Проверка статистических гипотез

Часть 3

Аббакумов Вадим Леонардович

Версия 4

Раздел 3

Популярные статистические критерии

Гипотеза о нормальности распределения случайной величины

- Гипотеза: Случайная величина имеет нормальное распределение, значения параметров распределения могут быть любыми.
- Конкурирующая гипотеза: Распределение случайной величины отличается от нормального.

Литература

Thode

Testing For Normality

CRC Press 2002   368с

Не дает рекомендаций по выбору критерия.

Критерий Шапиро-Уилка

- Критерий Шапиро-Уилка
- scipy.stats.shapiro   # В Python
- shapiro.test(data)   # В R

Другие критерии в Python

Критерий Anderson-Darling

scipy.stats.anderson

Критерий Lilliefors (Kolmogorov-Smirnov)

statsmodels.stats.diagnostic.lilliefors

В языке R пакет "nortest"

Критерий Anderson-Darling

```
ad.test(data)
```

Критерий Lilliefors (Kolmogorov-Smirnov)

```
lillie.test(x)
```

Число наблюдений

Если меньше 2000 (?) наблюдений, рекомендуется использовать критерий Шапиро-Уилка

если больше 2000 (?), то критерий Колмогорова-Смирнова.

А нужно ли проверять гипотезу нормальности?

- Методы, которые рассматриваются в курсе далее, работают когда
  - переменная имеет нормальное распределение
  - распределение несущественно отличается от нормального

гипотеза о нормальности распределения переменной отвергнута

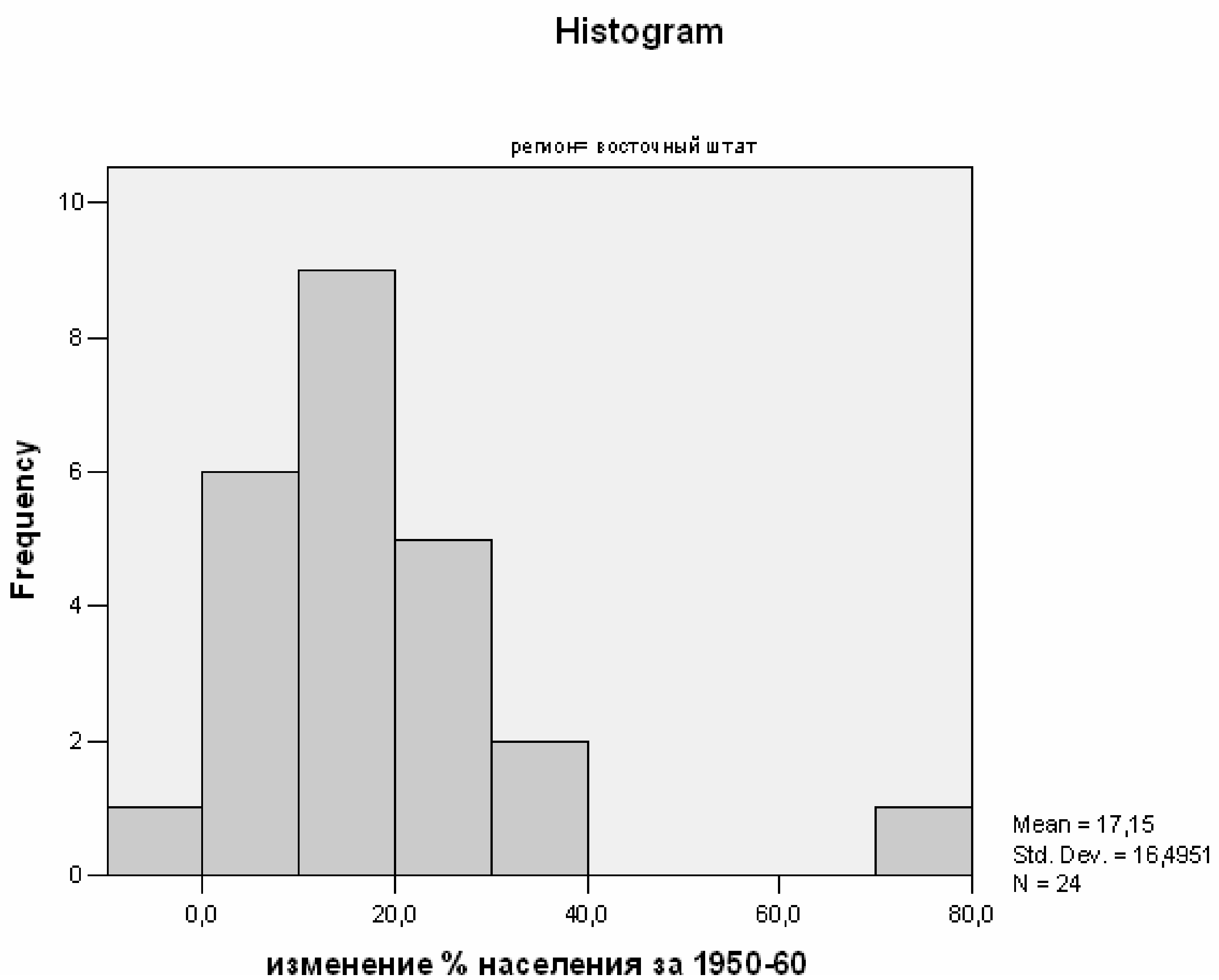
- Некоторые отклонения от нормальности существенные.
- Некоторые отклонения от нормальности несущественные.

Существенные отклонения

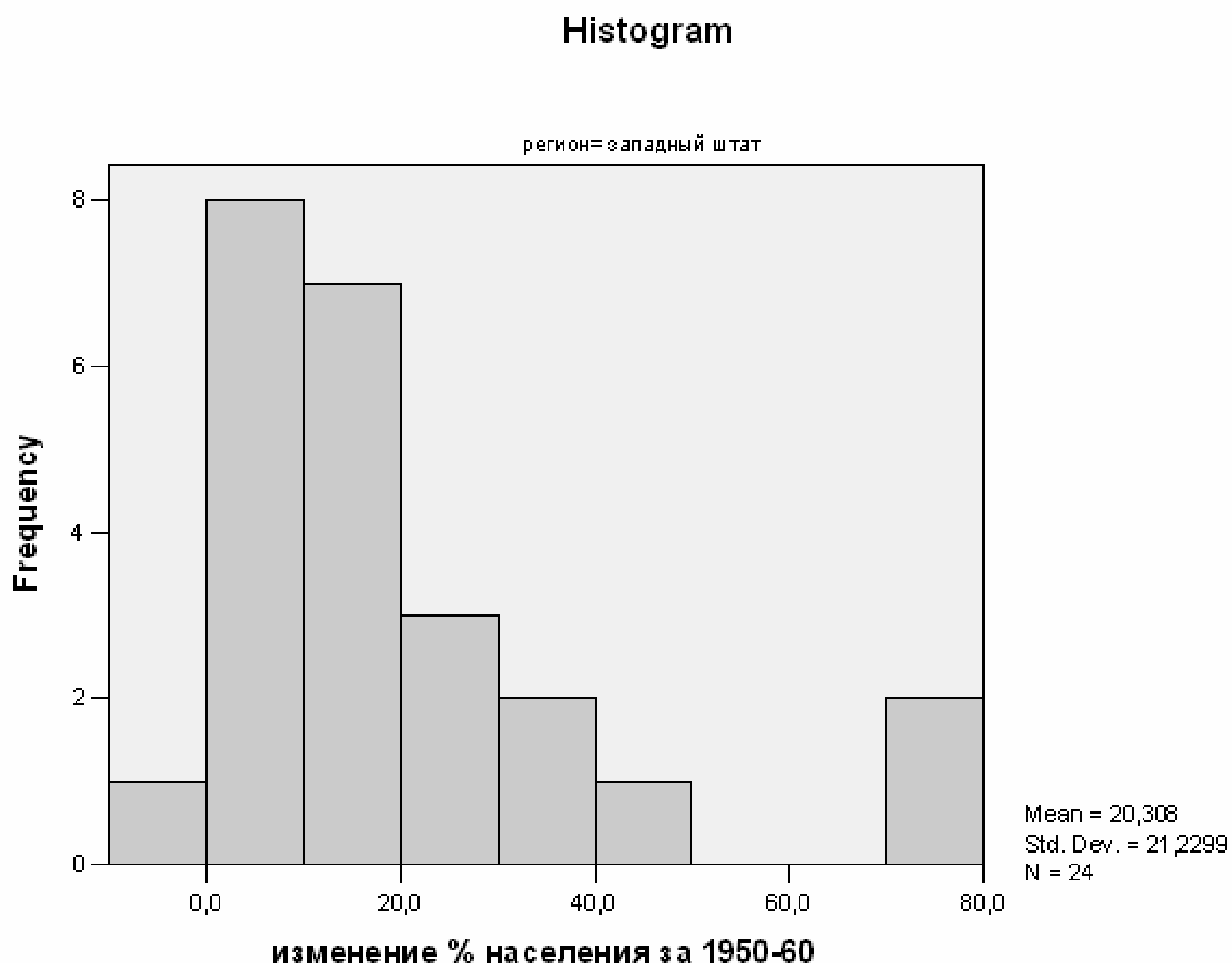
1. Наличие выбросов в данных.
2. Явная асимметрия гистограммы.
3. Очень сильное отклонение формы гистограммы от колоколообразной формы.

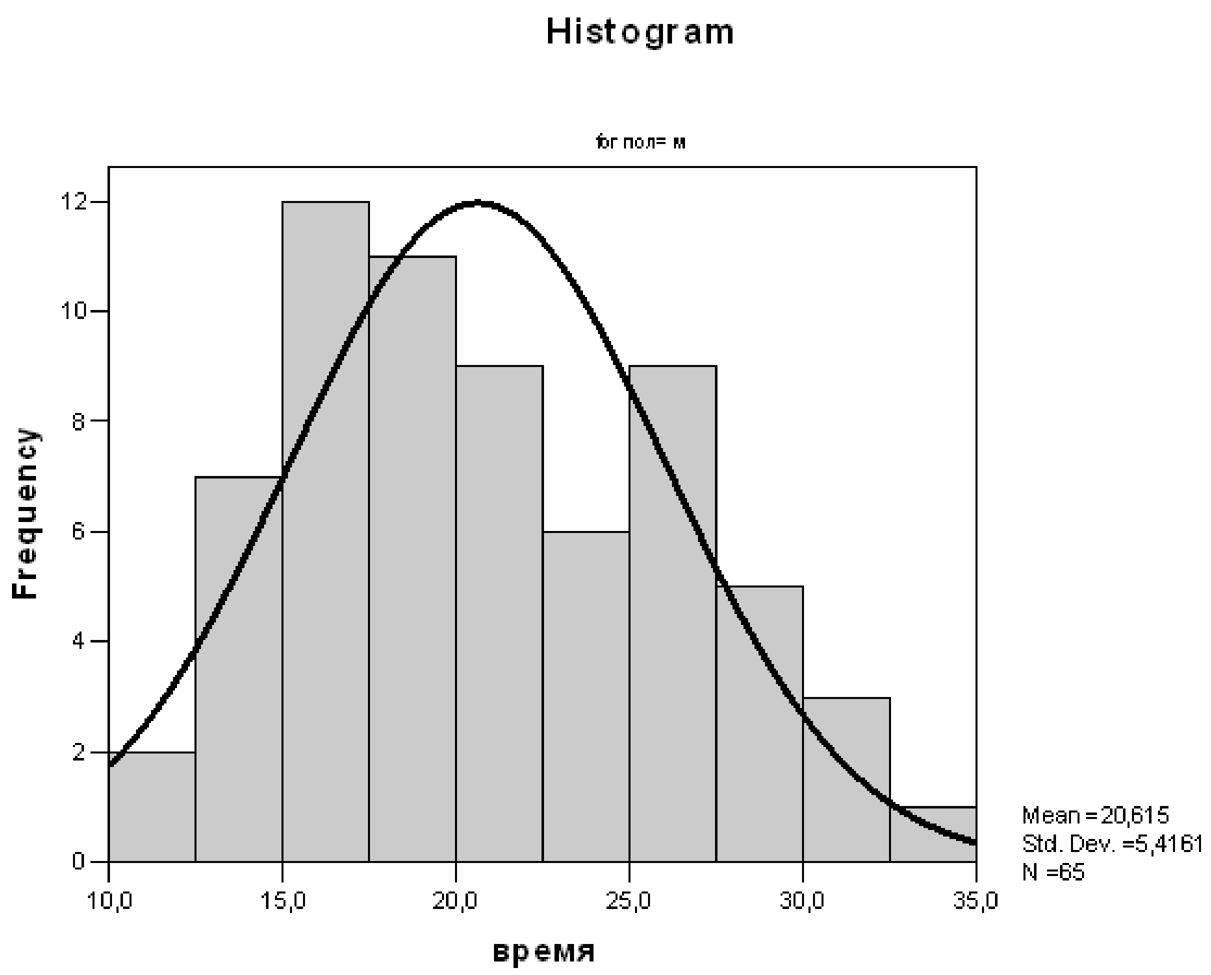
Рекомендуется относиться

- Строго - к присутствию выбросов,
- Снисходительно - к отклонениям от симметрии.
- Отношение к колоколообразной форме гистограммы зависит от числа наблюдений. Если меньше 30 наблюдений, наше отношение в высшей степени либерально, если число наблюдений находится между 30 и 150, мы относимся к отклонениям снисходительно, если имеется больше 150 наблюдений – строго.



Histogram
регион= восточный штат
10
8
6
4
2
0
Frequency
0,0
20,0
40,0
60,0
80,0
изменение % населения за 1950-60
Mean = 17,15
Std. Dev. = 16,4951
N = 24


Histogram
регион= западный штат
Frequency
8
6
4
2
0
0,0
20,0
40,0
60,0
80,0
изменение % населения за 1950-60
Mean = 20,308
Std. Dev. = 21,2299
N = 24



Histogram
for пол= м
Frequency
12
10
8
6
4
2
0
10,0
15,0
20,0
25,0
30,0
35,0
время
Mean =20,615
Std. Dev. =5,4161
N =65

Лекарства
Иногда они опаснее болезни...

- Выбросы — удаляем (осторожно!)
- Асимметрия — преобразуем данные (например, логарифмируем, или преобразование Бокса-Кокса)
- Бимодальность — разбиваем выборку на подвыборки

Пример 1

- Население городов России в 1959 году

- Исходные данные
- Логарифм населения

Пример 2

- Альбукерк – продажи домов

Гипотезы о типичных наблюдениях (о центрах распределений)

Сравнение центров распределений

- *Центр распределения* - то одно единственное число, которое описывает, характеризует выборку.
- В качестве центра чаще всего используют среднее арифметическое, медиану или усеченное среднее.
- Центр распределения и типичное наблюдение синонимы

Среднее или медиана?

- у человека в среднем меньше двух глаз

- x1 <- c(1, 1, 1, 1000, 1000)
  x2 <- c(2, 3, 4, 5, 6)
  Доходы x1 выше, x2 прибыльнее чаще

Другие методы оценки центра распределения

Andrews; Bickel; Hampel; Huber; Rogers, Tukey.

Robust estimates of location: survey and advances.

1972 Princeton University Press

Среднее арифметическое или медиана?

- Если распределение хотя бы одной из выборок существенно отличается от нормального, в качестве центра предлагается использовать медиану.
- В остальных случаях, то есть если распределение каждой выборки можно считать нормальным или несущественно отличающимся от нормального, в качестве центра предлагается использовать среднее арифметическое.

Выбор центра распределения

- Если центр распределения медиана, используем критерий Манна – Уитни-Вилкоксона или (редко) Mood's median критерий .
- Если центр распределения среднее арифметическое, используем одну из версий критерия Стьюдента.

Прагматичный подход

Применить оба теста.
Если выводы совпадают, ответ есть

Если выводы различны, начинаем разбираться.

Парные выборки

- В случае парных выборок имеются пары наблюдений (измерений) одного и того же объекта.
- Например: пары измерений делались в один и тот же момент.
- Например: два измерения одной переменной делались последовательно.

Независимые выборки

- В случае *независимых выборок* каждое наблюдение соответствует отдельному объекту, измеряются разные объекты.

Принадлежность объектов выборкам определяется по значениям дополнительной группирующей переменной.

В языке R и Pyton Независимые и парные выборки

- В питоне используем разные версии критериев

- Если выборки парные, используется опция paired = TRUE.
- Если выборки независимые, используется опция paired = FALSE.

Сравнение средних значений выборок

- H_0: Математические ожидания равны.
  - $EX = EY$
- H_1: Математические ожидания различны.
  - $EX \neq EY$

Т-критерий Стьюдента

В Python

scipy.stats.ttest_ind

scipy.stats.ttest_rel

В R

t.test(x, y, alternative = "two.sided", paired = FALSE, var.equal = FALSE)

Надо еще сравнить дисперсии - 1

Метод 1 — устарел, но еще встречается

F-test of equality of variances

Не рекомендуется, слишком чувствителен к отклонениям от нормальности. См.

http://en.wikipedia.org/wiki/F-test_of_equality_of_variances

В R

var.test(x, y)

Надо еще сравнить дисперсии - 2

Метод 2 — устарел, но еще встречается

Bartlett's test

Если данные нормально распределены, лучший вариант.

Не рекомендуется: чувствителен к отклонениям от нормальности;

Если данные не нормальны, часто дает "false positive" результат.

Надо еще сравнить дисперсии - 2

Метод 2

Bartlett's test

В R

```
bartlett.test(x, g, data=data.table)
bartlett.test(x~g, data=data.table)
```

Надо еще сравнить дисперсии - 3

- *Метод 3— устарел, но еще встречается*

- Levene's test
- Критерий Ливиня/Левена

Надо еще сравнить дисперсии - 3

- В R Levene's test

```
library(car)

leveneTest(x~g, data=data.table)
```

Надо еще сравнить дисперсии - 4

Fligner-Killeen test

Робастный, рекомендуется.

возможно Brown-Forsythe test еще лучше...

Надо еще сравнить дисперсии - 4

В Python

scipy.stats.fligner

В R

fligner.test(x~g, data=data.table)

Можно поступать лучше, но нельзя

В. В. Славова, Д. М. Чибисов, “О предварительном тестировании в задаче Беренса–Фишера”, Обозрение прикладной и промышленной математики, 16:2 (2009), 210–225

Если данные существенно отклоняются от нормальности

Но нужно сравнить средние

1. Все как в случае нормальности, если выборка большая (n> 300)
2. Преобразовать данные к нормальному виду

Анализ мощности

Можно начать изучение со статьи

A Gentle Introduction to Statistical Power and Power Analysis in Python

https://machinelearningmastery.com/statistical-power-and-power-analysis-in-python/

Сравнение медиан выборок

H_0: Медианы равны

H_1: Медианы различаются

Mood's median test в R

```
m <- median(c(x1,x2))     # joint median
f11 <- sum(x1>m)          # Pop.1 samples above median
f12 <- sum(x2>m)
f21 <- sum(x1<=m)         # Pop.1 samples below or at
   median
f22 <- sum(x2<=m)
# 2x2 contingency table
table <- matrix(c(f11,f12,f21,f22), nrow=2,ncol=2)
chisq.test(table)
```

Mood's median test в Python

scipy.stats.median_test

Mood's median test

Friedlin, B. & Gastwirth, J. L. (2000).

Should the median test be retired from general use?

The American Statistician, 54, 161–164.

Не рекомендуют. Большая ошибка 2 рода для малых выборок (по сравнению с другими тестами)

Можно применять только для больших выборок 3000+ наблюдений

Критерий Манна-Уитни

Mann-Whitney U-test

Mann–Whitney–Wilcoxon,

Wilcoxon rank-sum test,

Wilcoxon–Mann–Whitney test

Важно!

- Критерий Манна-Уитни на самом деле проверяет другую гипотезу.
- Имеются две выборки наблюдений случайных величин X и Y.
- H_0: $P\{X>Y\}=P\{X<Y\}$.
- H_1: $P\{X>Y\} \neq P\{X<Y\}$.

Goos, Meintrup
Statistics with JMP: Hypothesis Tests, Anova and Regression 2016

9 A Nonparametric Hypothesis Test for the Medians of Two Independent Samples

… The recommended test in that case is the Wilcoxon rank-sum test. ...

The Wilcoxon rank-sum test is also called the Mann–Whitney U-test.

9.1 The Hypotheses Tested

The rank-sum test is designed to determine whether the medians of two populations are equal. The test’s null hypothesis is

H0: Med 1 = Med 2

where Med 1 and Med 2 represent the medians of the first population and the second population, respectively.

Критерий Манна-Уитни не учитывает величины наблюдений, а только их ранги

Статистика критерия Манна-Уитни U

$$U1 = n1*n2 + \{n1 * (n1 + 1)/2\} — T1$$

$$U2 = n1*n2 + \{n2 * (n2 + 1)/2\} — T2$$

$$U = \min(U1, U2)$$

Ti — сумма рангов в объединенной выборке наблюдений из выборки i

n1 и n2 — размеры выборок

Статистика критерия Манна-Уитни вариант объяснения идеи метода

Обозначим одну выборку x, другую y.

Для каждого наблюдения из выборки x сосчитаем число тех наблюдений в выборке y, которые меньше его. (пока считаем, что совпадений нет).

Сложим все полученные числа.

Тогда причем тут медианы? Корректный ответ.

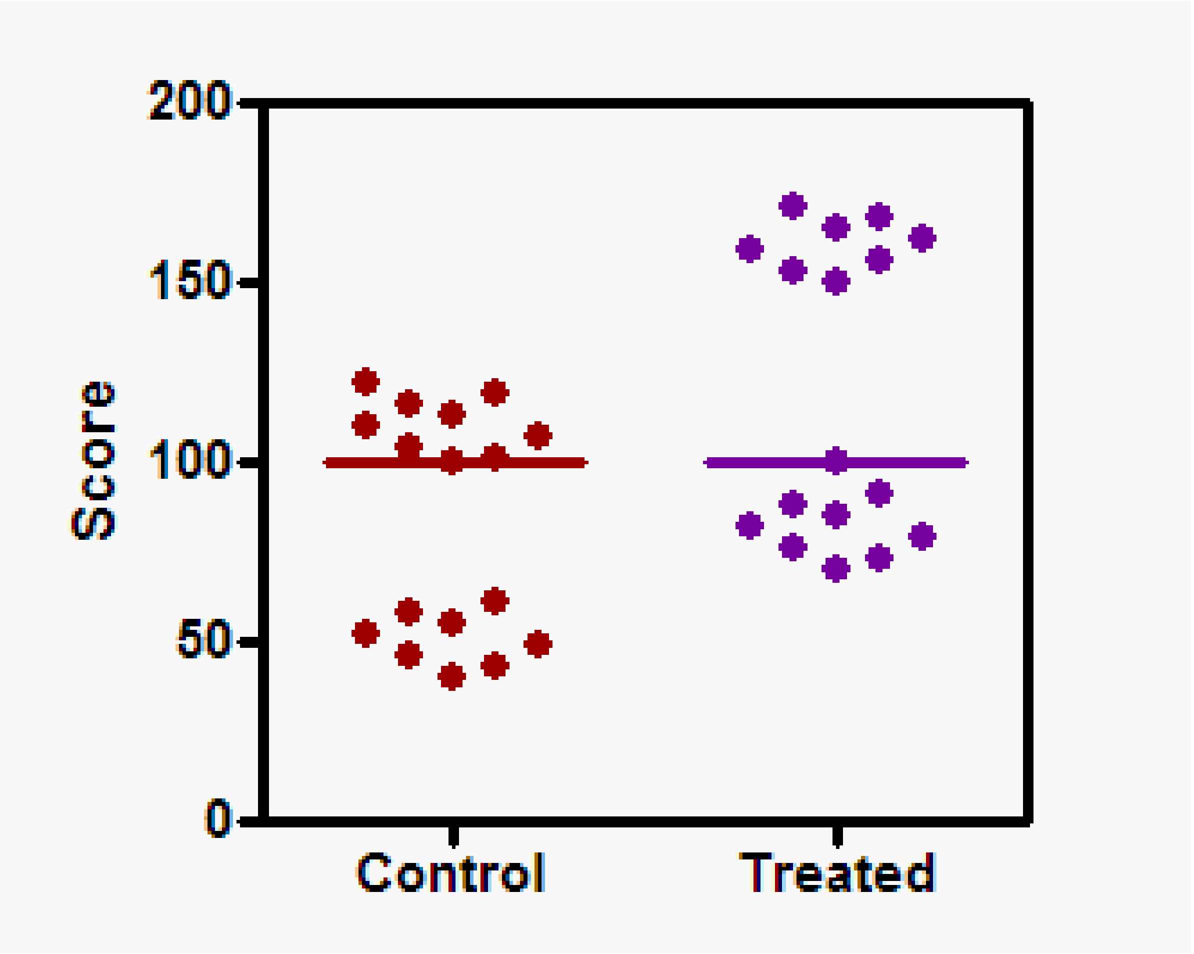
Дополнительные предположения

- Наблюдения случайных величин могут принимать любые значения в интервале
- Альтернативная гипотеза сдвига, т.е.

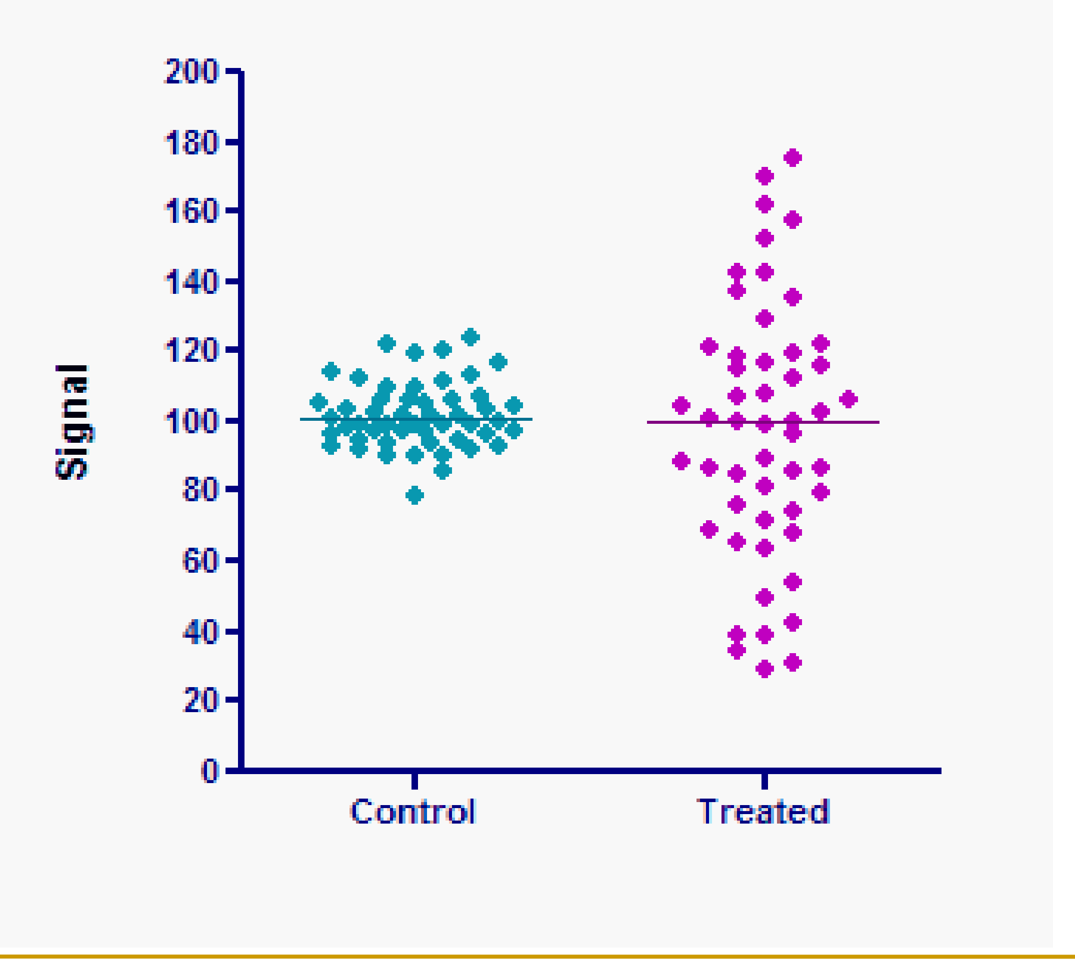
$$F_Y(t) = F_X(t + \Delta)$$

- Если критерий MWW отвергает гипотезу, то медианы различны

Гипотеза отвергается: p=0.0288



Гипотеза не отвергается: p=0.46



Критерий
Манна-Уитни-Вилкоксона в R

```
wilcox.test(x, y,
     alternative = "two.sided",
     paired = FALSE,
     exact = TRUE,
     correct = FALSE)
```

Критерий Манна-Уитни-Вилкоксона в Pyton

scipy.stats.mannwhitneyu

scipy.stats.wilcoxon

Примеры

- Время в магазинах
- Альбукерк
- Обучение менеджеров

Еще раз, кратко

Если оба распределения несущественно отличаются от нормального, то используем критерий Стьюдента

Если сравниваем медианы, и много наблюдений , то используем критерий Муда

Если сравниваем медианы и распределения одинаковы по форме, то используем критерий Манна-Уитни

Если сравниваем медианы и распределения НЕ одинаковы по форме, иногда нельзя применять критерий Манна-Уитни,

x1 <- c(1, 1, 1, 1000, 1000)

x2 <- c(2, 3, 4, 5, 6)

Доходы x1 выше, x2 прибыльнее чаще