Практическая работа № 10. Корреляционный анализ

10.1. О содержании и задачах практической работы

В данной практической работе обсуждается решение следующей задачи.

Наблюдаются две случайные величины X и Y. Проведено n опытов и получена выборка из совместного закона распределения вектора (X,Y). Требуется сделать заключение о наличии корреляционной связи между X и Y.

Решение задачи можно разбить на два этапа.

На первом этапе проводится первичная обработка двумерной выборки: построение диаграммы рассеивания, получение точечной оценки коэффициента корреляции Пирсона и/или Спирмена.

На втором этапе проводится проверка гипотез о возможных значениях коэффициента корреляции. В частности, если значение выборочного коэффициента корреляции оказалось по абсолютной величине небольшим, то выдвигается и проверяется гипотеза об отсутствии значимости коэффициента корреляции (равенстве его нулю). Эта гипотеза может быть интерпретирована как гипотеза о наличии или отсутствии линейной связи между переменными X и Y.

Реализация второго этапа исследования неоднозначна и определяется тем, можно ли считать, что случайные величины X и Y распределены по нормальному закону. Если нет основания отвергать предположение о нормальности законов распределения случайных величин X и Y, то можно построить доверительный интервал для коэффициента корреляции Пирсона, а затем при условии малых значений выборочного коэффициента корреляции проверить гипотезу о значимости коэффициента корреляции (об используемых при этом статистиках см. далее).

Если относительно законов распределения случайных величин X и Y мы можем утверждать только то, что они непрерывны, имеет смысл воспользоваться непараметрическими методами, и при проверке гипоте-

зы о значимости корреляционной связи использовать коэффициент корреляции Спирмена.

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

1. Зависимые и независимые случайные величины, условные законы распределения [1, с. 66 – 72, 86 – 89; 2, с. 85 – 95].

2. Числовые характеристики случайных векторов. Ковариация и коэффициент корреляции (Пирсона) [1, с. 82 – 86; 2, с. 85 – 95].

3. Многомерное нормальное распределение [1, с. 103 – 105; 2, с. 85 – 95].

4. Статистическая описание двумерного случайного вектора: корреляционное поле (диаграмма рассеивания), корреляционная таблица [1, с. 200 – 203; 2, с. 203 – 206].

5. Оценка параметров распределения двумерного вектора. Проверка гипотез о коэффициенте корреляции Пирсона [1, с. 203 – 207; 2, с. 246 – 247, 273 – 275].

6. Непараметричекие методы исследования связи между случайными величинами; ранговый коэффициент корреляции Спирмена, ранговый коэффициент корреляции Кендалла [2, с. 354 – 356].

10.3. 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
import seaborn
%matplotlib inline
```

Генерация многомерного нормального распределения в библиотеке numpy

Функция `np.random.multivariate_normal(mean, cov, n)`

Возвращает выборку объема n для многомерного нормального распределения с заданным вектором математических ожиданий `mean` и ковариационной матрицей `cov`.

Средства визуализации диаграммы рассеивания

Функция `plt.scatter(x, y)` модуля `matplotlib.pyplot`
Строит диаграмму рассеивания признаков `x` и `y`.
Параметры: `x, y` – два массива одинаковой длины.

Функция `seaborn.pairplot(data, vars=None, kind='scatter', diag_kind='hist', height=4)` пакета `seaborn`
Строит диаграммы рассеивания пар признаков из `vars`, а также визуализирует распределение отдельных признаков.
Параметры: – датафрейм; `vars` – список имен переменных из `vars`, которые будут использованы для вывода диаграммы (если не задан, используются все числовые колонки `data`); `kind` – тип диаграммы рассеяния (обычная `'scatter'` или с линией регрессии `'reg'`); `diag_kind` – тип диагональных графиков (`'auto'`, `'hist'`, `'kde'`); `height` – высота каждой фацеты (в дюймах).

Расчет выборочных характеристик двумерной выборки

Функция `np.cov(x, y=None, rowvar=True, bias=False, ddof=None)` библиотеки `numpy`
Вычисляет выборочную ковариационную матрицу.
Параметры: `x` – одномерный или двумерный массив. Если одномерный – вычисляется ковариация между `x` и `y`. Если двумерный – при значении `rowvar=True` (по умолчанию) вычисляется ковариация между строками массива `x`, при значении `rowvar=False` – между столбцами массива `x`; `bias` – признак, определяющий способ нормализации, по умолчанию (`False`) – производится деление на $n-1$, иначе на n; `ddof` – выполняет функцию, аналогичную признаку `bias`: при значении `ddof=1` производится деление на $n-1$, при значении `ddof=0` – деление на n.

Функция `np.corrcoef(x, y=None, rowvar=True)` библиотеки `numpy`

Вычисляет выборочную корреляционную матрицу.

Параметры: `x` – одномерный или двумерный массив. Если одномерный – вычисляется коэффициент корреляции между `x` и `y`. Если двумерный: при значении `rowvar=True` (по умолчанию) вычисляется коэффициент корреляции между строками массива `x`, при значении `rowvar=False` – между столбцами массива `x`.

Метод `data.corr(method='pearson')` библиотеки `pandas`

Параметры: `data` – объект `DataFrame`, `method` – задает вид коэффициента корреляции `'pearson'`, `'spearman'`, `' kendall'` (по умолчанию `'pearson'`).

Средства визуализация корреляционной матрицы

Функция `seaborn.heatmap(data, annot=None, fmt='.2g', linewidth=0, linecolor='white', cbar=True, cbar_kws=None, cbar_ax=None)` пакета `seaborn`

Принимает на вход прямоугольный массив данных и отображает данные с помощью цвета. Цветовая панель показывает соответствие цвета числовым значениям переменной.

Параметры: `data` – объект `DataFrame`, `annot` – признак: если `True`, в каждую ячейку карты выводится значение признака; `fmt` – строка: задает формат для случая `annot= True`, `linewidth`, `linecolor` – размер и цвет линий, разделяющих ячейки; `cbar`, `cbar_kws`, `cbar_ax` – информация о необходимости вывода, цвете и расположении цветовой панели.

Критерий значимости коэффициента корреляции Пирсона модуля scipy.stats

Функция `sts.pearsonr(x,y)`

Проверяет гипотезу об отсутствии значимой линейной связи.

Параметры: `x,y` - одномерные массивы одинаковой длины.

Возвращает `r` – выборочный коэффициент корреляции Пирсона, `p-value` – достигаемый уровень значимости.

Значение p-value вычисляется с использованием распределения Стьюдента.

Критерий значимости коэффициента корреляции Спирмена модуля scipy.stats

Функция sts.spearmanr(a,b=None, axis=0, nan_policy='propogate')

Проверяет гипотезу об отсутствии значимой монотонной связи.

Параметры: a, b – два одномерных или двумерных массива одинакового размера, nan_policy ('propagate', 'raise', 'omit') – задает способ обработки пропущенных (NaN) значений.

Возвращает r – выборочный коэффициент корреляции, p-value – достигаемый уровень значимости. Значение p-value вычисляется с использованием распределения Стьюдента.

10.4. Примеры для совместного обсуждения

Пример 1. Средняя температура июня в Москве и Ярославле в градусах Цельсия измерялась в течение 40 лет. Данные приведены в следующей таблице (первая строка – температура в Москве, вторая строка – температура в Ярославле):

x	*y*	*x*	*y*	*x*	*y*	*x*	*y*	*x*	*y*
12,0	10,8	13,9	10,1	15,0	13,8	17,2	13,9	18,1	16,0
12,0	11,3	14,2	10,0	15,0	16,0	16,9	14,8	18,4	17,8
12,0	12,0	14,0	10,0	15,5	13,9	16,9	15,0	19,2	15,0
12,0	13,0	14,0	12,0	15,9	14,7	17,0	16,0	19,3	16,1
12,8	10,9	13,9	12,4	16,0	13,0	16,8	17,0	20,0	17,0
13,8	10,0	15,0	11,0	15,9	15,0	17,5	16,0	20,1	17,7
13,1	11,5	14,9	13,0	16,0	16,0	18,0	14,0	14,0	14,8
13,0	13,0	14,9	14,2	16,9	12,9	18,0	14,8	14,0	15,2

Построим диаграмму рассеивания; для данных по Москве и Ярославлю построим гистограммы; найдем ковариационную и корреляционную матрицы Пирсона; найдем значение выборочного коэффициента корреляции Спирмена; визуализируем корреляционные матрицы с помощью тепловых карт.

Вариант решения примера приведен в приложении к практической реботе № 10, пример 1. Обратите внимание, что значения коэффициентов корреляции Пирсона и Спирмена оказались очень близкими друг к другу и довольно большими по значению (около 0,75). Это указывает на наличие тесной линейной зависимости между температурами в Москве и Ярославле.

Пример 2. Проведем несколько статистических экспериментов, чтобы научиться по диаграмме рассеивания считывать (на качественном уровне) информацию о значении выборочного коэффициента корреляции двумерного вектора. Сгенерируем выборку объема n двумерного нормального распределения с конкретными параметрами m_X, m_Y, σ_X, σ_Y, ρ, построим диаграмму рассеивания, вычислим корреляционную матрицу. Сопоставим значение ρ теоретического распределения с его выборочной оценкой и диаграммой рассеивания. Далее будем менять значения ρ в рамках отрезка $[-1,1]$ и наблюдать за изменением вида диаграммы. К каким выводам можно прийти?

В приложении к практической работе № 10, пример 2 приведены фрагменты выполнения задания.

Пример 3. Пусть генеральная совокупность имеет нормальный закон распределения. Что можно сказать о наличии корреляционной зависимости между случайными величинами – выборочной оценкой математического ожидания и исправленной выборочной дисперсией? Для выдвижения гипотезы используйте статистический эксперимент. Проверьте вашу гипотезу.

10.5. Задания для самостоятельного выполнения

Задание 1. Сгенерируйте выборку объема n двумерного нормального распределения с параметрами m_X, m_Y, σ_X, σ_Y, ρ и выполните следующие действия:

1) постройте диаграмму рассеяния;
2) постройте гистограммы компонент;
3) найдите выборочные характеристики компонент;
4) найдите выборочное значение коэффициента корреляции Пирсона;

5) постройте доверительный интервал для коэффициента корреляции Пирсона;

6) проверьте гипотезу о значимости коэффициента корреляции Пирсона.

Задание 2. Случайный вектор (X,Y) распределена по круговому нормальному закону ($\rho_{X,Y} = 0, \sigma_X = \sigma_Y = 1$). Пусть $V = X \cdot Y$, $W = 0,5\left(X^2 - Y^2\right)$.

1) Подтвердите или опровергните гипотезу о том, что случайные величины распределены по нормальному закону (для проверки используйте критерий Шапиро-Уилка, Q-Q график).

2) Методом статистического эксперимента исследуйте вопрос о корреляционной зависимости величин V и W (какой коэффициент корреляции корректно использовать?). Возникшую гипотезу проверьте с помощью подходящего критерия.