СЕМИНАРСКА РАБОТА ПО ПРЕДМЕТОТ БИЗНИС СТАТИСТИКА

Обработка на податочно множество во пакетот „R“

Студент: Теа Минова 213004
Професор: проф. д-р Верица Бакева Смиљкова
Факултет за информатички науки и компјутерско инженерство

Содржина

ВОВЕДНИ ИНФОРМАЦИИ

Извор на податочно множество: https://www.kaggle.com/datasets

Линк до податочно множество: https://www.kaggle.com/datasets/ankits29/used-car-price-data

Ова податочно множество, чија цел е анализа на цените на користени возила, се состои од 2237 единки (возила), кои се разгледувани според 8 обележја и тоа:

- продажна цена
- изминати километри
- година на производство
- број на претходни сопственици
- тип на гориво
- вид на трансмисија
- информации за осигурување
- состојба на возилото (оцена од 0 до 5)

ИНИЦИЈАЛНА ОБРАБОТКА

Поради големото влијание на екстремните вредности врз резултатите од анализата на ова податочно множество, направено е иницијално филтрирање - оттргнување на единките со екстремни вредности од множеството. На тој начин се добива множество од 2075 единки, додека бројот на обележја останува ист.

А. ПРВ ДЕЛ – ДЕСКРИПТИВНА СТАТИСТИКА

При следната анализа на податочното множество, се избрани две обележја и тоа продажната цена на возилата и нивните изминати километри, чии вредности, за полесна визуелизација и разработка се поделени со 1000.

ТАБЕЛИ ЗА РАСПРЕДЕЛБА НА ЧЕСТОТИ

Во продолжение се претставени табели за распределба на честоти врз основа на двете избрани обележја, при што податоците се распределени во 12 интервали со ширина од 14 (за километри) и 65 (за цена). За секој интервал соодветно се определени средна точка, честота, релативна честота и кумулативна честота.

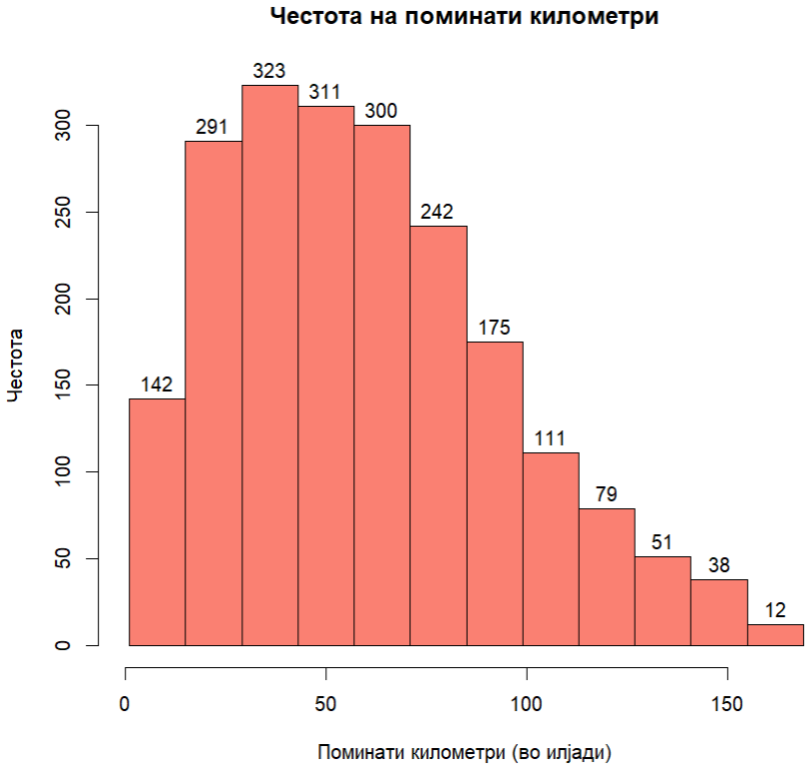
ИЗМИНАТИ КИЛОМЕТРИ (во илјади)

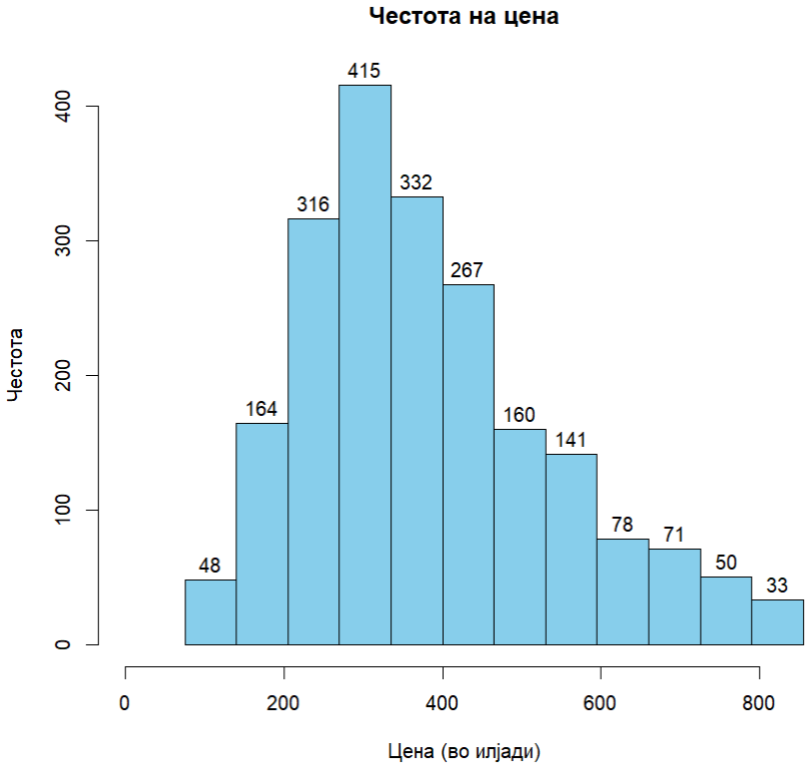
	Средни_точки1	Честоти1	Релативни_честоти1	Кумулативни_честоти1
[1,15)	8	142	0.0684	142
[15,29)	22	291	0.1402	433
[29,43)	36	323	0.1557	756
[43,57)	50	311	0.1499	1067
[57,71)	64	300	0.1446	1367
[71,85)	78	242	0.1166	1609
[85,99)	92	175	0.0843	1784
[99,113)	106	111	0.0535	1895
[113,127)	120	79	0.0381	1974
[127,141)	134	51	0.0246	2025
[141,155)	148	38	0.0183	2063
[155,169)	162	12	0.0058	2075

ПРОДАЖНА ЦЕНА (во илјади)

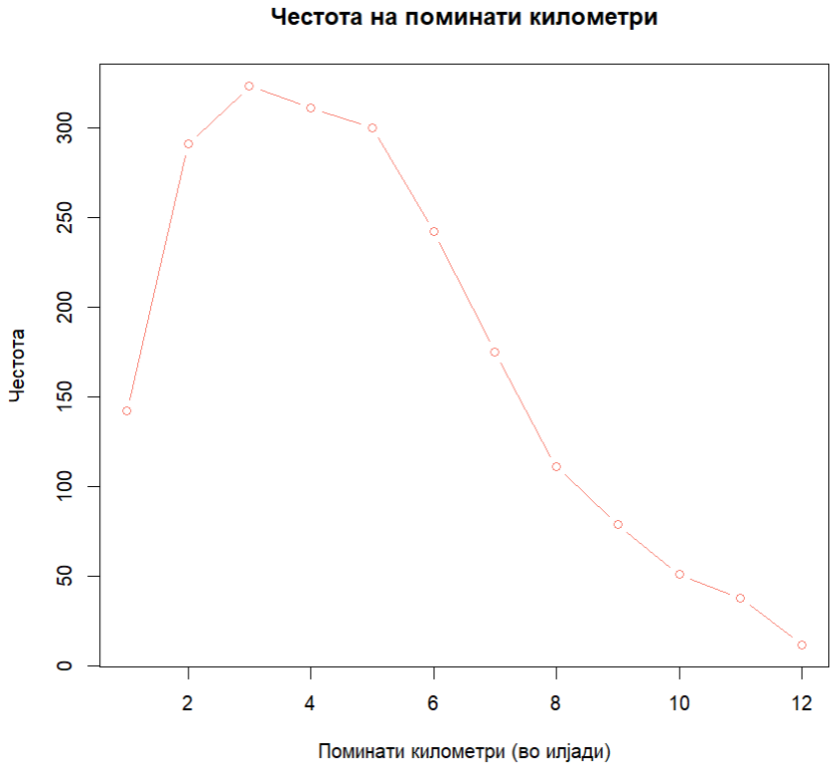
	Средни_точки2	Честоти2	Релативни_честоти2	Кумулативни_честоти2
[75,140)	107.5	48	0.0231	48
[140,205)	172.5	164	0.0790	212
[205,270)	237.5	316	0.1523	528
[270,335)	302.5	415	0.2000	943
[335,400)	367.5	332	0.1600	1275
[400,465)	432.5	267	0.1287	1542
[465,530)	497.5	160	0.0771	1702
[530,595)	562.5	141	0.0680	1843
[595,660)	627.5	78	0.0376	1921
[660,725)	692.5	71	0.0342	1992
[725,790)	757.5	50	0.0241	2042
[790,855)	822.5	33	0.0159	2075

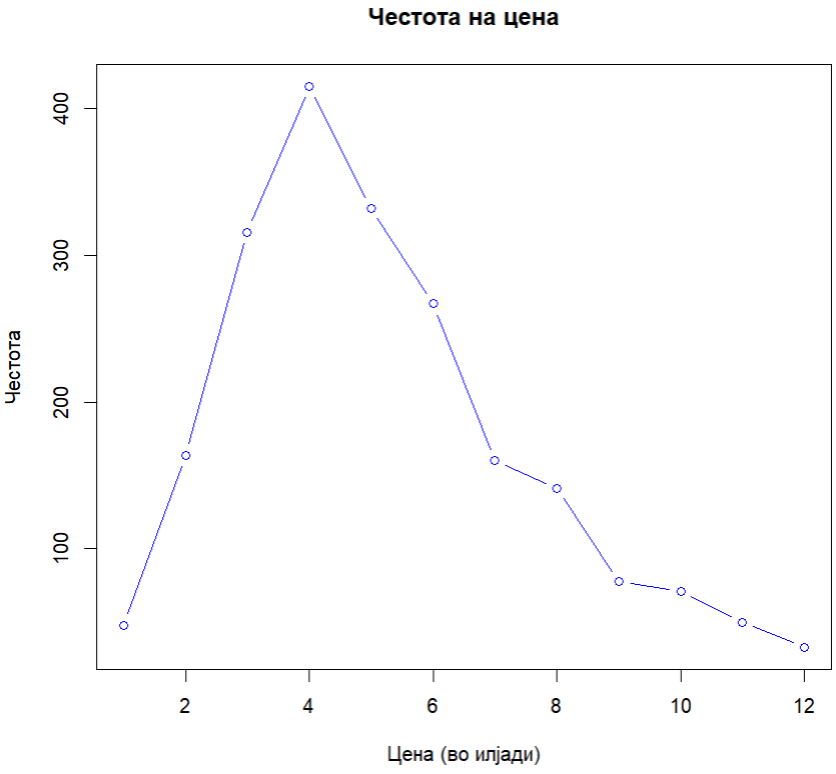
ГРАФИЧКО ПРЕТСТАВУВАЊЕ НА ПОДАТОЦИТЕ СО ХИСТОГРАМИ

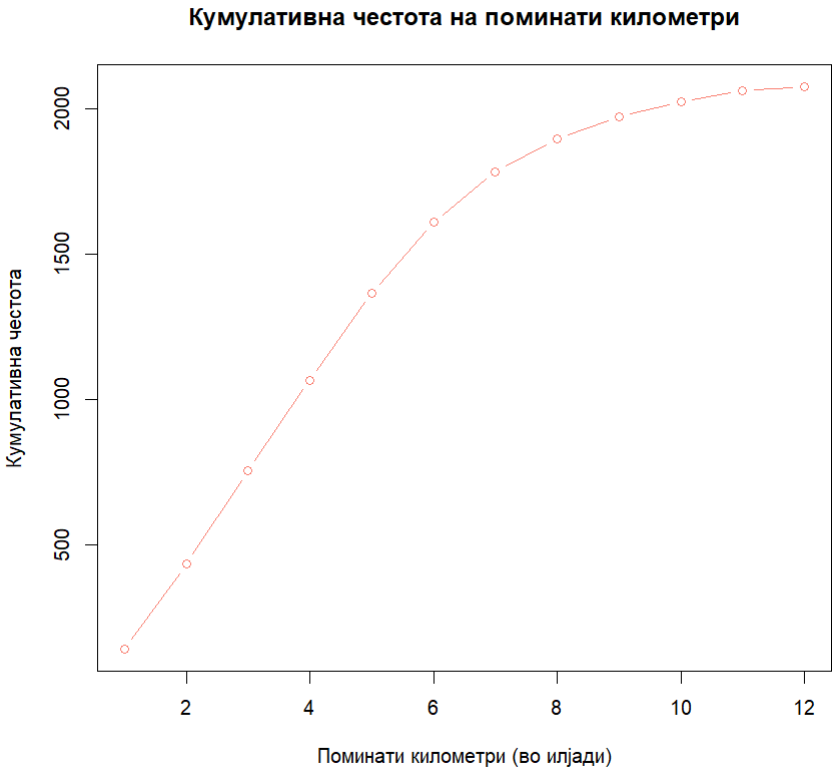


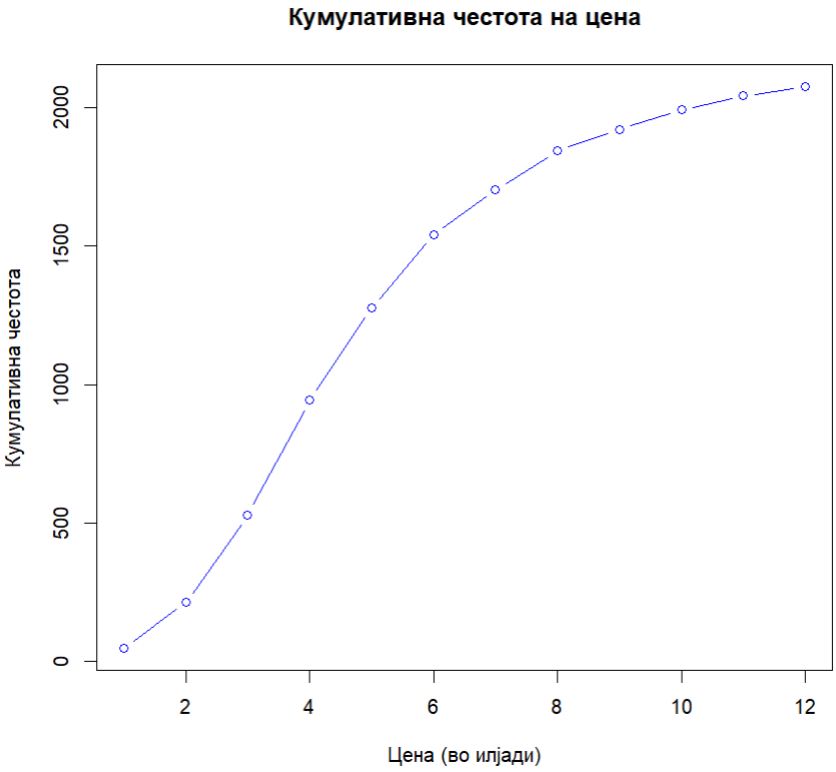


ГРАФИЧКО ПРЕТСТАВУВАЊЕ НА ПОДАТОЦИТЕ СО ПОЛИГОНИ






Кумулативна честота на поминати километри
Кумулативна честота
500
1000
1500
2000
2
4
6
8
10
12
Поминати километри (во илјади)


Кумулативна честота на цена
Кумулативна честота
0
500
1000
1500
2000
2
4
6
8
10
12
Цена (во илјади)

ПРЕТСТАВУВАЊЕ НА ПОДАТОЦИТЕ СО СТЕБЛО - ЛИСТ ДИЈАГРАМ

ИЗМИНАТИ КИЛОМЕТРИ (во илјади)

0 |
1122222333333344444444455555555666666667777777777888888888899999999999999
1 |
00000000000000000000011111111111222222222222333333333344444444444444455555555555555555566666666666666677777777777777788888888888888999999999999999999
2 |
00000000000000000000000111111111111111122222222222222222222222223333333333333333333333333333344444444444444444444444455555555555555555555566666666666666666777777777777777777777888888888888888888888888888888999999999999999999999
3 |
00000000000000000011111111111111111111222222222222222222222222233333333333333333333333333334444444444444444444444444455555555555555555555555566666666666666666777777777777777777778888888888889999999999999999999999999
4 |
00000000000000000000000001111111111111111111122222222222222222222222233333333333333334444444444444444444444444455555555555555555555555566666666666666666666666677777777777777777778888888888888888888888999999999999999999999
5 |
00000000000000000000000011111111111111111111222222222222222222222223333333333333333333333333344444444444444444444444455555555555555555555566666666666666777777777777777777778888888888888888888888999999999999999999999
6 |
00000000000000000000011111111111111111111222222222222222333333333333333333333344444444444444444444444444555555555555555555566666666666666666777777777777777777778888888888888888999999999999999999999
7 |
00000000000000000000111111111111111111111222222222222222233333333333333333333334444444444444444555555555555556666666666666666777777777777777777888888889999999999999
8 |
00000000000001111111111111111111112222222222222222233333333333333333333444444444444444444445555555555555555566666666677777777778888888888888888999999999999
9 |
000000000000001111111111222222222233333344444444444444455555555555566666666666777777777888888888888889999999999999
10 |
0000000000111111112222223333333344444444445555555555566666677777788888899999
11 |
000011111111112222233334445555555555566667777777888888999
12 |
00000001111112222333344444455555666667777778889999
13 |
000011111333333344445566666777777788899
14 |
011122222233444556677888999
15 |
01111111233455556677779
16 |
0

ПРОДАЖНА ЦЕНА (во илјади)

7 |
599
8 |
3
9 |
2259
10 |
1349
11 |
12244666789
12 |
001344556789
13 |
1111234445578
14 |
2233446777999
15 |
00011234556666788899
16 |
000001567888899
17 |
000001122333344455566666777777888899
18 |
001111111222222333466667777788999
19 |
000001233444455556677778889
20 |
00000001111122223334555556777778888999
21 |
000111111222222333333344555666778888899999999
22 |
0000011112222333333344455555667777788888888888999999
23 |
000001111222233333344455556666777888899
24 |
00000001112222222333333444455555566667777888888999
25 |
000000111222222233333334445555555556666667777888899999999
26 |
0000111111111222223344444444555555666666667777788999999999
27 |
000011111222222333444445555555666666777777888999999999
28 |
00000111122222223333334444455556666667777777888999999
29 |
000001111112222223333344444444555556666666677777778888
30 |
00000001111111112222222333333444455555555666666666667777777888888889999999999999
31 |
0000011111122222223333334444444455555555666666666777777788888888889999999
32 |
000111111222222223333333334444444445555566666777777788888889999
33 |
00000111111112222222233333333333344455555556666666666666777777788999999999999999
34 |
000000111112222233333334444444444455555555666666666777888889999999
35 |
000111111122233333344444455555566666667778888899
36 |
011122223334444455555666777788999999
37 |
00000011112223333333445666666777777788888889999999

38 |
00001111122223333444555555667778888999999
39 |
000000011111222222333444455566666777777788999999
40 |
00001122222223333444455566777777788888899999
41 |
000000011111122233333344444445555666667777888899
42 |
00000000000012222233333444444444566666777778888899999
43 |
000111222334444455555566788999
44 |
00001122223333444444456666777777777888
45 |
000011111122222333333444566666667899999
46 |
00011122233455566778899999
47 |
001112222233344455555667788 9
48 |
000011112222334455667777899
49 |
000133344555556778899
50 |
01122333444455566779
51 |
00111222333566778889
52 |
000001122223344444456666788899
53 |
001111134444555668888 9
54 |
000122234455555778999999
55 |
001112222333335556667789
56 |
001111123334455666668
57 |
00222233455667778899999
58 |
23444556667788899999
59 |
0133345557899
60 |
02333444444566677799
61 |
334556778899
62 |
011556
63 |
112333355789
64 |
0033446899
65 |
33335566788
66 |
145788889
67 |
01112334588999
68 |
12244456679
69 |
022455667789
70 |
0012666788999
71 |
0119
72 |
023333446899
73 |
0113347788
74 |
0124566699
75 |
133566899
76 |
12223599
77 |
338
78 |
033579
79 |
0136
80 |
11379
81 |
49999
82 |
00023488
83 |
0579
84 |
0113466

ГРАФИК НА РАСЕЈУВАЊЕ ЗА ПОДАТОЦИТЕ

Во продолжение е претставен график на расејување за податоците од претходно избраните две обележја – поминатите километри и продажната цена на возилата, со чија помош подоцна се дискутира односот помеѓу обележјата, односно се врши регресиона анализа на истите.

Засега, разгледувајќи го дијаграмот, може да се претпостави дека помеѓу обележјата километри и цена не постои поврзаност. Причината за тоа е што точките, кои го претставуваат секој подреден пар на поминати километри (во илјади) и цена (во илјади) на возилата, се значително распространети, при што е тешко да се одреди дали тие се подредени според некое правило.

МЕРКИ ЗА ЦЕНТРАЛНА ТЕНДЕНЦИЈА НА ПОДАТОЦИТЕ

		ЗНАЧЕЊЕ	ИЗМИНАТИ КИЛОМЕТРИ (во илјади)	ПРОДАЖНА ЦЕНА (во илјади)
ПРОСЕК		Аритметичка средина	59,51181	382,09783
МЕДИЈАНА		Стедина на подредени вредности	55	347
МОДА		Најчесто набљудувана вредност	34	336
КВАРТАЛИ	**ПРВ**	Вредност таква што приближно 25% од податоците во подредениот примерок се лево од неа	32	268
	ВТОР	Вредност таква што приближно 50% од податоците во подредениот примерок се лево од неа	55	347
	ТРЕТ	Вредност таква што приближно 75% од податоците во подредениот примерок се лево од неа	82	470

МЕРКИ ЗА ВАРИРАЊЕ НА ПОДАТОЦИТЕ

	ЗНАЧЕЊЕ	ИЗМИНАТИ КИЛОМЕТРИ (во илјади)	ПРОДАЖНА ЦЕНА (во илјади)
ОПСЕГ	Разликата помеѓу најголемата и најмалата вредност во примерокот	159	771
ИНТЕРКВАРТАЛЕН РАСПОН	Разлика помеѓу третиот и првиот квартал, се користи за определување на екстремни вредности	50	202
ДИСПЕРЗИЈА	Средна вредност на квадрираното растојание на дадените вредности од очекуваните (средини)	1192,4496	24980,7990
СТАНДАРДНА ДЕВИЈАЦИЈА	Варирање на податоците во примерокот околу просекот на примерокот	34,5319	158,0532
КОЕФИЦИЕНТ НА КОРЕЛАЦИЈА	Јачина на линеарна врска меѓу две квантитативни променливи	-0,23265	

Б. ВТОР ДЕЛ – АНАЛИТИЧКА СТАТИСТИКА

Со помош на претходно добиените резултати од дескриптивната статистика, сега може да се пристапи кон аналитичкиот дел, во кој некои предвидувања, прогнозирања и проценувања ќе бидат тестирани врз даденото податочно множество, со цел донесување на некаков општ заклучок.

ОПРЕДЕЛУВАЊЕ ИНТЕРВАЛ НА ДОВЕРБА ЗА МАТЕМАТИЧКО ОЧЕКУВАЊЕ

За одредување на интервал на доверба за математичко очекување, постојат повеќе фактори кои влијаат на изборот на метод за спроведување на постапката. Кога станува збор за претходно анализираното обележје – изминати километри, може да се контатира дека дисперзијата на општата популација користени возила во светот, за нас е непозната. Дополнително непознат е и видот на распределба кој е присутен кај овој тип податоци. Бидејќи бројот на единки во наштиот примерок е поголем од 30 и следствено на горенаведените констатации, во овој случај за одредување на интервалот на доверба за математичкото очекување на изминатите километри, ќе се користи следната формула:

$$\left(\bar{X} - z_{\alpha/2}\frac{S}{\sqrt{n}}\,,\qquad \bar{X} + z_{\alpha/2}\frac{S}{\sqrt{n}}\right)$$

каде $\bar{X}$ е просекот во дадениот примерок, $z_{\alpha/2}$ е фактор на доверба, S е стандардната девијација на примерокот и n е бројот на единки во примерокот.

Во овој случај,

$$\bar{X} = 59{,}51181,$$

$$S = 34{,}5319,$$

$$n = 2075,$$

а со ниво на доверба од 95%,

$$z_{\alpha/2} = 1{,}95996,$$

со што се добива дека интервалот на доверба за математичкото очекување е

$$(58{,}02601\ ;\ 60{,}99760)$$

Преку овој резултат може да се заклучи дека со сигурност од 95% вистинската просечна вредност за изминати километри на користени возила е помеѓу 58,02601 илјади и 60,99760 илјади, односно ако ја повторуваме постапката на земање примерок и за секој од примероците го определуваме интервалот на доверба, тогаш 95% од интервалите пресметани на овој начин, ќе ја содржат вистинската вредност на математичкото очекување.

ТЕСТИРАЊЕ ХИПОТЕЗИ ЗА МАТЕМАТИЧКО ОЧЕКУВАЊЕ

Врз основа на претходно добиените резултати, следно може да се пристапи кон тестирање на хипотези. Хипотезите кои ќе бидат тестирани во овој случај се следните:

Нулта хипотеза:

Очекуваната вредност на изминати километри (во илјади) е еднаква на 62.

Алтернативна хипотеза:

Очекуваната вредност на изминати километри (во илјади) не е еднакво на 62.

Како што е наведено и претходно, за дадениот случај дисперзијата и распределбата на изминати километри (во илјади) на користени возила се непознати и анализата се врши врз примерок со повеќе од 30 единки. Следствено, за тестирање на хипотезите, согласно централната гранична теорема, може да се користи Z – статистика, односно следната формула:

$$Z_0 = \frac{\bar{X} - \mu_0}{S}\sqrt{n}$$

каде $\bar{X}$ е просекот во дадениот примерок, μ_0 е претпоставеното математичко очекување, S е стандардната девијација на примерокот и n е бројот на единки во примерокот.

Во овој случај,

$\bar{X}$ = 59,51181,

μ_0 = 62,

S = 34,5319,

n = 2075,

со што се добива дека вредноста на тест-статистиката е:

$$Z_0 = -3.2823$$

Како следен чекор, се проверува дали добиената вредност припаѓа во т.н. критичен домен, кој лесно можеме да го определиме преку претходно добиената вредност на факторот на доверба $z_{\alpha/2}$. Критичниот домен го има следниот облик:

$C = \left(-\infty, -z_{\alpha/2}\right) \cup \left(z_{\alpha/2}, +\infty\right)$, односно $C = (-\infty;\ -1{,}95996) \cup (1{,}95996;\ +\infty)$,

преку што јасно може да се увиди дека Z_0 припаѓа на критичниот домен.

Оттука следи дека нултата хипотеза се отфрла, односно со сигурност од 95%, очекуваната вредност на изминати километри (во илјади) не е еднакво на 62.

ТЕСТ ЗА РАСПРЕДЕЛБА

Во овој дел од статистичката анализа ќе биде спроведен тест за проверка, дали вредностите за изминатите километри (во илјади) на користените возила од даденото податочно множество се во согласност со определената претпоставка за нормална распределба. Во суштина ќе бидат тестирани следните хипотези:

Нулта хипотеза:

Вредностите за изминатите километри (во илјади) на користените возила имаат нормална распределба.

Алтернативна хипотеза:

Вредностите за изминатите километри (во илјади) на користените возила немаат нормална распределба.

За тестирање на хипотезите, се спроведува Шапиро-Вилк тест. Тест-статистиката го има следниот облик:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

каде n е бројот на единки во примерокот, $x_{(i)}$ е ниво на i-тата статистика, x_i е i-тиот податок од примерокот, $\bar{x}$ е просекот во примерокот, а a_i е коефициент кој се пресметува со следната формула:

$$(a_1, \ldots, a_n) = \frac{m^T V^{-1}}{C}$$

Повеќе информации за Шапиро-Вилк тестот, може да се најдат на следниот линк:

https://en.wikipedia.org/wiki/Shapiro%E2%80%93Wilk_test

Со дополнителни пресметки се добива p-вредност < 2.2e-16, која претставува најмало ниво на значајност кое би водело до отфрлање на нултата хипотеза со дадените податоци. Бидејќи ваквата вредност е помала од 0.05, соодветно на одбраното ниво на доверба од 95%, нултата хипотеза се отфрла, а како заклучок се издвојува дека вредностите за изминатите километри (во илјади) на користените возила немаат нормална распределба.

ТЕСТИРАЊЕ ХИПОТЕЗИ ЗА НЕЗАВИСНОСТ

Како што е наведено на почетокот од оваа семинарска работа, две од 8-те обележја на податочното множество се квалитативните обележја број на претходни сопственици и тип на гориво на користените возила. Токму овие обележја ќе бидат од суштинско значење во тестирањето на хипотези за независност.

Категориите присутни за број на претходни сопственици на користените возила во ова податочно множество се следните:

- прв сопственик
- втор сопственик
- трет сопственик

Категориите присутни за тип на гориво на користените возила во ова податочно множество се следните:

- дизел
- бензин
- бензин + CNG
- бензин + LPG

Со цел да се тестира независноста на овие две обележја, се поставуваат следните хипотези:

Нулта хипотеза:

Бројот на претходни сопственици и типот на гориво на користените возила се независни обележја.

Алтернативна хипотеза:

Бројот на претходни сопственици и типот на гориво на користените возила не се независни обележја.

За полесно доаѓање до заклучокот, во продолжение е претставена табела на контингенција за обележјата.

	Diesel	Petrol	Petrol + CNG	Petrol + LPG	Total
First Owner	470	992	109	6	1577
Second Owner	90	295	38	2	425
Third Owner	11	56	6	0	73
Total	571	1343	153	8	2075

Понатаму се преминува кон Хи-квадрат тестот за нзависност, каде се користи следната формула:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{k} \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}}$$

каде r е број на категории за бројот на претходни сопственици, k е бројот на категории за типот на гориво, n_{ij} е фреквенција на бројот на индивидуи кои припаѓаат во *i*-тата категоријата на бројот на претходни сопственици и *j*-тата категоријата на типот на гориво, $n_{i.}$ е фреквенција на бројот на индивидуи кои припаѓаат во *i*-тата категоријата на бројот на претходни сопственици, $n_{.j}$ фреквенција на бројот на индивидуи кои припаѓаат во *j*-тата категоријата на типот на гориво и n е бројот на единки во примерокот.

Со пресметка на оваа формула се добива дека вредноста на оваа тест статистика е:

$$\chi^2 = 19.492$$

Следен чекор е определување на критичниот домен, кој го има следниот облик:

$$C = \left(\chi^2_{\alpha,(r-1)(k-1)}, +\infty\right), \qquad \text{односно} \qquad C = (12{,}59159\,;\, +\infty)$$

каде α е нивото на значајност на тестот (во овој случај 0,05), r е број на категории за бројот на претходни сопственици и k е бројот на категории за типот на гориво.

Оттука јасно може да се увиди дека χ^2 припаѓа на критичниот домен, поради што се отфрла нултата хипотеза.

Заклучокот што следи е дека бројот на претходни сопственици и типот на гориво на користените возила не се независни обележја.

РЕГРЕСИОНА АНАЛИЗА

Според претходно претставениот график на расејување, беше поставена претпоставката дека помеѓу обележјата изминати километри и продажна цена на користените возила од примерокот не постои поврзаност. За поддржување на таа претпоставка, дополнително претходно е определен и коефициентот на корелација на двете обележја со помош на следната формула:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\sum_{i=1}^{n}(x_i - \bar{x})^2)(\sum_{i=1}^{n}(y_i - \bar{y})^2)}}$$

каде n е бројот на единки во примерокот, x_i е вредност за изминати километри на i-тата единка, $\bar{x}$ е просечната вредност на изминати километри (во илјади) во примерокот, y_i е вредност за продажна цена на i-тата единка, $\bar{y}$ е просечната вредност на продажна цена (во илјади) во примерокот.

Со добиениот резултат $r = -0{,}23265$, бидејќи вредноста на коефициентот на корелација за овој примерок е поблиску до 0 отколку -1, може да се заклучи дека станува збор за послаба негативна линеарна поврзаност на изминатите километри и продажната цена на користените возила од примерокот.

За дополнителна анализа на моделот на линеарна регресија, во продолжение се определени непознатите параметри β_0 (пресек со y-оска) и β_1 (коефициент на правец) на моделот, со цел да се добие права која "најдобро одговара" на множеството набљудувани вредности.

За пресметка на параметрите се користат следните формули:

$$\beta_0 = \bar{y} - \frac{SS_{XY}}{SS_X}\bar{x} \qquad\qquad \beta_1 = \frac{SS_{XY}}{SS_X}$$

каде $\bar{x}$ е просечната вредност на изминати километри (во илјади) во примерокот, $\bar{y}$ е просечната вредност на продажна цена (во илјади) во примерокот и дополнително:

$$SS_X = \sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad SS_{XY} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Со пресметка на горенаведените податоци се добива равенката на правата од облик:

$$y = \beta_0 + \beta_1 x\,, \qquad \text{односно} \qquad y = 445{,}469 - 1{,}065x$$