Проект
по
Вероятност и Статистика

летен семестър, 2019/2020

Изготвил:
Иван Арабаджийски, ФН: 81631, 3 курс, 2 поток, 5 група

1. Описание на данните

Днаните са взети от проучване с участието на специалисти от Секция/Клиника по ендокринология и болести на обмяната при УМБАЛ „Св. Георги“ ЕАД, МУ-Пловдив и Катедра по клинична лаборатория към МУ-Пловдив.

Дизайнът на проучването включва изследване на 80 жени, в репродуктивна възраст, подписали доброволно писмено информирано съгласие за участие в проучването. Те са разделени в четири групи: жени с новодиагностициран ПКОС (n=29), жени с новодиагностициран захарен диабет тип 2 (n=10), жени с новодиагностициран метаболитен синдром (n=18) и съответни по възраст клинично здрави жени (n=26).

Жените, включени в проучването ще бъдат набирани сред хоспитализираните и амбулаторни пациентки на Клиниката по ендокринология към УМБАЛ „Св. Георги” ЕАД.

Предвижда се определяне на показателите малондиалдехид, супероксидна дисмутаза и глутатионова пероксидаза, LH, FSH, имунореактивен инсулин, общ холестерол, HDL-холестерол, LDL-холестерол, триглицериди, като проява на оксидативен стрес при пациенти с инсулинова резистентност – сднократно.

1.1 Основни въпроси, на които се отговаря чрез анализа на данните:

1. Съществува ли статистически значима разлика между серумната концентрация на глюкозата в групата на клинично здравите и в групата на диабетиците?
2. Съществува ли статистически значима разлика между серумната концентрация на холестерола в групата на клинично здравите и в групата на диабетиците?
3. Какво можем да кажем за серумната концентрация на глюкоза в останалите групи спрямо групата на диабетиците?
4. Има ли връзка между глюкоза и холестерол в глупата на клинично здравите?
5. Има ли връзка между глюкоза и холестерол в глупата на диабетиците?
6. Има ли връзка между триглицеридите и глюкозата в глупата на клинично здравите?
7. Има ли връзка между триглицеридите и глюкозата в глупата на новодиагностицираните с ПКОС?
8. Има ли връзка между триглицеридите и холестерола в глупата на клинично здравите?
9. Има ли връзка между триглицеридите и холестерола в глупата на новодиагностицираните с ПКОС?

1.2 Променливи, чрез които са представени данните:

- ➔ Горепосочените групи жени – за по-кратко група 1, 2, 3, 4
- ➔ **GLUC** – кръвна захар
- ➔ **CHOL**– общ холестерол
- ➔ **Tg** – триглицериди

1.3 Използвани статистически методи:

- Описателна (дескриптивна) статистика, включваща изчисляване на средна стойност, медиана, мода, стандартно отклонение и вариация (дисперсия).

Тъй като в стандартния пакет на R няма метод за изчисление на мода, е имплементирана допълнителна функция:

```
getMode <- function(values) {
uniqueValues <- unique(values)
uniqueValues[which.max(tabulate(match(values,
uniqueValues)))]
}
```

- Представяне на данните графично посредством хистограми

Определяне типа на всяко едно от изследваните разпределения на данните с цел последващ избор на статистически тест за сравняване на данните. Приложен е тест на Shapiro-Wilcoxon с нулева хипотеза H0 „Разпределението е нормално“ и алтернативна хипотеза „Разпределението не е нормално“ с равнище на значимост p = 0/05.

- Поставените задачи за изследване изискват сравнение на независими извадки. В случай на нормалност на двете сравнявани разпределения използваме едностранен t-test за сравнение на средните стойности. В случай, че поне едно от разпределенията на сравняваните извадки не е нормално, използваме непараметричния тест на Mann-Whitney-Wilcoxon.
- Приложен е корелационен анализ за установяване на зависимостта между нивата на холестерол и глюкоза, холестерол и триглицериди и глюкоза и триглицериди на клинично здравите и диабетиците.

2. Дескриптивни статистики

За целите на изследването данните са разделени в няколко отделни dataframe-ове, както следва:

1. group1CHOL – холестеролът на клинично здравата група
2. group2CHOL – холестеролът на новодиагностицираните с ПКОС
3. group3CHOL – холестеролът на новодиагностицираните с метаболитен синдром
4. group4CHOL – холестеролът на новодиагностицираните със захарен диабет тип 2

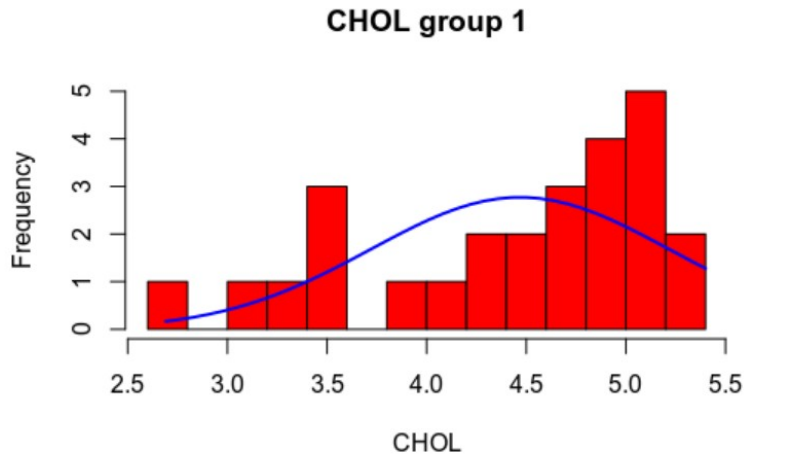
5. group1GLUC – Глюкозата на клинично здравата група
6. group2GLUC – Глюкозата на новодиагностицираните с ПКОС
7. group3GLUC – Глюкозата на новодиагностицираните с метаболитен синдром
8. group4GLUC – Глюкозата на новодиагностицираните със захарен диабет тип 2

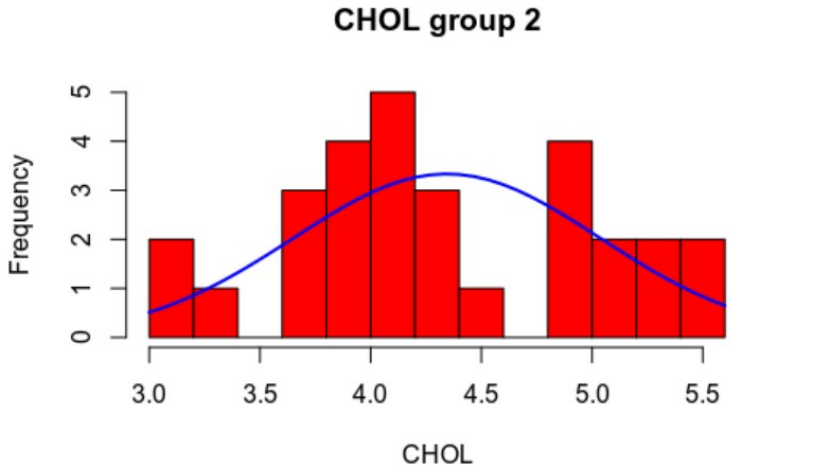
9. group1Tg – Триглицеридите на клинично здравата група
10. group2Tg – Триглицеридите на новодиагностицираните с ПКОС
11. group3Tg – Триглицеридите на новодиагностицираните с метаболитен синдром
12. group4Tg – Триглицеридите на новодиагностицираните със захарен диабет тип 2

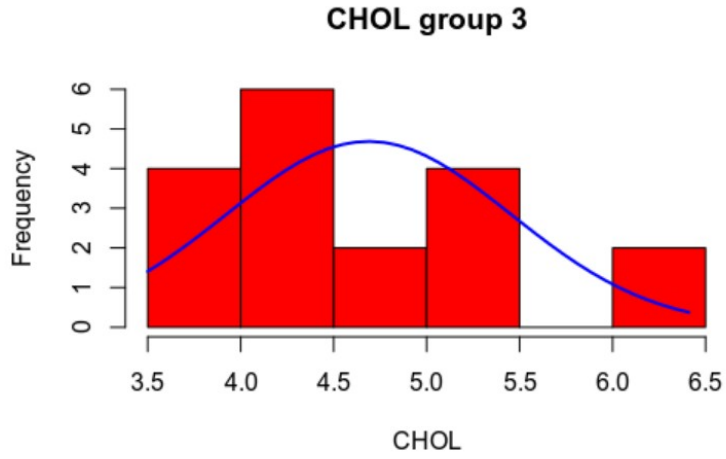
В следващите таблици са представени дескриптивните статистики за всеки един от горепосочените dataframe-ове.

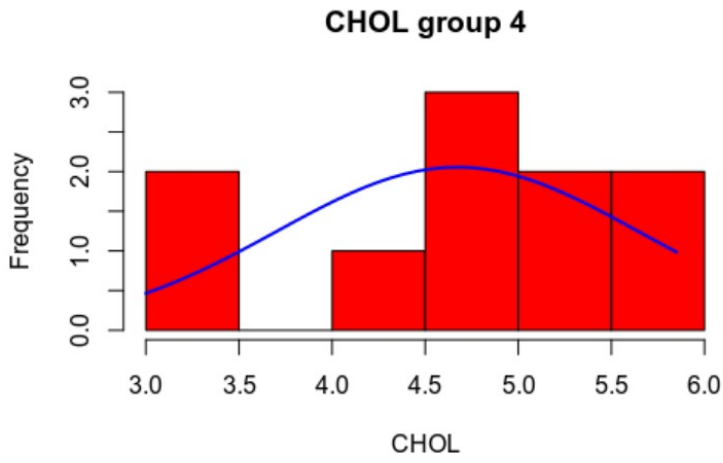
Извадка	Средна стойност	Медиана	Мода	Стандартно отклонение	Дисперсия
group1CHOL	4.468	4.715	4.9	0.7483085	0.5599655
group2CHOL	4.344	4.140	3.9	0.6939916	0.4816244
group3CHOL	4.688	4.500	3.9	0.766261	0.5871559
group4CHOL	4.673	4.900	5.3	0.9705331	0.9419344
group1GLUC	4.965	4.900	4.8	0.4267574	0.1821218
group2GLUC	4.893	5.000	5.3	0.4956461	0.245665
group3GLUC	5.744	5.850	5.9	0.3257972	0.1061438
group4GLUC	6.850	6.700	6.7	0.5835714	0.3405556
group1Tg	0.7038	0.6800	0.7	0.2368979	0.05612062
group2Tg	0.7166	0.6500	0.8	0.242447	0.05878054
group3Tg	1.026	0.920	1.38	0.3862523	0.1491908
group4Tg	1.501	1.280	0.9	0.6883547	0.4738322

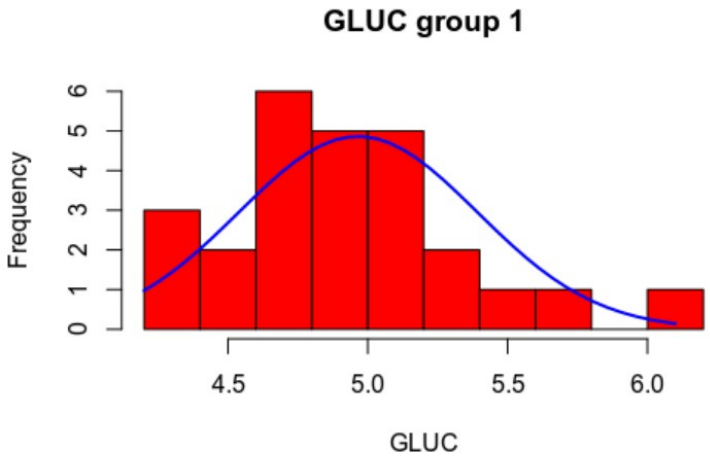
3. Графично представяне

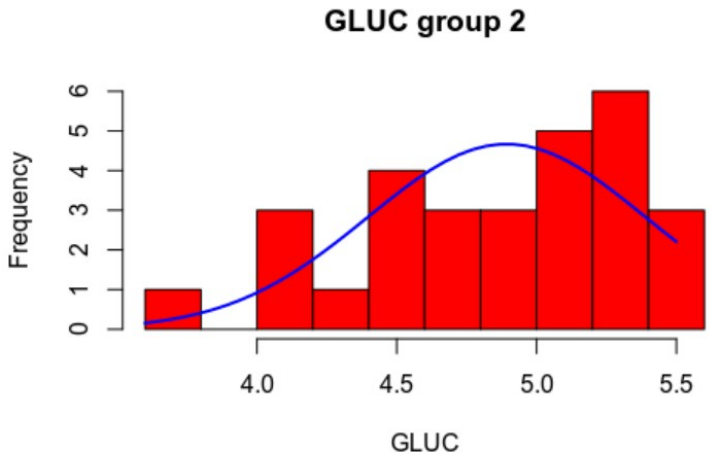


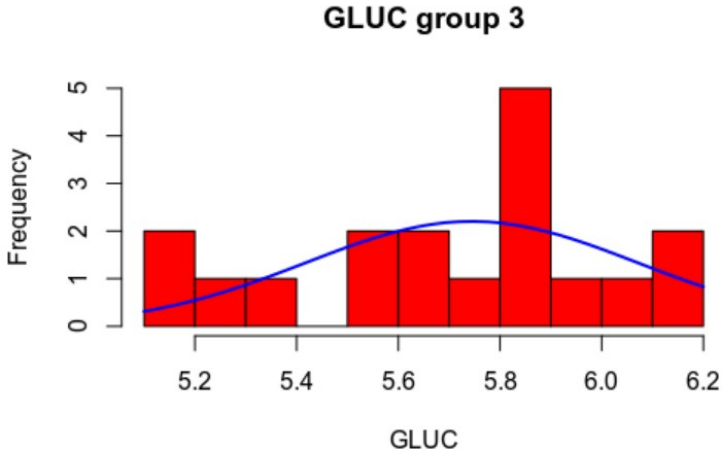


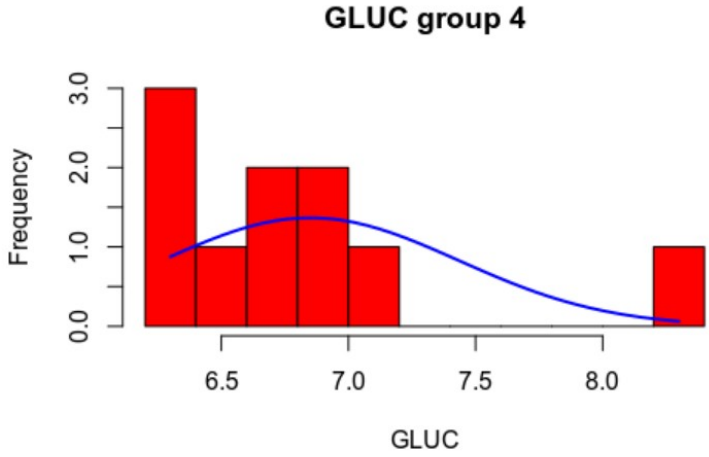


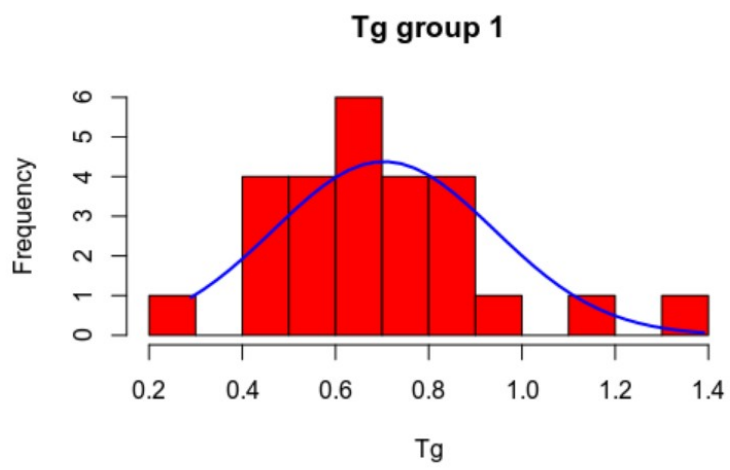


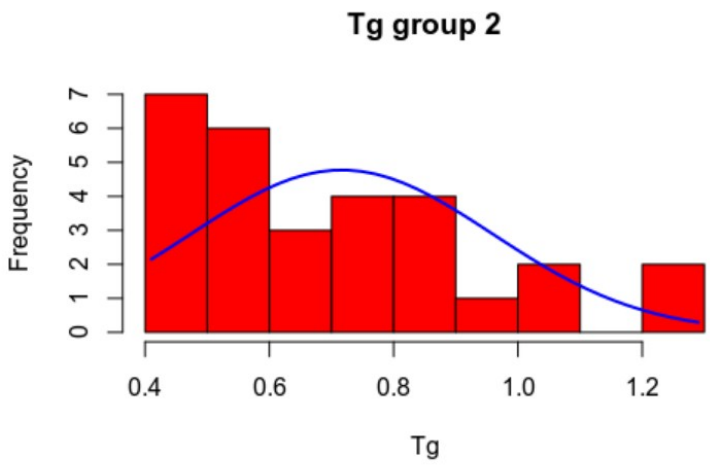

GLUC group 1
Frequency
0 1 2 3 4 5 6
4.5 5.0 5.5 6.0
GLUC

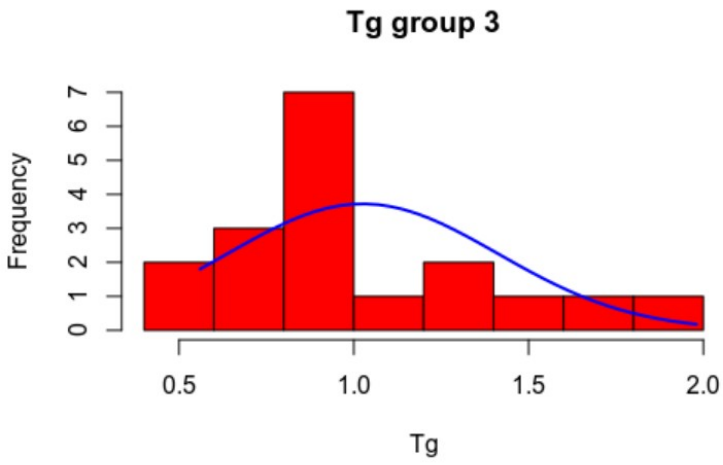

GLUC group 2
Frequency
0 1 2 3 4 5 6
4.0 4.5 5.0 5.5
GLUC

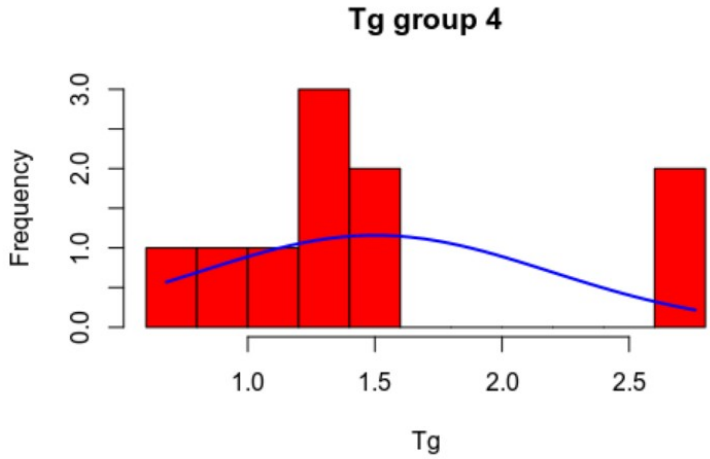

GLUC group 3
Frequency
0 1 2 3 4 5
5.2 5.4 5.6 5.8 6.0 6.2
GLUC


GLUC group 4
Frequency
0.0 1.0 2.0 3.0
6.5 7.0 7.5 8.0
GLUC


Tg group 1
Frequency
0 1 2 3 4 5 6
0.2 0.4 0.6 0.8 1.0 1.2 1.4
Tg


Tg group 2
Frequency
0 1 2 3 4 5 6 7
0.4 0.6 0.8 1.0 1.2
Tg


Tg group 3
Frequency
0 1 2 3 4 5 6 7
0.5 1.0 1.5 2.0
Tg


Tg group 4
Frequency
0.0 1.0 2.0 3.0
1.0 1.5 2.0 2.5
Tg

4. Определяне вида на разпределенията

Извадка	p-value:	Нормално ли е разпределението?
group1CHOL	0.02293	не
group2CHOL	0.2117	да
group3CHOL	0.344	да
group4CHOL	0.1965	да
group1GLUC	0.2761	да
group2GLUC	0.04414	не
group3GLUC	0.2576	да
group4GLUC	0.0194	не
group1Tg	0.1082	да
group2Tg	0.01227	не
group3Tg	0.05004	да
group4Tg	0.050002	да

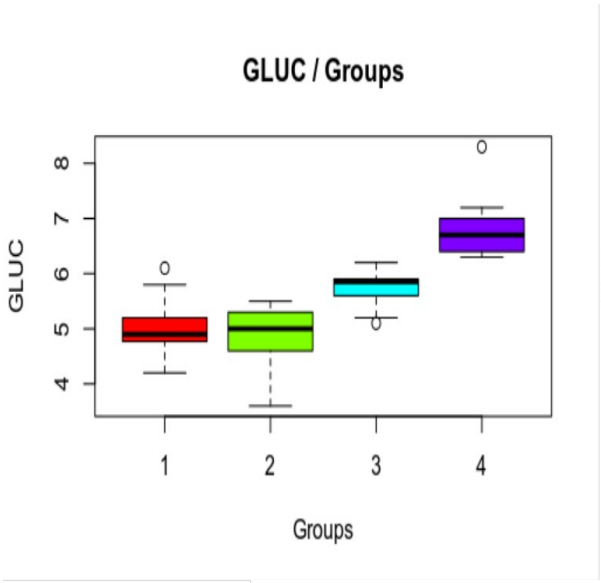
5. Сравняване на данните

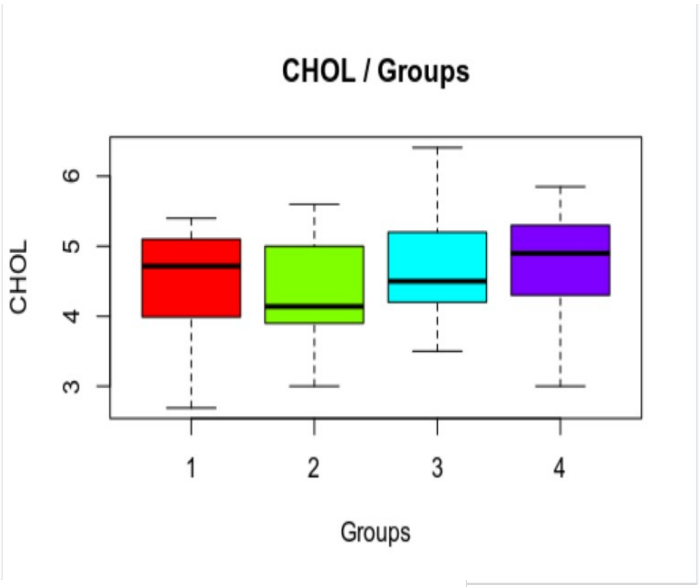
5.1 Сравняване на числовите данни по групи

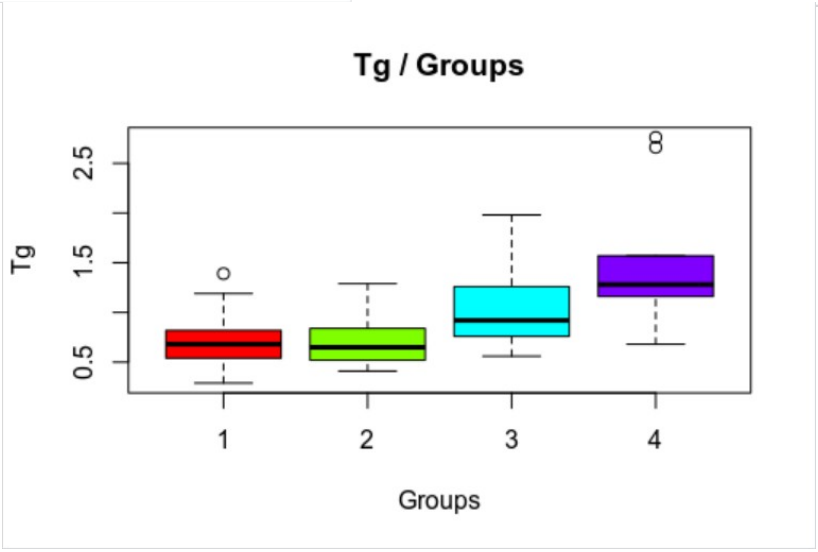
Тъй като имаме 4 групи ще изпозлваме теста на Крускал за да определим дали има статистически значима разлика между четирите групи.

Данни:	p-value:	Заключение:
GLUC vs Group	~0	Има статистически значима разлика
CHOL vs Group	0.3941	Няма статистически значима разлика
Tg vs Group	~0	Има статистически значима разлика

Подрепяме с boxplots:



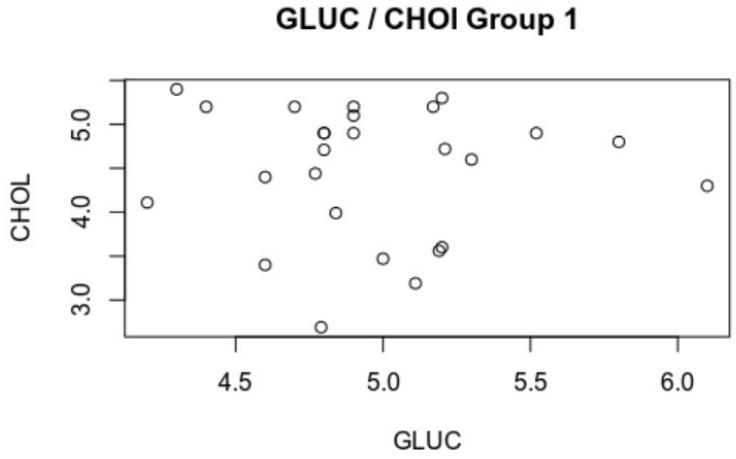


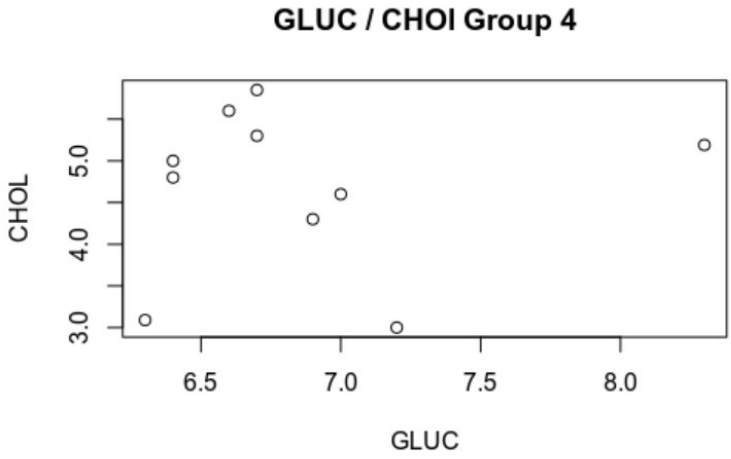


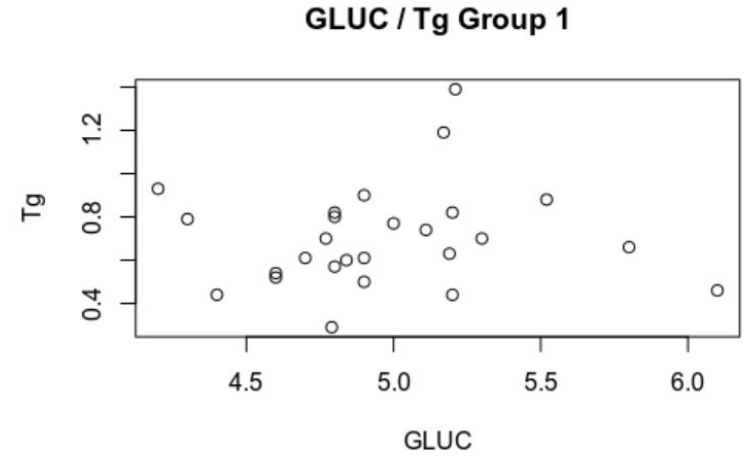
5.2 Търсим зависимост между кръвната захар и холестеролът в групата на клинично здравите и групата на диабетиците.

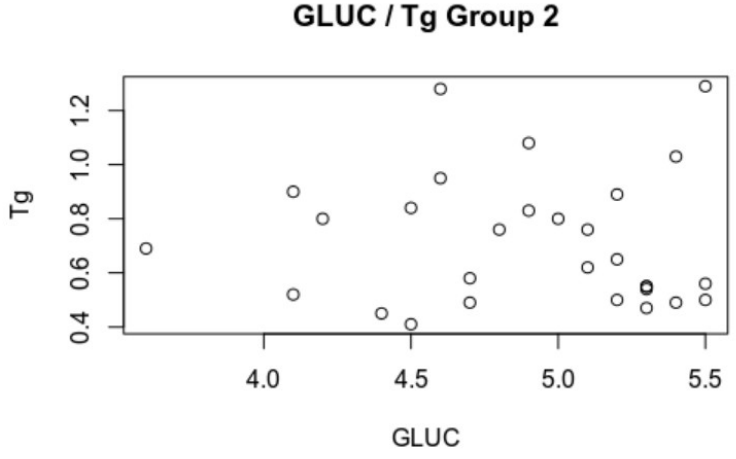
Извадка	Корелация
Group1CHOL vs group1GLUC	-0.05635747
Group4CHOL vs group4GLUC	-0.07317209
Group2GLUC vs group2Tg	-0.09456148
Group2CHOL vs group2Tg	0.3254389
Group2CHOL vs group2Tg	0.5453993

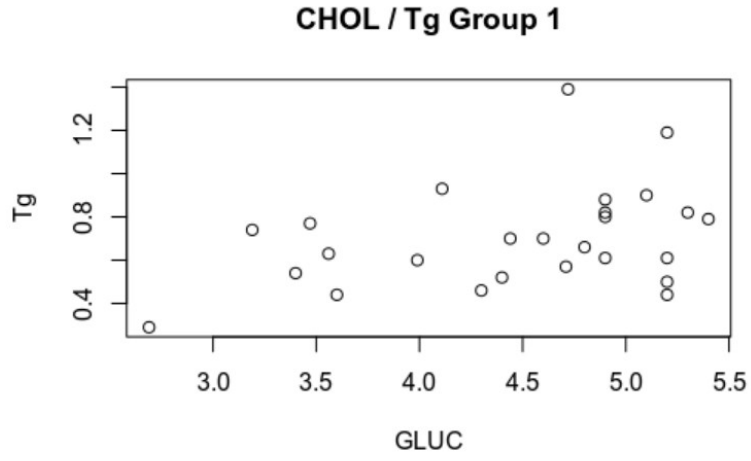
Подрепяме с dotplots:

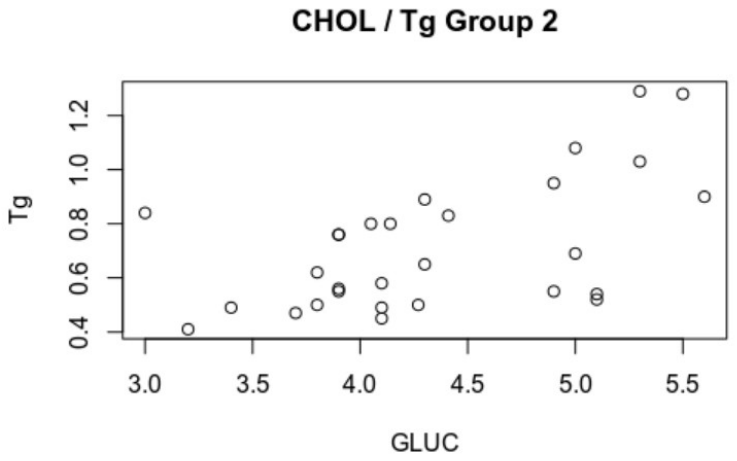












6.Отговори на някои важни въпроси

1. Съществува ли статистически значима разлика между серумната концентрация на глюкозата в групата на клинично здравите и в групата на диабетиците?

Разпределението в група 1 е нормално, но в група 4 не е, следователно ще приложим непараметричен тест за сравняване на независими извадки на Mann-Whitney-Wilcoxon. Използваме вградената функция wilcox.test(). Резултатът е:

```
> wilcox.test(groups.group1$GLUC.mmol.l, groups.group4$GLUC.mmol.l)
        Wilcoxon rank sum test with continuity correction

data:  groups.group1$GLUC.mmol.l and groups.group4$GLUC.mmol.l
W = 0, p-value = 4.71e-06
alternative hypothesis: true location shift is not equal to 0
```

Правим заключение, че има статистически значима разлика.

2. Съществува ли статистически значима разлика между серумната концентрация на холестерола в групата на клинично здравите и в групата на диабетиците?

По подобни на горните съображения прилагаме теста на Mann-Whitney-Wilcoxon и получаваме резултат:

```
> wilcox.test(groups.group1$CHOL.mmol.l, groups.group4$CHOL.mmol.l)

        Wilcoxon rank sum test with continuity correction

data:  groups.group1$CHOL.mmol.l and groups.group4$CHOL.mmol.l
W = 108, p-value = 0.4469
alternative hypothesis: true location shift is not equal to 0
```

p-value > 0.05 и заключаваме, че статистически значима разлика няма.

3. Сега ще анализираме данни между две групи с различни заболявания и ще видим съществува ли статистически значима разлика между серумните концентрации на глюкозата в групата на жените с новодиагностициран ПКОС и в групата на жените със захарен диабет тип 2.

```
> wilcox.test(groups.group2$GLUC.mmol.l, groups.group4$GLUC.mmol.l)

          Wilcoxon rank sum test with continuity correction

data:  groups.group2$GLUC.mmol.l and groups.group4$GLUC.mmol.l
W = 0, p-value = 3.256e-06
alternative hypothesis: true location shift is not equal to 0
p-value <= 0.05
```

Правим заключение, че има статистически значима разлика.
Последно правим същото сравнение между група 3 и 4.

```
> wilcox.test(groups.group3$GLUC.mmol.l, groups.group4$GLUC.mmol.l)
```

```
        Wilcoxon rank sum test with continuity correction

data:  groups.group3$GLUC.mmol.l and groups.group4$GLUC.mmol.l
W = 0, p-value = 1.663e-05
alternative hypothesis: true location shift is not equal to 0
```

и отново се оказва, че има статистически значима разлика. Така откриваме, че заболелите от диабет имат статистически значима разлика в изследванията си за кръвна захар не само спрямо контролната група на здравите, но и спрямо останалите групи.

4. Има ли връзка между серумните нива на гл/коза и холестерол в глупата на клинично здравите?

```
> cor(groups.group1$GLUC.mmol.l, groups.group1$CHOL.mmol.l, method = "spearman")
[1] -0.05635747
```

От това можем да заключим, че кимаме слаба отрицателна корелация между двете величини.

5. Има ли връзка между серумните нива на гл/коза и холестерол в глупата на диабетиците?

```
> cor(groups.group4$GLUC.mmol.l, groups.group4$CHOL.mmol.l, method = "spearman")
[1] -0.07317209
```

От това можем да заключим, че имаме слаба отрицателна корелация между двете величини.

6. Има ли връзка между серумните нива на триглицеридите и глюкозата в глупата на клинично здравите?

```
> cor(groups.group2$GLUC.mmol.l, groups.group2$Tg.mmol.l, method = "spearman")
[1] -0.09456148
```

От това можем да заключим, че имаме слаба отрицателна корелация между двете величини.

7. Има ли връзка между серумните нива на триглицеридите и глюкозата в глупата на жените с новодиагностициран ПКОС ?

```
> cor(groups.group1$GLUC.mmol.l, groups.group1$Tg.mmol.l, method = "spearman")
[1] 0.1631255
```

От това можем да заключим, че корелацията е слаба.

8. Има ли връзка между серумните нива на триглицеридите и холестерола в глупата на клинично здравите ?

```
> cor(groups.group1$CHOL.mmol.l, groups.group1$Tg.mmol.l, method = "spearman")
[1] 0.3254389
```

От това можем да заключим, че корелацията е умерена.

9. Има ли връзка между триглицеридите и холестерола в глупата на жените с новодиагностициран ПКОС ?

```
> cor(groups.group2$CHOL.mmol.l, groups.group2$Tg.mmol.l, method = "spearman")
[1] 0.5453993
```

От това можем да заключим, че корелацията е значителна.