

УДК 338.001.36

Т.Г. Апалькова, Е.А. Ашихмина, В.А. Дормидонтова

Финансовый университет при Правительстве Российской Федерации, г. Москва,
email: TGApalkova@fa.ru, kate.ashikhmina123@mail.ru, valeriadormi@gmail.com

ПРИМЕНЕНИЕ ЯЗЫКОВ ПРОГРАММИРОВАНИЯ R И PYTHON ДЛЯ ПРОВЕРКИ ГИПОТЕЗЫ О РАВЕНСТВЕ ДИСПЕРСИЙ ПО КРИТЕРИЮ ФИШЕРА В ЭКОНОМИЧЕСКИХ ИССЛЕДОВАНИЯХ

Ключевые слова: равенство дисперсий экономических признаков, разброс среднедушевых доходов, R, Python.

В статье рассматривается алгоритм решения типовой задачи о сравнении степени рассеяния признаков, характеризующих экономические объекты. В конкретном примере сравниваются уровни разброса среднедушевых доходов населения РФ в различных децильных группах во втором и третьем квартале 2022 года. В качестве инструментов проверки гипотезы о равенстве дисперсий признаков по критерию Фишера анализируются программные средства с общедоступным кодом – R и Python. Показано, что хотя оба программных средства дают одинаковые результаты, в R для проверки гипотезы достаточно одной базовой функции, в то время как в Python для этой цели необходимо запрограммировать пользовательскую функцию. Этим обосновывается утверждение о том, что два примерно одинаковых по популярности инструмента математико-статистического анализа не являются абсолютно взаимозаменяемыми. Для экономиста, имеющего базовую подготовку уровня бакалавриата более удобным представляется язык R, поскольку не требует специальных знаний для настройки теста.

T.G. Apalkova, E.A. Ashikhmina, V.A. Dormidontova

Financial University under the Government of the Russian Federation, Moscow,
email: TGApalkova@fa.ru, kate.ashikhmina123@mail.ru, valeriadormi@gmail.com

APPLICATION OF THE PROGRAMMING LANGUAGES R AND PYTHON TO TEST THE HYPOTHESIS OF EQUALITY OF VARIANCES ACCORDING TO THE FISHER CRITERION IN ECONOMIC RESEARCH

Keywords: equality of variances of economic characteristics, dispersion of per capita incomes, R, Python.

The article considers an algorithm for solving a typical problem of comparing the degree of dispersion of features that characterize economic objects. In a specific example, the levels of variation in the average per capita income of the population of the Russian Federation in different decile groups in the second and third quarters of 2022 are compared. As tools for testing the hypothesis of equality of feature variances by the Fisher criterion, software tools with publicly available code – R and Python – are analyzed. It is shown that although both tools give the same results, in R one basic function is enough to test a hypothesis, while in Python it is necessary to program a user-defined function for this purpose. This substantiates the assertion that two tools of mathematical and statistical analysis that are approximately the same in popularity are not absolutely interchangeable. For an economist with a basic bachelor's degree, the R language seems to be more convenient, since it does not require special knowledge to set up the test.

Нередко при изучении экономических явлений в теории и на практике, исследователь приходит к необходимости проверить некоторое умозаключение, сформировавшееся на основе наблюдений. Назовем такое нуждающееся в проверке заключение предположением. В силу того, что экономические объекты описываются множеством количественных признаков, подобные предположения часто выдвигаются относительно значений некоторых коли-

чественных характеристик. Для проверки состоятельности многие из таких предположений могут быть сформулированы как статистические гипотезы. Так, например, может возникнуть необходимость проверки истинности следующих заключений:

1. Рентабельность малых предприятий отрасли составляет в среднем 11,2%.
2. Доля брака в продукции поставщика N не превышает 0,0003 от объема партии.

3. Средняя производительность труда рабочих сборочного цеха существенно повысилась после запуска нового конвейера.

4. Новая технология обработки металла позволила снизить изменчивость (рассеяние) показателей качества комплектующих изделий.

5. Волатильность двух финансовых инструментов существенно не различается.

6. Среднедневной объем реализации готовой продукции соответствует запланированному уровню.

Эти и подобные им утверждения могут быть проверены математико-статистическими методами. Для этого сначала собираются сведения об интересующем признаке объекта (например, фиксируется дневной объем реализации товара магазином), и таким образом формируется выборка относительно небольшого, по сравнению с генеральной совокупностью, объема. Затем формулируется гипотеза относительно характеристики, интересующей в контексте проверяемого предположения. В приведенном выше перечне примеров проверка утверждений 1,3,6 требует постановки гипотез о математических ожиданиях, 2 – о равенстве генеральных долей, 4 и 5 – о равенстве дисперсий. Отметим также, что для проверки гипотез, соответствующих утверждениям 1,2 и 6 необходимо сформировать по одной выборке, остальные требуют по две выборки.

Проверка статистической гипотезы осуществляется при помощи критерия, определяемого характеристикой, относительно которой гипотеза формулируется, закона распределения совокупностей, из которых взяты выборки, их количеством.

Наиболее часто в источниках, описывающих практико-ориентированные задачи, можно встретить примеры проверки гипотез о равенстве математического ожидания некоторого признака гипотетическому значению, о равенстве математических ожиданий признака для двух генеральных совокупностей, о неизменности математического ожидания генеральной совокупности при повторном измерении, о равенстве математических ожиданий признака для более, чем двух генеральных совокупностей [2]. Математическое ожидание как числовая характеристика среднего значения слу-

чайной величины представляет интерес в первую очередь и при исследовании практически любого количественного признака.

Задачи из практики, требующие формулировки гипотез о дисперсии представляют собой более узкий класс. В экономических публикациях чаще всего дисперсия упоминается как характеристика вариабельности доходности, мера волатильности финансового инструмента [2].

Мы хотели бы привести другой пример, в котором дисперсия как характеристика рассеяния выступает показателем степени неравенства доходов населения.

Цель исследования

В настоящей статье преследуется цель показать наиболее технически простое и доступное любому экономисту-исследователю, прошедшему базовый уровень подготовки бакалавра, решение задачи проверки гипотезы о равенстве двух генеральных дисперсий по выборочным значениям экономических признаков.

Материалы и методы исследования

Исходными данными выступают сведения о среднедушевом доходе децильных групп населения. Под децильными доходными группами понимаются группы, каждая из которых содержит 10% от общей численности населения, и которые различаются уровнем дохода: в первую децильную группу входят граждане с наименьшим среднедушевым доходом, в десятую – с наибольшим (Таблица 1).

Принято считать, что характер распределения населения по уровню среднедушевых денежных доходов в генеральной и выборочной совокупностях соответствует логнормальному закону распределения [3]. Это предположение также может быть сформулировано как статистическая гипотеза и проверено в два этапа: сначала исходные данные логарифмируются по основанию экспонента, затем гипотеза о соответствии нормальному распределению проверяется при помощи критерия Колмогорова (например, в модификации Лилиефорса), критерия Шапиро-Уилка, или любого другого на усмотрение исследователя.

Таблица 1

Распределение среднедушевых доходов населения по 10-процентным группам населения во втором и 3 кварталах 2022 года [4]

	Среднедушевые денежные доходы, в месяц, рублей	
	2022 год, II квартал	2022 год, III квартал
Все население	44 374	46 025
в том числе в группировках по 10-процентным группам населения:		
1 группа (с наименьшими доходами)	8 865	9 377
2 группа	15 267	16 072
3 группа	20 097	21 094
4 группа	25 007	26 184
5 группа	30 422	31 789
6 группа	36 782	38 359
7 группа	44 776	46 591
8 группа	55 794	57 912
9 группа	73 800	76 352
10 группа (с наибольшими доходами)	132 925	136 520

В случае, если гипотеза о соответствии среднедушевых доходов логнормальному распределению не отвергается, имеет смысл проверять гипотезу о неизменности дисперсии среднедушевых доходов в отчетном периоде по сравнению с базовым. Опишем процедуру проверки ниже.

Проверяется гипотеза $H_0: \sigma_{i0}^2 = \sigma_{i1}^2$,

где σ_{i0}^2 – дисперсия среднедушевых доходов в децильных группах базового

года, σ_{i1}^2 – дисперсия среднедушевых доходов в децильных группах базового года. Формулировка альтернативной гипотезы H_1 зависит от априорных ожиданий в изменении равенства. Так, если предполагается, что принятые меры государственного регулирования должны были привести к снижению степени неравенства, альтернативная гипотеза бу-

дет иметь вид $H_0: \sigma_{i0}^2 > \sigma_{i1}^2$. Можно сна-

чала сформулировать H_1 как $\sigma_{i0}^2 \neq \sigma_{i1}^2$, а если H_0 отвергнута, альтернативную гипотезу можно конкретизировать.

Критерий, по которому проверяется гипотеза о равенстве двух дисперсий – это критерий Фишера-Снедекора. Статистика критерия выглядит следующим образом:

$$F_{\text{набл}} = \frac{s_{i0}^2}{s_{i1}^2}, \quad (1)$$

где s_{i0}^2 и s_{i1}^2 – выборочные оценки соответствующих дисперсий,

$F_{\text{набл}}$ – наблюдаемое значение статистики Фишера.

В случае, если $F_{\text{набл}}$ оказывается меньше единицы, нулевая гипотеза отвергается, когда выполняется неравенство $F_{\text{набл}} < F_{\text{крит.}}$ и не отвергается при $F_{\text{набл}} > F_{\text{крит.}}$. В противном случае нулевая гипотеза отвергается, когда выполняется неравенство $F_{\text{набл}} > F_{\text{крит.}}$ и не отвергается при $F_{\text{набл}} < F_{\text{крит.}}$.

В качестве инструментальных средств при проверке гипотезы о равенстве генеральных дисперсий может быть использован функционал надстройки «Анализ данных» MsExcel, или же, что более предпочтительно в современных условиях постоянно

ужесточающихся санкций стран Запада против РФ, – программное обеспечение с общедоступными исходными кодами. Наиболее распространенными примерами последнего являются языки R и Python.

Пример кода в R с пояснениями:

1) Сформируем выборки, содержащие среднедушевые доходы для децильных доходных групп населения в сравниваемых периодах:

```
Xq2=c(8865, 15267, 20097,
25007, 30422, 36782, 44776, 55794,
73800, 132925)
```

```
Xq3=c(9377, 16072, 21094,
26184, 31789, 38359, 46591, 57912,
76352, 136520)
```

Xq2 – доходы децильных групп во втором квартале 2022 года, Xq3 – доходы децильных групп в третьем квартале 2022 года.

2) В первую очередь проверяются гипотезы о логнормальном распределении среднедушевых доходов во 2м и 3м кварталах при помощи критерия Лиллиефорса.

Подключается библиотека DescTools:

```
library(DescTools)
```

применяется процедура проверки на «нормальность» к обоим массивам прологарифмированных доходов:

```
LillieTest(log(Xq2))
```

```
LillieTest(log(Xq3))
```

В результате выполнения последних двух операций получаем вывод на консоль:

```
> LillieTest(log(Xq2))
```

```
Lilliefors (Kolmogorov-Smirnov)
normality test
```

```
data: log(Xq2)
```

```
D = 0.061153, p-value = 1
```

```
> LillieTest(log(Xq3))
```

```
Lilliefors (Kolmogorov-Smirnov)
normality test
```

```
data: log(Xq3)
```

```
D = 0.06106, p-value = 1
```

Таким образом, можно считать, что логарифмы обеих выборок имеют нормальное распределение (p-value = 1 в обоих случаях), а следовательно, исходные выборки распределены по логнормальному закону.

3) Проверим гипотезы о равенстве дисперсий логарифмов среднедушевых доходов. Равенство дисперсий логарифмов будет означать равенство дисперсий среднедушевых доходов. Для проверки

гипотезы о равенстве дисперсий двух нормально распределенных выборок в языке R существует специальная базовая функция var.test():

```
var.test(log(Xq2),log(Xq3))
```

Результат выполнения команды:

F test to compare two variances

data: log(Xq2) and log(Xq3)

F = 1.0222, num df = 9, denom df = 9,
p-value = 0.9744

alternative hypothesis: true ratio of
variances is not equal to 1

95 percent confidence interval:

0.2539019 4.1154024

sample estimates:

ratio of variances

1.022208

Таким образом, вся процедура на языке R, вместе с проверкой гипотез о нормальности, составляет всего 5 строк кода. В языке Python подобные расчёты осуществляются сложнее, поскольку специальная функция для проверки гипотезы о равенстве двух дисперсий отсутствует.

Пример проверки гипотезы с помощью пользовательской функции в Python:

```
# исходные данные
```

```
Xq2=[8865, 15267, 20097,
25007, 30422, 36782, 44776, 55794,
73800, 132925]
```

```
Xq3=[9377, 16072, 21094,
26184, 31789, 38359, 46591, 57912,
76352, 136520]
```

Нам понадобятся две библиотеки: numpy для работы с массивами и scipy.stats для определения p-value

```
#подключаем библиотеки
```

```
import numpy as np
```

```
import scipy.stats as sts
```

Далее, в силу отсутствия готового решения, создаём функцию, которая вычисляет наблюдаемое значение статистики Фишера и затем определяет p-value (за основу взят и модифицирован скрипт ресурса Stack Overflow)

```
#определение функции F-test
```

```
def F_test(Xq2, Xq3):
```

```
Xq2 = np.array(Xq2)
```

```
Xq3 = np.array(Xq3)
```

```

F = np.var(np.log(Xq2), ddof=1)/
np.var(np.log(Xq3), ddof=1) #вычисление
статистики F
dfn = Xq2.size-1 #число степеней
свободы числителя
dfd = Xq3.size-1 #число степеней
свободы знаменателя
p = 2*(1-sts.f.cdf(F, dfn, dfd)) #нахож-
дение p-value
return F, p
#результаты применения F-test
F_test(Xq2, Xq3)
Обращение к функции позволяет по-
лучить наблюдаемое значение статисти-
ки и p-value:
#результаты применения F-test
F_test(Xq2, Xq3)
Непосредственно результат, который
выводится на экран, выглядит следую-
щим образом:

```

Out[29]: (1.0222077334186428, 0.9744353758845832)

Рис. 1. Результат, выведенный на экран

Первое число в скобках – это наблюдаемое значение статистики F , второе – p -value.

Результаты исследования и их обсуждение

Наблюдаемое значение F фактически представляет собой отношение выборочных дисперсий логарифмов доходов децильных групп во втором и в третьем кварталах 2022 года. Значение p -value = 0.9744 существенно больше уровня значимости 0,05 и позволяет заключить, что гипотеза о равенстве дисперсий логарифмов среднедушевых доходов не отвергается. Следовательно, статистически значимых различий между степенью неравенства распределения доходов по децильным группам во втором и в третьем квартале не выявлено. Необходимо отметить, что речь идет именно об отсутствии статистически значимых различий. Возможно, что с точки зрения экономиста-исследователя положительным сигналом окажется даже небольшое отклонение. Кроме того, проверка гипотезы о равенстве дисперсий позволяет делать вывод на известном уровне значимости только если исходные данные распределены нормально, т.е. для натуральных логарифмов доходов. А дисперсии логдоходов и доходов могут отличаться и увеличение степени разброса логдоходов может соответствовать снижению степени разброса исходных до-

ходов. Поэтому, применение статистических процедур должно осуществляться с осторожностью и дополнять, а не заменять экономический анализ.

Как видно из приведенных выше скриптов, реализация пользовательской функции Python даёт результаты идентичные полученным с помощью стандартной базовой функции `var.test()` в R, однако её создание отнимает больше времени, чем вызов стандартной функции и требует определенных знаний. Так, например, чтобы указать тип альтернативной гипотезы, в функции `var.test()` достаточно указать значение аргумента `alternative`, который по умолчанию предусматривает значение «two.sided», т.е. альтернативная гипотеза предпола-

гает, что $\sigma_{i0}^2 \neq \sigma_{i1}^2$. Возможность выбора типа альтернативной гипотезы в Python ещё более усложнит пользовательскую функцию.

Справедливости ради следует отметить, что существуют и другие критерии (Бартлетта, Левене), позволяющие проверить гипотезу о равенстве двух и более дисперсий (критерий Фишера приложен только для двувыворочных тестов). И процедуры проверки гипотез по этим критериям реализованы в Python через специальные методы библиотеки `scipy.stats`. Но эти критерии, в отличие от критерия Фишера, не всегда входят в программы курсов математической статисти-

стики для бакалавриата, таким образом, с ними знаком далеко не каждый рядовой экономист.

Выводы

Задача сравнительного анализа степени отклонения признака от среднего значения актуальна для экономических исследований в различных постановках и может быть решена при помощи математико-статистического метода проверки гипотез. Автоматизировать процедуру проверки позволяют современные вычислительные средства, в частности – языки программирования со свободным кодом R и Python. Мнение о взаимозаменяемости этих инструментальных средств не вполне соответствует действительности, что удалось продемонстрировать на примере задачи проверки гипотезы о равенстве дисперсий при помощи критерия Фишера. Стандартная процедура R позволяет проверить гипотезу с помощью одной команды, при этом возможна настройка функции для разных типов альтернативных гипотез.

Так, если в условиях сформулированной выше задачи о неравенстве доходов децильных групп для исследователя принципиально выяснить, снизилась ли степень различия доходов после реализации мер государственного регулирования, актуальна альтернативная гипотеза о снижении дисперсии в отчетном периоде по сравнению с базовым. Python не предусматривает возможности проведения проверки гипотезы о равенстве дисперсий по критерию Фишера с помощью какой-либо базовой функции, или функции из известных библиотек, процедуру надо программировать. Настройка в запрограммированной функции возможности выбирать тип альтернативной гипотезы делает код ещё более сложным и требует достаточно хорошего знания математической статистики, чтобы верно определить p-value в каждом из трех случаев альтернативных гипотез. Таким образом, в экономических исследованиях с применением математико-статистических методов R демонстрирует лучшую функциональность.

Библиографический список

1. Якушев А. Справочник контролера машиностроительного завода // Машиностроение. 1970. № 2. С. 471.
2. CFA – Проверка гипотез о равенстве (неравенстве) двух дисперсий // Финансовый учёт. 2019. [Электронный ресурс]. URL: <https://fin-accounting.ru/cfa/11/quantitative/cfa-tests-concerning-equality-inequality-two-variances> (дата обращения: 15.01.2023).
3. Социально-экономические индикаторы бедности в 2013-2020 гг. // Федеральная служба государственной статистики Росстат Москва. 2021. [Электронный ресурс]. URL: https://rosstat.gov.ru/storage/mediabank/Bul_Ind_bedn_2013-2020.pdf (дата обращения: 15.01.2023).
4. Неравенство и бедность // Официальный сайт Федеральной службы государственной статистики Росстат Москва. 2022. [Электронный ресурс]. URL: <https://rosstat.gov.ru/folder/13723> (дата обращения: 15.01.2023).
5. How do I do a F-test in python // Stack Overflow. 2014. [Электронный ресурс]. URL: <https://stackoverflow.com/questions/21494141/how-do-i-do-a-f-test-in-python/67885335#67885335> (дата обращения: 15.01.2023).