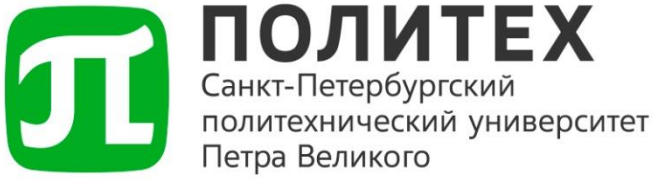
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ

ВЫСШЕГО ОБРАЗОВАНИЯ
«САНКТ-ПЕТЕРБУРГСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ПЕТРА ВЕЛИКОГО»
Институт компьютерных наук и технологий
Высшая школа искусственного интеллекта



ЛАБОРАТОРНАЯ РАБОТА №2

по дисциплине «Элементы теории вероятности и линейной алгебры»

Вариант №11

Выполнил
Студент 3540201/10301 группы Ф.М. Титов

Руководитель
доцент, к.т.н. А.В. Востров

Санкт-Петербург
2021 г.

Оглавление

Постановка задачи

Смоделировать две нормальные выборки со следующими параметрами: mx=3, my = mx + 1.5, Dx = 11, Dy = Dx + 3, объем первой выборки 50, объем второй выборки 100. Не засоряя первую выборку, проверить по вашему выбору все шесть описанных в подразд. 5.3 и 5.4 гипотез, приняв уровень значимости 0.1.

Теоретическая часть

Для проверки статистической гипотезы с использованием уровня значимости используют понятия нулевой и альтернативной гипотез. Проверяемая гипотеза называется нулевой гипотезой и обозначается чаще всего H0. Альтернативная (конкурирующая) гипотеза обозначается H1. Правило, по которому принимается решение о нулевой гипотезе, называется критерием. Все решения принимаются на основе выборки, следовательно, на основе какой-нибудь статистики. Эта статистика называется статистикой z критерия.

Далее выбирается уровень вероятности α, $\alpha > 0$, который называется уровнем значимости. Событие считается практически невозможным, если его вероятность меньше α.

Принцип принятия и отклонения гипотезы представлен на рис. 1-2. $z_в$ – выборочное значение статистики z. Тогда критерий формулируется следующим образом: гипотеза H0 отклоняется при $z_в \in \omega$ и принимается, если $z_в \in W \setminus \omega$. Множество ω всех значений статистики z, при которых гипотеза H0 отклоняется, называется критической областью; область $W \setminus \omega$ называется областью принятия решения. Границей между областями ω и $W \setminus \omega$ является критическая точка z_α. Уровень значимости α, таким образом, задает размеры критической области. Кроме того, эти размеры зависят область принятия вер- область отклонения еще и от формулировки альтернативной гипотезы. Критерий, основанный на использовании заранее заданного уровня значимости, называется критерием значимости.

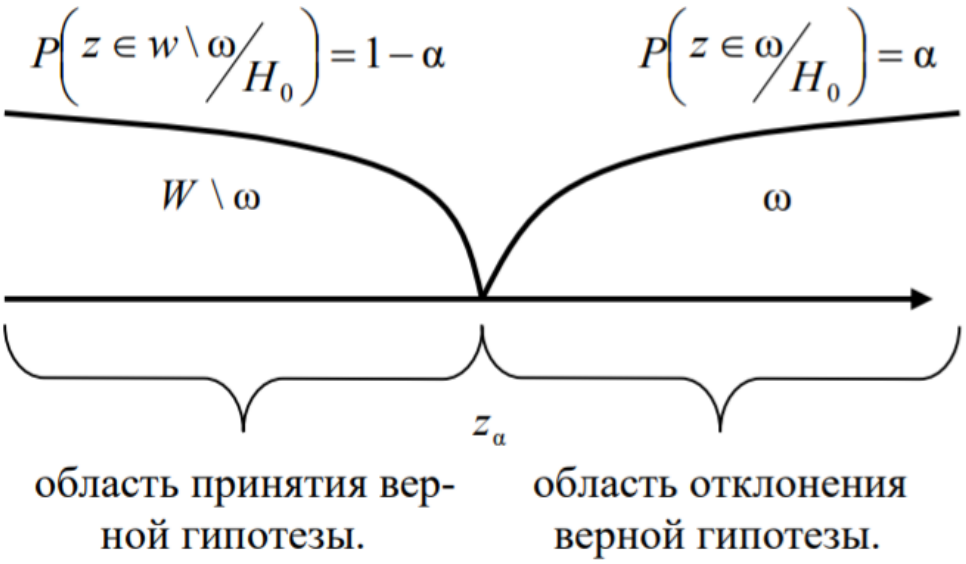


Рис. 1. Принцип принятия и отклонения гипотез.

Решение, принимаемое о гипотезе H_0 по выборке.	H_0 отвергается, H_1 принимается.	H_0 принимается, H_1 отвергается.
Нулевая гипотеза H_0 - верна.	$P\left(z \in \omega / H_0\right) = \alpha$ - отвергается верная гипотеза. Ошибка первого рода.	$P\left(z \in w \setminus \omega / H_0\right) = 1 - \alpha$ - принимается верная гипотеза.
Гипотеза H_0 неверна, то есть верна гипотеза H_1.	$P\left(z \in \omega / H_1\right) = 1 - \beta$ - отвергается неверная гипотеза. Мощность критерия.	$P\left(z \in w \setminus \omega / H_1\right) = \beta$ - принимается неверная гипотеза. Ошибка второго рода.

Рис. 2. Принятие и отклонение гипотез.

Уровень значимости α – вероятность допустить ошибку первого рода (отвергнуть верную гипотезу H0). Нулевая гипотеза отвергается, если вероятность меньше альфа.

β - вероятность сделать ошибку второго рода (принять неверную H0). 1-β – мощность критерия. Мощность – вероятность отклонения ложной нулевой гипотезы.

Проверяемые гипотезы:

1. Гипотеза о числовом значении математического ожидания нормального распределения при известной дисперсии.
   Нулевая гипотеза здесь H0 : mX = a.
   $$H_0 : m_X = a_0$$
   $$z = \frac{m_X - a_0}{\sqrt{\frac{D_X}{n}}}$$

2. Гипотеза о числовом значении математического ожидания нормального распределения при неизвестной дисперсии.
   $$H_0 : m_X = a_0$$
   $$z = \frac{m_X - a_0}{\sqrt{\frac{\hat{D}_X}{n}}}$$

3. Гипотеза о числовом значении дисперсии нормального распределения.

$$H_0 : D_X = D_0$$

$$z = \frac{\widehat{D}_X (n-1)}{D_0}$$

4. Гипотеза о равенстве математических ожиданий двух нормальных распределений с известными дисперсиями.

$$H_0 : m_X = m_Y$$

$$z = \frac{\left(m_X^* - m_Y^*\right)}{\sqrt{\frac{D_X}{n_1} + \frac{D_Y}{n_2}}} \in N(0,1).$$

5. Гипотеза о равенстве математических ожиданий двух нормальных распределений с неизвестными, но равными дисперсиями.

$$H_0 : m_X = m_Y$$

$$z = \frac{\left(m_X^* - m_Y^*\right) - \left(m_X - m_Y\right)}{\sqrt{\frac{D}{n_1} + \frac{D}{n_2}}}$$

6. Гипотеза о равенстве дисперсий двух нормальных распределений.

$$H_0 : D_X = D_Y$$

$$z = \frac{\widehat{D}_X}{\widehat{D}_Y} \in F_{n_1-1, n_2-1}$$

Реализация

В ходе работы были смоделированы две нормальные выборки с необходимыми параметрами с помощью программного модуля Python numpy.random (рис.3).

```
N_X = 50
N_Y = 100
MX = 3
MY = MX + 1.5
DX = 11
DY = DX + 3
alpha = 0.1

sigma_x = math.sqrt(DX)
sigma_y = math.sqrt(DY)

norm_dist_x = np.random.normal(MX, sigma_x, N_X)
norm_dist_y = np.random.normal(MY, sigma_y, N_Y)
```

Рис. 3. Моделирование нормальных выборок.

Далее была реализована вспомогательная функция plot_histogram, строящая гистограмму выборки (рис. 4). Для этого (и для всех других графиков) был использован модуль Python matplotlib. На вход функция принимает массив сгенерированных псевдослучайных чисел, мат.ожидание и стандартное отклонение.

```
def plot_histogram(dist_arr, mu, sigma):
    count, bins, ignored = plt.hist(dist_arr, density=True)
    plt.plot(bins, 1 / (sigma * np.sqrt(2 * np.pi)) *
             np.exp(- (bins - mu) ** 2 / (2 * sigma ** 2)),
             linewidth=2, color='r')
    plt.grid()
    plt.show()
```

Рис. 4. Листинг функции plot_histogram.

Далее для проверки каждой была реализована отдельная функция. Все функции схожи по реализации, на вход принимают массив сгенерированной выборки для нормального распределения, мат.ожидание, стандартное отклонение, а также параметр α.

Для проверки гипотезы «о числовом значении математического ожидания нормального распределения при известной дисперсии» была написана функция check_H1 (рис.5). В ней используется функция norm из пакета Python stats.

Для проверки гипотезы «о числовом значении математического ожидания нормального распределения при неизвестной дисперсии» была

написана функция check_H2 (рис.6). В ней используется функция t из пакета Python stats.

```
def check_H1(dist_arr, mu, sigma, alpha):
    z_right = stats.norm.ppf(1 - alpha / 2)
    z_left = stats.norm.ppf(alpha / 2)

    sample_mean = s.mean(dist_arr)

    z = (sample_mean - mu) / (sigma / math.sqrt(len(dist_arr)))

    print(z_left, z_right)
    print(z)

    return (z < z_right) * (z > z_left)
```

Рис. 5. Листинг функции check_H1.

```
def check_H2(dist_arr, mu, alpha):
    N = len(dist_arr)
    z_right = stats.t.ppf(1 - alpha / 2, N - 1)
    z_left = stats.t.ppf(alpha / 2, N - 1)

    sample_mean = s.mean(dist_arr)
    sample_variance = s.variance(dist_arr, sample_mean)

    z = (sample_mean - mu) / math.sqrt(sample_variance / N)

    print(z_left, z_right)
    print(z)

    return (z < z_right) * (z > z_left)
```

Рис. 6. Листинг функции check_H2.

Для проверки данной гипотезы «о числовом значении дисперсии нормального распределения» была написана функция check_H3 (рис.7). При поисках границ используется Хи квадрат. Соответствующая функция (chi2) импортируется из модуля Python stats.

```
def check_H3(dist_arr, sigma, alpha):
    N = len(dist_arr)
    z_right = stats.chi2.ppf(1 - alpha / 2, N - 1)
    z_left = stats.chi2.ppf(alpha / 2, N - 1)

    sample_mean = s.mean(dist_arr)
    sample_variance = s.variance(dist_arr, sample_mean)

    z = ((N - 1) * sample_variance) / sigma ** 2

    print(z_left, z_right)
    print(z)

    return (z < z_right) * (z > z_left)
```

Рис. 7. Листинг функции check_H3.

Для проверки гипотезы «о равенстве математических ожиданий двух нормальных распределений с известными дисперсиями» была написана функция check_H4 (рис.8).

```python
def check_H4(dist_arr_x, dist_arr_y, dx, dy, alpha):
    n_x = len(dist_arr_x)
    n_y = len(dist_arr_y)

    z_right = stats.norm.ppf(1 - alpha / 2)
    z_left = stats.norm.ppf(alpha / 2)

    sample_mean_x = s.mean(dist_arr_x)
    sample_mean_y = s.mean(dist_arr_y)

    z = (sample_mean_x - sample_mean_y) / math.sqrt((dx / n_x) + (dy / n_y))

    print(z_left, z_right)
    print(z)

    return (z < z_right) * (z > z_left)
```

Рис. 8. Листинг функции check_H4.

Для проверки гипотезы «о равенстве математических ожиданий двух нормальных распределений с неизвестными, но равными дисперсиями» была написана функция check_H5 (рис.9). Для поиска границ используется t функция.

Для проверки гипотезы «о равенстве дисперсий двух нормальных распределений» была написана функция check_H6 (рис.10).

```python
def check_H5(dist_arr_x, dist_arr_y, alpha):
    n_x = len(dist_arr_x)
    n_y = len(dist_arr_y)

    z_right = stats.t.ppf(1 - alpha / 2, n_x + n_y - 2)
    z_left = stats.t.ppf(alpha / 2, n_x + n_y - 2)

    sample_mean_x = s.mean(dist_arr_x)
    sample_mean_y = s.mean(dist_arr_y)

    sample_variance_x = s.variance(dist_arr_x, sample_mean_x)
    sample_variance_y = s.variance(dist_arr_y, sample_mean_y)

    nominator = sample_mean_x - sample_mean_y
    denominator = math.sqrt(
        ((1 / n_x) + (1 / n_y)) * ((sample_variance_x * (n_x - 1) + sample_variance_y * (n_y - 1)) / (n_x + n_y - 2)))
    z = nominator / denominator

    print(z_left, z_right)
    print(z)

    return (z < z_right) * (z > z_left)
```

Рис. 9. Листинг функции check_H5.

```python
def check_H6(dist_arr_x, dist_arr_y, alpha):
    n_x = len(dist_arr_x)
    n_y = len(dist_arr_y)

    z_right = stats.f.ppf(1 - alpha / 2, n_x - 1, n_y - 1)
    z_left = stats.f.ppf(alpha / 2, n_x - 1, n_y - 1)

    sample_mean_x = s.mean(dist_arr_x)
    sample_mean_y = s.mean(dist_arr_y)

    sample_variance_x = s.variance(dist_arr_x, sample_mean_x)
    sample_variance_y = s.variance(dist_arr_y, sample_mean_y)

    z = sample_variance_x / sample_variance_y

    print(z_left, z_right)
    print(z)

    return (z < z_right) * (z > z_left)
```

Рис. 10. Листинг функции check_H6.

Результаты

Для двух выборок были построены гистограммы (рис. 11-12).

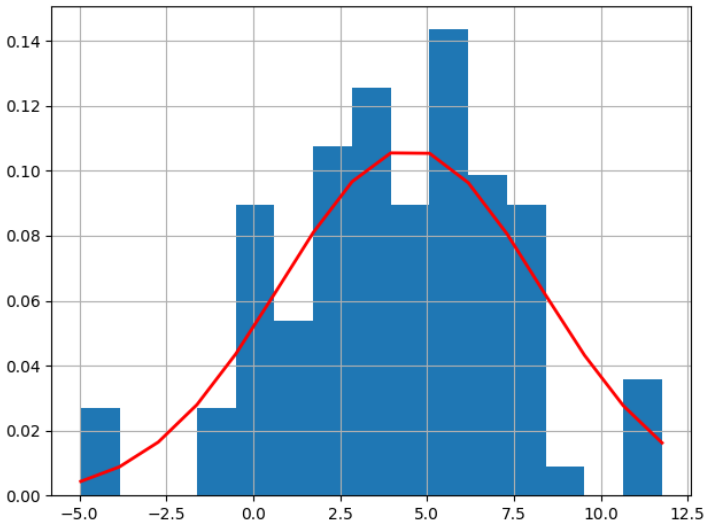


Рис. 11. Гистограмма для первой выборки n=50.

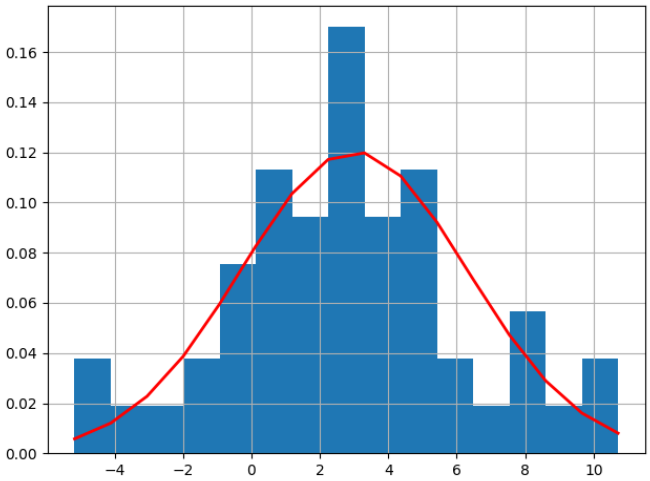


Рис. 12. Гистограмма для второй выборки n=100.

Результат проверки первой гипотезы представлен на рис. 13. Все четыре раза гипотеза была принята.

```
-1.6448536269514729 1.6448536269514722
1.3275551268233108
Гипотеза принята
----------------------------------------
-1.6448536269514729 1.6448536269514722
1.193536730973111
Гипотеза принята
----------------------------------------
-1.6448536269514729 1.6448536269514722
-0.7193704287651245
Гипотеза принята
----------------------------------------
-1.6448536269514729 1.6448536269514722
1.1956787436351997
Гипотеза принята
----------------------------------------
```

Рис. 13. Результат проверки гипотезы №1.

Результат проверки второй гипотезы представлен на рис. 14. Все четыре раза гипотеза была принята.

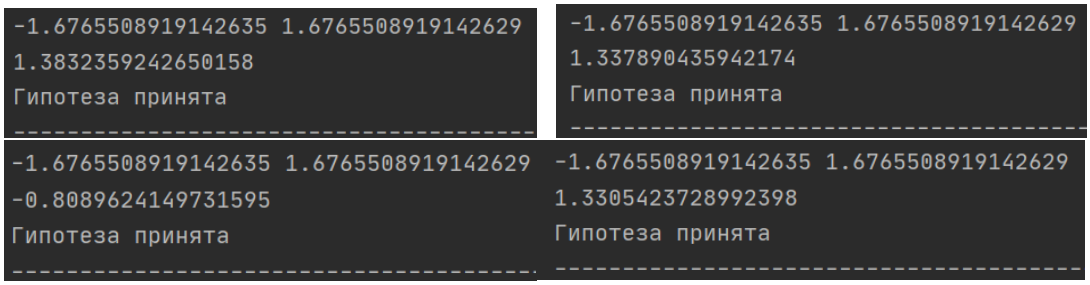


Рис. 14. Результат проверки гипотезы №2.

Результат проверки третьей гипотезы представлен на рис. 15. Все четыре раза гипотеза была принята.

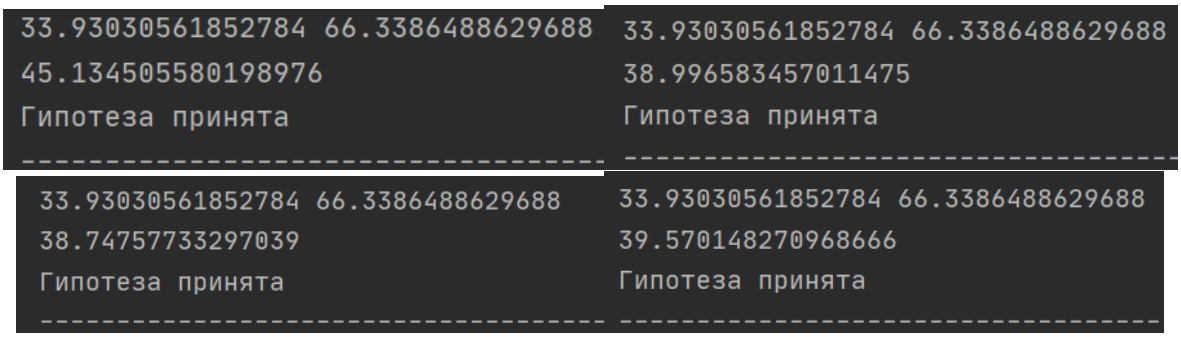


Рис. 15. Результат проверки гипотезы №3.

Результат проверки четвертой гипотезы представлен на рис. 16. Один раз гипотеза была принята, три – отвергнута.

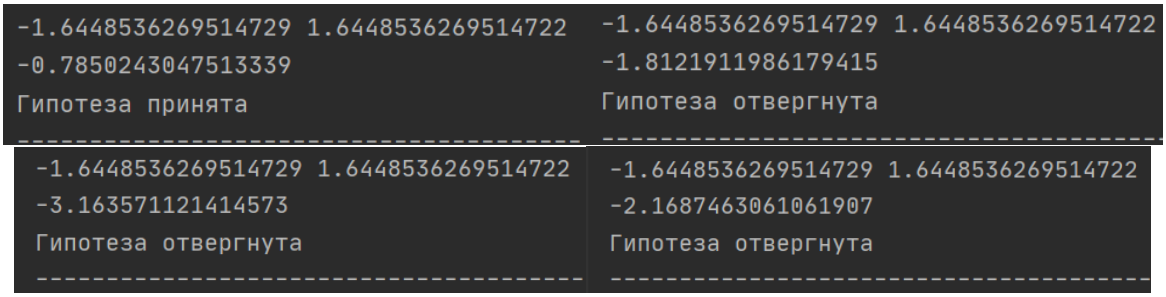


Рис. 16. Результат проверки гипотезы №4.

Результат проверки пятой гипотезы представлен на рис. 17. Один раз гипотеза была принята, три – отвергнута.

Результат проверки шестой гипотезы представлен на рис. 18. Один раз гипотеза была принята, три – отвергнута.

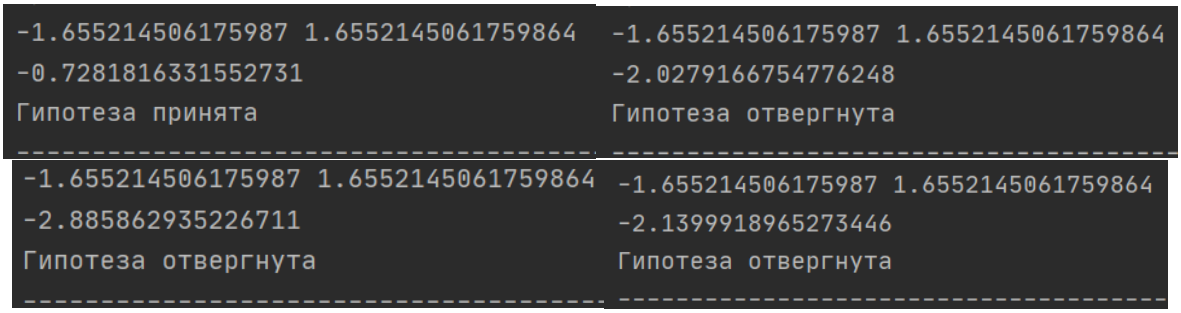


Рис. 17. Результат проверки гипотезы №5.

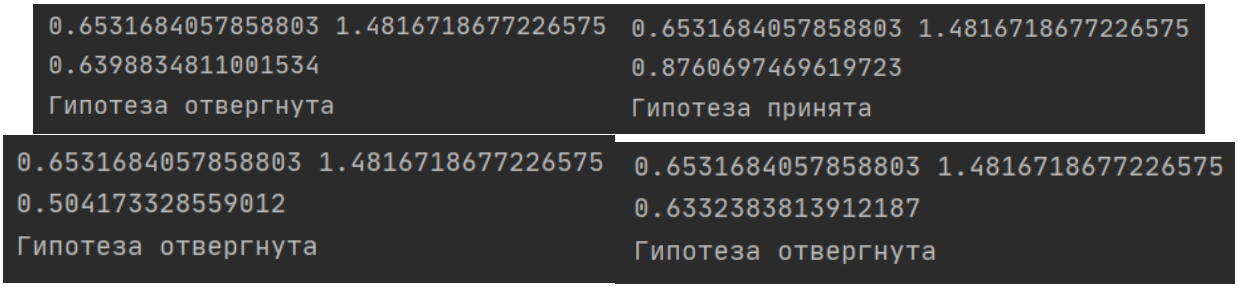


Рис. 18. Результат проверки гипотезы №6.

Заключение

В результате работы было смоделировано две нормальные выборки с необходимыми параметрами. Далее были проверены гипотезы (при заданном уровне значимости 0.1):

1. Гипотеза о числовом значении математического ожидания нормального распределения при известной дисперсии.
2. Гипотеза о числовом значении математического ожидания нормального распределения при неизвестной дисперсии.
3. Гипотеза о числовом значении дисперсии нормального распределения.
4. Гипотеза о равенстве математических ожиданий двух нормальных распределений с известными дисперсиями.
5. Гипотеза о равенстве математических ожиданий двух нормальных распределений с неизвестными, но равными дисперсиями.
6. Гипотеза о равенстве дисперсий двух нормальных распределений.

Первые три гипотезы четыре раза были приняты как верные. Ошибки второго рода, которые наблюдаются при проверке остальных трех гипотез (гипотезы 4-6, которые заведомо являются ложными, однако в некоторых случаях принимаются), обусловлены выбранным уровнем значимости. При увеличении значения уровня значимости вероятность ошибки первого рода будет возрастать, а вероятность ошибки второго рода – уменьшаться.