Звіт

до лабораторної роботи №2

з дисципліни “Аналіз даних”

на тему “Статистичне виведення”

Роботу виконали студенти групи КМ-01:

Бабич Ірина

Іваник Юрій

Романецький Микита

Суховій Ігор

Шолоп Любомир

Київ – 2023

Основна частина

I. Мотивація проведення дослідження

Дослідницькі питання залишилися незмінними у порівняння з першою лабораторною роботою, а саме:

1. Люди якого віку / статі / раси більше часу проводить в лікарні
2. Який вік / стать / раса людей найчастіше зустрічається в лікарні
3. Залежність вартості оплати від часу перебування/ віку/ ризику смертності.
4. Ризик смертності в залежності від віку / важкості захворювання

Гіпотези, які будуть тестуватися у ході виконання дослідження:

1. Гіпотези щодо рівностей вибіркових середніх для тривалості перебування у лікарні для чоловіків та жінок.
2. Гіпотези щодо рівностей вибіркових середніх для тривалості перебування у лікарні для представників різних рас попарно.
3. Гіпотези щодо рівностей вибіркових середніх для тривалості перебування у лікарні для представників різних вікових груп попарно між собою.
4. Гіпотези щодо відсоткових розподілів за расою, статтю та віковою групою у вибірці в лікарнях відносно відсотків у популяції штату на момент спостереження.
5. Гіпотеза щодо рівностей розподілів для різних важкостей захворювань по вікових групах для різних статей.
6. Гіпотеза щодо рівностей розподілів для різних важкостей захворювань по расових групах

II. Довірчі інтервали

Довірчі інтегрални обчислюються для атрибутів Length.of.Stay, Total.Charges, Total.Costs, Birth.Weight.

Статистики, для яких було побудовано довірчі інтервали:

1. 1-й квантиль
2. Медіана
3. 3-ій квантиль
4. Дисперсія
5. Коефіцієнт кореляції Пірсона

Для побудови довірчих інтегралів для квантилів та медіан було використано t-розподіл, але для оцінки дисперсії статистик було використано бутстреп з параметром R=100.

Таблиця для 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]	-

Табл.1

Таблиця для 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]	-

Табл. 2

Таблиця для 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

Назва статистики	Довірчий інтервал (95%)	Оцінка дисперсії через бутстреп
Вибіркове середнє	[15950.925, 16033.899]	-

Табл. 3

Таблиця для 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]	-

Табл. 4

Також, для порівняння, довірчі інтеграли для квантилів та медіан були побудовані через бутстреп, метод – basic.

	Median	**1st Quantile**	**3rd Quantile**
Length.of.Stay	All values equal to 3	All values equal to 2	All values equal to 6
Total.Charges	(23393, 23744)	(11933, 12150)	(46104, 46939)
Total.Costs	(8760, 8891)	(4668, 4749)	(16630, 16898)
Birth.Weight	All values equal to 3300	All values equal to 2900	All values equal to 3600

Табл. 5

Для обчислення коефіцієнтів кореляції Пірсона було застосовано бутстреп. При використанні бутстрепу було обрано параметр R=10, а довірчі інтервали було обчислено basic, percentile та studentized методами.

Значення кореляцій:

	Age.Group	Total.Charges	APR.Risk.of.Mortality	Length.of.Stay	APR.Severity.of.Illness.Code
Age.Group	1.0000000	0.1327187	0.5011360	0.1113223	0.3904089
Total.Charges	0.1327187	1.0000000	0.3130371	0.7060176	0.3308819
APR.Risk.of.Mortality	0.5011360	0.3130371	1.0000000	0.2983804	0.7507220
Length.of.Stay	0.1113223	0.7060176	0.2983804	1.0000000	0.3460253
APR.Severity.of.Illness.Code	0.3904089	0.3308819	0.7507220	0.3460253	1.0000000

Рис. 1 – Значення кореляцій

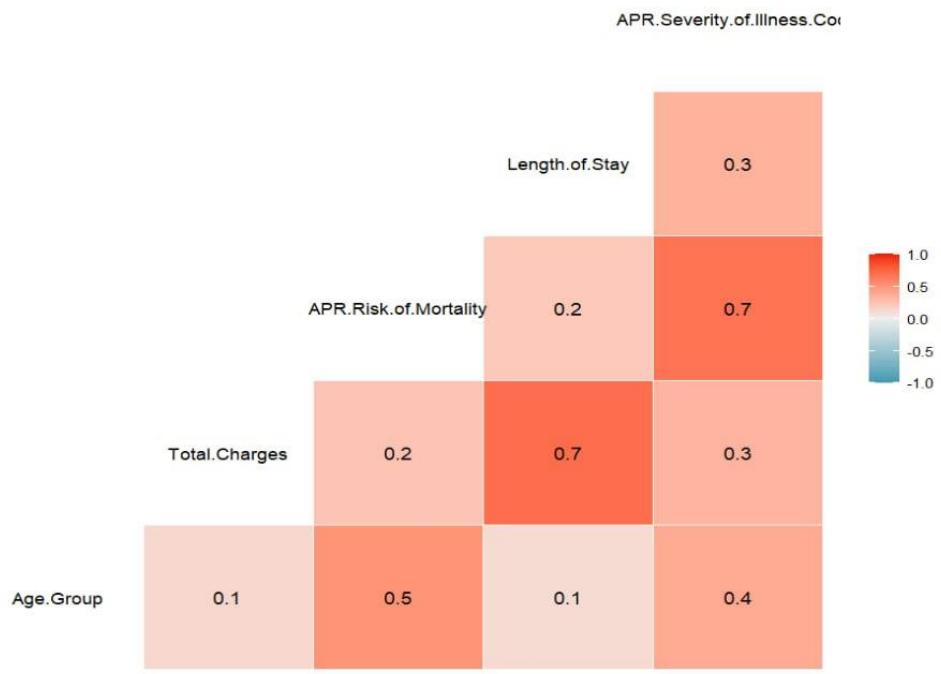

Рис. 2 – Графік кореляцій

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 важкостю захворювання – 0.7.

Довірчі інтеграли для Коефіцієнта кореляції Пірсона:

1. Basic confidence interval

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

Кореляція	Рівень довіри	a	b
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

Табл. 6

2. Percentile confidence interval

Кореляція	Рівень довіри	a	b
Age.Group Total.Charges	95%	0.1311	0.1360
Age.Group APR.Risk.of.Mortality	95%	0.5006	0.5020
Age.Group Length.of.Stay	95%	0.1097	0.1128
Age.Group APR.Severity.of.Illness.Code	95%	0.3898	0.3914
Total.Charges APR.Risk.of.Mortality	95%	0.3097	0.3150
Total.Charges Length.of.Stay	95%	0.7016	0.7090
Total.Charges APR.Severity.of.Illness.Code	95%	0.3285	0.3350
APR.Risk.of.Mortality Length.of.Stay	95%	0.2978	0.3003
APR.Risk.of.Mortality APR.Severity.of.Illness.Code	95%	0.7497	0.7512

Кореляція	Рівень довіри	a	b
Length.of.Stay APR.Severity.of.Illness.Code	95%	0.3456	0.3479

Табл. 7

3. Studentized confidence interval

Кореляція	Рівень довіри	a	b
Age.Group Total.Charges	95%	0.1292	0.1342
Age.Group APR.Risk.of.Mortality	95%	0.5004	0.5015
Age.Group Length.of.Stay	95%	0.1101	0.1128
Age.Group APR.Severity.of.Illness.Code	95%	0.3889	0.3916
Total.Charges APR.Risk.of.Mortality	95%	0.3094	0.3149
Total.Charges Length.of.Stay	95%	0.7030	0.7125

Кореляція	Рівень довіри	a	b
Total.Charges APR.Severity.of.Illness.Code	95%	0.3263	0.3334
APR.Risk.of.Mortality Length.of.Stay	95%	0.2941	0.2994
APR.Risk.of.Mortality APR.Severity.of.Illness.Code	95%	0.7503	0.7519
Length.of.Stay APR.Severity.of.Illness.Code	95%	0.3422	0.3470

Табл. 8

Кореляції між *Age.Group та Total.Charges, Age.Group та APR.Risk.of.Mortality, Age.Group та Length.of.Stay, Age.Group та Length.of.Stay, Total.Charges та APR.Risk.of.Mortality, Total.Charges та Length.of.Stay, Total.Charges та APR.Severity.of.Illness.Code, APR.Risk.of.Mortality та Length.of.Stay, APR.Risk.of.Mortality та APR.Severity.of.Illness.Code, Length.of.Stay та APR.Severity.of.Illness.Code* (тобто усі кореляції) знаходяться в межах відповідних довірчих інтервалів.

III. Тестування гіпотез

1) Гіпотези 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

Табл. 9

Нульова гіпотеза була відхилена, що свідчить про статистично значущу різницю в середніх значеннях.

2) 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

Ознака 1	Ознака 2	H0 гіпотеза	Оцінена різниця середніх	Довірчий інтервал	Оцінена стандартна помилка	T-статистика	P-значення
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

Табл. 10

За результатами тесту, нульові гіпотези для всіх порівнянь були відхилені, що свідчить про статистично значущу різницю в середніх тривалості перебування у лікарні для представників різних рас.

3) 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

Порівняння вікових груп	Нульова гіпотеза	Оцінена різниця середніх	Довірчий інтервал	Оцінена стандартна похибка	t-статистика	P-значення
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

Табл. 11

Отже, на підставі проведених тестів Вальда можна стверджувати, що існують статистично значимі відмінності у тривалості перебування у лікарні між різними віковими групами. В порівнянні вікових груп 50-69 років з 18-29 років, 30-49 років та 70 і більше років, нульові гіпотези про рівність середніх тривалостей перебування у лікарні відхиляються. Це свідчить про наявність статистично значимих відмінностей у тривалості перебування у лікарні між цими групами.

4) 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

Ознака	H0 гіпотеза	Оцінена ймовірність	Ймовірність для порівняння	Довірчий інтервал	Оцінена стандартна помилка	T-статистика	P-значення
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

Табл. 12

Нульові гіпотези для всіх ознак були відхиленими. Це означає, що імовірності розподілів (за статтю, расою та віковою групою) у вибірці лікарень статистично значущі та відрізняються від ймовірностей для порівняння. Тобто, вибірки мають статистично значимі розбіжності за цими ознаками.

Оцінена ймовірність – це відсотковий розподіл частки ознак(наприклад частки жінок) в лікарні, а ймовірність для порівняння – відсотковий розподіл ознаки в популяції штату.

5) 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

Вікова група	Ознака	H0 гіпотеза	Хі-квадрат	df	P-значення
0 to 17	male	Відхилена	164.09405	NA	0.0004997501
0 to 17	female	Відхилена	164.09405	NA	0.0004997501

Табл. 13

Згідно з результатами тесту, нульові гіпотези для всіх вікових груп та ознак були відхилені, що свідчить про статистично значущі різниці в розподілах важкості захворювань між статями в кожній віковій групі.

6) 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

Табл. 14

Згідно з результатами тесту, нульові гіпотези для всіх розглянутих пар расових груп та ознак були відхилені, що свідчить про статистично значущі різниці в розподілах важкості захворювань між расовими групами.

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.