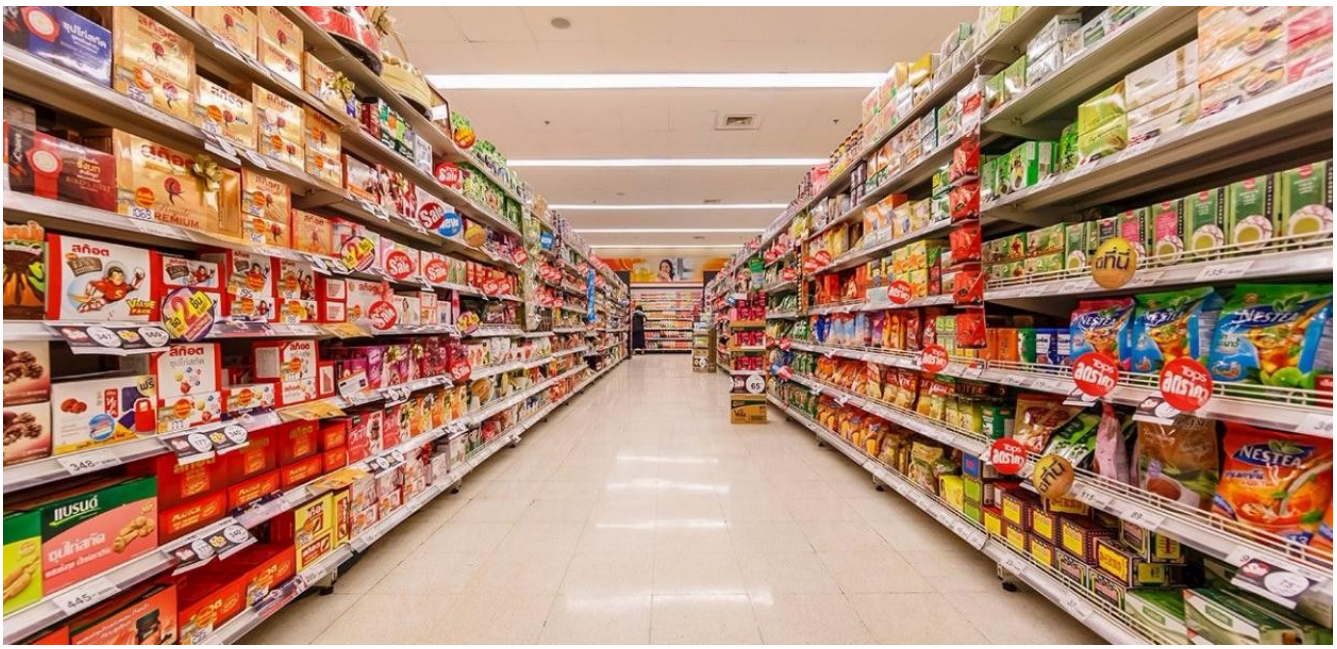
Универзитет „Св. Кирил и Методиј“ во Скопје
ФАКУЛТЕТ ЗА ИНФОРМАТИЧКИ НАУКИ И КОМПЈУТЕРСКО ИНЖЕНЕРСТВО

СЕМИНАРСКА РАБОТА ПО БИЗНИС СТАТИСТИКА



Тема: Статистичка анализа на продажбата на супермаркетите во САД и нејзината зависност од областа на продавниците и бројот на достапните артикли

Ментор:
д-р Методија Јанчески

Изработил:
Тодор Јовановски

јули, 2022 година
Скопје

Вовед

Супермаркет е продавница за самопослужување која што нуди широк спектар на храна, пијалоци и производи за домаќинството притоа организирани во разни целини и секции. Овој вид на продавница е значително голем и има широка селекција од горенаведените намирници, но е помала и поограничена од хипермаркетите или пазарите на големо. Меѓутоа во секојдневната употреба во САД, „продавница за храна (grocery store)“ е синоним за супермаркет и не често се однесува на други видови продавници кои нудат разни намирници.

Во оваа семинарска работа е извршена статистичка анализа на продажбата во 896 различни супермаркет компании во САД и нејзниата зависност од големината на продавницата и дневниот број на клиенти. Предмет на анализата се следните обележја:

- ID на продавницата: (индекс) ID на одредената продавница
- Површина: физичка површина на продавницата изразена во јарди квадратни
- Достапни артикли: број на различни производи и добра кои ги се достапни во соодветната продавница
- Дневен број на клиенти: просек на дневни клиенти кои го посетувале супермаркетот во текот на месецот
- Продажба: приход на продавницата изразен во американски долари (US $)

Извор на податочното множество: Supermarket store branches sales analysis - kaggle
Линк: https://www.kaggle.com/datasets/surajjha101/stores-area-and-sales-data?resource=download

Првиот дел од оваа анализа се состои од табеларно и графички претставување на податочните множества користејќи табели на честоти, хистограми, графици на расејување и дескриптивни статистики. За исиот во главно се користени бројот на артиклите достапни во маркетите како квантитативно дискретно обележје како и измерената површина на продавницата во јадри квадратни како нумеричо обележје од непреброив карактер. Во вториот дел од анализата одредуваме интервали на доверба, поставуваме хипотези и правиме тестови за распределба и крајно, извршена е регресиона анализа на избраните обележја.

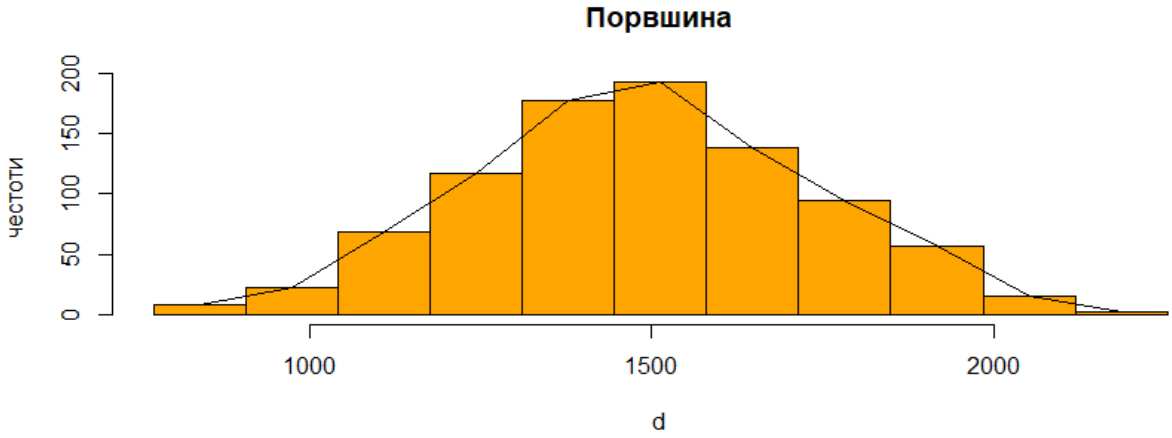
ПРВ ДЕЛ

Табела на честоти и графички прикази

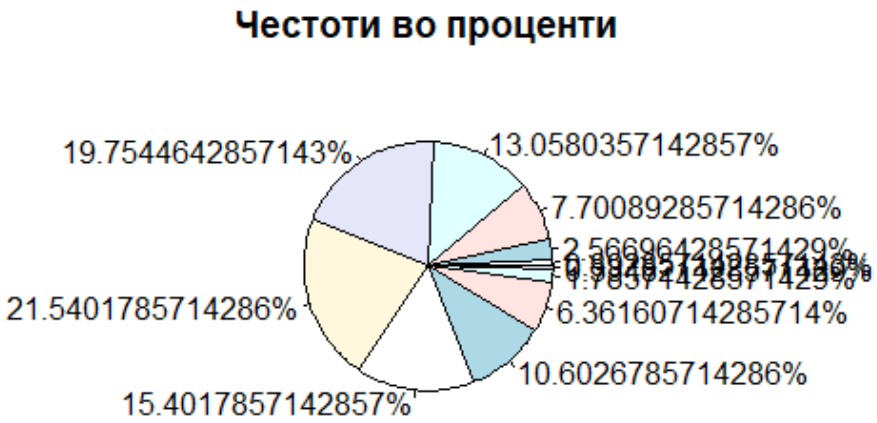
Најпрво ќе разгледаме табела на честоти за податоците изведени при мерењето на површината на избраните супермаркети. Распределбата е извршена во 11 интервали секој со по 135 единки.

Површина на продавницата изразена во јарди квадратни						
Интервал	Средна точка	Честота	Релативна честота	Релативна честота (%)	Кумулативна честота	Релативна кумулативна честота (%)
[770, 905)	837.5	8	0.01	0.89	8	0.89
[905, 1040)	972.5	23	0.03	2.57	31	3.46
[1040, 1175)	1107.5	69	0.08	7.70	100	11.16
[1175, 1310)	1242.5	117	0.13	13.06	217	24.22
[1310, 1445)	1377.5	177	0.20	19.75	394	43.97
[1445, 1580)	1512.5	193	0.22	21.54	587	65.51
[1580, 1715)	1647.5	138	0.15	15.40	725	80.92
[1715, 1850)	1782.5	95	0.11	10.60	820	91.52
[1850, 1985)	1917.5	57	0.06	6.36	877	97.88
[1985, 2120)	2052.5	16	0.02	1.79	893	99.67
[2120, 2255)	2187.5	3	0.00	0.33	896	100

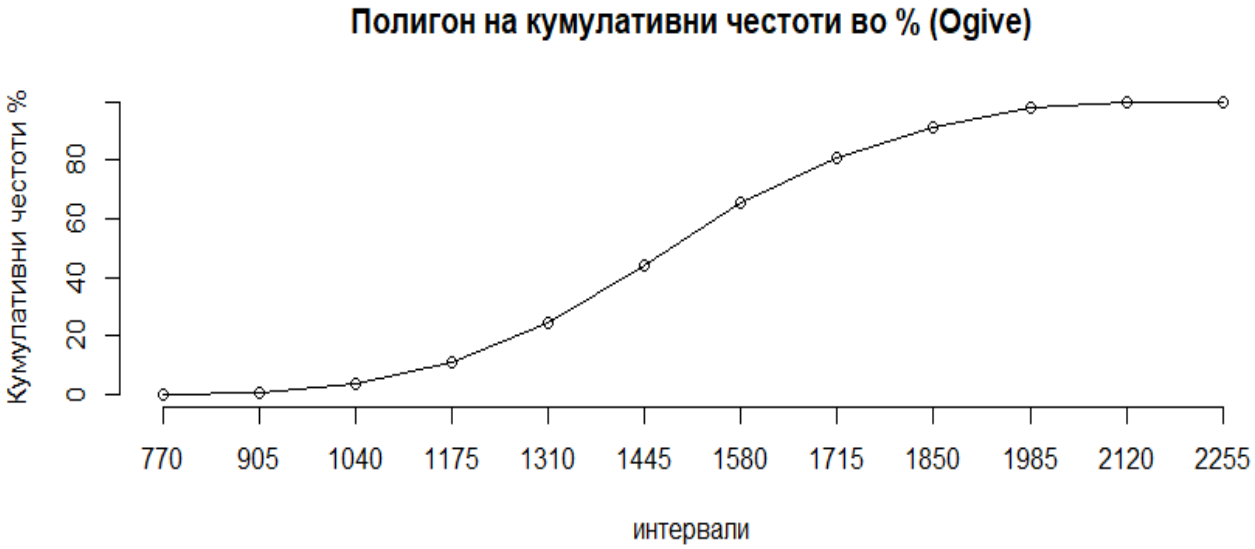
Во продолжение следи хистограм изведен од податоците за површината на продавниците. Можеме да забележиме дека столбчињата одлично претставуваат нормална распределба. Исто така даден е и полигон на фреквенции во кој се поврзуваат средните точки на горните хоризонтални страни на секој од правоаголниците со искршена линија.



Следната слика претставува пита изработена од честотите во процети за површината на продавницата изразена во јарди квадратни.



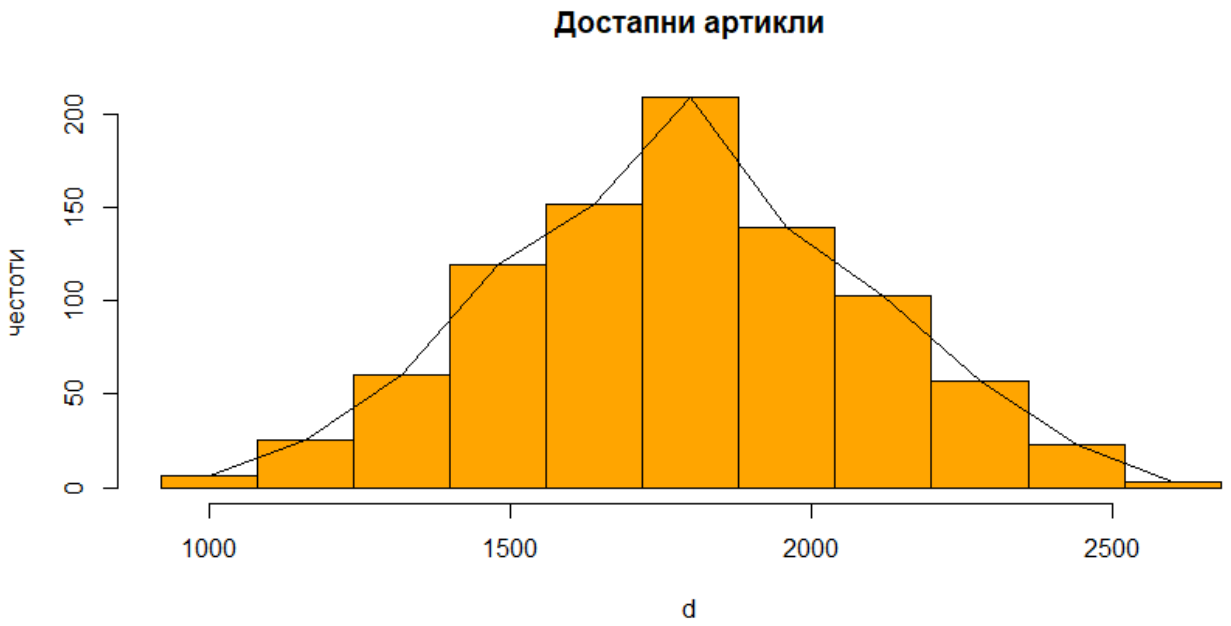
Во прилог имаме и полигон на кумулативни честоти во % (Ogive) кој ги поврзува границите на интервалите (не средините) со соодветната кумулативна честота изразена во проценти.



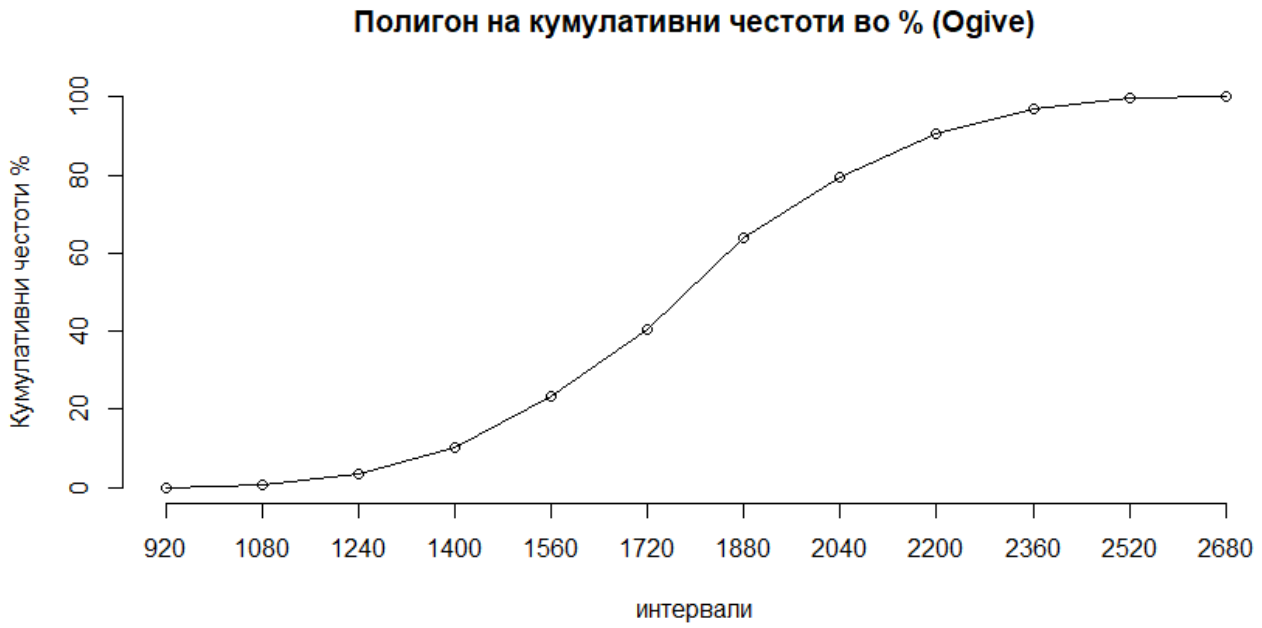
Следно обележје врз кое градиме табела на честоти е бројот на достапни артикли во секоја од продавниците. Составена е од 11 интервали во кои секој содржи 160 единици.

Достапни артикли						
Интервал	Средна точка	Честота	Релативна честота	Релативна честота (%)	Кумулативна честота	Релативна кумулативна честота (%)
[920, 1080)	1000	6	0.01	0.67	6	0.67
[1080, 1240)	1160	25	0.03	2.79	31	3.46
[1240, 1400)	1320	60	0.07	6.70	91	10.16
[1400, 1560)	1480	119	0.13	13.28	210	23.44
[1560, 1720)	1640	152	0.27	16.96	362	40.40
[1720, 1880)	1800	209	0.23	23.33	571	63.73
[1880, 2040)	1960	139	0.16	15.51	710	79.24
[2040, 2200)	2120	103	0.11	11.50	813	90.74
[2200, 2360)	2280	57	0.06	6.36	870	97.10
[2360, 2520)	2440	23	0.03	2.57	893	99.67
[2520, 2680)	2600	3	0.00	0.33	896	100

Хистограмот за достапните производи има многу сличен изглед како и претходното. Повторно, одредени се средните точки на секој од интервалите и истите се поврзани со соодветните честоти.

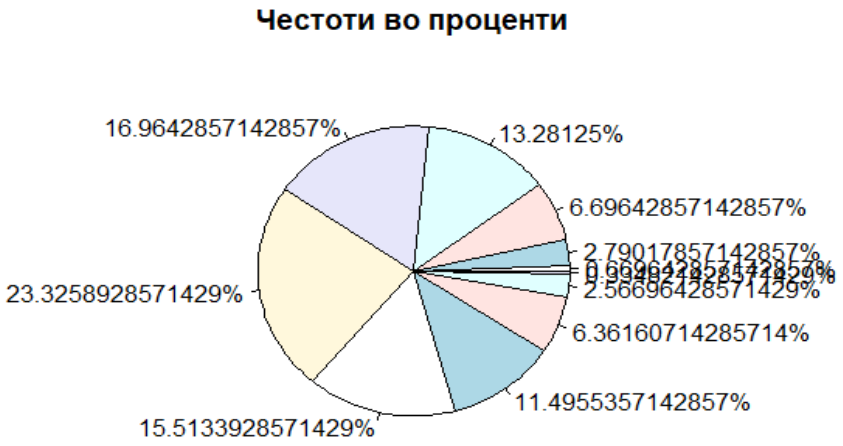


Воочливо е дека овие две обележја имаат силна поврзаност. И за ова подмножество е изработен Ogive полигонот кој ги поврзува границите на интервалите со кумулативните честоти во проценти.



Граници на интервалите (*Не средини*)

Слично изгледа и питата за честоти во проценти за бројот на артикли достапни во секој од супермаркетите.



Стебло лист дијаграм

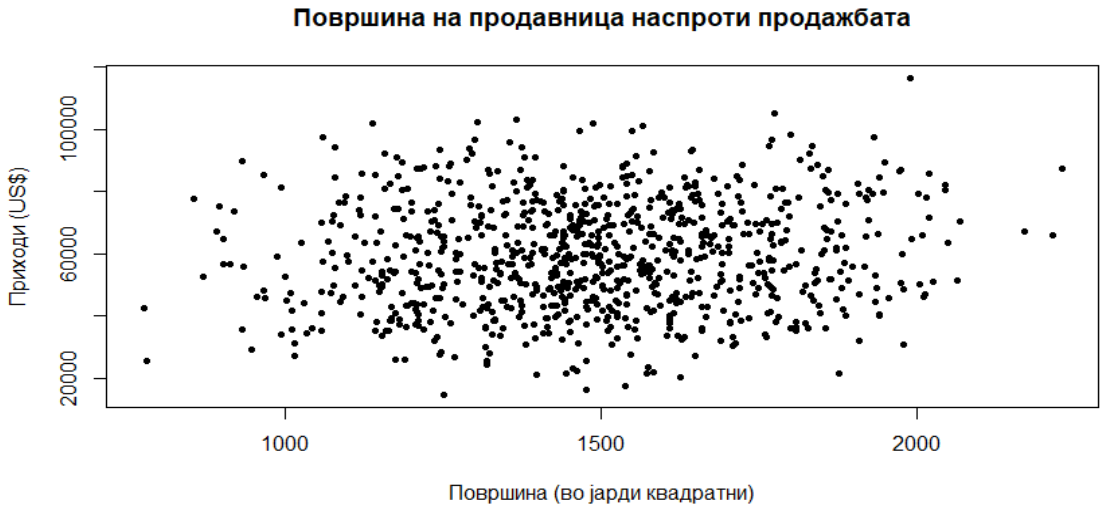
За составување на стебло лист дијаграмот во продолжение, користено е множество на податоци за најефтините електрични коли достапни на пазарот. До истото може да се пристапи преку следниов линк: https://www.kaggle.com/datasets/kkhandekar/cheapest-electric-cars.
За конкретниот стебло лист дијаграм употребено е подмножеството на податоци кои се добиени со мерење на времето потребно за секој од избраните автомобили да постигне брзина од 0 – 100 kmh (километри на час).

Децималната запирка се наоѓа кај |

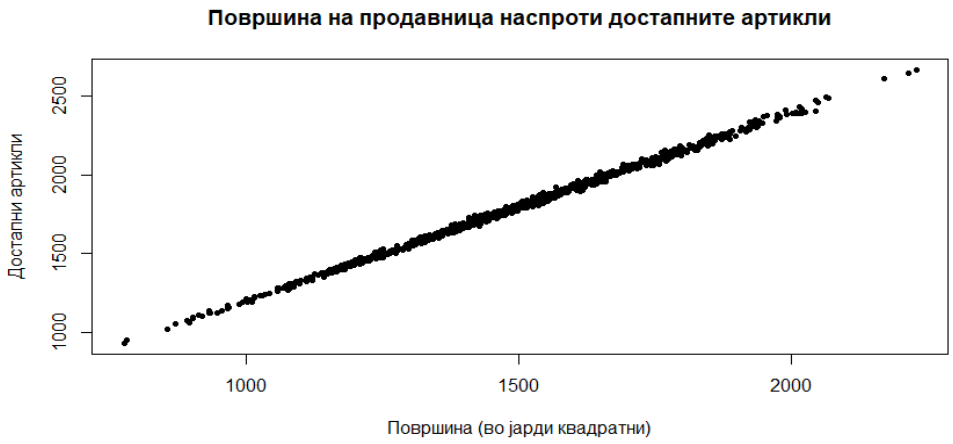
2 | 15888
3 | 0245578
4 | 0005566789
5 | 00111567779
6 | 00233568889
7 | 000333333555556899999
8 | 11235578
9 | 000000055677899
10 | 0000
11 | 44699
12 | 3367
14 | 0
22 | 4

График на расејување

Од графикот на расејување добиен за соодносот на површината на супермаркетите и приходите на истите, можеме да заклучиме дека не постои поврзаност помеѓу избраните обележја. Приходите на компанијата не зависат во ниедна мера од површината на продавницата.



Од друга страна, при набљудување на графикот на расејување кој ги прикажува достапните артикли во секој од супермаркетите и нивната површина, очигледна е линерната врска помеѓу двете обележја. Забележуваме сила позитивна поврзаност. Со зголемување на површина на супермаркетот паралелно се ослободува место за повеќе производи.



Дескриптивни статистики

Малку погоре видовме како од необработените податоци се добиваат табели на фреквенција, хистограми и други визуелни прикази. Сепак, поцелосно разбирање на податоците може да се постигне ако за податоците се пресметаат одредени бројни карактеристики кои се вредности на таканаречените дескриптивни статистики. Во следната табела се пресметани и мерките на централна тенденција (мода, медијана, просек) и варирање (опсег, интерквартален распон, дисперзија, стандардна девијација, коефициент на варирање).

Централна тенденција		
	Површина	Артикли
Мода	1458	1858
Медијана	1477	1774
Просек	1485.41	1782.04
Варирање		
Опсег	1454	1735
Прв квартал (Q1)	1317	1576
Втор квартал (Q2)	1477	1774
Трет квартал (Q3)	1654	1983
Интерквартален распон	337	407
Дисперзија	62618.56	89923.25
Стандардна девијација	250.24	299.87
Коефициент на варирање	16.85%	16.83%

Коефициент на корелација	0.9988908

ВТОР ДЕЛ

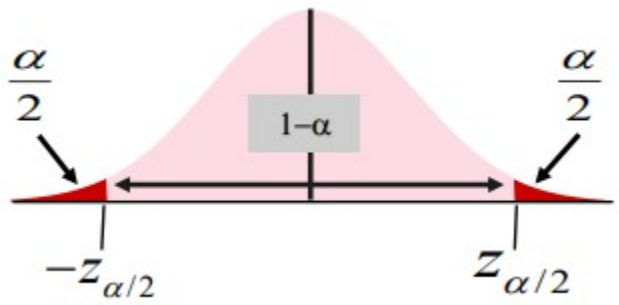
Интервал на доверба

Вообичаен начин да се воведе точност на оценувачот е преку интервално оценување. Се формира интервал (базиран на примерокот) кој со одредена веројатност ја содржи точната, но непозната вредност на параметарот. Овој интервал е познат како интервал на доверба (ИД).

Интервал на доверба

$$\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

Ниво на доверба	*Ниво на доверба,* $1-\alpha$	$Z_{\alpha/2}$ *вредност*
80%	.80	1.28
90%	.90	1.65
95%	.95	1.96
98%	.98	2.33
99%	.99	2.58
99.8%	.998	3.08
99.9%	.999	3.27



$$e = ME = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

error

Просечниот број на дневни потрошувачи на примерок од 896 супермаркети во САД изнесува 786.35. Под предпоставка дека бројот на дневни клиенти има приближно нормална распределба со стандардна девијација од 265.39 клиенти, да се определи 95% интервал на доверба за очекуваниот број на потрошувачи во текот на денот.

$\overline{X}$ = 786.35; n = 896; σ = 265.69; $z_{a/2}$ = 1.96; e = 17.38;

Значи, 95% интервал на доверба за μ е (768.98, 803.73)

Тестирање хипотези

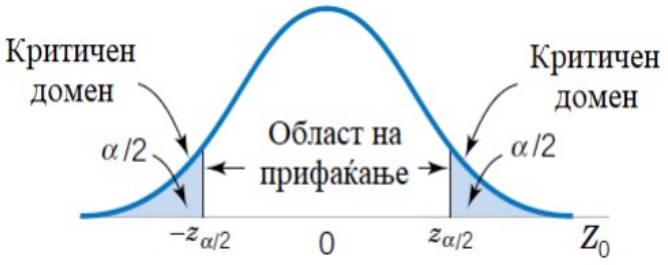
Познато е дека очекуваниот број на дневни клиенти во продавниците во САД изнесува 786.35 со стандардна девијација од 265.69 потрошувачи. При случаен избор на 60 супермаркети во Калифорнија утврдено е дека просечниот број на клиентите кои ја посетуваат продавницата во текот на денот изнесува 820.98. Дали при ниво на значајност $\alpha = 0.05$, можеме да заклучиме дека очекуваниот број на дневни клиенти во маркетите во Калифорнија отстапува од стандардите? Се претпоставува дека просечниот број на клиентите има приближно нормална распределба.

$H_0 \rightarrow \mu_0 = 786.35;$
$H_a \rightarrow \mu_0 \neq 786.35;$

$\overline{X} = 820.98;$
$n = 896;$
$\sigma = 265.69$
$z_{a/2} = 1.96;$

$z_0 = 3.906$
$c = (-\infty, -1.96) \cup (1.96, +\infty)$
$z_0 \in c \rightarrow$ ја отфрламе H_0 како грешна

Заклучуваме дека очекуваниот број на клиенти кои ги посетуваат маркетите во Калифорнија во текот на денот, отстапува од стандардите.

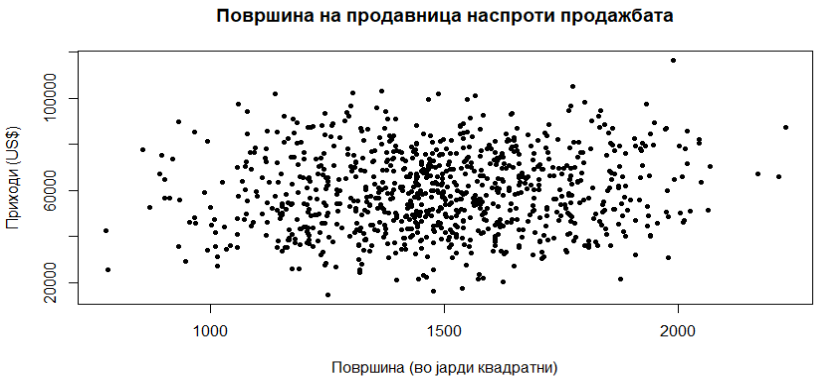


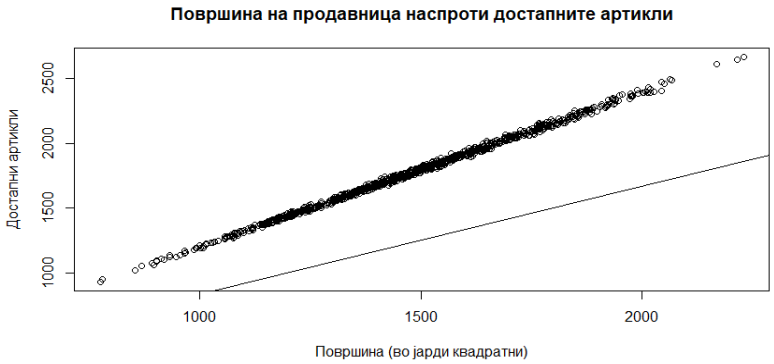
	H_0 е точна	H_0 не е точна
Тестот ја отфрла H_0	**Грешка од I тип**	Нема грешка
Тестот не ја отфрла H_0	Нема грешка	**Грешка од II тип**

Регресиона анализа

Честопати во пракса се сретнуваме со мерења на повеќе обележја на исти единки од популацијата. Вакви мерења најчесто се прават за да се провери како промените на одредени обележја (независни променливи) влијаат на промените на обележјето од интерес (зависна променлива). Регресијата е една од најчесто користените статистички методи.

Да ги искористиме горенаведените графици на расејување.





Да ја испитаме врската помеѓу обележјата и во двата случаи користејќи го коефициентот на корелација. Коефициентот на корелацијата го мери правецот и јачината на линеарната врска помеѓу две квантитативни променливи. Во првиот случај забележуваме дека речиси нема никаква поврзаност помеѓу двете обележја, па природно, пресметаниот коефицинет на корелација за истиот е поблизу до нулата – 0.098. Од друга страна, коефициентот на корелација во вториот случај изнесува 0.99889, што укажува на многу силна позитивна линеарна врска помеѓу обележјата.

Изработил: Тодор Јовановски (213133)