Семинарска работа по предметот Бизнис статистика

Изработено од Маја Вуевска индекс: 203007

Од следниот извор: Find Open Datasets and Machine Learning Projects | Kaggle го избрав следното податочно множество Latest Worldwide Covid-19 Vaccine Data | Kaggle
Податочното множество кое ќе се разгледува се работи за статистиките на најновите светски Ковид-19 вакцинации. Во него има категориски податоци кои ги претставуваат државите во светот, конкретно во ова податочно множество се 182 и има нумерички податоци со кои се претставени неколку обележја на државите, а тоа се: број на вакцини кои се администрирани на секои 100 луѓе, вкупен број на дози кои се администрирани, број на луѓе вакцинирани со една доза (изразен во проценти) и број на луѓе вакцинирани со две дози (изразен во проценти).

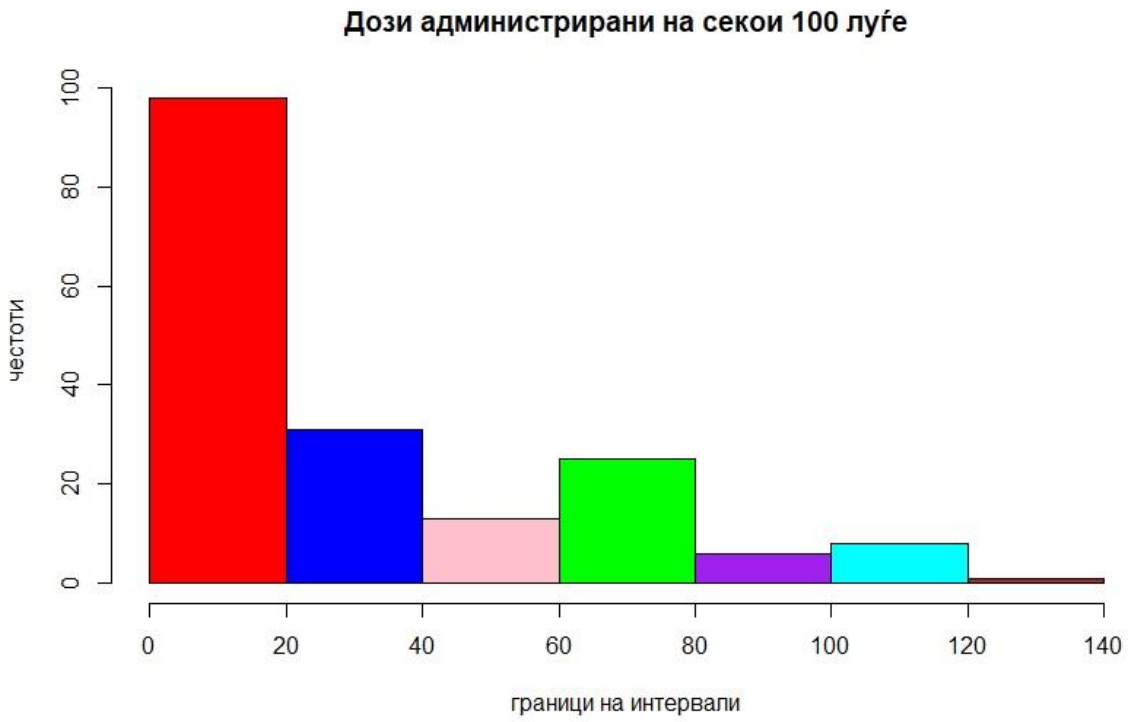
Дел А

1. Организација и претставување на податоци

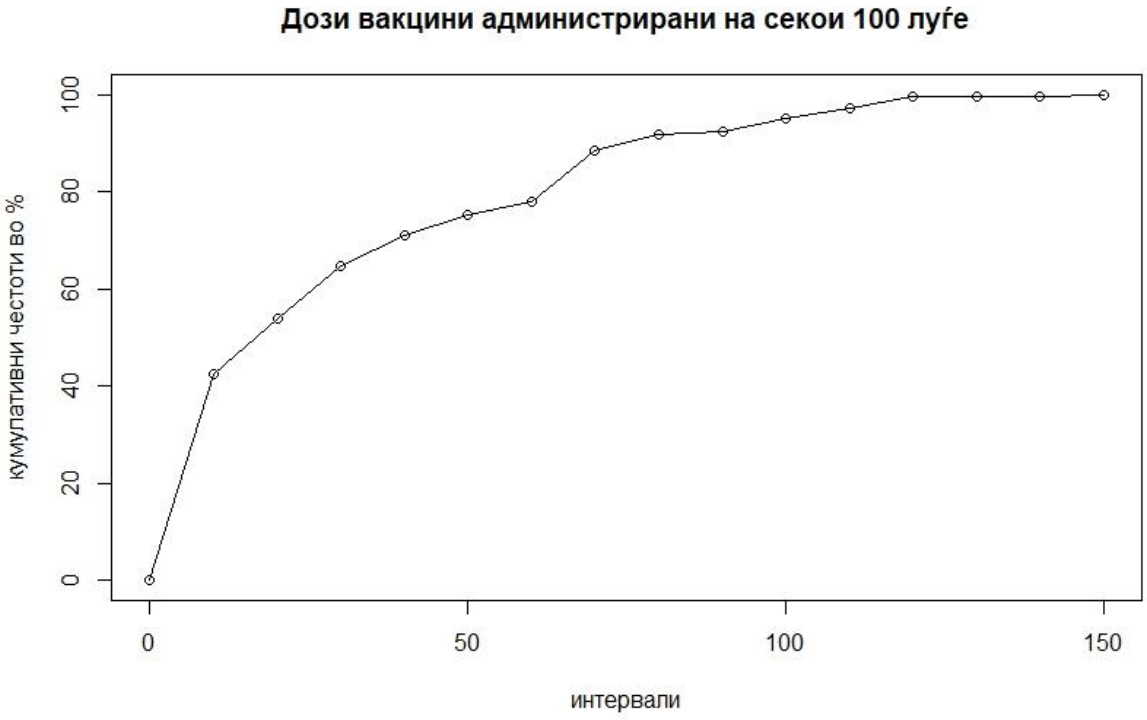
Го бирам првото обележје, односно "број на вакцини администрирани на секои 100 луѓе". Бидејќи во податочното множество има 182 држави, тоа значи дека обемот на примерокот е 182, односно n = 182. Најголемиот податок е 140, најмалиот е 0, а бидејќи рангот на податоците е разликата меѓу најголемиот и најмалиот податок, рангот е 140 – 0 = 140. Бројот на интервали (k) во кои треба да се групираат податоците е приближно квадратен корен од обемот на примерокот, односно $\sqrt{182}$ кое е еднакво на 13,49 а тоа потоа ќе се заокружи на 14. Исто така бројот на интервали најчесто треба да е број меѓу 5 и 15, а нашиот број припаѓа во тој интервал. Ширината на интервалите (w) се пресметува со формулата w ≥ R / k, односно во нашиот пример е 140 / 14 = 10. Горната граница од интервалот се зема како поголема од најголемиот податок со цел да може да се опфатат и вредностите кои се наоѓаат во последниот интрервал, односно се проширува до 150. Во R се добиваат следните резултати:

Интервали	Средни точки	Честоти	Релативни честоти	Кумулативни честоти	Релативни кумулативни честоти	Релативни честоти во %	Релативни кумулативни честоти во %
[0,10)	5	77	0.42	77	0.42	42.0	42
[10,20)	15	21	0.12	98	0.54	12.0	54
[20,30)	25	20	0.11	118	0.65	11.0	65
[30,40)	35	11	0.06	129	0.71	6.0	71
[40,50)	45	8	0.04	137	0.75	4.0	75
[50,60)	55	5	0.03	142	0.78	3.0	78
[60,70)	65	19	0.10	161	0.88	10.0	88
[70,80)	75	6	0.03	167	0.92	3.0	92
[80,90)	85	1	0.01	168	0.92	1.0	92
[90,100)	95	5	0.03	173	0.95	3.0	95
[100,110)	105	4	0.02	177	0.97	2.0	97
[110,120)	115	4	0.02	181	0.99	2.0	99
[120,130)	125	0	0.00	181	0.99	0.0	99
[130,140)	135	0	0.00	181	0.99	0.0	99
[140,150)	145	1	0.01	182	1.00	1.0	100
Вкупно		182	1			100.0	

Хистограм за бројот на дози администрирани на секои 100 луѓе е следниот:



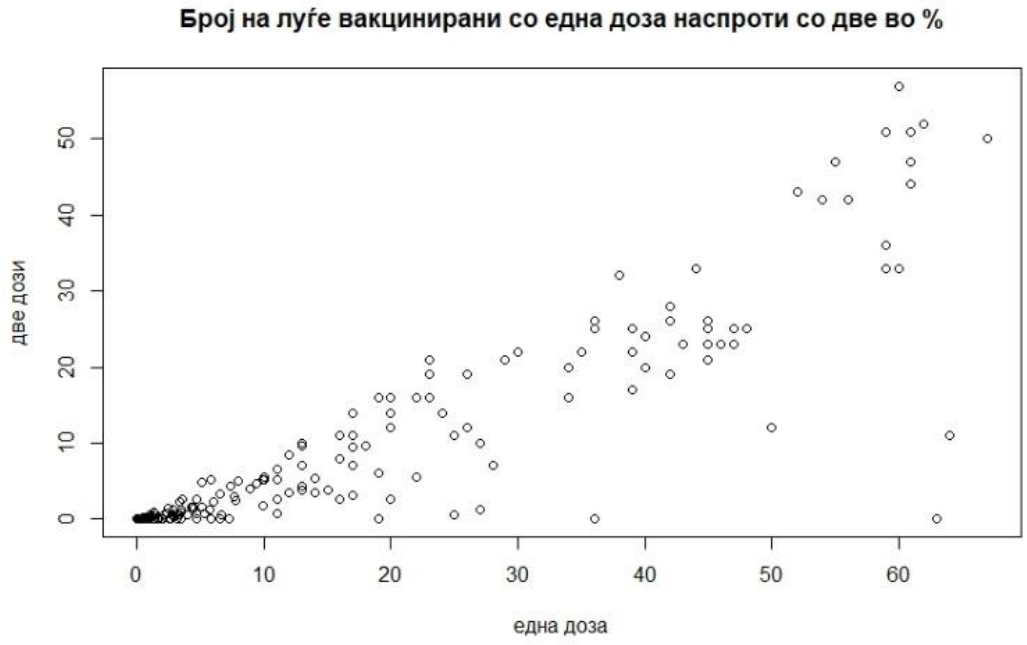
Полигон на кумулативни честоти во проценти за истото обележје:



2. Стебло-лист дијаграм од обележјето “процент на жители од популацијата вакцинирани со една доза“, каде стеблото претставува првата цифра од бројот, а втората цифра е неговиот лист и таа се наоѓа десно од “|“.

```
0 | 000000000011111111111111111111122222222333333333333333334444
0 | 5555555666677777788899
1 | 00001111223333344
1 | 5666777778999
2 | 00002233334
2 | 55667789
3 | 044
3 | 56668999
4 | 00222234
4 | 55556778
5 | 024
5 | 56999
6 | 00111234
6 | 7
```

3. График на расејување на обележјата процент на лица од популацијата вакцинирани само со една доза наспроти со две дози. Точките се натрупиваат околу бројот 1, што значи дека има силна линеарна поврзаност меѓу овие две обележја, односно тие се поврзани.



4. Просек на обележјето “бројот на вакцини администрирани на секои 100 луѓе“ е 29. Медијана за истото обележје е 16.
   Бидејќи во R не е дефинирана функција за пресметување на мода, формирав своја функција. Тоа е следната:
   my_mode <- function(x) {
   + unique_x <- unique(x)
   + tabulate_x <- tabulate(match(x, unique_x))
   + unique_x[tabulate_x == max(tabulate_x)]
   + }

И потоа ја повикав за обележјето кое го разгледував и за просекот и за медијаната. И добив две моди: 1.7 и 62.

5. Квартили за истото обележје од четвртата точка се:

   $Q_1 = P_{25} = 3.1$

   $Q_2 = P_{50} = 16$

   $Q_3 = P_{75} = 48.8$

Ранг или опсег, односно разликата меѓу најголемиот и најмалиот податок е 140 – 0 = 140

Интерквартални распон, односно разликата меѓу третиот и првиот квартил е $Q_3 - Q_1 = 48.8 – 3.1 = 45.7$, а ова R го заокружува на 46.

6. Дисперзија на истото обележје е 1040, а стандардна девијација е 32
7. Коефициент на корелација меѓу процентот на лица вакцинирани со една доза и со две е 0.89 значи има силна позитивна линеарна поврзаност

Дел Б

1. Обележје кое ќе се разгледува во овој дел е процентот на луѓе од популацијата вакцинирани со две дози односно целосно вакцинирани, и за него параметарот кој ќе го разгледуваме е математичкото очекување.
   За да се одреди интервалот на доверба прво треба да се одреди обемот на примерокот. Во нашиот случај тој е 182 што е поголемо од 30 значи примерокот е голем. Бидејќи разгледуваме конкретно обележје од примерокот, позната е дисперзијата односно и стандардната девијација на обележјето. Со помош на z-тест ние го одредуваме интервалот на доверба за математичкото очекување.
   Формулата за тоа е следната: $\bar{X} \pm z_{\frac{a}{2}} \frac{\sigma}{\sqrt{n}}$
   Стандардната девијација на обележјето е 13. Прво ја пресметуваме стандардната грешка $\frac{\sigma}{\sqrt{n}}$, а потоа маргината на грешка се пресметува како производ од стандардната грешка и z нормирано за 95% ниво на доверба. Со математичко очекување кое е 9.7, интервалот на доверба се добива дека е (7.8, 11.6).
2. Според претходната точка, познато е дека очекуваниот процент на луѓе во популацијата на секоја држава вакцинирани со две дози е 9.7 со стандардна девијација 13. На случаен начин се бираат 124 држави и се пресметува просекот на истото обележје од тој примерок и за него се добива 10.04. Со ниво на значајност од 0.05 треба да се провери дали очекуваниот процент на луѓе вакцинирани со две дози отстапува од стандардите. Се поставува нулта хипотеза H_0: $\mu_0 = 9.7$ и се тестира дали е точна. Се спроведува двостран тест односно алтернативната хиппотеза е H_a: $\mu_0 \neq 9.7$. Се користи централна гранична теорема односно $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$. За z се добива дека е 0.35. Критичниот домен е следниот (-1.96, +1.96). Па бидејќи z тестот припаѓа во критичниот домен, нултата хипотеза се отфрла. Односно со ниво на значајност од $\alpha = 0.05$, очекуваниот процент на луѓе вакцинираи со две дози е различен од 9.7 односно отстапува од стандардите.
3. Тест за распределба. Ќе тестираме дали обележјето X – процент на луѓе од популацијата вакцинирани со две дози, има нормална распределба. Се поставуваат

следните хипотези:

H_0: X има нормална распределба

H_a: X нема нормална распределба

Се користи Shapiro-Wilk тест за нормална распределба. Со ниво на значајност од 0.05, за p-вредноста се добива дека е 4.27×10^{-13} . Па бидејќи p-вредноста е помала од алфа, нултата хипотеза се отфрла. Односно обележјето X нема нормална распределба.

4. Треба да се провери дали две обележја се зависни. Ги земаме двете последни обележја, процент на луѓе вакцинирани со една доза наспроти со две. Каде
   X – процент на луѓе вакцинирани со една доза од популацијата на една држава, а
   Y – процент на луѓе вакцинирани со две дози од популацијата на една држава. И ги поставуваме хипотезите:

   H_0: X и Y се независни обележја

   H_a: X и Y се зависни обележја

   Се користи Пирсонов Хи-квадрат тест. За Хи-квадрат се добива вредност 7.427, за ниво на значајност алфа 0.05 се добива p-вредност 6×10^{-13}, и бидејќи p-вредноста е помала од алфа, нултата хипотеза се отфрла односно двете обележја се зависни.
5. Правиме регресиона анализа на обележјата процент на луѓе вакцинирани со една доза, наспроти со две. И гледаме дали има линеарна поврзаност меѓу нив. Правата на регресија е следна: y = 5.3 + 1.3x. Коефициентот на детерминираност е 0.89192. Тој е број меѓу 0 и 1, што значи дека има послаба линеарна поврзаност меѓу x и y. Зависноста меѓу обележјата е прикажана и визуелно.