Практическая работа № 7. Проверка гипотез о законе распределения, проблема нормализации выборки

7.1. О содержании и задачах практической работы

В настоящей практической работе обсуждаются четыре задачи.

1. Преобразование данных в формат, удобный для работы с инструментами языка программирования Python.
2. Проверка гипотез о законах распределения.
3. Визуальные средства оценки близости генеральной совокупности теоретическому закону.
4. Преобразование в выборку данных, имеющих распределение, близкое к нормальному.

О первой задаче. В реальной работе результаты наблюдений чаще всего хранятся в файлах формата csv или excel. Прежде чем заняться обработкой и анализом данных, их приходится экспортировать. В языке программирования Python имеется специальная библиотека Pandas для работы с таблицами – объектами DataFrame. Начиная с этого практикума, решение многих задач будет предваряться созданием объекта DataFrame на основе csv или excel-файла.

О второй задаче. Существует много методов проверки соответствия закона распределения генеральной совокупности теоретическому закону того или иного типа. В данной практической работе рассмотрим универсальный критерий согласия Пирсона (хи-квадрат), а также критерий Шапиро – Уилка, предназначенный исключительно для проверки непротиворечивости данных предположению о нормальном законе распределения генеральной совокупности.

О третьей задаче. Первоначальную информацию о структуре выборки мы получаем, анализируя гистограмму. В частности, анализ гистограммы может натолкнуть нас на правдоподобные предположения о типе закона распределения генеральной совокупности. Возникшую гипотезу можно «проверить», используя, например, критерий согласия Пирсона. Но можно также оценить степень близости эмпирического и предполагаемого теоретического распределения графически, а именно

построить график квантиль-квантиль (Q – Q-график). При построении такого графика на оси абсцисс откладываются квантили предполагаемого теоретического закона, на оси ординат – выборочные квантили набора данных (если оба набора квантилей соответствуют одному и тому же закону распределения, то график будет близок к прямой линии).

О четвертой задаче. Многие методы статистического анализа основаны на предположении о нормальном законе распределения генеральной совокупности. Но на практике гипотеза о нормальности подтверждается далеко не всегда. В некоторых случаях, применив к каждому элементу выборки некоторую функцию, удается получить новый набор данных, лучше согласующийся с предположением о нормальном распределении. Никакого однозначного руководства, как преобразовывать данные и имеет ли смысл это делать вообще, не существует. Чаще всего используют какое-нибудь степенно́е преобразование. Иногда имеет смысл прологарифмировать данные. В языке программирования Python реализовано однопараметрическое семейство преобразований Бокса-Кокса, которое включает в себя и то и другое (параметр подбирается для достижения наилучшего приближения к нормальному распределению).

7.2. Математические понятия и утверждения: краткая информация и ссылки на источники

1. Проверка гипотез о законе распределения. Критерий согласия хи-квадрат (Пирсона) [1, с. 193 – 198; 2, с. 286 – 289].

Критерии согласия позволяют оценить вероятность того, что полученная выборка не противоречит сделанному предположению о законе распределения рассматриваемой случайной величины. Для этого выбирается некоторая величина γ, являющаяся мерой расхождения статистического и теоретического законов распределения, и определяется такое значение γ_α, чтобы $P\{\gamma > \gamma_\alpha\} = \alpha$, где α – малая величина (уровень значимости), значение которой устанавливается из практических соображений. Если полученное на опыте значение γ^* меры расхождения γ больше γ_α (выполняется неравенство $\gamma^* > \gamma_\alpha$), то отклонение от теоретического закона считается значимым, и предположение о виде закона распределения должно быть отвергнуто (вероятность отвергнуть

правильное предположение в этом случае равна α). Если значение $\gamma^* \le \gamma_\alpha$, то отклонение считается не значимым, т.е. данные опыта не противоречат сделанному предположению о виде закона распределения.

Проверку гипотезы о законе распределения можно вести, как и случае проверки гипотез о параметрах распределения, с использованием p-значения: по значению γ^* определить вероятность $p = P\{\gamma > \gamma^*\}$. Если $\alpha > p$, то отклонения значимые; если $\alpha \le p$, то отклонения не значимые. Значение p, весьма близкие к 1 (очень хорошее согласие), соответствует событию, имеющему весьма малую вероятность, что может указывать на недоброкачественность выборки (например, из первоначальной выборки без основания выброшены элементы, дающие большое отклонение от среднего).

Известно несколько вариантов критериев согласия, отличающихся выбором меры расхождения статистического и теоретического законов распределения. В настоящей практической работе, как было сказано выше, будем использовать критерий согласия хи-квадрат (Пирсона).

2. Критерий Шапиро – Уилка.

Критерий предназначен для проверки гипотезы о том, что набор данных принадлежит нормальной генеральной совокупности. Альтернативная гипотеза – отрицание основной.

Пусть $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ – вариационный ряд, построенный по эмпирической выборке. Тогда выборочное значение статистики рассчитывается по формуле

$$w = \frac{1}{g^2}\left(\sum_i \left(a_{n-i+1}(x_{(n-i+1)} - x_{(i)})\right)\right)^2,\ g^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2,$$

где индекс i изменяется от 1 до $n/2$ или от 1 до $(n-1)/2$ при четном и нечетном n соответственно.

Коэффициенты a_{n-i+1} табулированы (в источниках приводятся значения коэффициентов для выборок объема $8 \le n \le 99$).

Информации об аналитическом распределении статистики в литературе нет.

3. Однопараметрическое преобразование Бокса-Кокса.

Преобразование Бокса-Кокса используется для преобразования данных в выборку, имеющую распределение, близкое к нормальному.

Преобразование выполняется по правилу:

$$y = \begin{cases} \dfrac{x^{\lambda} - 1}{\lambda}, & \text{если } \lambda > 0, \\ \ln x, & \text{если } \lambda = 0. \end{cases}$$

Оценку $\hat{\lambda}$ для оптимального значения параметра λ (обеспечивающего «наилучшее» приближение к нормальному распределению) находят методом максимального правдоподобия.

Для заданного значения уровня значимости α строится доверительный интервал для λ вида:

$$\left|\mathrm{llf}\left(\hat{\lambda}\right) - \mathrm{llf}\left(\lambda\right)\right| < \frac{1}{2}\chi^2_{1-\alpha}\left(1\right),$$

где llf – логарифм функции правдоподобия; $\chi^2_{1-\alpha}\left(1\right)$ – квантиль распределения хи-квадрат с одной степенью свободы порядка $1-\alpha$.

7.3. Библиотечные инструменты языка программирования Python

Загрузка основных модулей

```
import numpy as np
import scipy.stats as sts
import matplotlib.pyplot as plt
import pandas as pd
%matplotlib inline
```

Экспорт данных из csv или Excell- файла в объект DataFrame библиотеки Pandas

Функция `pd.read_excel('Data.xlsx', sep=',' header = 'infer', index_col=None)`

Создает объект `DataFrame`.

Параметры: `Data.xlsx` – строка с указанием пути к файлу; `sep` – разделитель (по умолчанию `,`); `header` – строка, содержащая имена столбцов (по умолчанию `'infer'` – в качестве имен используется первая строка данных); `index_col` – указывает, какой столбец в файле использовать в качестве индекса, если установить `index_col= False`, первый столбец данных в качестве индекса использоваться не будет.
Функция `pd.read_csv()` имеет аналогичные параметры.

Метод `DataFrame.replace('-',np.nan)` позволяет заменить в `DataFrame` одни элементы (в данном случае`'-'`) на другие (в данном случае `np.nan`).

Метод `DataFrame.dropna()` удаляет строки (целиком), в которых есть пропущенные элементы.

Средства группировки выборки библиотеки numpy

Функция `np.histogram(a, bins=10, range=None, weights=None, density= None)`
Возвращает два массива: `hist` – массив высот столбцов гистограммы; `bin_edges` – массив границ интервалов.
Параметры: `a` – одномерный массив (выборка); `bins` – число интервалов группировки (по умолчанию – 10) или последовательность, задающая границы интервалов; если `bins='auto'` – число интервалов выбирается как максимальное из величин, получаемых по правилу Стерджесса и Фридмана-Диакониса;
`range` – начальная и конечная границы интервалов (если параметр не определен, то в качестве границ берутся минимальный и максимальный элементы выборки), элементы выборки вне области `range` игнорируются; `density` – если `True` – строится гистограмма относительных частот (суммарная площадь прямоугольников равна 1); `weights` – массив весов той же формы, что и `a`.

Реализация критерия хи-квадрат проверки гипотезы о законе распределения в модуле scipy.stats

Функция `sts.chisquare (f_obs, f_exp = None, ddof = 0, axis = 0)`

Возвращает наблюдаемое значение статистики и p-значение (т.е. максимальное значение уровня значимости, при котором основная гипотеза принимается).

Параметры: `f_obs` – наблюдаемые частоты (n_i); `f_exp` – частоты гипотетического (согласно основной гипотезе) распределения (np_i) (по умолчанию равные между собой); `ddof` – число параметров гипотетического распределения, оцениваемых по выборке. На вход функции можно подавать многомерный массив. Критерий будет применяться к каждому столбцу массива (если `axis = 1`, то к строке).

Условие использования: все наблюдаемые и гипотетические частоты должны быть не менее 5 ($n_i \geq 5, np_i \geq 5$).

Реализация критерия Шапиро-Уилка в модуле scipy.stats

Функция `sts.shapiro (x, a=None, reta=False)`

Возвращает наблюдаемое значение статистики p-значение; массив параметров (присутствует, если `reta=True`).

Параметры: `x` – одномерный массив (выборка); `a` – массив внутренних параметров (если не заданы, вычисляется самой функцией); `reta` – признак, нужно ли возвращать вычисленные параметры).

Преобразование Бокса-Кокса модуля scipy.stats

Функция `sts.boxcox(x, lmbda=None, alpha=None)`

Возвращает: (1) `boxcox` – массив, результат преобразования Бокса-Кокса; (2) если параметр `lmbda=None` , то второй возвращаемый параметр `maxlog` – значение `lmbda`, максимизирующее логарифм функции правдоподобия; (3) если `lmbda=None` и `alpha` не `None`, возвращается кортеж, содержащий границы доверительного интервала.

Параметры: `x` – входной одномерный массив положительных чисел (выборка); `lmbd` – если не `None`, преобразование выполняется для этого значения; `alpha` – если не `None`, то функция возвращает в качестве третьего аргумента $100(1-\alpha)\%$ -й доверительный интервал для параметра `lmbda`.

Построение графика квантиль-квантиль в модуле scipy.stats

`Функция sts.probplot (x, sparams=[m,s], dist='norm', plot=plt)`

Строит график связи между наблюдаемыми значениями переменной и теоретическими квантилями.

Параметры: `x` – выборка данных, для которых строится график; `sparams` – кортеж параметров (для примера взяты параметры нормального распределения); `dist` - наименование распределения (для примера указано нормальное, так же будет по умолчанию).

7.4. Примеры для совместного обсуждения

В файле «Data_7» приведены массивы результатов наблюдения двух случайных величин. Будем рассматривать каждый из массивов 1 и 2 как случайную выборку из генеральной совокупности с неизвестным законом распределения (для каждой выборки генеральная совокупность своя). Начнем с экспорта данных в таблицу `DataFrame`.

Пример 1. Для выборки 1 построим гистограмму относительных частот. Ее вид (см. приложение к практической работе № 7, пример 1) позволяет выдвинуть гипотезу о равномерном законе распределения генеральной совокупности. Проверим эту гипотезу, используя критерий согласия хи-квадрат. Проанализируем результаты (вариант выполнения задания приведен в приложении к практической работе № 7).

Пример 2. Для выборки 2 построим гистограмму относительных частот. Ее вид очевидно не согласуется с гипотезой о нормальном распределении генеральной совокупности (см. приложение к практической работе № 7, пример 2).

Применим к выборке 2 логарифмическое преобразование (используем формулу $y = \ln x$). Вновь построим гистограмму. Теперь можно рассматривать гипотезу о нормальном распределении.

С помощью критерия (на этот раз воспользуемся критерием Шапиро – Уилка) проверим гипотезу о распределении преобразованных данных по нормальному закону. Проанализируем результаты.

Оценим с помощью Q-Q – графика близость распределения преобразованных данных к нормальному закону. Проанализируем результаты и сформулируем выводы (вариант выполнения задания приведен в приложении к практической работе № 7, пример 2).

7.5. Задания для самостоятельного выполнения

Задание 1. В файле «Data_7_1» приведены массивы результатов наблюдения нескольких случайных величин. Будем рассматривать каждый из массивов как случайную выборку из генеральной совокупности с неизвестным законом распределения (для каждого массива генеральная совокупность своя). Экспортируйте данные. Для каждой из выборок постройте гистограмму относительных частот и на основе визуального качественного анализа гистограмм отберите две выборки: выборку A, позволяющую выдвинуть гипотезу о принадлежности нормальному распределению генеральной совокупности; выборку B, позволяющую выдвинуть гипотезу о показательном распределении генеральной совокупности. Если среди массивов данных нет ни одного похожего на выборку из генеральной совокупности с показательным законом распределения, то попробуйте преобразовать один из них таким образом, чтобы к полученной выборке гипотеза о показательном распределении генеральной совокупности подходила.

1) Проверьте гипотезу о том, что выборка A взята из генеральной совокупности, имеющей нормальное распределение, с помощью критерия согласия хи-квадрат. Решение должно быть подробным, с выполнением всех шагов алгоритма и отслеживанием корректности применения. Сделайте выводы.

2) Проверьте гипотезу о том, что выборка B (исходная или полученная путем преобразования) взята из генеральной совокупности, имеющей показательное распределение. Используйте критерий согласия хи-квадрат. Решение должно быть подробным, с выполнением всех шагов алгоритма и отслеживанием корректности применения. Сделайте выводы.

Задание 2.

1) Преобразуйте данные, применив к выборкам A, B преобразование Бокса-Кокса (далее AA, BB – преобразованные данные).

2) Постройте гистограммы A и AA, B и BB. Сопоставьте гистограммы прообразов и образов. Результаты прокомментируйте.

3) С помощью критерия Шапиро – Уилка (используйте функцию `shapiro()` модуля `scipy.stats`) проверьте для всех четырех выборок гипотезы о том, что выборки принадлежат нормально распределенным генеральным совокупностям. Результаты образов и прообразов сопоставьте и прокомментируйте.

4) Для всех четырех выборок постройте Q-Q – график. Сопоставьте графики прообразов и образов. Результаты прокомментируйте и сопоставьте с результатами п. 3).