Семинарска работа по предметот Бизнис статистика

Податочно множество

Податочно множество: https://www.kaggle.com/datasets/burnoutminer/heights-and-weights-dataset

Ова податочно множество содржи мерења за висината и тежината на 25000 различни 18 годишни лица.

Обем на примерокот: 25000 единки

Обележја:

1. Индекс: Уникатна бројка на поединецот (Категориски номинален податок)
2. Висина (инчи): Измерена висина на поединецот (Квантитативен непрекинат податок)
3. Тежина (либри): Измерена висина на поединецот (Квантитативен непрекинат податок)

Резултати

Прв дел

Задача 1

Табели на честоти

Дел од табелата за висина

	Interval	Freq	Midpoint	Rel_freq	Cum_freq
1	(60.28,60.37]	0	60.325	0.00000	0
2	(60.37,60.47]	0	60.420	0.00000	0
3	(60.47,60.56]	0	60.515	0.00000	0
4	(60.56,60.65]	1	60.605	0.00004	1
5	(60.65,60.75]	0	60.700	0.00000	1
6	(60.75,60.84]	1	60.795	0.00004	2
7	(60.84,60.94]	2	60.890	0.00008	4
8	(60.94,61.03]	0	60.985	0.00000	4
9	(61.03,61.13]	0	61.080	0.00000	4
10	(61.13,61.22]	0	61.175	0.00000	4
11	(61.22,61.31]	1	61.265	0.00004	5
12	(61.31,61.41]	1	61.360	0.00004	6
13	(61.41,61.5]	0	61.455	0.00000	6
14	(61.5,61.6]	2	61.550	0.00008	8
15	(61.6,61.69]	0	61.645	0.00000	8
16	(61.69,61.78]	0	61.735	0.00000	8
17	(61.78,61.88]	1	61.830	0.00004	9
18	(61.88,61.97]	4	61.925	0.00016	13
19	(61.97,62.07]	2	62.020	0.00008	15
20	(62.07,62.16]	0	62.115	0.00000	15
21	(62.16,62.25]	1	62.205	0.00004	16
22	(62.25,62.35]	2	62.300	0.00008	18
23	(62.35,62.44]	9	62.395	0.00036	27
24	(62.44,62.54]	15	62.490	0.00060	42
25	(62.54,62.63]	10	62.585	0.00040	52
26	(62.63,62.72]	16	62.675	0.00064	68
27	(62.72,62.82]	9	62.770	0.00036	77
28	(62.82,62.91]	12	62.865	0.00048	89
29	(62.91,63.01]	15	62.960	0.00060	104
30	(63.01,63.1]	14	63.055	0.00056	118
31	(63.1,63.19]	14	63.145	0.00056	132
32	(63.19,63.29]	20	63.240	0.00080	152
33	(63.29,63.38]	23	63.335	0.00092	175
34	(63.38,63.48]	34	63.430	0.00136	209
35	(63.48,63.57]	31	63.525	0.00124	240
36	(63.57,63.67]	39	63.620	0.00156	279
37	(63.67,63.76]	43	63.715	0.00172	322
38	(63.76,63.85]	50	63.805	0.00200	372
39	(63.85,63.95]	51	63.900	0.00204	423
40	(63.95,64.04]	44	63.995	0.00176	467
41	(64.04,64.14]	69	64.090	0.00276	536
42	(64.14,64.23]	72	64.185	0.00288	608
43	(64.23,64.32]	84	64.275	0.00336	692
44	(64.32,64.42]	86	64.370	0.00344	778

Дел од табелата за тежина

	Interval	Freq	Midpoint	Rel_freq	Cum_freq
1	(78.01,78.6]	1	78.305	0.00004	1
2	(78.6,79.19]	0	78.895	0.00000	1
3	(79.19,79.78]	0	79.485	0.00000	1
4	(79.78,80.37]	0	80.075	0.00000	1
5	(80.37,80.95]	0	80.660	0.00000	1
6	(80.95,81.54]	0	81.245	0.00000	1
7	(81.54,82.13]	0	81.835	0.00000	1
8	(82.13,82.72]	1	82.425	0.00004	2
9	(82.72,83.3]	1	83.010	0.00004	3
10	(83.3,83.89]	1	83.595	0.00004	4
11	(83.89,84.48]	2	84.185	0.00008	6
12	(84.48,85.07]	1	84.775	0.00004	7
13	(85.07,85.65]	1	85.360	0.00004	8
14	(85.65,86.24]	1	85.945	0.00004	9
15	(86.24,86.83]	2	86.535	0.00008	11
16	(86.83,87.42]	2	87.125	0.00008	13
17	(87.42,88]	1	87.710	0.00004	14
18	(88,88.59]	2	88.295	0.00008	16
19	(88.59,89.18]	1	88.885	0.00004	17
20	(89.18,89.77]	2	89.475	0.00008	19
21	(89.77,90.35]	7	90.060	0.00028	26
22	(90.35,90.94]	4	90.645	0.00016	30
23	(90.94,91.53]	4	91.235	0.00016	34
24	(91.53,92.12]	5	91.825	0.00020	39
25	(92.12,92.71]	7	92.415	0.00028	46
26	(92.71,93.29]	9	93.000	0.00036	55
27	(93.29,93.88]	8	93.585	0.00032	63
28	(93.88,94.47]	13	94.175	0.00052	76
29	(94.47,95.06]	11	94.765	0.00044	87
30	(95.06,95.64]	8	95.350	0.00032	95
31	(95.64,96.23]	19	95.935	0.00076	114
32	(96.23,96.82]	17	96.525	0.00068	131
33	(96.82,97.41]	17	97.115	0.00068	148
34	(97.41,97.99]	23	97.700	0.00092	171
35	(97.99,98.58]	23	98.285	0.00092	194
36	(98.58,99.17]	22	98.875	0.00088	216
37	(99.17,99.76]	21	99.465	0.00084	237
38	(99.76,100.3]	47	100.030	0.00188	284
39	(100.3,100.9]	48	100.600	0.00192	332
40	(100.9,101.5]	45	101.200	0.00180	377
41	(101.5,102.1]	51	101.800	0.00204	428
42	(102.1,102.7]	54	102.400	0.00216	482
43	(102.7,103.3]	52	103.000	0.00208	534

Средни точки

Средните точки на интервалите се добиени со собирање на долната и горната граница на интервалот и делење со 2.

Релативни фреквенции

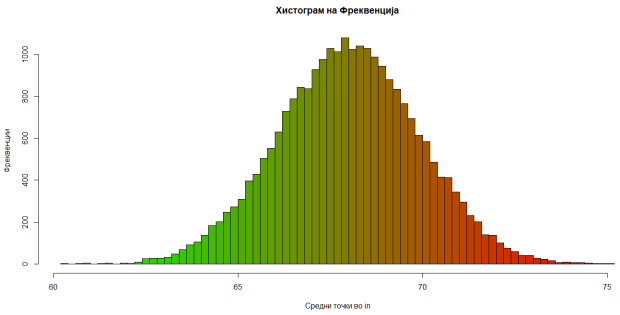
Релативните фреквенции се добиваат со делење на фреквенцијата со обемот на примерокот.

Кумулативни фреквенции

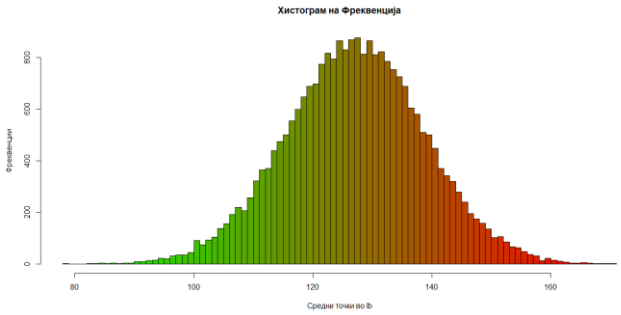
Кумулативните фреквенции се добиваат со собирање на последователните фреквенции (фреквенцијата на сегашниот интервал и интервалите пред него)

Хистограми

Хистограм за висина

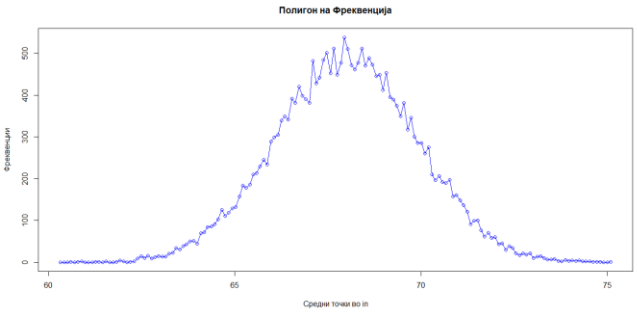


Хистограм за тежина

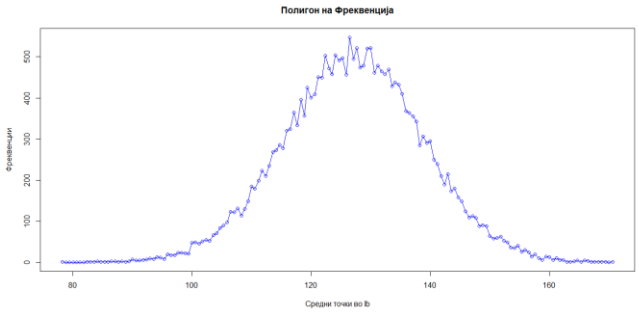


Полигони

Полигон за висина



Полигон за тежина



Задача 2

Стебло-лист дијаграм на висината

```
60 | 36899
61 | 346689999
62 | 012333444444444455555555555556666666666666777777777777777888888999+4
63 | 000000000000000011111111111112222222222222223333333333333333333333+247
64 | 000000000000000000000000000000000000000000000011111111111111111111+906
65 | 000000000000000000000000000000000000000000000000000000000000000000000+2017
66 | 000000000000000000000000000000000000000000000000000000000000000000000+3712
67 | 000000000000000000000000000000000000000000000000000000000000000000000+4846
68 | 000000000000000000000000000000000000000000000000000000000000000000000+5023
69 | 000000000000000000000000000000000000000000000000000000000000000000000+3781
70 | 000000000000000000000000000000000000000000000000000000000000000000000+2228
71 | 000000000000000000000000000000000000000000000000000000000000000000000+973
72 | 000000000000000000000000000000000000000000000000000000000111111111111+258
73 | 000000000000011111111111112222222222222233333333333444444444555555+6
74 | 000000122223333445556 78
75 | 12
```

Стебло-лист дијаграм на тежината

```
 78 | 06
 80 |
 82 | 413
 84 | 3493
 86 | 057048
 88 | 028268
 90 | 002233577923346889
 92 | 12234667778901113445688889
 94 | 00112222233455566788900011223456777777889 9
 96 | 0000111123444555566677778888990000000112334455556666667788888999 9
 98 | 000001111233333444455556677777999900001111112222223334445566667777778
100 | 00000001111111122222222223333333333444444444444555555666666666777777+81
102 | 00000000001111111122222223333333444444444455555555555566666666666+116
104 | 000000000000011111111112222222222333333333333333344444444444444445555+211
106 | 000000000000011111111111111111112222222222222222222233333333333333+331
108 | 000000000000000011111111111111111111111111222222222222222222333333+382
110 | 000000000000000000000000000011111111111111111111111111111111111222+598
112 | 000000000000000000000000000000011111111111111111111111111111111111+726
114 | 000000000000000000000000000000011111111111111111111111111111111+888
116 | 000000000000000000000000000000000000000000000000000000000000001111111+1058
118 | 000000000000000000000000000000000000000000000000000000000000000000000+1256
120 | 000000000000000000000000000000000000000000000000000000000000000000000+1394
122 | 000000000000000000000000000000000000000000000000000000000000000000000+1512
124 | 000000000000000000000000000000000000000000000000000000000000000000000+1631
126 | 000000000000000000000000000000000000000000000000000000000000000000000+1654
128 | 000000000000000000000000000000000000000000000000000000000000000000000+1592
130 | 000000000000000000000000000000000000000000000000000000000000000000000+1568
132 | 000000000000000000000000000000000000000000000000000000000000000000000+1457
134 | 000000000000000000000000000000000000000000000000000000000000000000000+1335
136 | 000000000000000000000000000000000000000000000000000000000000000000111+1122
138 | 000000000000000000000000000000000000000000000000000011111111111111111+927
140 | 000000000000000000000000000000000000000000000000000111111111111111111+747
142 | 000000000000000000000000000000000011111111111111111111111111122222+581
144 | 000000000000000000000000000000000011111111111111111111111111122222+447
146 | 000000000000000011111111111111111111111222222222222333333333333333+294
148 | 000000001111111111111111111122222222222222222223333333333333344444444+211
150 | 000000000000000111111122222222333333333334444444455555555555566666+129
152 | 000000111111111112222233333344444444444555555555556666666677777+71
154 | 000000001112222233333344444455555556666666777888888999999990000000+31
156 | 000001111122223333444445555566677788990111123333444445556777889999
158 | 000022345556888011223334444567789 9
160 | 000001124457789900001334679 99
162 | 1124459399
164 | 566747789
166 | 23388
168 | 291
```

Задача 3

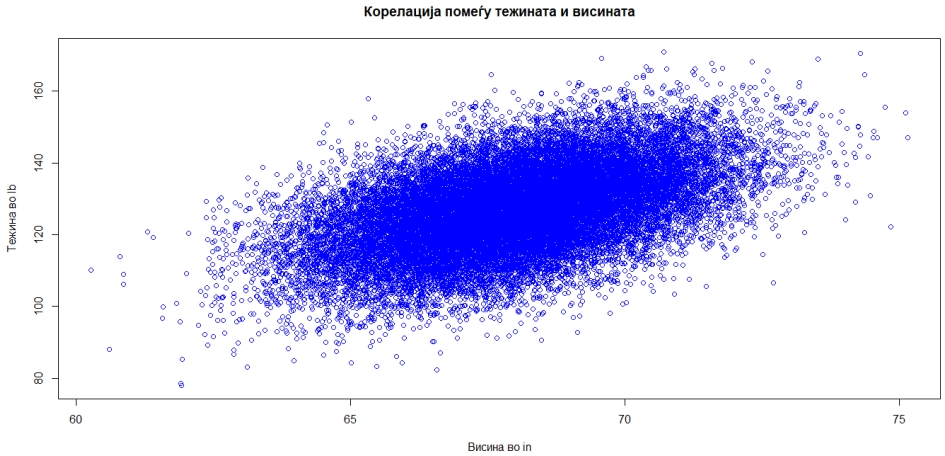


График на расејување

Позитивна корелација

Графикот ни покажува позитивна корелација меѓу висината и тежината. Како се зголемува висината, така и тежината има тенденција да се зголеми.

Густина на податоците

Графикот е густо пополнет во средината, покажувајќи дека повеќето индивидуи во податочното множество припаѓаат во одреден опсег на висина и тежина.

Аутлаери

Има одредени лица кои се наоѓаат надвор од главниот куп. Овие аутлаери се лицата кои се или многу повисоки, пониски, полесни или потешки од просечната популација.

Варијација

Покрај тоа што има позитивен тренд, графикот покажува дека има доста варијација. Не сите индивидуи со дадена висина имаат иста тежина. Варијацијата може да е поради повеќе фактори, како возраст, пол, композиција на телото и стилот на животот

Задача 4

Мода

Модата е вредноста која има најголема честота. Еден примерок може да нема мода, но може да има и повеќе моди.

Мода на висината:	Мода на тежината:
67.9406	124.7975

Медијана

Медијана е вредноста која се наоѓа на средината на подреден примерок. Ако бројот на податоци е непарен, се зема вредноста во средината. Ако бројот на податоци е парен, се зема просекот на податоците во средината.

Медијана на висината:	Медијана на тежината:
67.9957	127.1577

Просек

Просекот е збирот на набљудуваните вредности поделен со големината на примерокот. Недостаток на оваа мерка за централна тенденција е тоа што е многу осетлива на екстремни вредности.

Просек на висината:	Просек на тежината:
67.99311	127.0794

Задача 5

Квартали

Кварталите се вид на перцентили кои го делат примерокот на четири дела со приближно еднаква големина.

1. Q_1 - Првиот квартал е вредност таква што приближно 25% од податоците во подредениот примерок се лево од него, а приближно 75% се десно. Познат е и како долен квартал.
2. Q_2 - Вториот квартал е медијана на примерокот, така што приближно 50% од податоците во подредениот примерок се лево од него, а приближно 50% се десно.
3. Q_3 - Третиот квартал е вредност таква што приближно 75% од податоците во подредениот примерок се лево од него, а приближно 25% се десно. Познат е и како горен квартал.

Кај висината:	Кај тежината:
$Q_1 = 66.7044$	$Q_1 = 119.3087$
$Q_2 = 67.9957$	$Q_2 = 127.1577$
$Q_3 = 69.2730$	$Q_3 = 134.8929$

Опсег

Опсегот е разликата помеѓу најголемата и најмалата набљудувана вредност.

Опсег на висината:	Опсег на тежината:
14.87444	92.90924

Интерквартален опсег

Интерквартилниот опсег е разликата помеѓу третиот и првиот квартал.

Интерквартален опсег на висината:	Интерквартален опсег на тежината:
2.56856	15.58418

Задача 6

Дисперзија

Дисперзијата е мерка на расејување. Ја пресметува средната вредност на квадрираното растојание на дадените вредности од очекуваните.

Дисперзија на висината:	Дисперзија на тежината:
3.616382	135.9765

Стандардна девијација

Стандардната девијација е позитивниот корен на дисперзијата. Ако таа има мала вредност тоа упатува дека сите елементи од групата се многу блиску до средната вредност, додека големата вредност упатува дека групата е широка, со широк опсег на вредности.

Стандардна девијација на висината:	Стандардна девијација на тежината:
1.901679	11.6609

Задача 7

Коефициент на корелација

Коефициентот на корелација го мери правецот и јачината на линеарната врска помеѓу две квантитативни променливи. Секогаш има вредност помеѓу -1 и 1. Ако вредноста е блиску до 0 се зборува за слаба линеарна врска. Јачината на линеарната врска расте со растење на вредноста кон 1 или опаѓање кон -1. Ако има вредност 1 или -1 тоа значи дека точките лежат на една права.

Коефициент на корелација:
0.5028585

Ова значи дека има умерена позитивна врска помеѓу висината и тежината.

Втор дел

Задача 1

Интервал на доверба

Интервал на доверба претставува интервал каде што со одредена веројатност може да кажеме дека оценуваниот параметар се наоѓа во него.

За нашиот примерок ние ќе ја оцениме вредноста на математичкото очекување со интервал на доверба од 90%

Интервал на доверба:
(67.973330493858 , 68.012896699742)

Задача 2

Хипотези

H_0: EX = 68
H_A: EX ≠ 68

Тест

Z_0 = -0.5725657
α = 0.1
c = 1.644854

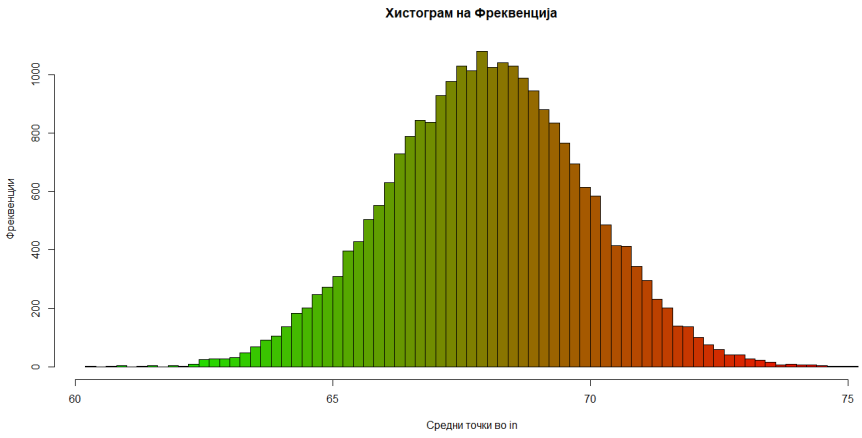
C = (-∞ , -1.644854) ∪ (1.644854 , ∞)
$Z_0 \notin C$

Нултата хипотеза не се отфрла, односно EX = 68

Задача 3

Визуелна интерпретација

Ако податоците се нормално распределени, хистограмот би требало да има облик на ѕвонче.



По набљудување на хистограмот, тој дефинитивно има облик на ѕвонче.

Хипотези

H_0: Обележјето има нормална распределба
H_A: Обележјето нема нормална распределба

Тест

Ќе ја добиеме p-вредноста преку Kolmogorov-Smirnov тест. Иако има некои вредности кои се исти, тие имаат минимално влијание на крајниот резултат гледајќи го тоа колку е обемен примерокот.
p-вредност го претставува најмалото ниво на значајност кое би водело до отфрлање на нултата хипотеза.

α = 0.1
p-вредност = 0.9785271

α е помала од p-вредноста, тоа значи дека нултата хипотеза не се отфрла, односно обележјето има нормална распределба.

Задача 4

Хипотези

H_0: Висината и тежината се независни
H_A: Висината и тежината се зависни

Тест

Ќе ја добиеме p-вредноста преку Karl Pearson тест.

α = 0.1
p-вредност = 0

α е поголема од p-вредноста, тоа значи дека нултата хипотеза се отфрла, односно висината и тежината се зависни

Задача 5

Регресиона анализа

Регресиона анализа се користи за одредување на видот на врската на обележјата и главната цел кога се користи овој метод е да се предвиди или процени вредноста на едната променлива за дадена вредност на другата променлива.

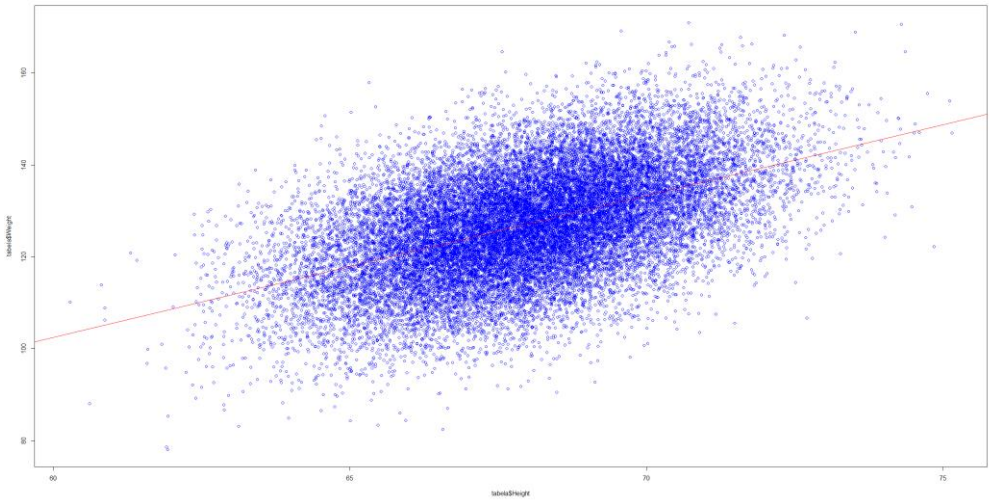


График на корелација

Права на регресија на тежина по висина

$$y = -82.576 + 3.083x$$

Пример:

При проверување на очекуваната тежина за висина од 60 инчи, се добива 102.404 либри.