СТАТИСТИЧНЕ ВИВЕДЕННЯ

Лабораторна робота №2

Виконали студенти КМ-01:
Бабич Ірина
Іваник Юрій
Романецький Микита
Суховій Ігор
Шолоп Любомир

Дослідницькі питання

I. Люди якого віку / статі / раси більше часу проводить в лікарні
II. Який вік / стать / раса людей найчастіше зустрічається в лікарні
III. Залежність вартості оплати від часу перебування/ віку/ ризику смертності.
IV. Ризик смертності в залежності від віку / важкості захворювання

Опис даних

- Обраний датасет містить дані щодо надання лікарських послуг пацієнтам у штаті Нью-Йорк за 2015 рік.
- Він не містить даних, які би могли вказувати на причетність до них окремих осіб. Також, з цього файлу було виключено вторинні діагнози та процедури, а також коди оплати послуг. Один рядок відповідає інформації про рівно 1 клінічний випадок.
- Датасет після очистки містить 2342182 записів та 33 змінних.
- Вік пацієнтів представлений у таких вікових групах:
  - від 0 до 17 років
  - від 18 до 29 років
  - 30 до 49 років
  - від 50 до 69 років
  - і від 70 років і старше

Змінні датасету

- Кожна колонка датасету відповідає одній змінній.
- Більшість змінних в датасеті є категорійними
- Числовими і впорядкованими змінними в датасеті є:
  - *Length.of.Stay*
  - *Birth.Weight (через природу даних містить найбільшу кількість NA значень)*
  - *Total.Charges*
  - *Total.Costs*

Обчислення довірчих інтервалів

- Довірчі інтеграли обчислюються для атрибутів:
  - *Length.of.Stay*
  - *Total.Charges*
  - *Total.Costs*
  - *Birth.Weight*
- Статистики, для яких було побудовано довірчі інтервали:
  - *1-й квантиль*
  - *Медіана*
  - *3-ій квантиль*
  - *Дисперсія*
  - *Коефіцієнт кореляції Пірсона*

Length.of.Stay

Назва статистики	Довірчий інтервал (95%)	Оцінка дисперсії через бутстреп
1-й квантиль	[1.989695, 2.010305]	64.71375
Медіана	[2.989695, 3.010305]	64.7155
3-й квантиль	[5.989695, 6.010305]	64.76851
Вибіркове середнє	[5.47058, 5.49119]	-

Total. Charges

Назва статистики	Довірчий інтервал (95%)	Оцінка дисперсії через бутстреп
1-й квантиль	[11931.16, 12137.25]	6476188303
Медіана	[23398.39, 23604.49]	6493849750
3-й квантиль	[46542.3, 46748.39]	6468228348
Вибіркове середнє	[43128.3963, 43334.48775]	-

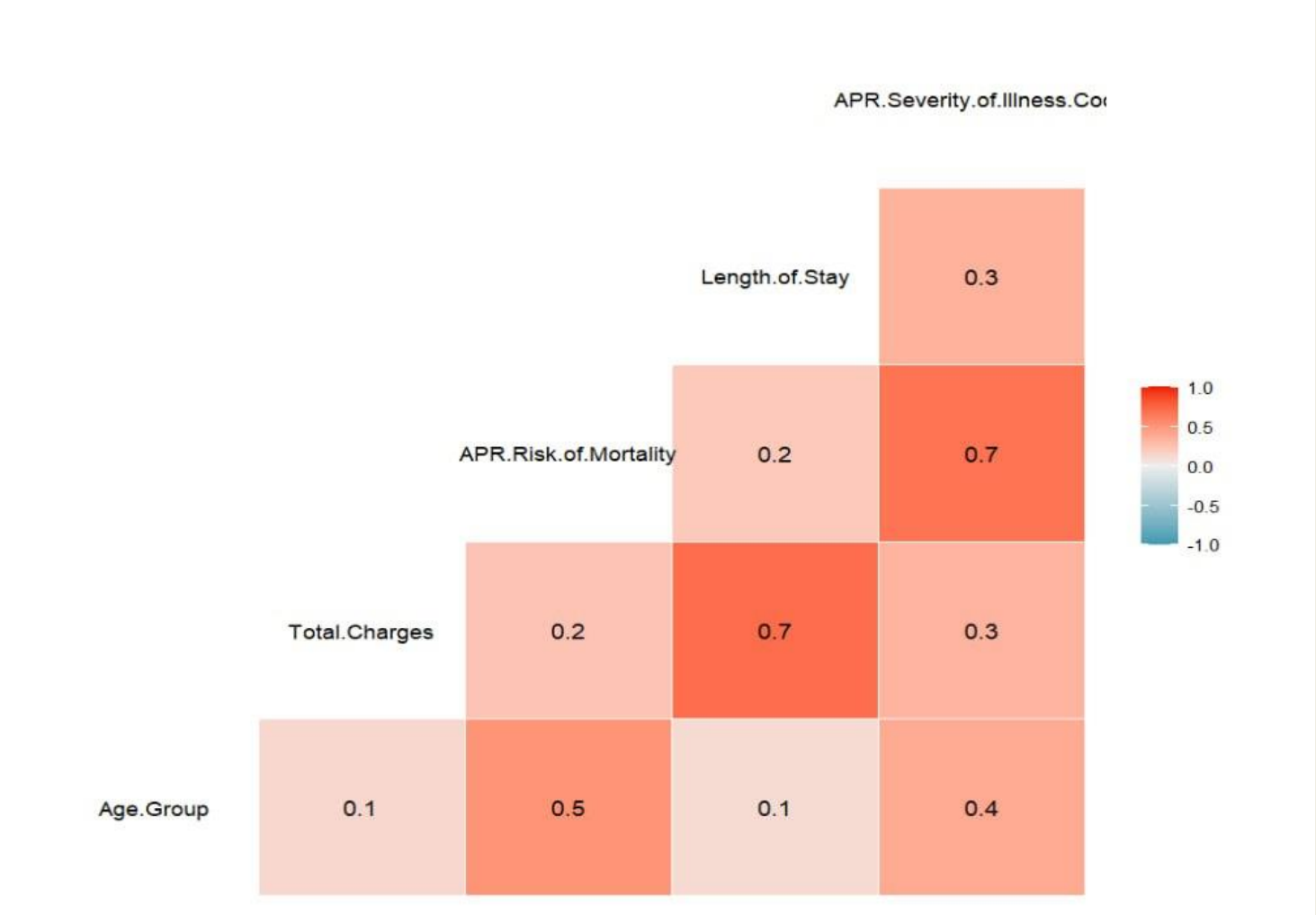
Total.Costs

Назва статистики	Довірчий інтервал (95%)	Оцінка дисперсії через бутстреп
1-й квантиль	[4684.433, 4767.407]	1043783579
Медіана	[8755.638, 8838.612]	1047760380
3-й квантиль	[16803.3, 16886.28]	1050874509
Вибіркове середнє	[15950.925, 16033.899]	-

Birth.Weight

Назва статистики	Довірчий інтервал (95%)	Оцінка дисперсії через бутстреп
1-й квантиль	[2899.163, 2900.837]	427423.2
Медіана	[3299.163, 3300.837]	427606
3-й квантиль	[3599.163, 3600.837]	427489.6
Вибіркове середнє	[3257.0177, 3262.3105]	-

Корелограма



Отже, ми отримали що залежність між довжиною перебування та загальними витратами – 0.7, залежність між загальними витратами та віковою групою – 0.1, залежність між загальними витратами та ризиком смертності – 0.2. Залежність між ризиком смертності та віковою групою – 0.5, між ризиком смертності та важкістю захворювання – 0.7.

Довірчі інтеграли для Коефіцієнта кореляції Пірсона

Кореляція	Рівень довіри	a	b
Age.Group Total.Charges	95%	0.1294	0.1344
Age.Group APR.Risk.of.Mortality	95%	0.5003	0.5017
Age.Group Length.of.Stay	95%	0.1099	0.1129
Age.Group APR.Severity.of.Illness.Code	95%	0.3894	0.3911
Total.Charges APR.Risk.of.Mortality	95%	0.3097	0.3150
Total.Charges Length.of.Stay	95%	0.7031	0.7104
Total.Charges APR.Severity.of.Illness.Code	95%	0.3267	0.3333
APR.Risk.of.Mortality Length.of.Stay	95%	0.2965	0.2990
APR.Risk.of.Mortality APR.Severity.of.Illness.Code	95%	0.7503	0.7517
Length.of.Stay APR.Severity.of.Illness.Code	95%	0.3442	0.3465

Усі кореляції знаходяться в межах відповідних довірчих інтервалів.

Тестування гіпотез

- Гіпотези, які будуть тестуватися у ході виконання дослідження:

I. Гіпотези щодо рівностей вибіркових середніх для тривалості перебування у лікарні для чоловіків та жінок.

II. Гіпотези щодо рівностей вибіркових середніх для тривалості перебування у лікарні для представників різних рас попарно.

III. Гіпотези щодо рівностей вибіркових середніх для тривалості перебування у лікарні для представників різних вікових груп попарно між собою.

IV. Гіпотези щодо відсоткових розподілів за расою, статтю та віковою групою у вибірці в лікарнях відносно відсотків у популяції штату на момент спостереження.

V. Гіпотеза щодо рівностей розподілів для різних важкостей захворювань по вікових групах для різних статей.

VI. Гіпотеза щодо рівностей розподілів для різних важкостей захворювань по расових групах

Гіпотеза щодо рівностей вибіркових середніх для тривалості перебування у лікарні для чоловіків та жінок.

H0 гіпотеза	Оцінена різниця середніх	Довірчий інтервал	Оцінена стандартна помилка	T-статистика	P-значення
Відхилена	0.847125	(0.834787, Inf)	0.007501	112.935	0

Нульова гіпотеза відхилена, отже, чоловіки в середньому проводять більше часу в лікарні ніж жінки

Гіпотези щодо рівностей вибіркових середніх для тривалості перебування у лікарні для представників різних рас попарно

Ознака 1	Ознака 2	H0 гіпотеза	Оцінена різниця середніх	Довірчий інтервал	Оцінена стандартна помилка	T-статистика	P-значення
White	Other Race	Відхилена	0.235633	(0.221438, 0.249828)	0.007243	32.534622	0.000000
White	Black/African American	Відхилена	-0.763672	(-0.779000, -0.748344)	0.007821	-97.647792	0.000000
White	Multi-racial	Відхилена	-0.263188	(-0.277932, -0.248445)	0.007522	-34.987674	0.000000
Other Race	Black/African American	Відхилена	-0.999305	(-1.015203, -0.983406)	0.008112	-123.191842	0.000000
Other Race	Multi-racial	Відхилена	-0.498821	(-0.514157, -0.483485)	0.007825	-63.751024	0.000000
Black/African American	Multi-racial	Відхилена	0.500484	(0.484093, 0.516874)	0.008363	59.848385	0.000000

Гіпотези щодо рівностей вибіркових середніх для тривалості перебування у лікарні для представників різних вікових груп попарно між собою.

Порівняння вікових груп	Нульова гіпотеза	Оцінена різниця середніх	Довірчий інтервал	Оцінена стандартна похибка	t-статистика	P-значення
50-69 і 18-29	Відхилено	1.527026	(1.512320, 1.541732)	0.007503	203.516374	0.000000
50-69 і 30-49	Відхилено	1.223896	(1.209075, 1.238716)	0.007562	161.856147	0.000000
50-69 і 70 або старше	Відхилено	-0.337316	(-0.352149, -0.322482)	0.007568	-44.571257	0.000000
50-69 і 0-17	Відхилено	2.097017	(2.081594, 2.112439)	0.007869	266.501291	0.000000
18-29 і 30-49	Відхилено	-0.303130	(-0.316655, -0.289605)	0.006901	-43.927063	0.000000
18-29 і 70 або старше	Відхилено	-1.864341	(-1.877880, -1.850803)	0.006908	-269.891276	0.000000
18-29 і 0-17	Відхилено	0.569991	(0.555809, 0.584173)	0.007236	78.772324	0.000000
30-49 і 70 або старше	Відхилено	-1.561211	(-1.574875, -1.547548)	0.006971	-223.952898	0.000000
30-49 і 0-17	Відхилено	0.873121	(0.858820, 0.887422)	0.007296	119.663240	0.000000
70 або старше і 0-17	Відхилено	2.434332	(2.420018, 2.448646)	0.007303	333.328894	0.000000

Гіпотези щодо відсоткових розподілів за расою, статтю та віковою групою у вибірці в лікарнях відносно відсотків у популяції штату на момент спостереження

Ознака	H0 гіпотеза	Оцінена ймовірність	Ймовірність для порівняння	Довірчий інтервал	Оцінена стандартна помилка	T-статистика	P-значення
Gender: F	Відхилена	0.556570	0.514470	(0.555934, 0.557207)	0.000325	129.695	0
Gender: M	Відхилена	0.443430	0.485530	(0.442793, 0.444066)	0.000325	-129.695	0
Race: White	Відхилена	0.569138	0.703000	(0.568504, 0.569772)	0.000324	-413.703	0
Race: Black/African American	Відхилена	0.189487	0.176000	(0.188986, 0.189989)	0.000256	52.67	0
Race: Other Race	Відхилена	0.231875	0.097000	(0.231334, 0.232415)	0.000276	489.100	0
Race: Multi-racial	Відхилена	0.009500	0.024000	(0.009375, 0.009624)	0.000063	-228.773	0
Age.Group: 0 to 17	Відхилена	0.150297	0.212710	(0.149839, 0.150754)	0.000234	-267.288	0
Age.Group: 18 to 29	Відхилена	0.104954	0.173760	(0.104561, 0.105346)	0.000200	-343.57	0
Age.Group: 30 to 49	Відхилена	0.194881	0.263360	(0.194373, 0.195388)	0.000259	-264.578	0
Age.Group: 50 to 69	Відхилена	0.275286	0.249550	(0.274714, 0.275858)	0.000292	88.181	0
70 or Older	Відхилена	0.274583	0.100620	(0.274011, 0.275154)	0.000292	596.535	0

Гіпотеза щодо рівностей розподілів для різних важкостей захворювань по вікових групах для різних статей.

Вікова група	Ознака	H0 гіпотеза	Хі-квадрат	df	P-значення
50 to 69	male	Відхилена	1826.2071	NA	0.0004997501
50 to 69	female	Відхилена	1826.2071	NA	0.0004997501
18 to 29	male	Відхилена	6781.0239	NA	0.0004997501
18 to 29	female	Відхилена	6781.0239	NA	0.0004997501
30 to 49	male	Відхилена	13403.975	NA	0.0004997501
30 to 49	female	Відхилена	13403.975	NA	0.0004997501
70 or Older	male	Відхилена	3028.8116	NA	0.0004997501
70 or Older	female	Відхилена	3028.8116	NA	0.0004997501
0 to 17	male	Відхилена	164.09405	NA	0.0004997501
0 to 17	female	Відхилена	164.09405	NA	0.0004997501

Перевірка рівностей розподілів для різних важкостей захворювань по расових групах

Расова група	H0 гіпотеза	Хі-квадрат	df	P-значення
White vs. Other Race	Відхилена	25601.004	3	< 2.2204e-16
White vs. Black/African Am.	Відхилена	4673.4048	3	< 2.2204e-16
White vs. Multi-racial	Відхилена	50.78586	3	5.433933e-11
Other Race vs. Black/African Am.	Відхилена	5405.3655	3	< 2.2204e-16
Other Race vs. Multi-racial	Відхилена	1282.9428	3	< 2.2204e-16
Black/African Am. vs. Multi-racial	Відхилена	259.24932	3	< 2.2204e-16

Висновки

- За допомогою тестування гіпотез було підтверджено, що існує різниця в тривалості перебування в лікарні для різних вікових груп (*гіпотеза 3*), чим старшою є людина, тим в середньому більше днів вони проводить в лікарні під час лікування, порівняно з іншими віковими групами(вийнятком є групи 0-17 та 18-29) .
- Також було виявлено, що жінки в середньому частіше потрапляють в лікарню, ніж чоловіки. Хоча чоловіки в середньому перебувають в лікарні довше за жінок. Представники білої раси потрапляють в лікарні менше, ніж можна було б очікувати за розподілом населення, а представники інших рас (тобто люди азійського походження та корінні жителі Америки)- більше. Найчастіше в лікарню потрапляють люди вікової категорії 50-69 та 70+.
- Представники білої раси проводять в середньому в лікарні більше часу ніж люди азійського походження та корінні жителі Америки, але менше за представників чорної та мульти рас. Представники чорної раси проводять в лікарні в середньому найбільше часу, найменше - люди азійського походження та корінні жителі Америки.
- Представники інших рас мають більший відсоток випадків з ризиком смерті minor, ніж інші 3 категорії, відповідно, відсоток важчих випадків менший.