Лабораторная работа №1

Александрова А.А.

14 12 2020

# Раздел I.

В этом разделе необходимо собрать данные по социально-экономическим показателям в регионах РФ и изучить их взаимосвязи.

**Необходимо построить:**

* гистограммы на одном полотне, отразив теоретические плотности распределения;
* таблицу со статистикой теста Андерсона-Дарлинга на нормальность;
* точечные графики взаимного разброса с полупрозрачными маркерами;
* графическое представление корреляционной матрицы.

### Обозначения количественных показателей:

* *SrDD.2017* - Среднедушевые денежные доходы населения(строковая переменная)
* *VPR.2017* - ВРП на душу населения
* *ObRT.2017* - Оборотной торговли на душу населения
* *RKBS.2016* - Расходы консолидированных бюджетов субъектов Российской Федерации: на социальную политику
* *KMP.2017* - Число малых предприятий на 10000 человек населения

## Импорт данных:

Файл с данными содержит 94 строк и 8 столбцов.

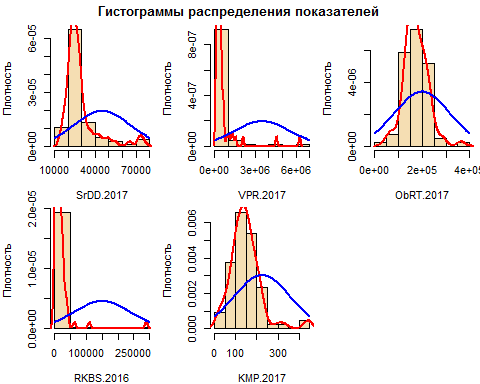
## Описательная статистика

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | SrDD.2017 | VPR.2017 | ObRT.2017 | RKBS.2016 | KMP.2017 |
| Среднее | 28973.9 | 545495.6 | 174486.8 | 19458.6 | 148.0 |
| Стандартное отклонение | 12077.5 | 837076.4 | 49804.8 | 33350.1 | 72.7 |
| Коэффициент вариации, % | 41.7 | 153.5 | 28.5 | 171.4 | 49.1 |

**Вывод:** показатели неоднородны.

## Анализ распределения данных

Построим гистограммы с теоретической нормальной плотностью распределения.



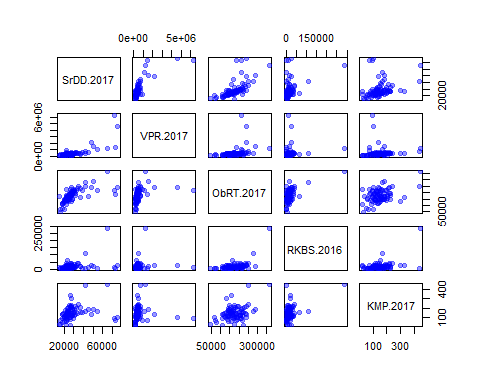
**Вывод:** распределения показателей ассиметричны по сравнению с нормальным законом(это видно у всех графиков кроме *“ObRT.2017”*) Кроме того, показатели имеют более островершинные распределения по сравнению с нормальным законом

## Тест Андерсона-Дарлинга на нормальность распределения

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | SrDD.2017.A | VPR.2017.A | ObRT.2017.A | RKBS.2016.A | KMP.2017.A |
| Статистика теста | 7.3505 | 16.9644 | 0.5570 | 13.8394 | 1.8153 |
| Критическое значение | 0.0000 | 0.0000 | 0.1461 | 0.0000 | 0.0001 |

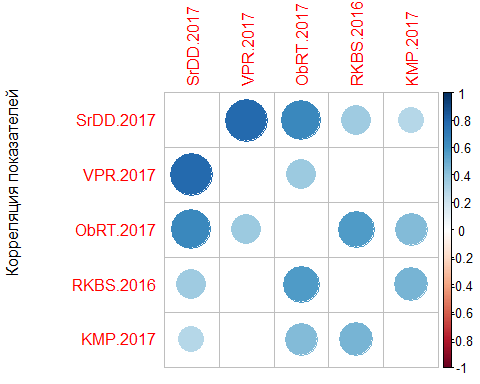
**Вывод:** Проверяя гипотезу нормального распределения для переменных на уровне значимости 0.05, мы отвергаем все кроме *“ObRT.2017”*, так как остальные p-значения < 0.05 .

## Графики разброса



**Вывод:** по первой строке графиков попытаемся определить наличие связи по первой строке графиков попытаемся определить наличие связи между *“SrDD.2017”* и остальными показателями. Вероятно, что связь отсутствует со следующими показателями: *“VPR.2017”* и *“RKBS.2016”*. С *“ObRT.2017”* и *“KMP.2017”* возможны слабые положительные линейные связи. Чтобы проверить эти предположения, рассчитаем корреляционную матрицу и проверим коэффициенты на значимость.

## Корреляционная матрица



**Вывод:** Взаимосвязь SrDD.2017 значима со всеми показателями, кроме *“KMP.2017”*

# Раздел II.

В этом разделе необходимо прологарифмировать данные из первого раздела и провести с новыми данными те же махинации.

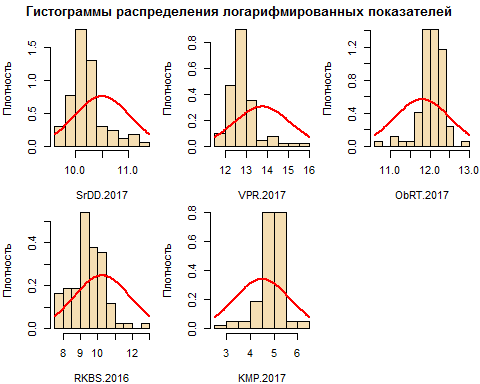
## Описательная статистика логарифмированных данных:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | SrDD.2017 | VPR.2017 | ObRT.2017 | RKBS.2016 | KMP.2017 |
| Среднее | 10.213 | 12.859 | 12.025 | 9.401 | 4.873 |
| Стандартное отклонение | 0.329 | 0.677 | 0.315 | 0.904 | 0.540 |
| Коэффициент вариации, % | 3.221 | 5.265 | 2.620 | 9.616 | 11.081 |

**Вывод:** показатели считаются однородными, т.к. коэффициенты вариации не превышают 12%.

## Анализ распределения данных

Построим гистограммы с их фактическими плотностями распределений и наложим сверху теоретическую нормальную плотность распределения.



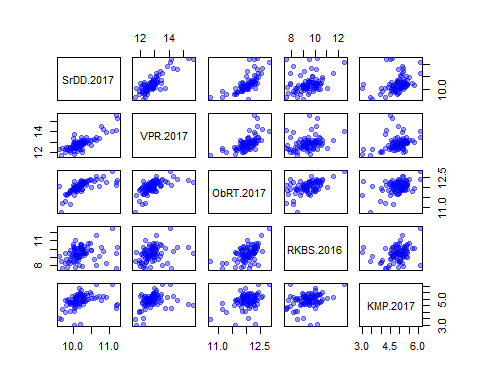
**Вывод:** Судя по гистограммам, распределения всех показателей ассиметричны по сравнению с нормальным законом.(менее заметно это у *“RKBS.2016”*) Кроме того, показатели имеют более островершинные распределения по сравнению с нормальным законом.

## Тест Андерсона-Дарлинга на нормальность распределения

|  |  |  |
| --- | --- | --- |
|  | p | W |
| SrDD.2017 | 0.0000 | 2.8509 |
| VPR.2017 | 0.0000 | 2.5482 |
| ObRT.2017 | 0.0020 | 1.3061 |
| RKBS.2016 | 0.3259 | 0.4159 |
| KMP.2017 | 0.0000 | 2.0376 |

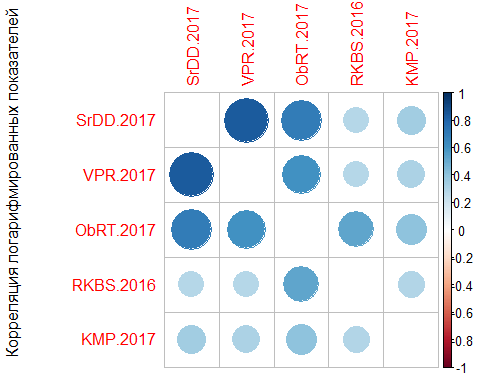
**Вывод:** Проверяя гипотезу нормального распределения для переменных на уровне значимости 0.05, мы отвергаем все, кроме *“RKBS.2016”*, так как остальные p-значения < 0.05 .

## Графики разброса



**Вывод:** по первой строке графиков попытаемся определить наличие связи между *“SrDD.2017”* и остальными показателями. Можно предположить, что со всеми показателями существует слабая положительная связь. А с *“RKBS.2016”* нелинейная. Чтобы проверить эти предположения, расчитаем корреляционную матрицу и проверим коэффициенты на значимость.

## Корреляционная матрица



**Вывод:** Взаимосвязь SRDD.2017 значима c VRP.2017, ObRT.2016. Корреляция между этими факторами стала сильнее. И между другими установилась взаимосвязь