих хипотеза. Дакле, доказаћемо да је област

$$C = \left\{(x_1, x_2, \ldots, x_n) | \quad \frac{L(\theta_0; x_1, x_2, \ldots, x_n)}{L(\theta_1; x_1, x_2, \ldots, x_n)} \le k\right\}, \tag{5.3}$$

где је $k > 0$ изабрано тако да буде задовољен услов (5.2). Такво k у апсолутно непрекидном случају увек постоји док у дискретном случају може да се деси да нам је потребан и помоћни експеримент за који је

$$\frac{L(\theta_0; x_1, x_2, \ldots, x_n)}{L(\theta_1; x_1, x_2, \ldots, x_n)} = k,$$

да бисмо постигли $R(\theta_0, C) = R(\theta_1, C)$.

Доказ. Да бисмо доказали да је C дефинисана у (5.3) минимаксно решење, посматрајмо произвољну другу критичну област $A \subset R^n$ величине α за коју је

$$R(\theta_0, C) \geq R(\theta_0, A). \tag{5.4}$$

Области A у којима је $R(\theta_0, C) < R(\theta_0, A)$ не треба ни разматрати, јер би у том случају очигледно било

$$R(\theta_0, C) = R(\theta_1, C) < \max\{R(\theta_0, A), R(\theta_1, A)\}.$$

Услов (5.4) је еквивалентан са

$$\mathcal{L}(\theta_0, \theta_1) \int_C dF_{\theta_0}(\mathbf{x}) \geq \mathcal{L}(\theta_0, \theta_1) \int_A dF_{\theta_0}(\mathbf{x}).$$

Отуда

$$\int_C dF_{\theta_0}(\mathbf{x}) \geq \int_A dF_{\theta_0}(\mathbf{x}).$$

Како је

$$\alpha = \int_C dF_{\theta_0}(\mathbf{x}),$$

то је

$$\alpha = \int_A dF_{\theta_0}(\mathbf{x}).$$

Међутим, према Нејман–Пирсоновој теореми, C је најбоља критична област величине α, дакле, она критична област величине α за коју је β најмање, тј.

$$\int_{C^c} dF_{\theta_1}(\mathbf{x}) \leq \int_{A^c} dF_{\theta_1}(\mathbf{x}).$$

Отуда,

$$\mathcal{L}(\theta_1, \theta_0) \int_{C^c} dF_{\theta_1}(\mathbf{x}) \leq \mathcal{L}(\theta_1, \theta_0) \int_{A^c} dF_{\theta_1}(\mathbf{x}).$$

Што значи да је

$$R(\theta_1, C) \leq R(\theta_1, A),$$

односно,

$$\max\{R(\theta_0, C), R(\theta_1, C)\} \leq \max\{R(\theta_0, A), R(\theta_1, A)\}$$

што је и требало доказати. □

5.2 Бајесово одлучивање

Код Бајесовог оцењивања се параметар θ посматра као случајна величина. Дакле, за θ се дефинише простор вероватноћа $(\boldsymbol{\Theta}, \mathcal{S}, Q)$, где је $\mathcal{S}$ – Борелово σ–поље на $\boldsymbol{\Theta}$, а Q – мера, тј. вероватноћа на $(\boldsymbol{\Theta}, \mathcal{S})$. Меру Q називамо **априорном расподелом** параметра θ. При томе се претпоставља да априорна расподела припада некој фамилији априорних расподела $\mathcal{H}$. За свако фиксирано θ са $f(\mathbf{x}; \theta)$ означаваћемо густину расподеле узорка која одговара расподели вероватноћа P, а са $h(\theta)$ априорну густину расподеле која одговара расподели вероватноћа Q, са функцијом расподеле H. Дакле, $f(\mathbf{x}; \theta)$ је условна густина расподеле за $\mathbf{X}$ под условом да је права вредност параметра баш θ. Она је $\mathcal{S}$–измерљива по θ за сваки $\mathbf{x} \in \mathcal{X}$. Тада је

$$g(\mathbf{x}; \theta) = f(\mathbf{x}; \theta)h(\theta), \; \mathbf{x} \in \mathcal{X}, \; \theta \in \boldsymbol{\Theta}$$

густина заједничке расподеле за вишедимензиону случајну променљиву $(\mathbf{X}, \theta)$ на простору $(\mathcal{X} \times \boldsymbol{\Theta}, \mathcal{B} \times \mathcal{S})$. Условна густина расподеле

$$h(\theta|\mathbf{X} = \mathbf{x}) = \frac{f(\mathbf{x}; \theta)h(\theta)}{\int_{\boldsymbol{\Theta}} f(\mathbf{x}; \tau)dH(\tau)}$$

за свако $\mathbf{x} \in \mathcal{X}$ за које је $f(\mathbf{x}) = \int_{\boldsymbol{\Theta}} f(\mathbf{x}; \tau)dH(\tau) > 0$, зове се **апостериорна густина расподеле** параметра θ .

Нека је $F_\theta(\mathbf{x})$ одговарајућа функција расподеле за густину $f(\mathbf{x}; \theta)$. Означимо са $H(\theta)$, $\theta \in \boldsymbol{\Theta}$, априорну функцију расподеле на $\boldsymbol{\Theta}$ којој одговара априорна густина расподеле $h(\theta)$. Нека је са $H(\theta|\mathbf{X} = \mathbf{x})$ означена апостериорна расподела (функција расподеле) параметра θ при задатом $\mathbf{X} = \mathbf{x}$. Маргинална (безусловна) густина за $\mathbf{X}$ се тада може посматрати као

$$f_H(\mathbf{x}) = \int_{\boldsymbol{\Theta}} f(\mathbf{x}; \theta)dH(\theta)\,,$$

са одговарајућом безусловном функцијом расподеле у ознаци $F_H(\mathbf{x})$, а **апостериорна густина расподеле** за θ при задатом $\mathbf{X} = \mathbf{x}$ има облик

$$h(\theta|\mathbf{X} = \mathbf{x}) = \frac{f(\mathbf{x};\theta)h(\theta)}{f_H(\mathbf{x})}.$$

Надаље ћемо размотрити неке особине Бајесове оцене за функцију губитка облика $\mathcal{L}(\theta, d) = \lambda(\theta)W(|d - \theta|)$, $d = w(\mathbf{x})$. Истакнимо поново да је функција ризика одлуке d:

$$R(\theta, w) = \lambda(\theta)\int_{\mathcal{X}} W(|w(\mathbf{x}) - \theta|)dF_\theta(\mathbf{x}).$$

С обзиром да је $\mathcal{L}(\theta, d)$ ограничена функција по θ за сваки d, то је и $R(\theta, d)$ ограничена за сваки d. С друге стране, ако уведемо претпоставку о априорној расподели параметра θ, **априорни ризик оцене** d **у односу на априорну функцију расподеле** H је априорно очекивање функције ризика дефинисано са

$$R(H, w) = \int_{\boldsymbol{\Theta}} R(\theta, w)dH(\theta) = E(R(\theta, w)).$$

Док је априорни ризик условно математичко очекивање функције ризика под условом да је учињена одлука d, дотле је **апостериорни ризик** функције $w(\mathbf{x}) = d$ при заданом $\mathbf{X} = \mathbf{x}$:

$$\int_{\boldsymbol{\Theta}} \lambda(\theta)W(|w(\mathbf{x}) - \theta|)dH(\theta|\mathbf{x}).$$

Дакле, апостериорни ризик је очекивање функције губитка у односу на апостериорну расподелу параметра. Отуда је очигледно да је априорни ризик $R(H, d)$ очекивање апостериорног ризика у односу на $F_H(\mathbf{x})$ – расподелу узорка $\mathbf{X}$ при априорној расподели H. Заиста,

$$R(H, w) = \int_{\boldsymbol{\Theta}} R(\theta, w)dH(\theta) = \int_{\boldsymbol{\Theta}} \left(\int_{\mathcal{X}} \mathcal{L}(\theta, w(\mathbf{x}))dF_\theta(\mathbf{x})\right) dH(\theta).$$

Уколико је у питању апсолутно непрекидна расподела обележја X, али и апсолутно непрекидна расподела (априорна и апостериорна) непознатог параметра θ, априорни ризик постаје:

$$R(H, w) = \int_{\boldsymbol{\Theta}} R(\theta, w)h(\theta)d\theta = \int_{\boldsymbol{\Theta}} \int_{\mathcal{X}} \mathcal{L}(\theta, w(\mathbf{x}))f_H(\mathbf{x})h(\theta|\mathbf{x})d\mathbf{x}d\theta =$$

$$= \int_{\mathcal{X}} \left(\int_{\mathbf{\Theta}} \mathcal{L}(\theta, w(\mathbf{x})) h(\theta|\mathbf{x}) d\theta \right) f_H(\mathbf{x}) d\mathbf{x}.$$

С бајесовске тачке гледишта, пошто је $\mathbf{X}$ реализовано, најпогоднија функција за даље разматрање је апостериорни ризик, а не априорни. Значи, **бајесовска оцена** параметра θ у односу на априорну расподелу H је таква оцена (одлука) из D која минимизира апостериорни ризик при задатом $\mathbf{X} = \mathbf{x}$. Ако означимо са $\hat{\theta}_H(\mathbf{X})$ бајесовску оцену, онда је, за раније коментарисану функцију губитка,

$$\int_{\mathbf{\Theta}} \lambda(\theta) W(|\hat{\theta}_H(\mathbf{x}) - \theta|) dH(\theta|\mathbf{x}) = \inf_{d \in D} \int_{\mathbf{\Theta}} \lambda(\theta) W(|d - \theta|) dH(\theta|\mathbf{x}).$$

За задато H, Бајесова оцена не мора бити јединствена. Она ће бити јединствена у случају да је функција губитка строго конвексна. Бајесова оцена минимизира и априорни ризик (следи из леме Фатуа). Многи аутори је и дефинишу као онај елемент из D који минимизира априорни ризик. Обе дефиниције воде ка истом решењу.

Пример 89. Вратимо се примеру 85. Нека је априорна расподела за θ:

$$P\{\theta = 0\} = \frac{3}{4}, \quad P\{\theta = 1\} = \frac{1}{4}.$$

Нека је $\alpha = 1$, тада је априорна функција ризика

$$\begin{aligned} R(\theta, w) &= E\{\mathcal{L}(\theta, w(\mathbf{X}))|\theta\} = \mathcal{L}(0, w(\mathbf{x})) \cdot \frac{3}{4} + \mathcal{L}(1, w(\mathbf{x})) \cdot \frac{1}{4} = \\ &= |d| \cdot \frac{3}{4} + |1 - d| \cdot \frac{1}{4} = \frac{1}{2} d + \frac{1}{4}. \end{aligned}$$

Дакле,

$$\inf_{\mathrm{d} \in \mathrm{D}} R(\theta, \mathrm{d}) = \frac{1}{4},$$

тј. јединствено Бајесово решење је $w(\mathbf{x}) = 0$. Скренимо пажњу на чињеницу да у случају $D = \{d | 0 < d \leq 1\}$ инфимум функције ризика био би исти, $1/4$, али ниједна одлука $\mathrm{d} \in \mathrm{D}$ не би била Бајесово решење.

Нека је сада $\alpha > 1$. Тада је функција ризика

$$R(\theta, \mathrm{d}) = \frac{3}{4} \mathrm{d}^{\alpha} + \frac{1}{4} \cdot (1 - \mathrm{d})^{\alpha}.$$

Функција ризика има минимум за одлуку

$$\mathrm{d} = \left(1 + 3^{\frac{1}{\alpha-1}}\right)^{-1} \in (0,1)\,,$$

која јесте Бајесово решење. $\triangle$

Може се говорити и о довољним статистикама у Бајесовом смислу, о чыему овде неће бити речи.

На Бајесовом принципу се заснивају не само тачкасте, већ и оцене параметара областима. Овде ће бити речи само о оцени једнодимензионог параметра θ, тј. о интервалу као оцени једнодимензионог параметра. Дакле, треба одредити две статистике, тј. две процедуре $w_1(\mathbf{X})$ и $w_2(\mathbf{X})$ такве да је

$$P\{w_1(\mathbf{X}) \le \theta \le w_2(\mathbf{X}) | \mathbf{X} = \mathbf{x}\} = \gamma = 1 - \alpha$$

за унапред задати ниво поверења γ. То значи да

$$\int_{[w_1(\mathbf{x}),\ w_2(\mathbf{x})]} dH(\theta|\mathbf{x}) = \gamma\,. \tag{5.5}$$

Пример 90. На основу узорка $(X_1, X_2, \ldots, X_n)$ одредити Бајесов интервал поверења за математичко очекивање обележја X чија расподела припада фамилији допустивих расподела $\{\mathcal{N}(\theta, \sigma^2),\ -\infty < \theta < +\infty\}$ под претпосравком да је априорна расподела параметра θ такође нормална: $\mathcal{N}(\mu,\ \nu^2)$. Параметри σ^2, μ и ν^2 су познати.

Ако као процедуру w изаберемо функцију средине узорка, $w(\mathbf{X}) = \varphi(\overline{X}_n)$, означимо, са $Y = \overline{X}_n$, треба донети две одлуке $w_1(\mathbf{x})$ и $w_2(\mathbf{x})$, тј $\varphi_1(y)$ и $\varphi_2(y)$ које ће задовољити горе наведени услов (5.5). Условна расподела за Y под условом θ је $\mathcal{N}(\theta,\ \frac{\sigma^2}{n})$ што ће дати одговарајућу условну густину расподеле

$$f(y;\theta) = \sqrt{\frac{n}{2\pi\sigma^2}} \exp\left(-\frac{(y-\theta)^2}{\frac{2\sigma^2}{n}}\right).$$

Заједничка густина расподеле за вектор (Y, θ) је тада:

$$g(y;\theta) = \frac{\sqrt{n}}{2\pi\nu\sigma} \exp\left(-\frac{(\theta-\mu)^2}{2\nu^2} - \frac{n(y-\theta)^2}{2\sigma^2}\right),$$

тј. посматрани вектор има расподелу

$$\mathcal{N}\left(\mu\,,\,\mu\,,\,\nu^2 + \frac{\sigma^2}{n}\,,\,\nu^2\,,\,\frac{\nu}{\sqrt{\nu^2 + \frac{\sigma^2}{n}}}\right),$$

где је

$$\frac{\nu}{\sqrt{\nu^2 + \frac{\sigma^2}{n}}}$$

коефицијент корелације компонената.

Апостериорна густина расподеле за θ под условом $\mathbf{X} = \mathbf{x}$, односно, $Y = y$ биће:

$$h(\theta|Y = y) = \frac{1}{2\pi(\nu^2 + \frac{\sigma^2}{n})} \exp\left(-\frac{(y-\mu)^2}{2(\nu^2 + \frac{\sigma^2}{n})}\right),$$

што значи да је апостериорна расподела нормална

$$\mathcal{N}\left(\frac{\mu\sigma^2 + ny\nu^2}{n\nu^2 + \sigma^2},\ \frac{\nu^2\sigma^2}{n\nu^2 + \sigma^2}\right).$$

Коначно, интервална Бајесова оцена за θ одређује се из

$$P\left(\frac{\mu\sigma^2 + nY\nu^2}{n\nu^2 + \sigma^2} - z\sqrt{\frac{\nu^2\sigma^2}{n\nu^2 + \sigma^2}} < \theta < \frac{\mu\sigma^2 + nY\nu^2}{n\nu^2 + \sigma^2} + z\sqrt{\frac{\nu^2\sigma^2}{n\nu^2 + \sigma^2}}\right) = \gamma,$$

$z \in R$. △

Бајесов концепт тестирања статистичких хипотеза изложићемо на примеру тестирања просте нулте хипотезе против алтернативне такође просте хипотезе. То је случај дводимензионог параметарског простора $\mathbf{\Theta} = \{\theta_0, \theta_1\}$. Дакле, случајна промењлива θ је дискретног типа и

$$h(\theta_0) + h(\theta_1) = 1.$$

У том случају је маргинална густина за $\mathbf{X}$:

$$\sum_{\mathbf{\Theta}} f(\mathbf{x};\theta)h(\theta) = f(\mathbf{x};\theta_0)h(\theta_0) + f(\mathbf{x};\theta_1)h(\theta_1),$$

односно, апостериорна густина за θ је:

$$h(\theta|\mathbf{X} = \mathbf{x}) = \frac{h(\theta)f(\mathbf{x};\theta)}{f(\mathbf{x};\theta_0)h(\theta_0) + f(\mathbf{x};\theta_1)h(\theta_1)}.$$

Тестираћемо хипотезу $H_0 : \theta = \theta_0$ против алтернативне $H_1 : \theta = \theta_1$. Бајесово решење ће бити функција одлуке $w(\mathbf{x})$ за коју је апостериорни ризик минималан. Дакле, треба минимализовати

$$E(\mathcal{L}(\theta, w(\mathbf{X}))|\mathbf{X} = \mathbf{x}).$$

С обзиром да је скуп могућих одлука двочлан, то треба размотрити само две могуће вредности апостериорног ризика и то:

1. $w = \theta_0$

$$\Rightarrow E(\mathcal{L}(\theta, w(\mathbf{X}))|\mathbf{X} = \mathbf{x}) = E(\mathcal{L}(\theta, \theta_0)|\mathbf{X} = \mathbf{x}) = \\ = \frac{\mathcal{L}(\theta_1, \theta_0)h(\theta_1)f(\mathbf{x};\theta_1)}{f(\mathbf{x};\theta_0)h(\theta_0) + f(\mathbf{x};\theta_1)h(\theta_1)}$$

2. $w = \theta_1$

$$\Rightarrow E(\mathcal{L}(\theta, w(\mathbf{X}))|\mathbf{X} = \mathbf{x}) = E(\mathcal{L}(\theta, \theta_1)|\mathbf{X} = \mathbf{x}) = \\ = \frac{\mathcal{L}(\theta_0, \theta_1)h(\theta_0)f(\mathbf{x};\theta_0)}{f(\mathbf{x};\theta_0)h(\theta_0) + f(\mathbf{x};\theta_1)h(\theta_1)} .$$

Према томе, прихватићемо хипотезу $H_i : \theta = \theta_i, \ i = 0, 1$ ако и само ако је

$$\frac{\mathcal{L}(\theta_j, \theta_i)h(\theta_j)f(\mathbf{x};\theta_j)}{f(\mathbf{x};\theta_0)h(\theta_0) + f(\mathbf{x};\theta_1)h(\theta_1)} < \frac{\mathcal{L}(\theta_i, \theta_j)h(\theta_i)f(\mathbf{x};\theta_i)}{f(\mathbf{x};\theta_0)h(\theta_0) + f(\mathbf{x};\theta_1)h(\theta_1)},$$

за $j \neq i$. Или што је исто

$$\frac{f(\mathbf{x};\theta_j)}{f(\mathbf{x};\theta_i)} < \frac{\mathcal{L}(\theta_i, \theta_j)h(\theta_i)}{\mathcal{L}(\theta_j, \theta_i)h(\theta_j)} \quad , \quad i, j = 0, 1 \, , \, i \neq j \, .$$

Уколико се догоди да се уместо неједнакости јавља једнакост, морали бисмо да потражимо додатне информације, односно да обавимо неки помоћни експеримент који би нам помогао да донесемо одлуку.

Уочимо важну чињеницу да се последњом неједнакошћу која је горе наведена описује баш најбоља критична област према тврђењу Нејман–Пирсонове теореме с обзиром да се на левој страни неједнакости налази количник веродостојности за који се ова теорема везује, јер је

$$f(\mathbf{x};\theta) \equiv L(\theta; x_1, x_2, \ldots, x_n) \quad , \quad \mathbf{x} = (x_1, x_2, \ldots, x_n) \in R^n \, ,$$

функција веродостојности.

Глава 6

Додатак

6.1 Важније расподеле вероватноћа

6.1.1 Расподеле дискретног типа

1. **Бернулијева расподела**
   $X : \mathcal{B}(1, p)\,, \quad 0 < p < 1$

$$P\{X = 0\} = 1 - p \qquad \text{и} \qquad P\{X = 1\} = p$$

$$E(X) = p$$
$$D(X) = p(1 - p)$$

2. **Биномна расподела**
   $X : \mathcal{B}(n, p)\,, \quad 0 < p < 1\,, \quad n \in N$

$$P\{X = k\} = \binom{n}{k} p^k (1 - p)^{n-k} \quad , \quad 0 \leq k \leq n$$

$$E(X) = \sum_{k=0}^{n} kP\{X = k\} = np$$
$$D(X) = np(1 - p)$$

3. **Пуасонова расподела**

   $X : \mathcal{P}(\lambda)\,, \quad \lambda > 0$

$$P\{X = k\} = \frac{\lambda^k}{k!} e^{-\lambda} \quad , \quad k = 1, 2, \ldots$$

$$E(X) = \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} = \lambda$$

$$D(X) = \lambda$$

4. **Геометријска расподела**

   $0 < p < 1$

$$P\{X = k\} = (1 - p)^{k-1} p \quad , \quad k = 1, 2, \ldots$$

$$E(X) = \sum_{k=1}^{\infty} k P\{X = k\} = \frac{1}{p}$$

$$D(X) = \frac{1}{p^2}$$

6.1.2 Расподеле апсолутно непрекидног типа

1. **Нормална (Гаусова) расподела**

   $X : \mathcal{N}(m, \sigma^2)\,, \quad -\infty < m < +\infty\,, \quad \sigma^2 > 0$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-m)^2}{2\sigma^2}} \quad , \quad -\infty < x < +\infty$$

$$F(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} e^{-\frac{(t-m)^2}{2\sigma^2}} dt$$

$$E(X) = m$$
$$D(X) = \sigma^2$$

2. **Лог-нормална расподела**

   $X = \exp\{Y\}\,, \quad Y : \mathcal{N}(m, \sigma^2)\,, \quad -\infty < m < +\infty\,, \quad \sigma^2 > 0$

$$f(x) = \begin{cases} \frac{1}{x\sigma\sqrt{2\pi}} \exp\{-\frac{(\ln x - m)^2}{2\sigma^2}\}\,, & x > 0 \\ 0\,, & x \leq 0 \end{cases}$$

$$E(X) = \exp\{m + \frac{1}{2}\sigma^2\}$$
$$D(X) = \exp\{2m + \sigma^2\}\left(\exp\{\sigma^2\} - 1\right)$$

3. **Тропараметарска лог–нормална расподела**

   $X = \exp\{Y\} + x_0\,, \quad Y : \mathcal{N}(m, \sigma^2)\,, \quad -\infty < x_0 < +\infty$

$$f(x) = \begin{cases} \frac{1}{(x-x_0)\sigma\sqrt{2\pi}} \exp\left\{-\frac{(\ln(x-x_0)-m)^2}{2\sigma^2}\right\}, & x > x_0 \\ 0, & x \leq x_0 \end{cases}$$

4. **Униформна расподела**

   $X : \mathcal{U}(a, b)\,, \quad a, b \in R\,, \quad a < b$

$$f(x) = \begin{cases} \frac{1}{b-a}, & x \in (a, b) \\ 0, & \text{иначе} \end{cases}$$

$$F(x) = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & a \le x < b \\ 1, & x \ge b \end{cases}$$

$$E(X) = \frac{a+b}{2}$$

$$D(X) = \frac{(b-a)^2}{12}$$

5. **Вејбулова расподела**

   $a > 0\,, \quad b > 0$

$$f(x) = \begin{cases} \frac{b}{a}(\frac{x}{a})^{b-1} \exp\{-(\frac{x}{a})^b\}, & x > 0 \\ 0, & x \le 0 \end{cases}$$

$$E(X) = a^{-\frac{1}{b}} \Gamma\left(1 + \frac{1}{b}\right)$$

$$D(X) = a^{-\frac{2}{b}} \left(\Gamma\left(1 + \frac{2}{b}\right) - \Gamma^2\left(1 + \frac{1}{b}\right)\right)$$

6. **Експоненцијална расподела**

   $X : \mathcal{E}(\lambda)\,, \quad \lambda > 0$

$$f(x) = \begin{cases} \lambda \exp\{-\lambda x\}\,, & x > 0 \\ 0\,, & x \le 0 \end{cases}$$

$$E(X) = \frac{1}{\lambda}$$

$$D(X) = \frac{1}{\lambda^2}$$

7. **Двојна експоненцијална расподела**

$$f(x) = \exp\left\{-e^{-x}\right\}, \quad -\infty < x < +\infty$$

8. **Двострана експоненцијална расподела**

$\lambda > 0$

$$f(x) = \frac{\lambda}{2}\exp\{-\lambda|x|\}, \quad -\infty < x < +\infty$$

9. **Кошијева расподела**

$\alpha > 0$

$$f(x) = \frac{\alpha}{\pi(x^2+\alpha^2)}, \quad -\infty < x < +\infty$$

10. **Лапласова расподела**

$\lambda > 0, \quad -\infty < m < +\infty$

$$f(x) = \frac{1}{2\lambda}\exp\left\{-\frac{|x-m|}{\lambda}\right\}, \quad -\infty < x < +\infty$$

11. **Гама расподела**

$\alpha \geq 0, \quad \beta > 0$

$$f(x) = \begin{cases} \frac{1}{\beta^\alpha\Gamma(\alpha)}x^{\alpha-1}\exp\{-\frac{x}{\beta}\}, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

$$E(X) = \alpha\beta$$
$$D(X) = \alpha\beta^2$$

12. **Тропараметарска гама расподела**

$\alpha > 0\,,\quad \beta > 0\,,\quad -\infty < x_0 < +\infty$

$$f(x) = \begin{cases} \frac{1}{\beta^\alpha\Gamma(\alpha)}(x - x_0)^{\alpha-1}\exp\left\{-\frac{x-x_0}{\beta}\right\}, & x > x_0 \\ 0\,, & x \le x_0 \end{cases}$$

13. **Лог–Пирсон** III **расподела**

$\alpha > 0\,,\quad \beta > 0\,,\quad -\infty < x_0 < +\infty$

$$f(x) = \begin{cases} \frac{1}{\beta^\alpha x\Gamma(\alpha)}(\ln x - x_0)^{\alpha-1}\exp\left\{-\frac{\ln x-x_0}{\beta}\right\}, & x > x_0 \\ 0\,, & x \le x_0 \end{cases}$$

14. **Хи квадрат (χ^2) расподела**

$X : \chi^2_\nu\,,\quad \nu \in N$

$$f(x) = \begin{cases} \dfrac{x^{\frac{\nu}{2}-1}}{2^{\frac{\nu}{2}}\Gamma\left(\dfrac{\nu}{2}\right)}\exp\{-\frac{x}{2}\}, & x > 0 \\ 0\,, & x \le 0 \end{cases}$$

$$E(\chi^2_\nu) = \nu$$
$$D(\chi^2_\nu) = 2\nu$$

15. **Фишерова расподела**

$X : F_{\nu_1,\nu_2}\,, \quad \nu_1\,,\nu_2 \in N$

$$X = \frac{\frac{\chi^2_{\nu_1}}{\nu_1}}{\frac{\chi^2_{\nu_2}}{\nu_2}}$$

и $\chi^2_{\nu_1}$ и $\chi^2_{\nu_2}$ су независне случајне променљиве.

$$f(x) = \begin{cases} \dfrac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)}\left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}} x^{\frac{\nu_1}{2}-1}\left(1+\frac{\nu_1}{\nu_2}x\right)^{-\frac{\nu_1+\nu_2}{2}}, & x > 0 \\ 0\,, & x \le 0 \end{cases}$$

$$E(X) = \frac{\nu_2}{\nu_2 - 2} \quad , \quad \nu_2 > 2$$

$$D(X) = \frac{2{\nu_2}^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-4)(\nu_2-2)^2} \quad , \quad \nu_2 > 4$$

16. **Студентова (t) расподела**

$t : t_\nu\,, \quad \nu \in N$

$$t_\nu = \frac{X}{\sqrt{\frac{\chi^2_\nu}{\nu}}}\,, \quad \text{где je} \quad X : \mathcal{N}(0,1)$$

и X и χ^2_ν су независне случајне променљиве.

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu\pi}}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} \quad , \quad -\infty < t < +\infty$$

$$E(X) = 0$$

$$D(X) = \frac{\nu}{\nu-2} \quad , \quad \nu > 2$$

17. **Гумбелова расподела**

$\alpha > 0\,, \quad \beta > 0$

$$f(x) = \alpha \exp\{-\alpha(x-\beta) - \exp(-\alpha(x-\beta))\} \quad , \quad -\infty < x < +\infty \quad ,$$

$Z = \alpha(X - \beta)$

$$f(z) = \exp\{-z - \exp\{-z\}\} \quad , \quad -\infty < z < +\infty$$

18. **Парето расподела**

$\alpha > 0\,, \quad \beta > 0$

$$f(x) = \begin{cases} \alpha\beta^\alpha x^{-(\alpha+1)}\,, & x > \beta \\ 0\,, & x \le \beta \end{cases}$$

6.2 Функција генератриса момената

Функција генератриса момената случајне променљиве X је функција по t, $t \in R$ у ознаци

$$M_X(t) = E(e^{tX})$$

тј.

$$M_X(t) = \sum_{j=1}^{\infty} e^{tx_j} f(x_j) \quad \text{у дискретном случају}$$

и

$$M_X(t) = \int_{-\infty}^{+\infty} e^{tx} f(x) dx \quad \text{у апсолутно непрекидном случају}$$

за оне вредности t за које наведено математичо очекивање постоји.

Пример 91.

$$I_A = \begin{cases} 1, & x \in A \\ 0, & \text{иначе} \end{cases} \quad , \quad P(I_A = 1) = p \quad , \quad 0 \le p \le 1$$

$$M_{I_A}(t) = E(e^{I_A t}) = pe^t + 1 - p\,.\triangle$$

Пример 92.

$$X : \mathcal{N}(0,1)$$

$$M_X(t) = E(e^{tX}) = \int_{-\infty}^{+\infty} e^{tx} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = e^{\frac{t^2}{2}}\,.\triangle$$

Пример 93.

$$X : \mathcal{N}(m, \sigma^2)$$

$$\begin{aligned} M_X(t) &= E(e^{tx}) = \int_{-\infty}^{+\infty} e^{tx} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-m)^2}{2\sigma^2}} dx = \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{+\infty} \exp\left\{-\left(\frac{x^2}{2\sigma^2} - \left(t + \frac{m}{\sigma^2}\right) x + \frac{m^2}{2\sigma^2}\right)\right\} dx = \\ &= \exp\left\{\left(tm + \frac{\sigma^2 t^2}{2}\right)\right\}\,.\triangle \end{aligned}$$

Из чињенице да математичко очекивање не постоји увек, тј. да има случајних променљивих које немају математичко очекивање, јасно је да функција генератриса момената не постоји за сваку случајну променљиву. Функција генератриса момената се, с тога дефинише на следећи начин.

Дефиниција 46. Претпоставимо да постоји позитиван број h такав да за $-h < t < h$ очекивање $E\left(e^{tX}\right)$ постоји, тада функцију $M_X(t) = E\left(e^{tX}\right)$ за $t \in (-h, h)$ зовемо *функција генератриса момената случајне променљиве* X. $\diamondsuit$

Пример 94. Нека случајна променљива X има густину расподеле

$$f(x) = \begin{cases} \frac{6}{\pi^2 x^2}, & x = 1, 2, 3, \ldots \\ 0, & \text{иначе} \end{cases} .$$

Ако би постојала функција генератриса момената случајне променљиве X, било би:

$$M(t) = E(e^{tX}) = \sum_x e^{tx} f(x) = \sum_{x=1}^{\infty} \frac{6e^{tx}}{\pi^2 x^2} .$$

Дакле, треба испитати конвергенцију реда чији је општи члан $a_n = \frac{6e^{tn}}{\pi^2 n^2}$. Очигледно

$$\lim_{n\to\infty} \frac{a_{n+1}}{a_n} = e^t > 1 \quad \text{за} \quad t > 0 .$$

То значи да за свако $t > 0$, $M(t)$ дивергира тј. не постоји позитиван број х такав да за $-h < t < h$ очекивање $M(t)$ постоји.$\triangle$

Чињеница да функција генератриса момената не постоји увек не умањује значај ове функције због широке класе расподела за коју она постоји.

Лако је показати да је Мак–Лоранов развој ове функције:

$$M_X(t) = 1 + E(X)t + E(X^2)\frac{t^2}{2!} + \ldots + E(X^r)\frac{t^r}{r!} + \ldots \quad ,$$

што разјашњава и порекло њеног имена – функција генератриса момената, јер следи да је

$$E(X^r) = \frac{d^r}{dt^r} M_X(t)|_{t=0} \quad , \quad r = 0, 1, 2, \ldots \quad .$$

Значај функције генератриса момената објашњава следећа теорема коју наводимо без доказа.

Теорема 22. *Претпоставимо да су X и Y случајне променљиве које имају функције генератриса момената $M_X(t)$ и $M_Y(t)$ редом. Тада X и Y имају исту расподелу вероватноћа ако и само ако су $M_X(t)$ и $M_Y(t)$ идентички једнаке.*

Следеће теореме су оперативног значаја.

Теорема 23. *Ако је $M_X(t)$ функција генератриса момената случајне променљиве X и $a \neq 0$ и b су константе, тада је функција генератриса момената случајне променљиве $Y = aX + b$*

$$M_Y(t) = e^{bt} M_X(at).$$

Доказ. Директно следи из особина експоненцијалне функције и дефиниције функције генератриса момената. □

Теорема 24. *Функција генератриса момената коначне суме независних случајних променљивих једнака је производу функција генератриса сабирака.*

Доказ. Нека су $X_1, \ldots, X_n$ независне случајне променљиве. Функција генератриса момената случајне променљиве $Y = \sum_{i=1}^{n} X_i$ је

$$\begin{aligned} M_Y(t) &= E\left(e^{tY}\right) = E\left(\exp\left(t\sum_{i=1}^{n} X_i\right)\right) = E\left(\prod_{i=1}^{n} e^{tX_i}\right) = \prod_{i=1}^{n} E\left(e^{tX_i}\right) = \\ &= \prod_{i=1}^{n} M_{X_i}(t)\,.\square \end{aligned}$$

Пример 95. Нека су $X_1, \ldots, X_n$ независне случајне променљиве са расподелама $X_i : \mathcal{N}(m, \sigma^2)$, $i = 1, 2, \ldots, n$. Применом функције генератриса момената показати да случајна променљива $U = \sum_{i=1}^{n} a_i X_i$, a_i су константе, има $\mathcal{N}\left(\sum_{i=1}^{n} a_i m_i\,, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right)$ расподелу.

С обзиром да X_i има нормалну расподелу са параметрима m_i и σ_i^2, то је њена функција генератриса момената

$$M_{X_i}(t) = \exp\left(m_i t + \frac{\sigma_i^2 t^2}{2}\right).$$

Према томе случајна променљива $a_i X_i$ има функцију генератриса момената

$$M_{a_i X_i}(t) = M_{X_i}(a_i t) = \exp\left(m_i a_i t + \frac{\sigma_i^2 a_i^2 t^2}{2}\right).$$

Најзад, с обзиром на независност случајних променљивих X_i, $i=1,2,\ldots,n$, функција генератриса момената њихове линеарне комбинације је

$$M_U(t)=\prod_{i=1}^{n} M_{X_i}(a_i t)=\exp\left(t\sum_{i=1}^{n} a_i m_i+\frac{t^2}{n}\sum_{i=1}^{n} a_i^2\sigma_i^2\right),$$

што је функција генератриса момената нормално расподељене случајне променљиве са параметрима

$$E(U)=\sum_{i=1}^{n} a_i m_i \qquad \text{и} \qquad D(U)=\sum_{i=1}^{n} a_i^2\sigma_i^2\,.\triangle$$

За дводимензиону случајну променљиву (X,Y) функција генератриса момената, уколико постоји, дефинише се као функција од два аргумента:

$$M(t,s)=E\left(e^{tX+sY}\right).$$

Наравно, све наведене теореме важе и када су у питању вишедимензионе случајне променљиве, чиме се овде нећемо посебно бавити, али ћемо ипак илустровати једним добро познатим примером.

Пример 96. Наћи функцију генератриса момената дводимензионог вектора нормално расподељених случајних променљивих

$$(X,Y):\mathcal{N}\left(m_X,m_Y,\sigma_X^2,\sigma_Y^2,\rho\right).$$

Из дефиниције следи да је

$$M(t,s)=\frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\cdot$$

$$\cdot\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} e^{tx+sy}e^{-\frac{1}{2(1-\rho^2)}\left(\frac{(x-m_X)^2}{\sigma_X^2}-2\rho\frac{(x-m_X)(y-m_Y)}{\sigma_X\sigma_Y}+\frac{(y-m_Y)^2}{\sigma_Y^2}\right)}dxdy=$$

$$=\exp\left\{m_X t+m_Y s+\frac{\sigma_X^2t^2+2\rho\sigma_X\sigma_Y ts+\sigma_Y^2s^2}{2}\right\}.\triangle$$

Пример 97. Нека је $(X_1,Y_1),(X_2,Y_2),\ldots,(X_n,Y_n)$ прост узорак обима n из дводимензионе нормалне расподеле $\mathcal{N}(m_X,m_Y,\sigma_X^2,\sigma_Y^2,\rho)$. Наћи заједничку расподелу двеју статистика

$$\overline{X}=\sum_{i=1}^{n}\frac{X_i}{n} \qquad \text{и} \qquad \overline{Y}=\sum_{i=1}^{n}\frac{Y_i}{n}\,.$$

С обзиром да свака од статистика $\overline{X}$ и $\overline{Y}$ има нормалну расподелу и то

$$\overline{X} : \mathcal{N}(m_X, \frac{\sigma_X^2}{n}) \qquad \text{и} \qquad \overline{Y} : \mathcal{N}(m_Y, \frac{\sigma_Y^2}{n}),$$

то је функција генератриса момената дводимензионе статистике $(\overline{X}\,, \overline{Y})$ дата са

$$M(t,s) = E\left(e^{t\overline{X}+s\overline{Y}}\right) = E\left(\exp\left\{\frac{1}{n}\sum_{i=1}^{n}(tX_i + sY_i)\right\}\right).$$

Како је узорак прост, то је

$$M(t,s) = \prod_{i=1}^{n} E\left(e^{\frac{t}{n}X_i + \frac{s}{n}Y_i}\right).$$

На основу примера 96 следи да је тражена функција

$$\begin{aligned} M(t,s) &= \prod_{i=1}^{n} \exp\left\{\frac{m_X t}{n} + \frac{m_Y s}{n} + \frac{\sigma_X^2(\frac{t}{n})^2 + 2\rho\sigma_X\sigma_Y\frac{t}{n}\frac{s}{n} + \sigma_Y^2(\frac{s}{n})^2}{2}\right\} = \\ &= \exp\left\{m_X t + m_Y s + \frac{\frac{\sigma_X^2}{n}t^2 + 2\rho\frac{\sigma_X\sigma_Y}{n}ts + \frac{\sigma_Y^2}{n}s^2}{2}\right\} \end{aligned}$$

што је функција генератриса момената дводимензионе нормалне расподеле $\mathcal{N}(m_X\,, m_Y\,, \frac{\sigma_X^2}{n}\,, \frac{\sigma_Y^2}{n}\,, \rho)$. $\triangle$

6.3 Тачкасто оцењивање параметара обележја коначне популације

Код бесконачне популације (која теоријски дозвољава избор узорака бесконачног обима) примењују се граничне теореме теорије вероватноће. Међутим, на коначној популацији ова теорија нема у потпуности оправдања. У оквиру овог поглавља ћемо се у кратким цртама упознати са оцењивањем параметара обележја задатог на популацији коначног обима. Разматраћемо оцене очекивања и дисперзије.

6.3.1 Оцене математичког очекивања и дисперзије

Подсетимо се да је средина узорка непристрасна оцена математичког очекивања без обзира на расподелу обележја.

Нека је дата популација Ω која је коначан скуп од N елемената. Посматраћемо избор узорка обима n, $n \leq N$, без враћања из популације Ω. Тај узорак је случајна величина – вектор $(X_1, \ldots, X_n)$. Сваки елемент популације има једнаку вероватноћу избора у узорак у моменту извлачења. Дакле, $\omega \in s$ у првом извлачењу са вероватноћом $\frac{1}{N}$, у другом $\frac{1}{N-1}$, итд. С обзиром да је популација коначна, обележје X је дискретног типа тј.

$$P\{X = x_j\} = \frac{N_j}{N} \quad , \quad j = 1, \ldots, k\,,$$

где су x_j, $j = 1, \ldots, k$ све могуће вредности обележја X, а N_j је укупан број елемената популације код којих је вредност посматраног обележја X баш x_j. Лако је показати да случајне променљиве X_i, $1 \leq i \leq n$ имају исту расподелу као и X иако су међусобно зависне. Наиме,

$$P\{X_i = x_j\} = P\{\omega : X_i(\omega) = x_j\} = P\{\omega_1^{(i)}, \ldots, \omega_{N_j}^{(i)}\}\,,$$

где је са $\omega_l^{(i)}$, $l = 1, 2, \ldots, N_j$ означен елемент популације, $\omega_l^{(i)} \in \Omega$, код кога је вредност обележја X баш x_j, а изабран је у узорак у i–том извлачењу. Тада тражену вероватноћу одређујемо као

$$P\{\omega_1^{(i)}, \ldots, \omega_{N_j}^{(i)}\} = \sum_{l=1}^{N_j} P(\omega_l^{(i)}),$$

где је $P(\omega_l^{(i)})$ вероватноћа да је у i–том опиту изабран баш $\omega_l^{(i)}$, што значи да се у претходних $i-1$ опита није реализовао, јер је узорак без враћања. Дакле,

$$P(\omega_l^{(i)}) = \frac{\binom{N-1}{1}}{\binom{N}{1}} \cdot \frac{\binom{N-2}{1}}{\binom{N-1}{1}} \cdot \cdots \cdot \frac{\binom{N-i}{1}}{\binom{N-i+1}{1}} \cdot \frac{\binom{1}{1}}{\binom{N-i}{1}} = \frac{1}{N} \quad , \quad \forall i = 1, 2, \ldots, n\,.$$

То значи да је

$$P\{X_i = x_j\} = \sum_{l=1}^{N_j} \frac{1}{N} = \frac{N_j}{N}$$

за свако $i = 1, 2, \ldots, n$, тј. да X_i има исту расподелу као и X. Код узорка са враћањем је ова чињеница очигледна.

Из чињенице о расподели случајних променљивих X_i, $i = 1, 2, \ldots, n$, тј. једнакости те расподеле са расподелом самог обележја X следи да је $E(X_i) = E(X)$, па је средина узорка (као и код простог узорка са враћањем) једнака

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i\,,$$

а раније смо показали да је она непристрасна оцена математичког очекивања обележја X.

С обзиром да је посматрана оцена непристрасна, средње квадратно одступање, као критеријум ваљаности, даје

$$E(\overline{X}_n - EX)^2 = D(\overline{X}_n)\,.$$

Како су случајне променљиве $X_1, \ldots, X_n$ код узорка без враћања зависне, биће

$$D(\sum_{i=1}^{n} X_i) = E\left(\sum_{i=1}^{n} X_i - nE(X)\right)^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} E(X_i - E(X))(X_j - E(X)) =$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} Cov(X_i, X_j) = nD(X_i) + n(n-1)Cov(X_1, X_2)\,,$$

јер сви парови (X_i, X_j) имају исту расподелу као и пар (X_1, X_2).

Може се показати да је $P\{(X_i, X_j) = (x_l, x_r)\} = \frac{1}{N(N-1)}$ за свако $i, j = 1, 2, \ldots, n$ и свако $l, r = 1, 2, \ldots, k$. То указује да расподела за (X_i, X_j) не зависи од i, j или n па следи

$$Cov(X_i, X_j) = Cov(X_1, X_2)\,.$$

Означимо ову коваријансу са $c(N)$. Дакле,

$$D(\overline{X}_n) = \frac{1}{n^2} D(\sum_{i=1}^{n} X_i) = \frac{1}{n^2}(nD(X) + n(n-1)c(N)) =$$
$$= \frac{1}{n}(D(X) + (n-1)c(N))\,.$$

Величину $c(N)$ могуће је одредити из услова да је $n = N$, тј. да узорком без враћања исцрпимо целу популацију. Тада је

$$D(\overline{X}_N) = \frac{1}{N}(D(X) + (N-1)c(N))\,.$$

Међутим, тада је $\overline{X}_N$ тачна вредност очекивања $E(X)$, јер је

$$\overline{X}_N = \frac{1}{N}(X(\omega_1) + \ldots + X(\omega_N)) = E(X)\,,$$

па је $D(\overline{X}_N) = 0$. Отуда $D(X) + (N-1)c(N) = 0$, односно $c(N) = -\frac{1}{N-1}D(X)$. Најзад

$$D(\overline{X}_N) = \frac{1}{n}(D(X) - \frac{n-1}{N-1}D(X)) = \frac{D(X)}{n}\left(1 - \frac{n-1}{N-1}\right) = \frac{D(X)}{n} \cdot \frac{N-n}{N-1}\,.$$

Подсетимо се да је грешка исте оцене код узорка са враћањем

$$E\left(\overline{X}_N - E(X)\right)^2 = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i - E(X)\right)^2 = \frac{1}{n}D(X)\,,$$

што значи да је грешка оцене код узорка без враћања мања у односу на ону код узорка са враћањем.

Фактор $\dfrac{N-n}{N-1} = 1 - \dfrac{n-1}{N-1}$, који је код популације великог обима N и обима узорка n који је мали у односу на обим популације приближно једнак $1 - \frac{n}{N}$, назива се *корекција због коначности популације*.

Када је реч о оцени дисперзије обележја $D(X)$, већ смо показали да је код узорка са враћањем

$$\widetilde{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

непристрасна оцена дисперзије обележја. Међутим, ако је обим узорка велики, количник $\frac{n}{n-1}$ је приближно једнак 1, па се за оцену дисперзије може узети и узорачка дисперзија

$$\overline{S}_n^{\,2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

за коју је $E(\overline{S}_n^{\,2}) = \frac{n-1}{n}D(X)$.

Размотримо пристрасност узорачке дисперзије за случај узорка без враћања из коначне популације. Тада је

$$\begin{aligned} E(\overline{S}_n^{\,2}) &= E\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2 - \overline{X}_n^{\,2}\right) = \frac{1}{n}\sum_{i=1}^{n}E(X_i^2) - E(\overline{X}_n^{\,2}) = \\ &= E(X^2) - \frac{D(X)}{n}\cdot\frac{N-n}{N-1} - (EX)^2 = D(X)\left(1 - \frac{1}{n}\cdot\frac{N-n}{N-1}\right) = \\ &= D(X)\frac{(n-1)N}{n(N-1)} = \frac{n-1}{n}D(X)\frac{N}{N-1}\,. \end{aligned}$$

Како је обим популације N најчешће велики, то је $\frac{N}{N-1}$ приближно 1 па се као непристрасна оцена дисперзије обележја може поново узети $\widetilde{S}_n^2$ код узорка малог обима n (рецимо $n \le 30$), односно $\overline{S}_n^{\,2}$ код узорка великог обима.

6.3.2 Оцене математичког очекивања и дисперзије код стратификованог узорка

Нека је поново реч о коначној популацији са обележјем X чија је расподела

$$X : \begin{pmatrix} x_1 & x_2 & \dots & x_k \\ \frac{M_1}{N} & \frac{M_2}{N} & \dots & \frac{M_k}{N} \end{pmatrix},$$

$$M_1 + M_2 + \ldots + M_k = N\,, \quad 0 \le M_i \le N\,, \quad i = 1, 2, \ldots, k\,.$$

Означимо са $X^{(h)}$ обележје X посматрано на h–том стратуму за који ћемо претпоставити да је обима N_h. Претпоставимо, такође, да је популација подељена на L стратума, $L < N$. Тада случајна променљива $X^{(h)}$ има следећу расподелу

$$X^{(h)} : \left(\begin{array}{cccc} x_1 & x_2 & \ldots & x_k \\ \frac{M_1^{(h)}}{N_h} & \frac{M_2^{(h)}}{N_h} & \ldots & \frac{M_k^{(h)}}{N_h} \end{array} \right) ,$$

$$\sum_{i=1}^{k} M_i^{(h)} = N_h \,, \quad 0 \leq M_i^{(h)} \leq N_h \,, \quad i = 1, 2, \ldots, k \,, \quad h = 1, 2, \ldots, L \,.$$

Његово очекивање је

$$E\left(X^{(h)}\right) = \sum_{j=1}^{k} x_j \frac{M_j^{(h)}}{N_h} \,,$$

а очекивање обележја X је

$$\begin{aligned} E(X) &= \sum_{j=1}^{k} x_j \frac{M_j}{N} = \frac{1}{N} \sum_{j=1}^{k} x_j \left(M_j^{(1)} + \ldots + M_j^{(L)} \right) = \\ &= \frac{1}{N} \left(\sum_{j=1}^{k} x_j M_j^{(1)} + \ldots + \sum_{j=1}^{k} x_j M_j^{(L)} \right) = \\ &= \frac{N_1}{N} \sum_{j=1}^{k} x_j \frac{M_j^{(1)}}{N_1} + \ldots + \frac{N_k}{N} \sum_{j=1}^{k} x_j \frac{M_j^{(L)}}{N_L} = \\ &= \frac{N_1}{N} E\left(X^{(1)}\right) + \ldots + \frac{N_k}{N} E\left(X^{(L)}\right) = \\ &= \frac{1}{N} \sum_{i=1}^{L} N_i E\left(X^{(i)}\right) = \sum_{i=1}^{L} w_i E\left(X^{(i)}\right) \,, \end{aligned}$$

где смо са $w_i = \frac{N_i}{N}$ означили удео i–тог стратума у популацији. Релација $E(X) = \sum_{i=1}^{L} w_i E\left(X^{(i)}\right)$ представља везу између математичког очекивања обележја и очекивања по стратумима.

Размотримо сада проблем оцењивања на основу узорка. Из укупно L стратума од којих j–ти има тачно N_j елемената популације, бира се по $n_j \leq N_j$, $j = 1, 2, \ldots, L$, елемената у узорак. Претпоставка је да су извлачења из различитих стратума независна.

Део узорка из h–тог стратума је $(X_1^{(h)}, \ldots, X_{n_h}^{(h)})$, а припадајућа узорачка средина и дисперзија су

$$\overline{X}_{n_h} = \frac{1}{n_h} \sum_{i=1}^{n_h} X_i^{(h)}$$

$$\overline{S}_{n_h}^2 = \frac{1}{n_h} \sum_{i=1}^{n_h} (X_i^{(h)} - \overline{X}_{n_h})^2 \,,$$

односно поправљена дисперзија узорка

$$\widetilde{S}_n^2 = \frac{n_h}{n_h - 1} \overline{S}_{n_h}^2 .$$

При томе је $\overline{X}_{n_h}$ непристрасна оцена средње вредности обележја $X^{(h)}$ у h–том стратуму.

Статистика

$$\widetilde{X} = \frac{1}{N} \sum_{i=1}^{L} N_i \overline{X}_{n_i} = \sum_{i=1}^{L} w_i \overline{X}_{n_i}$$

није средина узорка, али је ипак непристрасна оцена метеметичког очекивања обележја X без обзира да ли је узорак са враћањем или без враћања. Заиста, како је

$$E\left(\overline{X}_{n_h}\right) = EX^{(h)} \,,$$

то је

$$E(\widetilde{X}) = \frac{1}{N} \sum_{i=1}^{L} N_i E(\overline{X}_{n_i}) = \frac{1}{N} \sum_{i=1}^{L} N_i E(X^{(i)}) = E(X) \,.$$

Средњеквадратна грешка ове оцене је за узорак са враћањем

$$E(\widetilde{X} - E(X))^2 = D(\widetilde{X}) = \sum_{i=1}^{L} w_i^2 \frac{D(\overline{X}_{n_i})}{n_i} \,,$$

а за узорак без враћања

$$E(\widetilde{X} - E(X))^2 = D(\widetilde{X}) = \sum_{i=1}^{L} w_i^2 D(\overline{X}_{n_i}) \approx \sum_{i=1}^{L} w_i^2 \frac{D(X)}{n_i} \left(1 - \frac{n_i}{N_i}\right) .$$

6.3.3 Оцена параметра p биномне расподеле

У овом поглављу задатак ће нам бити да на основу узорка обима n оценимо вероватноћу p да поједини елемент популације $\omega \in \Omega$ има својство A. У ту сврху уочићемо индикатор догађаја A

$$I_A(\omega) = \begin{cases} 1, & \omega \in A \\ 0, & \text{иначе} \end{cases}$$

Тада је $E(I_A) = p$, па се оцењивање параметра p своди на оцењивање математичког очекивања индикаторског обележја. Ако претпоставимо да у узорку обима n има n_1 елемената са својством A, тада је

$$\overline{I}_{A,n} = \frac{1}{n} n_1 = \frac{n_1}{n}$$

релативна учестаност својства A у узорку и она представља непристрасну оцену параметра p. Дакле, $\hat{p} = \overline{I}_{A,n}$. Средњеквадратна грешка ове оцене је

а) код узорка са враћањем

$$E(\overline{I}_{A,n} - p)^2 = D(\overline{I}_{A,n}) = \frac{p(1-p)}{n}\,,$$

б) код узорка без враћања

$$E(\overline{I}_{A,n} - p)^2 = D(\overline{I}_{A,n}) = \frac{D(I_A)}{n} \cdot \frac{N-n}{N-1} = \frac{p(1-p)}{n} \cdot \frac{N-n}{N-1}\,,$$

што је за велики обим популације N приближно $\dfrac{p(1-p)}{n} \cdot \left(1 - \dfrac{n}{N}\right)$.

Један од раније наведених критеријума ваљаности оцене био је количник $\dfrac{(\hat{\theta} - \theta)^2}{\theta}$, где је $\hat{\theta}$ статистика којом оцењујемо параметар θ. Искористимо овај критеријум за одређивање обима узорка који би обезбедио да се постигне унапред задата тачност оцене параметра p. Дакле, за унапред задато $\varepsilon > 0$ и $0 < \alpha < 1$ тражимо n за који

$$P\{Q_k \le \varepsilon\} = 1 - \alpha\,, \quad \text{за који важи}, \quad Q_k = \sqrt{\frac{1}{k}\sum_{i=1}^{k} \frac{(\hat{\theta}_i - \theta_i)^2}{\theta_i}}\,.$$

Пример 98. Нека је обележје X дискретног типа са расподелом $P\{X = x_i\} = p_i$, $i = 1, \ldots, k$. Задатак је оценити параметре p_i ове расподеле на основу простог случајног узорка.

Ако са n_i означимо број појављивања вредности x_i у узорку обима n, тада је $f_i = \frac{n_i}{n}$ релативна учестаност вредности x_i у узорку. За оцену вероватноће p_i, као што смо показали узима се $\hat{p}_i = f_i$. За меру одступања те оцене узима се

$$\frac{(f_i - p_i)^2}{p_i}$$

односно, заједничка просечна мера за цео узорак биће

$$Q_k^2 = \frac{1}{k}\sum_{i=1}^{n}\frac{(f_i - p_i)^2}{p_i} = \frac{1}{k}\sum_{i=1}^{k}\frac{(n_i - np_i)^2}{n^2 p_i} = \frac{1}{n}\frac{1}{k}\sum_{i=1}^{k}\frac{(n_i - np_i)^2}{np_i}\,.$$

Као што је познато, статистика

$$\sum_{i=1}^{k}\frac{(n_i - np_i)^2}{np_i}$$

има приближно χ^2–расподелу са k степени слободе, што значи

$$knQ_k^2 \sim \chi^2_{k-1}\,, \quad \text{када} \quad n \longrightarrow \infty.$$

Дакле,

$$P\{Q_k \le \varepsilon\} = 1 - \alpha$$
$$P\{knQ_k^2 \le kn\varepsilon^2\} = 1 - \alpha$$
$$P\{\chi^2_{k-1} \le kn\varepsilon^2\} = 1 - \alpha\,.$$

Следи да је $kn\varepsilon^2$ квантил реда $1-\alpha$ случајне промењиве са χ^2_{k-1} расподелом одакле се добија величина $kn\varepsilon^2 = \chi^2_{k-1,1-\alpha}$ (што се чита из таблице за χ^2 расподелу) и

$$n = \frac{\chi^2_{k-1,1-\alpha}}{k\varepsilon^2}\,.$$

$\triangle$

6.4 Густине расподела статистика поретка апсолутно непрекидног типа

Нађимо маргиналне густине статистика поретка у случају обележја X апсолутно непрекидног типа. Нека је густина расподеле обележја X означена са f и нека је она строго позитивна на интервалу (a, b), за неке $a, b \in \overline{R}$ и $a < b$, тј. $f(x) > 0$ за $x \in (a, b)$. Тада је

$$F(x) = \begin{cases} 0, & x < a \\ \int_a^x f(w)dw, & a \leq x < b \\ 1, & x \geq b \end{cases} .$$

Очигледно, за $a < x < b$ је:

$$1 - F(x) = F(b) - F(x) = \int_x^b f(w)dw.$$

Дакле,

$$g_n(y_n) = \int_a^{y_n} dy_{n-1} \int_a^{y_{n-1}} dy_{n-2} \dots \int_a^{y_2} (n!) f(y_1) f(y_2) \dots f(y_n) dy_1 \,, \qquad a < y_n < b\,.$$

Како је

$$\int_a^{y_2} f(y_1) dy_1 = F(y_2)\,,$$

добијамо

$$\int_a^{y_3} F(y_2) f(y_2) dy_2 = \frac{(F(y_3))^2}{2}\,,$$

јер је $F(a) = 0$. Тада је:

$$g_n(y_n) = n! \frac{(F(y_n))^{n-1}}{(n-1)!} f(y_n)\,,$$

односно

$$g_n(y_n) = \begin{cases} n(F(y_n))^{n-1} f(y_n), & a < y_n < b \\ 0, & \text{иначе} \end{cases}$$

и слично

$$g_1(y_1) = \begin{cases} n(1-F(y_1))^{n-1} f(y_1), & a < y_1 < b \\ 0, & \text{иначе}\,. \end{cases}$$

Коначно, за произвољно y_j, $j = 1, \ldots, n$ уочимо да је

$$\int_a^x (F(w))^{\alpha-1} f(w) dw = \frac{(F(x))^\alpha}{\alpha} \quad , \quad \alpha > 0$$

и

$$\int_y^b (1-F(w))^{\beta-1} f(w) dw = \frac{(1-F(y))^\beta}{\beta} \quad , \quad \beta > 0$$

па се добија

$$g_j(y_j) = \begin{cases} \frac{n!}{(n-j)!(j-1)!} (1-F(y_j))^{n-j} (F(y_j))^{j-1} f(y_j), & a < y_j < b \\ 0, & \text{иначе}\,. \end{cases}$$

Маргиналну густину вектора (Y_i, Y_j), $Y_i \leq Y_j$, за било које $1 \leq i < j \leq n$ израчунавамо на следећи начин:

$$\begin{aligned} g_{ij}(y_i, y_j) = \int_a^{y_i} dy_{i-1} \int_a^{y_{i-1}} dy_{i-2} \ldots \int_a^{y_2} dy_1 \int_{y_i}^{y_j} dy_{i+1} \int_{y_{i+1}}^{y_j} dy_{i+2} \ldots \\ \ldots \int_{y_{j-2}}^{y_j} dy_{j-1} \int_{y_j}^b dy_{j+1} \int_{y_{j+1}}^b dy_{j+2} \ldots \int_{y_{n-1}}^b (n!) f(y_1) \ldots f(y_n) dy_n\,. \end{aligned}$$

Како је за $\gamma > 0$

$$\int_x^y (F(y)-F(w))^{\gamma-1} f(w) dw = -\frac{(F(y)-F(w))^\gamma}{\gamma}\Big|_x^y = \frac{(F(y)-F(x))^\gamma}{\gamma}\,,$$

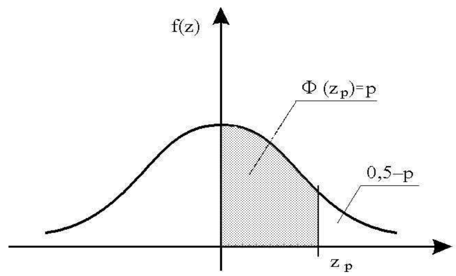
то је коначно маргинална густина вектора (Y_i, Y_j) дата са

$$g_{ij}(y_i, y_j) =$$

$$= \begin{cases} \frac{n!}{(i-1)!(j-i-1)!(n-j)!} (F(y_i))^{i-1} (F(y_j)-F(y_i))^{j-i-1} (1-F(y_j))^{n-j} f(y_i) f(y_j), \\ \qquad\qquad a < y_i < y_j < b \\ 0\,, \qquad\qquad \text{иначе}\,. \end{cases}$$

Статистичке таблице

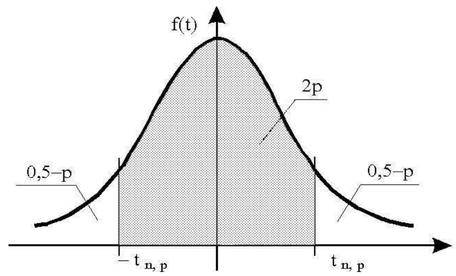
1. Нормална расподела



$\Phi(z_p) = p$

z	0	1	2	3	4	5	6	7	8	9
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0754
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2258	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2518	.2549
0.7	.2580	.2612	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2996	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998
3.5	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998
3.6	.4998	.4998	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.7	.4998	.4998	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.8	.4998	.4998	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.9	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000

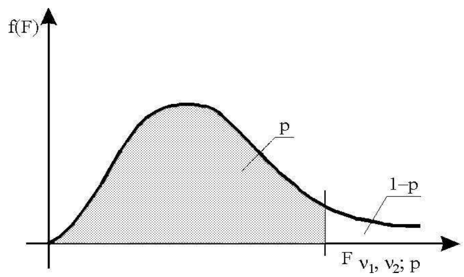
2. Студентова расподела



$P\{|t_n| < t_{n,p}\} = 2p$

$n \setminus p$	0.100	0.200	0.300	0.400	0.450	0.475	0.490	0.495
1	.325	.727	1.376	3.078	6.314	12.706	31.821	63.657
2	.289	.617	1.061	1.886	2.920	4.303	6.965	9.925
3	.277	.584	.978	1.638	2.353	3.182	4.541	5.841
4	.271	.569	.941	1.533	2.132	2.776	3.747	4.604
5	.267	.559	.920	1.476	2.015	2.571	3.365	4.032
6	.265	.553	.906	1.440	1.943	2.447	3.143	3.707
7	.263	.549	.896	1.415	1.895	2.365	2.998	3.499
8	.262	.546	.889	1.397	1.860	2.306	2.896	3.355
9	.261	.543	.883	1.383	1.833	2.262	2.821	3.250
10	.260	.542	.879	1.372	1.812	2.228	2.764	3.169
11	.260	.540	.876	1.363	1.796	2.201	2.718	3.106
12	.259	.539	.873	1.356	1.782	2.179	2.681	3.055
13	.259	.538	.870	1.350	1.771	2.160	2.650	3.012
14	.258	.537	.868	1.345	1.761	2.145	2.624	2.977
15	.258	.536	.866	1.341	1.753	2.131	2.602	2.947
16	.258	.535	.865	1.337	1.746	2.120	2.583	2.921
17	.257	.534	.863	1.133	1.740	2.110	2.567	2.898
18	.257	.534	.862	1.330	1.734	2.101	2.552	2.878
19	.257	.533	.861	1.328	1.729	2.093	2.539	2.861
20	.257	.533	.860	1.325	1.725	2.086	2.528	2.845
21	.257	.532	.859	1.323	1.721	2.080	2.518	2.831
22	.256	.532	.858	1.321	1.717	2.074	2.508	2.819
23	.256	.532	.858	1.319	1.714	2.069	2.500	2.807
24	.256	.531	.857	1.318	1.711	2.064	2.492	2.797
25	.256	.531	.856	1.316	1.708	2.060	2.485	2.787
26	.256	.531	.856	1.315	1.706	2.056	2.479	2.779
27	.256	.531	.855	1.314	1.703	2.052	2.473	2.771
28	.256	.530	.855	1.313	1.701	2.048	2.467	2.763
29	.256	.530	.854	1.311	1.699	2.045	2.045	2.462
30	.256	.530	.854	1.310	1.697	2.042	2.457	2.750
40	.255	.529	.851	1.303	1.684	2.021	2.423	2.704
60	.254	.527	.848	1.296	1.671	2.000	2.390	2.660
120	.254	.526	.845	1.289	1.658	1.980	2.358	2.617
∞	.253	.524	.842	1.282	1.645	1.960	2.326	2.576

3. Фишерова расподела



$P\{F_{\nu_1,\nu_2} < F_{\nu_1,\nu_2;\ p}\} = p$

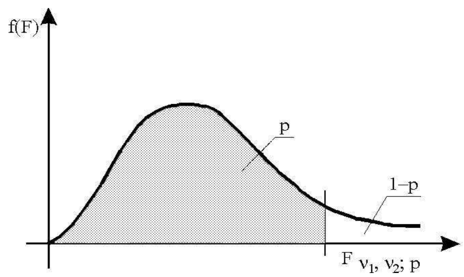
3. a) p = 0,990

$\nu_2 \setminus \nu_1$	1	2	3	4	5	6	7	8	9	10
1	4052	5000	5403	5625	5764	5859	5928	5981	6023	6056
2	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4
3	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2
4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5
5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1
6	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87
7	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62
8	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81
9	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26
10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10
14	8.86	6.51	5.56	5.04	4.70	4.46	4.28	4.14	4.03	3.94
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69
17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17
25	7.77	5.57	4.68	4.18	3.86	3.63	3.46	3.32	3.22	3.13
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.28	3.12	3.03
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32

наставак 3. а)

12	15	20	24	30	40	60	120	ν_1/ν_2
6106	6157	6209	6235	6261	6287	6313	6339	1
99.4	99.4	99.4	99.5	99.5	99.5	99.5	99.5	2
27.1	26.9	26.7	26.6	26.5	26.4	26.3	26.2	3
14.4	14.2	14.0	13.9	13.8	13.7	13.7	13.6	4
9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	5
7.72	7.56	7.40	4.31	7.23	7.14	7.06	6.97	6
6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	7
5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	8
5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	9
4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	10
4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	11
4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	12
3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	13
3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	14
3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	15
3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	16
3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	17
3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	18
3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	19
3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	20
3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	21
3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	22
3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	23
3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	24
2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	25
2.96	2.82	2.66	2.58	2.50	2.42	2.33	2.23	26
2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	27
2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	28
2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	29
2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	30
2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	40
2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	60
2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	120
2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	∞

Фишерова расподела



$P\{F_{\nu_1,\nu_2} < F_{\nu_1,\nu_2;\ p}\} = p$

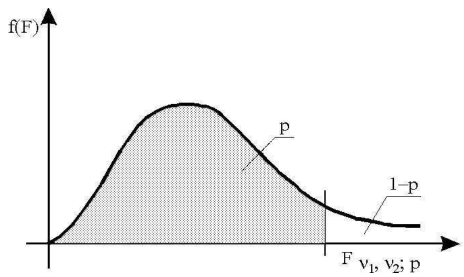
3. б) $p = 0,975$

$\nu_2 \setminus \nu_1$	1	2	3	4	5	6	7	8	9	10
1	648	799	864	900	922	937	948	957	963	969
2	38.5	39.0	39.2	39.2	39.3	39.3	39.4	39.4	39.4	39.4
3	17.4	16.0	15.4	15.1	14.9	14.7	14.6	14.5	14.5	14.5
4	12.2	10.6	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84
5	10.0	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77
21	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70
23	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64
25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61
26	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59
27	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63	2.57
28	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55
29	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59	2.53
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39
60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27
120	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16
∞	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05

наставак 3. б)

12	15	20	24	30	40	60	120	ν_1/ν_2
977	985	993	997	1001	1006	1010	1014	1
39.4	39.4	39.4	39.5	39.5	39.5	39.5	39.5	2
14.3	14.3	14.2	14.1	14.1	14.0	14.0	13.9	3
8.75	8.65	8.56	8.51	8.46	8.41	8.36	8.31	4
6.52	6.43	6.33	6.28	6.23	6.18	6.12	6.07	5
5.37	5.27	5.17	5.12	5.07	5.01	4.96	4.90	6
4.67	4.57	4.47	4.42	4.36	4.31	4.25	4.20	7
4.20	4.10	4.00	3.95	3.89	3.84	3.78	3.71	8
3.87	3.77	3.67	3.61	3.56	3.51	3.45	3.39	9
3.62	3.52	3.42	3.37	3.31	3.26	3.20	3.14	10
3.43	3.33	3.23	3.17	3.12	3.06	3.00	2.94	11
3.28	3.18	3.07	3.02	2.96	2.91	2.85	2.79	12
3.15	3.05	2.95	2.89	2.84	2.78	2.72	2.66	13
3.05	2.95	2.84	2.79	2.73	2.67	2.61	2.55	14
2.96	2.86	2.76	2.70	2.64	2.59	2.52	2.46	15
2.89	2.79	2.68	2.63	2.57	2.51	2.45	2.38	16
2.82	2.72	2.62	2.56	2.50	2.44	2.38	2.32	17
2.77	2.67	2.56	2.50	2.44	2.38	2.32	2.26	18
2.72	2.62	2.51	2.45	2.39	2.33	2.27	2.20	19
2.68	2.57	2.46	2.41	2.35	2.29	2.22	2.16	20
2.64	2.53	2.42	2.37	2.31	2.25	2.18	2.11	21
2.60	2.50	2.39	2.33	2.27	2.21	2.14	2.08	22
2.57	2.47	2.36	2.30	2.24	2.18	2.11	2.04	23
2.54	2.44	2.33	2.27	2.21	2.15	2.08	2.01	24
2.51	2.41	2.30	2.24	2.18	2.12	2.05	1.98	25
2.49	2.39	2.28	2.22	2.16	2.09	2.03	1.95	26
2.47	2.36	2.25	2.19	2.13	2.07	2.00	1.93	27
2.45	2.34	2.23	2.17	2.11	2.05	1.98	1.91	28
2.43	2.32	2.21	2.15	2.09	2.03	1.96	1.89	29
2.41	2.31	2.20	2.14	2.07	2.01	1.94	1.87	30
2.29	2.18	2.07	2.01	1.94	1.88	1.80	1.72	40
2.17	2.06	1.94	1.88	1.82	1.74	1.67	1.58	60
2.05	1.94	1.82	1.76	1.69	1.61	1.53	1.43	120
1.94	1.83	1.71	1.64	1.57	1.48	1.39	1.27	∞

Фишерова расподела



$$P\{F_{\nu_1,\nu_2} < F_{\nu_1,\nu_2;\ p}\} = p$$

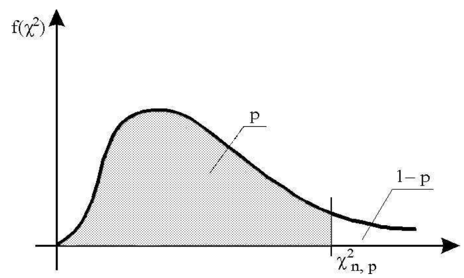
3. в) $p = 0,950$

$\nu_2 \setminus \nu_1$	1	2	3	4	5	6	7	8	9	10
1	161	200	216	225	230	234	237	239	241	242
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19
29	4.18	3.33	2.93	2.70	2.55	2.48	2.35	2.28	2.22	2.18
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.87	1.83

наставак 3. в)

12	15	20	24	30	40	60	120	ν_1/ν_2
244	246	248	249	250	251	252	253	1
19.4	19.4	19.4	19.5	19.5	19.5	19.5	19.5	2
8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	3
5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	4
4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	5
4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	6
3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	7
3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	8
3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	9
2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	10
2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	11
2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	12
2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	13
2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	14
2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	15
2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	16
2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	17
2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	18
2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	19
2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	20
2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	21
2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	22
2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	23
2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	24
2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	25
2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	26
2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	27
2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	28
2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	29
2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	30
2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	40
1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	60
1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	120
1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	∞

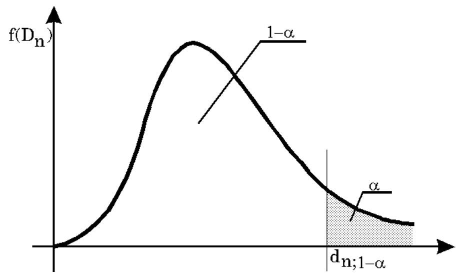
4. χ^2 расподела



$P\{\chi_n^2 < \chi_{n;\ p}^2\} = p$

$n \setminus p$	0.005	0.010	0.025	0.050	0.95	0.975	0.990	0.995
1	.0000	.0002	.0010	.0039	3.84	5.02	6.63	7.88
2	.0100	.0201	.0506	.103	5.99	7.38	9.21	10.6
3	.0717	.115	.216	.352	7.81	9.35	11.3	12.8
4	.207	.297	.484	.711	9.49	11.1	13.3	14.9
5	.412	.554	.831	1.15	11.1	12.8	15.1	16.7
6	.676	.872	1.24	1.64	12.6	14.4	16.8	18.5
7	.989	1.24	1.69	2.17	14.1	16.0	18.5	20.3
8	1.34	1.65	2.18	2.73	15.5	17.5	20.1	22.0
9	1.73	2.09	2.70	3.33	16.9	19.0	21.7	23.6
10	2.16	2.56	3.25	3.94	18.3	20.5	23.2	25.2
11	2.60	3.05	3.82	4.57	19.7	21.9	24.7	26.8
12	3.07	3.57	4.40	5.23	21.0	23.3	26.2	28.3
13	3.57	4.11	5.01	5.89	22.4	24.7	27.7	29.8
14	4.07	4.66	5.63	6.57	23.7	26.1	29.1	31.3
15	4.60	5.23	6.26	7.26	25.0	27.5	30.6	32.8
16	5.14	5.81	6.91	7.96	26.3	28.8	32.0	34.3
17	5.70	6.41	7.56	8.67	27.6	30.2	33.4	35.7
18	6.26	7.01	8.23	9.39	28.9	37.5	34.8	37.2
19	6.84	7.63	8.91	10.1	30.1	32.9	36.2	38.6
20	7.43	8.26	9.59	10.9	31.4	34.2	37.6	40.0
21	8.03	8.90	10.3	11.6	32.7	35.5	38.9	41.4
22	8.64	9.54	11.0	12.3	33.9	36.8	40.3	42.8
23	9.26	10.2	11.7	13.1	35.2	38.1	41.6	44.2
24	9.89	10.9	12.4	13.8	36.4	39.4	43.0	45.6
25	10.5	11.5	13.1	14.6	37.7	40.6	44.3	46.9
26	11.2	12.2	13.8	15.4	38.9	41.9	45.6	48.3
27	11.8	12.9	14.6	16.2	40.1	43.2	47.0	49.6
28	12.5	13.6	15.3	16.9	41.3	44.5	48.3	51.0
29	13.1	14.3	16.0	17.7	42.6	45.7	49.6	52.3
30	13.8	15.0	16.8	18.5	43.8	47.0	50.9	53.7
40	20.7	22.2	24.4	26.5	55.8	59.3	63.7	66.8
50	28.0	29.7	32.4	34.8	67.5	71.4	76.2	79.5
60	35.5	37.5	40.5	43.2	79.1	83.3	88.4	92.0
70	43.3	45.4	48.8	51.7	90.5	95.0	100	104
80	51.2	53.5	57.2	60.4	102	107	112	116
90	59.2	61.8	65.6	69.1	113	118	124	128
100	67.3	70.1	74.2	77.9	124	130	136	140

5. Расподела Колмогорова



$P\{D_n \geq d_{n;\ 1-\alpha}\} = \alpha$

$n \setminus \alpha$	0.200	0.150	0.100	0.050	0.010
1	.900	.925	.950	.975	.995
2	.684	.726	.776	.842	.929
3	.565	.597	.642	.708	.829
4	.494	.525	.564	.624	.734
5	.446	.474	.510	.563	.669
6	.410	.436	.470	.521	.618
7	.381	.405	.438	.486	.577
8	.358	.381	.411	.457	.543
9	.339	.360	.388	.432	.514
10	.322	.342	.368	.409	.486
11	.307	.326	.352	.391	.468
12	.295	.313	.338	.375	.450
13	.284	.302	.325	.361	.433
14	.274	.292	.314	.349	.418
15	.266	.283	.304	.338	.404
16	.258	.274	.295	.328	.391
17	.250	.266	.286	.318	.380
18	.244	.259	.278	.309	.370
19	.237	.252	.272	.301	.361
20	.231	.246	.264	.294	.352
25	.210	.220	.240	.264	.320
30	.190	.200	.220	.242	.290
35	.180	.190	.210	.230	.270
40				.190	.230
50				.190	.230
60				.170	.210
70				.160	.190
80				.150	.180
90				.140	
100				.140	
асимптотска формула	$\frac{1.07}{\sqrt{n}}$	$\frac{1.14}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.63}{\sqrt{n}}$

6. Случајни бројеви

51772	74640	42331	29044	46621
24033	23491	83587	06568	21960
45939	60173	52078	25424	11645
30586	02133	75797	45406	31041
03585	79353	81938	82322	96799
64937	03355	98683	20790	65304
15630	64759	51135	98527	62586
09448	56301	57683	30277	94623
21631	91157	77331	60710	52290
91097	17480	29414	06829	87843
62898	93582	04186	19640	87056
21387	76105	10863	97453	90581
55870	56974	37428	93507	94271
86707	12973	17169	88116	42187
85659	36081	50884	14070	74950
55189	00745	65253	11822	15804
41889	25439	88036	24034	67283
85418	68829	06652	41982	49159
16835	48653	71590	16159	14676
28195	27279	47152	35683	47280

Литература[1]

[1] Anděl J.: **Matematická statistika**, SNTZ/Alfa, Praha, 1985

[2] Anderson T.: **The statistical analysis of time series**, John Wiley & Sons, New York, 1971

[3] Боровков А.: **Математическая статистика**, Наука, Москва, 1984

[4] Box G., Jenkins G.: **Time series analysis**, Holden-Day, 1970

[5] Brockwell P., Davis R.: **Time series: Theory and methods**, Springer-Verlag, New York, 1987

[6] Brownlee K.: **Statistical theory and methodology in science and engineering**, John Wiley & Sons, New York, 1977

[7] Cochran W.: **Sampling techniques**, John Wiley & Sons, New York, 1963

[8] Дэйвид Г.: **Порядковые статистики**, Наука, Москва, 1979 (превод са енглеског)

[9] DeGroot M.: **Optimal statistical decisions**, McGrow-Hill Co., New York, 1970

[10] Ермаков С. и други: **Математическая теорија планирования эксперимента**, Наука, Москва, 1983

[11] Хаџић О.: **Нумеричке и статистичке методе у обради експерименталних података**, Универзитет у Новом Саду, Институт за математику, Нови Сад, 1992

[1] У навођењу литературе је поштован редослед "абц".

[12]Hogg R., Craig A.: **Introduction to mathematical statistics**, McMillan Co., New York, 1965

[13]Hogg R., McKean J., Craig A.:**Introduction to mathematical statistics**, Pearson, 2005

[14]Ивченко Г. И., Медведев Ю. И.: **Математическая статистика**, Вюшая школа, Москва, 1984

[15]Kendall M. G., Steward A.: **The Advanced Theory of Statistics**, vol. 3: Design and Analysis, and Time Series, Griffin company, London, 1968

[16]Малишић J.: **Временске серије**, Математички факултет, Београд, 2002

[17]Малишић J., Јевремовић В.: **Случајни процеси и временске серије**, Математички факултет, Београд, 2008

[18]Поповић Б.: **Математичка статистика и статистичко моделовање**, Природно–математички факултет, Ниш, 2003

[19]Поповић Б., Благојевић Б.: **Математичка статистика са применама у хидротехници**, Универзитет у Нишу, Ниш, 2003

[20]Поповић Б., Ристић М.: **Статистика у психологији**, Мрљеш, Београд, 2001

[21]Рао С.Р.: **Линейные статистические методы и их применения**, Наука, Москва, 1968 (превод са енглеског)

[22]Spiegel M. R.: **Probability and Statistics**, McGrow–Hill Book company, New York, 1975

[23]Стојановић С.: **Математичка статистика**, Научна књига, Београд, 1979

[24]Шметерер Л.: **Введеные в математическую статистику**, Наука, Москва, 1976 (превод са немачког)

[25]Thompson S.: **Sampling**, John Wiley & Sons, New York, 1992

[26]Wackerly D., Mendenhall W., Scheaffer R.: **Mathematical statistics with applications**, Duxbury Press, Belmont, 1996

[27] Zacks S.: **The theory of statistical inference**, John Wiley & Sons, New York, 1971

О аутору

Др Биљана Поповић је рођена у Нишу 1954. године. Академски назив дипломирани математичар стекла је на Филозофском факултету у Нишу 1976. године. Магистарску тезу (1982) и докторску дисертацију (1990) одбранила је на Математичком факултету ПМФ-а у Београду.

Универзитетску каријеру почиње на Филозофском факултету у Нишу 1979. године. По издвајању Природно-математичког факултета из Филозофског, постаје наставник новооснованог Природно-математичког факултета. Данас је редовни професор и предаје више предмета везаних за математичку статистику. У својој наставничкој каријери била је ангажована на Природно-математичком факултету Универзитета у Приштини (данас са седиштем у Косовској Митровици), на основним и последипломским студијама на Грађевинском факултету Универзитета у Нишу (данас Грађевинско–архитектонском), као и на Филозофском факултету, Студијска група за психологију, Универзтета у Нишу.

Од 1983. године је ангажована на реализацији научних пројеката, где је била и руководилац научног задатка на потпројекту "Стохастичка анализа". Била је координатор међународног научног пројекта "Asset and Liability Management".

Била је руководилац међународног наставног пројекта "Advanced Mathematics".

Објавила је већи број научних радова и учествовала у раду и организовању више научних скупова како домаћих тако и међународних. Објавила је више уџбеника и једну збирку задатака.

Рецензирала је више уџбеника и једну монографију. Рецензент је више научних часописа. Реферише за "Mathematical Reviews".

Члан је уредништва у четири инострана научна часописа.

Члан је Статистичког друштва Србије, ЈУПИМ-а, Друштва математичара Србије, Српског научног математичког друштва и

American Mathematical Society. Руководилац је Статистичке радионице Института за математику ПМФ-а у Нишу.

Ангажује се на популаризацији математичке статистике као широко примењливе научне дисциплине залажући се при томе за њену правилну и адекватну примену са јасним циљевима и строгим принципима статистичког закључивања. Зачетник је бављења математичком статистиком као научном дисциплином на Универзитету у Нишу.

Индекс појмова