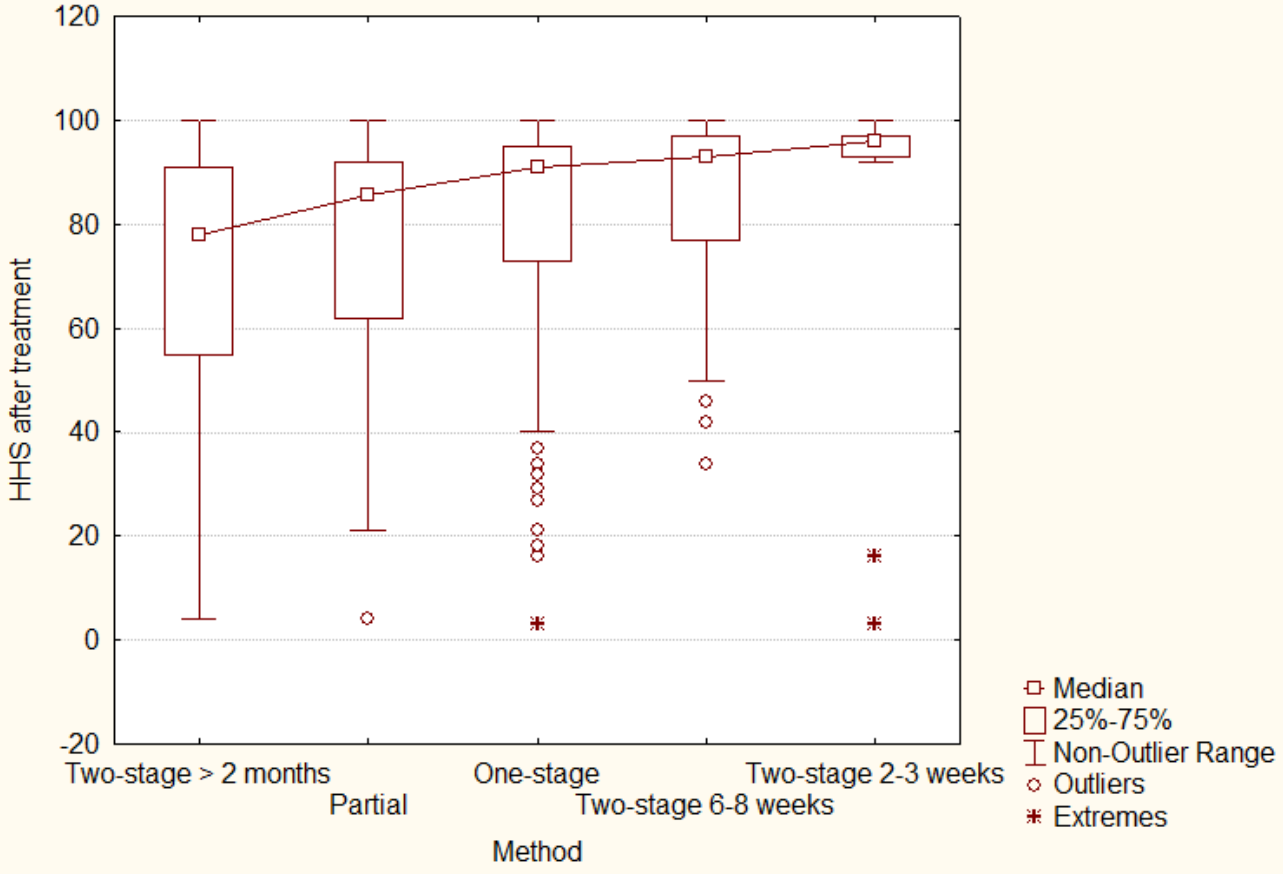
УНИВЕРСИТЕТ ИТМО

В. Н. Леоненко

ВЕРОЯТНОСТНЫЕ МЕТОДЫ АНАЛИЗА ДАННЫХ

Учебно–методическое пособие



Санкт–Петербург
2021

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

УНИВЕРСИТЕТ ИТМО

В. Н. Леоненко

ВЕРОЯТНОСТНЫЕ МЕТОДЫ АНАЛИЗА ДАННЫХ

Учебно–методическое пособие

по выполнению лабораторных работ

РЕКОМЕНДОВАНО К ИСПОЛЬЗОВАНИЮ В УНИВЕРСИТЕТЕ ИТМО

по направлению подготовки 11.04.02 Инфокоммуникационные технологии и системы связи в качестве учебно–методического пособия для реализации основных профессиональных образовательных программ высшего образования магистратуры

УНИВЕРСИТЕТ ИТМО

Санкт–Петербург
2021

Леоненко В.Н. Вероятностные методы анализа данных. Учебно–методическое пособие по выполнению лабораторных работ.

Учебно–методическое пособие. – СПб: Университет ИТМО, 2021. – 28 с.

Рецензент: Литневский Андрей Леонидович, к.ф.-м.н., доцент кафедры физики Физико-механического института ФГАОУ ВО «Санкт-Петербургский политехнический университет Петра Великого».

Настоящее учебно–методическое пособие составлено в соответствии с ОС Университета ИТМО 11.04.02 – Инфокоммуникационные технологии и системы связи.

Пособие содержит учебно–методические разработки, предназначенные для выполнения лабораторных работ по следующим темам: введение в вероятностные методы анализа данных; первичный и разведочный анализ; проверка статистических гипотез; корреляционные зависимости; линейная регрессия; анализ влияния факторов.

Учебно–методическое пособие «Методические рекомендации по выполнению лабораторных работ» предназначено для студентов, обучающихся по направлению 11.04.02 «Инфокоммуникационные технологии и системы связи».

УНИВЕРСИТЕТ ИТМО

Университет ИТМО – ведущий вуз России в области информационных и фотонных технологий, один из немногих российских вузов, получивших в 2009 году статус национального исследовательского университета. С 2013 года Университет ИТМО – участник программы повышения конкурентоспособности российских университетов среди ведущих мировых научно–образовательных центров, известной как проект «5 в 100». Цель Университета ИТМО – становление исследовательского университета мирового уровня, предпринимательского по типу, ориентированного на интернационализацию всех направлений деятельности.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

В настоящем пособии представлены методические указания к выполнению лабораторных работ по дисциплине «Вероятностные методы анализа данных» магистерской программы, разработанной в соответствии с Образовательным стандартом Университета ИТМО по направлению подготовки 11.04.02 «Инфокоммуникационные технологии и системы связи». Целью выполнения лабораторных работ является закрепление знаний, приобретение умений и владений методами анализа данных для формирования компетенции по использованию знаний в области профессиональной деятельности с применением цифровых технологий, а также организации и осуществлению научных исследований.

Целью изучения дисциплины является формирование у обучающегося профессиональных знаний в области методов организации и статистического анализа результатов научного эксперимента и математического моделирования. В задачи дисциплины входит изучение методов современной математической статистики и получение навыков проведения статистического анализа с использованием языка программирования Python.

По итогам выполнения всех работ студент должен получить теоретические знания об основах вероятностных методов анализа данных как дисциплины, о способах первичного и разведочного анализа данных, проверки гипотез и способах поиска связей между переменными.

По итогам выполнения лабораторной работы студенту необходимо подготовить письменный отчет, содержащий основные результаты работы. Отчет должен быть оформлен согласно требованиям ГОСТ и включать титульный лист и основную часть. Образец титульного листа представлен в Приложении 1.

Обязательным элементом сдачи отчета является устная защита лабораторной работы, в рамках которой студент отвечает на вопросы из контрольного списка.

Оборудование и программное обеспечение, необходимое для выполнения лабораторной работы:

Студент индивидуально выполняет типовой вариант работы на персональном компьютере с операционной системой Windows 7 или выше. Высокоуровневый язык программирования выбирается студентом.

Письменный отчет

Отчет по лабораторной работе представляется в печатном виде в формате, предусмотренном методическим пособием и шаблоном отчета по ЛР (приложение 1).

Критерии оценивания

Защита отчета проходит в форме доклада студента по выполненной работе и ответов на вопросы преподавателя.

Если оформление отчета и поведение студента во время защиты соответствуют указанным требованиям, студент получает максимальное количество баллов.

Отчет не может быть принят и подлежит доработке в случае:

- отсутствия необходимых разделов;
- отсутствия необходимого графического материала;
- некорректной обработки результатов расчёта и т.п.

Максимальное число баллов – 5, минимальное – 2.

Лабораторная работа № 1
Введение в вероятностные методы анализа данных. Числовые характеристики выборки

Цель работы

Целью лабораторной работы является получение студентом основных сведений о типах данных и числовых характеристиках выборок.

Краткие теоретические сведения

Анализ данных (data analysis) – это процесс исследования, очистки, преобразования и моделирования данных с целью обнаружения нужной информации, получения информированных выводов и поддержки принятия решений.

Исходные данные, анализ которых проводится, представляются в виде наборов x_1, x_2, …, x_n; y_1, y_2, …, y_m, z_1, z_2,…, z_l, и т.д., которые отражают значения переменных или признаков x,y,z объектов изучения. Например, если объектом изучения являются пациенты больницы, в качестве x и y могут выступать рост и вес пациента. Тогда каждый пациент i характеризуется парой значений (x_i, y_i).

Пусть рассматривается один признак x, характеризующий объект исследования. При вероятностном анализе данных целью исследователя является установление свойств генеральной совокупности объектов путём анализа ограниченной выборки измерений. Генеральной совокупностью называется множество (конечное или бесконечное) всех сходных объектов, представляющих интерес для исследователя. Выборкой называется подмножество генеральной совокупности x_1, x_2, …, x_n, выбранное для представления генеральной совокупности в статистическом анализе.

Переменная x может иметь один из следующих распространённых типов.

Категориальная переменная содержит данные с ограниченным числом уникальных значений или категорий, в текстовом или числовом формате. Такие данные также называют качественными.

- *Номинальный тип* категориальной переменной представляет собой категории без естественного упорядочения, например, название компании сотрудника, регион, почтовый индекс.
- *Порядковый тип* категориальной переменной представляет категории с естественным упорядочением, например, состояние пациента (от крайне плохого до отличного).

Непрерывная (интервальная) переменная – это переменная, принимающая числовые значения на некотором промежутке. Для таких переменных есть и порядок значений. Примерами интервальных переменных являются рост и вес, температура воздуха.

Для выборок с непрерывными значениями возможно посчитать числовые характеристики. Самыми популярными в практических приложениях являются следующие характеристики:

- Выборочное среднее $\bar{x}$
- Выборочная дисперсия s_x^2
- Выборочное стандартное отклонение s_x
- Медиана h_x

Команды на языке Python для вычисления соответствующих характеристик выборки с использованием библиотеки numpy:

- numpy.mean / numpy.nanmean
- numpy.var / numpy.nanvar
- numpy.std / numpy.nanstd
- numpy.median / numpy.nanmedian

Варианты команд с префиксом nan игнорируют значения NaN (пропуски в данных).

Пример листинга программы с вычислением и выводом выборочного среднего – см. листинг 1.

```
import numpy as np
a = np.array([1,2,3,4,5])
print(np.mean(a))
```

Листинг 1. Вычисление и вывод выборочного среднего.

Ход работы

1. Найти в открытых источниках или выдумать выборку непрерывных величин
2. Реализовать скрипт в Python для загрузки данных (соответствующие команды найти самостоятельно), посчитать и вывести основные числовые характеристики выборки (среднее, дисперсию, стандартное отклонение, медиану).

Вопросы:

1. Задачи анализа данных. Генеральная совокупность и выборка.
2. Типы переменных.
3. Основные числовые характеристики выборки.

Литература

1. Гмурман, В. Е. Теория вероятностей и математическая статистика — 12-е изд. — Москва : Издательство Юрайт, 2014. — 479 с.
2. Hastie, Trevor, Tibshirani, Robert, Friedman, Jerome. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition. Springer, 2009.
3. Кельберт М.Я., Сухов Ю.М. Вероятность и статистика в примерах и задачах. Т. 1: Основные понятия теории вероятностей и математической статистики. М.: МЦНМО, 2007

Лабораторная работа № 2
Первичный и разведочный анализ данных

Цель работы

Целью лабораторной работы является получение студентом навыков первичного и разведочного анализа данных, включая обнаружение выбросов, графический анализ, выдвижение гипотез о наличии связи между переменными.

Краткие теоретические сведения

Первичный анализ данных может включать в себя следующие этапы:

- Просмотр таблицы данных – в случае, если объём данных невелик.
- Графический анализ данных – построение линейчатых графиков, гистограмм, диаграмм рассеяния и диаграмм размаха.

Первичный анализ даёт возможность оценить характер данных и зафиксировать наиболее очевидные ошибки в данных, вызванные неправильным вводом или импортом данных (например, текстовые значения вместо численных), а также пропуски в данных.

Графический анализ позволяет оценить вид распределений величин в выборках (гистограммы), диапазон значений переменных (диаграммы размаха), обнаружить выбросы в данных (диаграммы размаха, линейные графики), оценить наличие и характер связи между парами переменных (диаграммы рассеяния).

Выбросом (грубой ошибкой, англ. outlier) называется результат измерения, выделяющийся из общей выборки. Наличие выбросов в данных может привести к некорректным результатам их статистического анализа, поэтому заблаговременное нахождение и удаление выбросов является важной задачей.

На листинге 2 приведён код для построения линейного графика. Результат выполнения кода приведён на рисунке 1.

```
import matplotlib.pyplot as plt

x = [1,2,3,4,5,6,7,8,9,10]
y = [10,20,34,38,358,59,71,82,89,101]

plt.plot(range(len(y)),y,'-')
plt.xlabel('x')
plt.ylabel('y')
```

```
plt.show()
plt.close()
```

Листинг 2. Код для построения линейного графика

На рисунке видно наличие выброса – элемента выборки у номер 5. Требуется проверить данный элемент на корректность значения (при наличии информации об исходных данных и методе их сбора) и, возможно, удалить его из выборки.

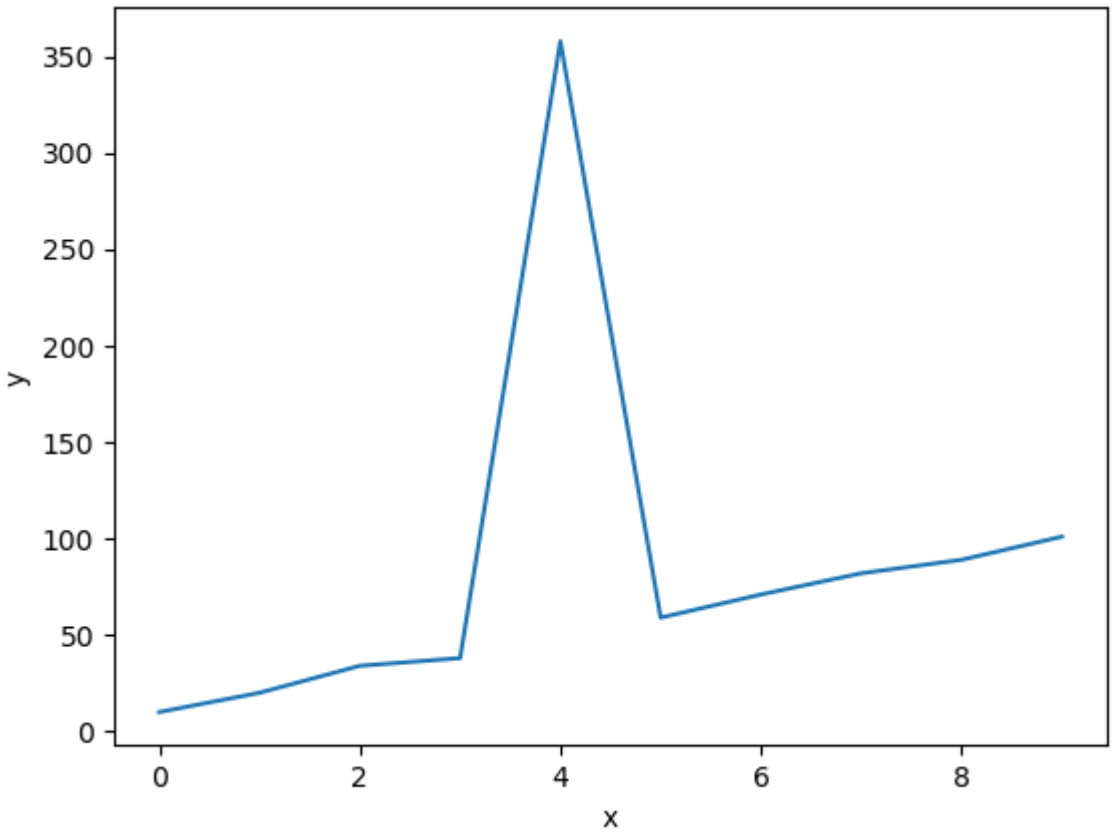


Рисунок 1. Выявление выброса с помощью графического анализа

Простейшим способом обнаружения выбросов в данных является использование межквартильного расстояния (interquartile range, IQR). Для произвольной случайной величины оценка интерквантильное расстояние рассчитывается по формуле IQR = Q3-Q1, где Q1 и Q3 – первый и третий квартили соответственно. Ошибками считаются точки, находящиеся вне диапазона [Q1-k*IQR, Q3+k*IQR], k выбирается равным фиксированному значению – наиболее часто используются значения 1.5 и 3.

Для нахождения IQR на Python используются команды поиска процентилей. Например, для нахождения 75-процентиля (0.75-квантиля) применяется команда numpy.percentile(data, 75) библиотеки numpy. Аналогично находится 25-процентиль. Пример кода для поиска интерквантильного

расстояния и границ отрезка, внутри которого должны лежать точки, не являющиеся ошибками, приведён на листинге 3.

```
import numpy

data = [10,20,34,38,357,59,71,82,89,101]
IQR = numpy.percentile(data, 0.75) - numpy.percentile(data, 0.25)
upper_limit = numpy.percentile(data, 0.75) + (IQR * 1.5)
upper_limit_extreme = numpy.percentile(data, 0.75) + (IQR * 3)
print(upper_limit, upper_limit_extreme)

lower_limit = numpy.percentile(data, 0.25) - (IQR * 1.5)
lower_limit_extreme = numpy.percentile(data, 0.25) - (IQR * 3)
print(lower_limit, lower_limit_extreme)
```

Листинг 3. Пример кода для поиска границ отрезка, внутри которого должны лежать корректные данные

Для попарного графического анализа двух переменных и поиска потенциальной связи между ними удобно использовать диаграмму рассеяния (облако точек). По форме полученного облака можно делать предварительные выводы о наличии или отсутствии связи, а также о её характере. На листинге 4 приведён код на Python для построения облака точек двух переменных. Следует обратить внимание на то, что размеры выборок данных для обеих переменных должны совпадать. На рисунках 2-4 приведены примеры графиков для различных вариантов данных. Судя по виду графика, на рисунке 2 – демонстрируется наличие явной зависимости ординат точек от их абсцисс – можно предположить наличие линейной связи между отображаемыми переменными a и b. На рисунке 3 также прослеживается связь, но она не является линейной. На рисунке 4 наличие связи между переменными маловероятно, поскольку точки с разными ординатами (значение элемента выборки y) могут иметь одинаковые абсциссы (значение элемента выборки x).

```
import matplotlib.pyplot as plt

a = [1,2,3,4,5,6,7,8,9,10]
b = [10,20,34,38,58,59,71,82,89,101]

plt.plot(a,b,'o')
plt.xlabel('a')
plt.ylabel('b')
plt.show()
plt.close()
```

Листинг 4. Код для построения облака точек двух переменных

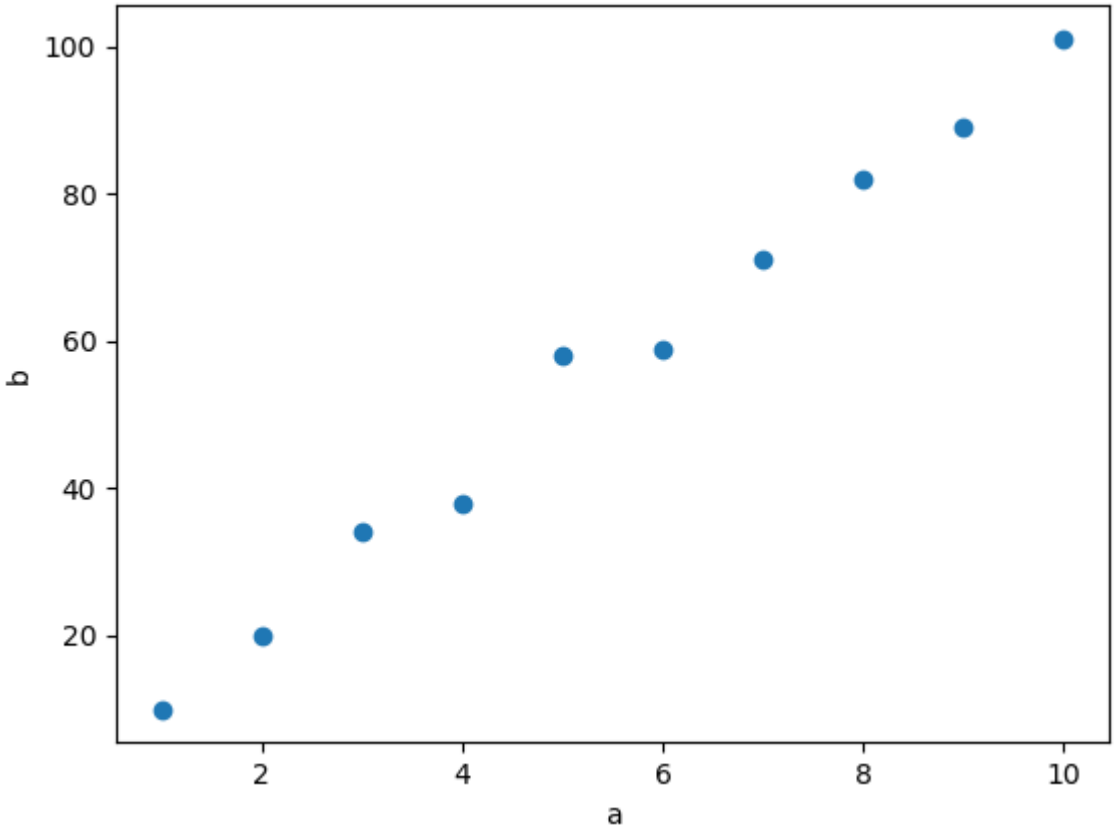


Рисунок 2. Облако точек с предполагаемой линейной связью между переменными a и b.

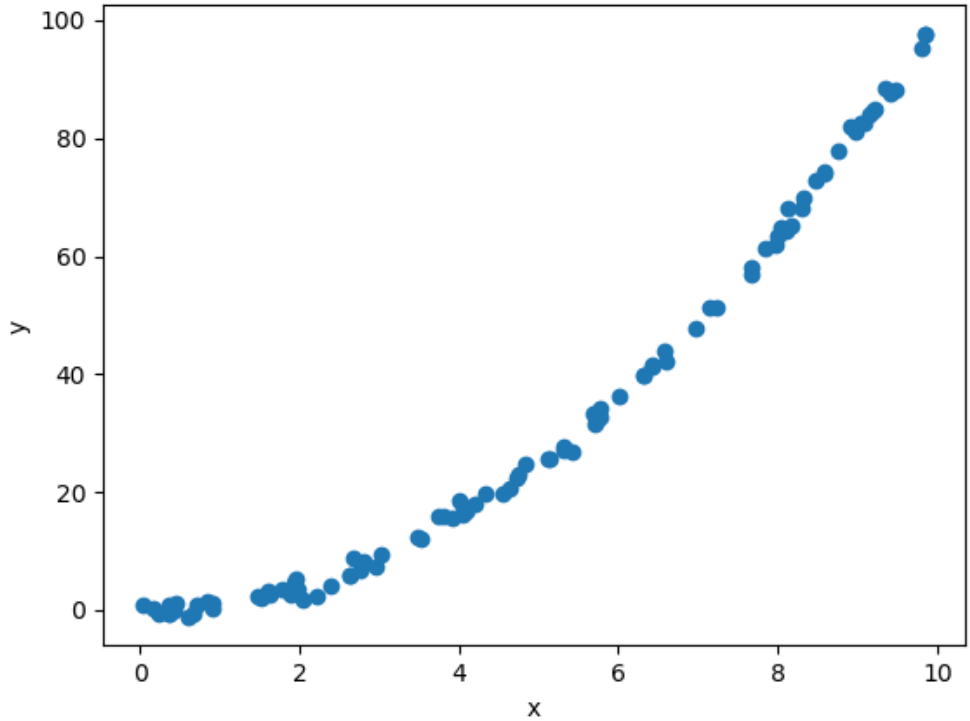


Рисунок 3. Облако точек с предполагаемой нелинейной связью между переменными x и y.

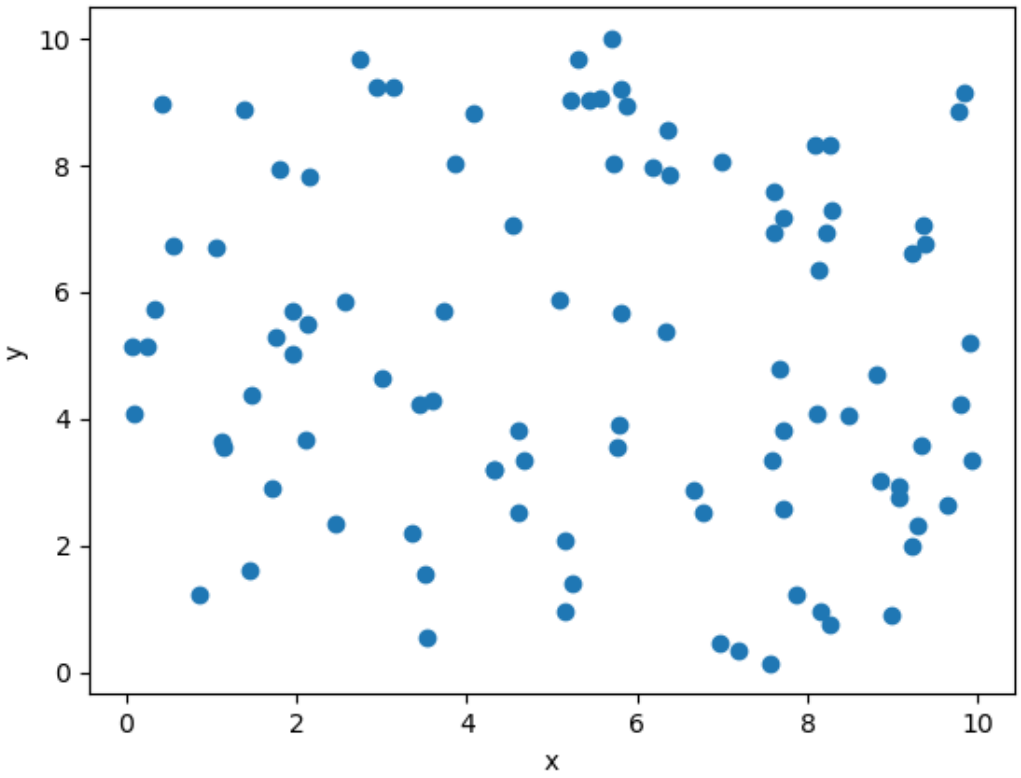


Рисунок 4. Облако точек с предполагаемым отсутствием связи между переменными x и y.

Поиск связи между переменными графическим анализом с помощью построения облаков точек может также производиться в исключительных случаях для трёх переменных, при этом облако точек становится трёхмерным. Однако визуальный анализ таких облаков обычно затруднён.

При необходимости поиска связи между множеством пар переменных в большом массиве данных анализ облаков точек нецелесообразен из-за большой трудоёмкости этой операции. В таких случаях следует опираться на другие методы, в частности, на корреляционный анализ.

Ход работы

1. Найти в открытых источниках или придумать две выборки непрерывных измерений x, y (размер выборки n >20).
2. Отобразить значения графически на одномерном интервале и проверить выборки на наличие грубых ошибок.
3. Построить облако пар точек (x,y) и выдвинуть гипотезу о наличии или отсутствии связи между переменными.

Вопросы

1. Основные способы обнаружения выбросов.
2. Способы графического анализа.

3. Выдвижение гипотез о наличии связи между переменными.

Литература

1. Гмурман, В. Е. Теория вероятностей и математическая статистика — 12-е изд. — Москва : Издательство Юрайт, 2014. — 479 с.
2. Крамер Г. Математические методы статистики. М.: Мир, 1975. - 648 с
3. Ллойд Э., Ледерман У. (Ред.). Справочник по прикладной статистике. В 2-х тт. Пер. с англ. Т.1-2. 1990.

Лабораторная работа № 3
Проверка статистических гипотез

Цель работы

Целью лабораторной работы является получение студентом навыков выдвижения и проверки статистических гипотез на примере применения критерия хи-квадрат для проверки гипотезы о законе распределения случайной величины.

Краткие теоретические сведения

При проведении научного исследования исследователь, как правило, имеет некоторые предварительные представления о данных перед началом их сбора и обработки, которые могут быть сформулированы в виде определённой гипотезы об изучаемом объекте. Эта гипотеза может возникать как результат теоретических исследований или являться следствием предшествующих этапов работы по изучению объекта. Пусть выдвигаемая гипотеза обозначается H_0.
Для принятия или отклонения гипотезы H_0 на основе имеющихся данных, представленных в виде выборки, требуется ответить на вопрос, согласуется ли гипотеза H_0 с рассматриваемой выборкой. Если ответ на данный вопрос положительный, то гипотеза принимается. В противном случае можно принять некую альтернативную гипотезу H_1. Статистические методы, которые позволяют ответить на указанный вопрос, называются методами проверки гипотез.

При проверке гипотезы с помощью статистических методов на основании имеющихся данных последние позволяют получить так называемое p-значение. Данное значение отражает вероятность того, что имеющуюся выборку данных возможно наблюдать при условии, что гипотеза H_0 верна, т.е. вероятность согласования наблюдений с первоначальным представлением об объекте. Предполагается, что все события, вероятность наступления которых очень мала, в практических приложениях можно считать невозможными. В таком случае малое значение p означает необходимость отвергнуть гипотезу H_0. Пороговое значение α, отделяющее возможные события от невозможных, называется уровнем значимости критерия, по которому проверяется гипотеза H_0. На практике обычно берут $\alpha \cdot 100\%$ = 5%, 1%, 0,1%, что соответствует значениям α=0,05; 0,01; 0,001. Если окажется, что $p \geq \alpha$, то гипотеза H0 принимается. Если окажется, что $p < \alpha$, то гипотеза H_0 отклоняется в пользу некоторой альтернативной гипотезы H_1.

В процессе проверки гипотез с помощью специальных статистических методов возможны ошибки, обусловленные конечностью выборочных данных.

Если верная гипотеза H_0 была отклонена статистическим методом, соответствующая ошибка называется ошибкой I рода. Если же неверная гипотеза H_0 была принята, возникает ошибка II рода. Вероятность ошибки I рода совпадает с уровнем значимости критерия α. Число $\alpha \cdot 100\%$ задает риск исследователя отклонить верную гипотезу H_0 на основании обработки исходных данных. Отклонение гипотезы H_0 на уровне значимости $\alpha \cdot 100\% = 5\%$ называют значимым, на уровне $\alpha \cdot 100\% = 1\%$ – статистически значимым, на уровне $\alpha \cdot 100\% = 0{,}1\%$ – высоко статистически значимым.

Примером статистического критерия является критерий хи-квадрат проверки гипотезы о том, что рассматриваемая выборка извлечена из генеральной совокупности с заданным законом распределения. Общая идея критерия хи-квадрат состоит в следующем. Отрезок, содержащий выборочные значения, разбивается на непересекающиеся интервалы (для непрерывных случайных величин) либо на отдельные категории (для дискретных случайных величин). При ручном выборе разбиения важно обеспечить число элементов выборки в каждом интервале, не меньшее пяти. Изучается соответствие наблюдаемого числа попаданий ожидаемому согласно выдвинутой гипотезе о законе распределения. Если вероятность того, что наблюдаемое распределение числа попаданий по интервалам возможно при предполагавшемся законе распределения, превышает значение уровня значимости α, то гипотеза H_0 принимается. В противном случае гипотеза отвергается и предполагается, что выборка извлечена из генеральной совокупности с каким-то другим распределением.

В данной лабораторной работе предлагается в качестве примера выдвижения и проверки статистической гипотезы применить критерий хи-квадрат проверки, извлечена ли выборка из генеральной совокупности с заданным распределением. Для этого требуется совершить следующие шаги:

- Найти произвольную выборку данных.
- Выдвинуть гипотезу H_0 относительно закона распределения, например, «выборка извлечена из генеральной совокупности со стандартным нормальным распределением».
- Построить эмпирическую и теоретическую оценки плотности распределения. На листинге 5 и рисунке 5 приведён пример для нормального и асимметричного нормального распределения.
- Зафиксировать уровень значимости критерия α.
- Применить критерий хи-квадрат (можно использовать встроенные функции в Python или Excel) и получить p-value.
- Сравнить p-value с выбранным уровнем значимости α.

- Отклонить или принять гипотезу H_0 в зависимости от знака неравенства $p \geq \alpha$ ($p < \alpha$).

```
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as stats

m =
[328,454,312,625,609,546,502,736,485,766,429,313,328,344,360,453,563,343
,375,28,312,361,297,437,328,328,328,297,359,328,361,703,500,344,329,312,
328,547,314,328,439,359,126,408,360,346,328,392,453,359]

samples=np.array(m)
mean=np.mean(samples)
var=np.var(samples)
std=np.sqrt(var)

x=np.linspace(min(samples), max(samples),12)
y_pdf=stats.norm.pdf(x,mean,std)
y_skew_pdf=stats.skewnorm.pdf(x,*stats.skewnorm.fit(samples))
l1,=plt.plot(x,y_pdf, label='Нормальное распределение')
l2,=plt.plot(x,y_skew_pdf, label='Асимметричное нормальное
распределение')
n,
bins,patches=plt.hist(samples,12,density=True,facecolor='g',edgecolor='b
lack', alpha=0.75)

plt.xlabel('Значение величины')
plt.ylabel('Частота')

plt.legend((l1,l2),(l1.get_label(), l2.get_label()), loc='upper right')

plt.axis([126, 766, 0, 0.01])
plt.show()
```

Листинг 5. Код для отрисовки теоретических оценок плотностей распределений и эмпирической функции плотности

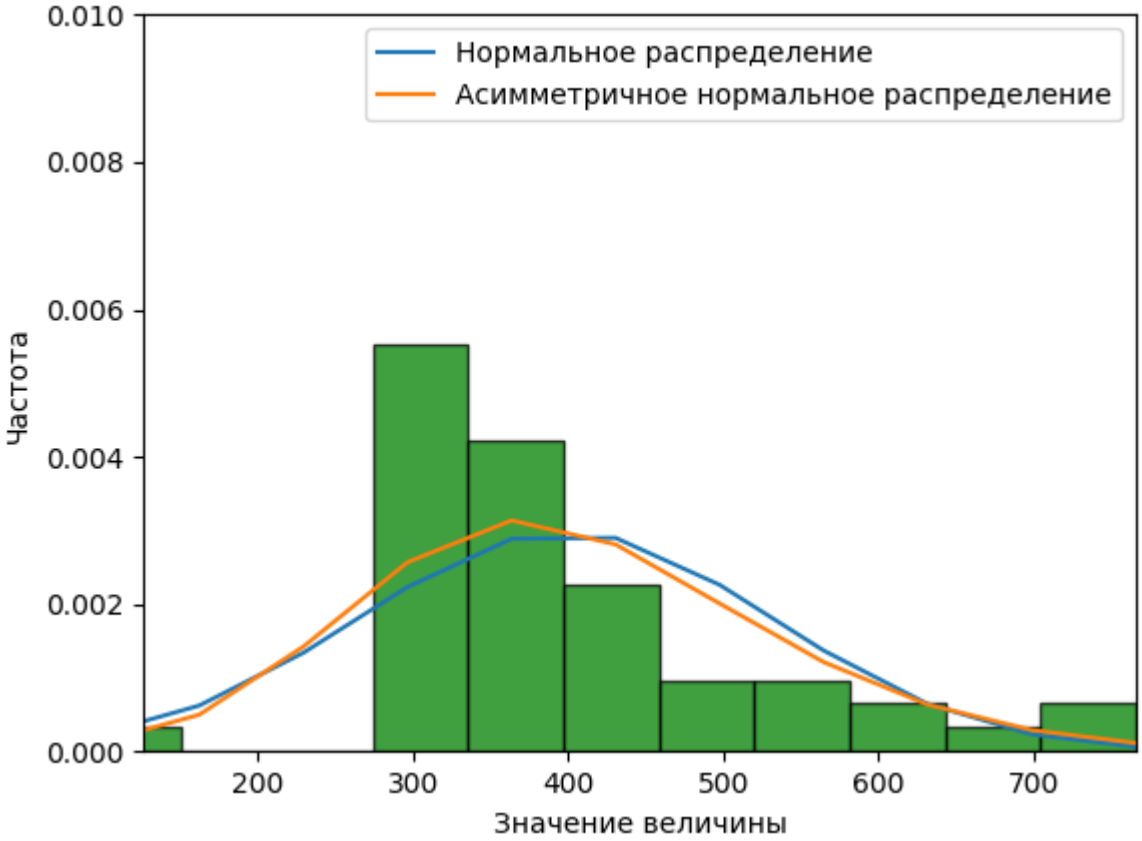


Рисунок 5. График теоретических оценок плотностей распределений и эмпирической функции плотности

Ход работы

1. Построить гистограмму плотности случайной величины и выдвинуть три разные гипотезы о законе её распределения.
2. Проверить выдвинутые гипотезы с помощью критерия хи-квадрат.
3. Самостоятельно найти функцию для проверки гипотезы о законе распределения случайной величины с помощью критерия Колмогорова-Смирнова в библиотеке scipy.stats. Применить её для рассматриваемых данных и сравнить результаты.

Вопросы

1. Проверка статистических гипотез.
2. Применение критерия хи-квадрат для проверки гипотезы о законе распределения случайной величины.

Литература

1. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика, Основы моделирования и первичная обработка данных. М.: 1983.

2. Гмурман, В. Е. Теория вероятностей и математическая статистика — 12-е изд. — Москва : Издательство Юрайт, 2014. — 479 с.
3. Королюк В.С., Портенко Н.И., Скороход А.В. Турбин А.Ф. Справочник по теории вероятностей и математической статистике. М.: 1985.

Лабораторная работа № 4
Корреляционные зависимости

Цель работы

Целью лабораторной работы является получение студентом навыков поиска связи между выборками и оценки силы найденной связи с помощью коэффициентов корреляции.

Краткие теоретические сведения

Рассмотрим переменные x и y, отражающие свойства некоторого объекта исследования. Будем говорить, что переменные x и y являются зависимыми, если допустимые значения пары (x; y) согласованы друг с другом по некоторому закону. Зависимости могут иметь скрытый характер и сложный вид, в силу этого их приходится искать в приближённой форме.

Для оценки наличия и силы зависимости между переменными используются выборочные коэффициенты корреляции. Коэффициенты корреляции могут принимать значения от -1 до 1, знак коэффициента указывает на направление связи (прямая или обратная). В случае если коэффициент корреляции близок к нулю, связь между переменными отсутствует. Значимость коэффициентов корреляции (значимое отличие их от нуля) может быть проверено с помощью специальных статистических критериев. В прикладных задачах обычно выбирается некоторое пороговое значение коэффициента корреляции (например, 0,6), которое позволяет говорить о наличии значимой связи между переменными. Выбор порогового значения зависит от области исследования и интерпретации рассматриваемого параметра.

Существует несколько разных коэффициентов корреляции, выбор между которыми осуществляется исходя из задачи исследователя. Если между некоторыми переменными x и y интервального или порядкового типа предполагается связь в монотонной форме, то существует некоторая монотонная функция $y \approx g(x)$. Для установления или отрицания факта такой связи используется понятие ранговой корреляции по Спирмену. Формула для вычисления выборочного коэффициента ранговой корреляции по Спирмену имеет следующий вид:

$$\hat{r}_s = 1 - \frac{6}{n^3 - n} \sum_{i=1}^{n} (\hat{x}_i - \hat{y}_i)^2,$$

где n – объём выборок (должен совпадать для x и y), $\hat{x}_i$ и $\hat{y}_i$ – ранги i-х элементов выборок x и y.

Для установления или отрицания факта связи между x и y интервального типа в линейной форме, т.е. в виде y≈a + bx, где a, b - некоторые константы, используется понятие корреляции по Пирсону. Значение выборочного коэффициента корреляции по Пирсону находится с помощью следующей формулы:

$$\bar{r}_{xy} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{s_x s_y},$$

где n – объём выборок, $\bar{x}$, $\bar{y}$ – выборочные средние, s_x, s_y – среднеквадратичные отклонения для выборок x и y соответственно.

Поскольку коэффициент корреляции по Пирсону менее устойчив к выбросам в данных, чем коэффициент корреляции по Спирмену, и может быть рассчитан только для переменных интервального типа, в практических задачах обычно имеет смысл отдавать предпочтение коэффициенту корреляции по Спирмену.

Команды на языке Python для вычисления коэффициентов корреляции по Пирсону и Спирмену с использованием библиотеки scipy.stats:

- scipy.stats.pearsonr
- scipy.stats.spearmanr

Ход работы

1. Рассмотреть следующие две выборки:
   X = {1,2,3,4,5,6,7,8,9,10}
   Y = {10, 34, 46, 55, 61, 76, 84, 99, 103, 110}
2. Построить облако точек для переменных X и Y.
3. Найти значения коэффициентов корреляции по Пирсону и по Спирмену между указанными переменными.

Вопросы

1. Зависимость между переменными. Формулы зависимости.
2. Ранговый коэффициент корреляции по Спирмену.
3. Коэффициент корреляции по Пирсону.

Литература

1. Гмурман, В. Е. Теория вероятностей и математическая статистика — 12-е изд. — Москва : Издательство Юрайт, 2014. — 479 с.
2. Hastie, Trevor, Tibshirani, Robert, Friedman, Jerome. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition. Springer, 2009.

Лабораторная работа № 5
Линейная регрессия

Цель работы

Целью лабораторной работы является получение студентом навыков поиска формулы линейной связи между выборками x и y на основе формулы парной регрессии.

Краткие теоретические сведения

Рассмотрим переменные x и y, отражающие свойства некоторого объекта исследования (например, рост и вес пациентов). В случае, если переменные x и y, являются зависимыми и связь между ними имеет линейный вид, возможен поиск конкретной формулы этой связи на основе линейной регрессии. Форма связи между зависимой переменной y и объясняющей переменной x в виде парной регрессии имеет следующий вид:

$$y = a + bx + u,$$

где a, b – параметры регрессии, u – скрытая переменная, отражающая влияние неизвестных факторов на y. Предполагается, что математическое ожидание переменной u равно нулю, а дисперсия D(u) является постоянной величиной.

Значения параметров регрессии a и b могут быть найдены на основе значений элементов выборок x и y с помощью метода наименьших квадратов. После нахождения оценочных значений параметров a^* и b^* формируется выборка остатков на основе следующей формулы:

$$e_i = y_i - (a^* + b^* x_i), i = 1,2, \dots, n.$$

Выборка остатков отражает значения скрытой переменной u и требуется для проверки вышеуказанных априорных предположений относительной этой переменной.

Код с примером расчёта параметров парной линейной регрессии и построения прогноза с её помощью приведён на листинге 6.

```
import numpy as np
from sklearn.linear_model import LinearRegression

X = np.array([[1, 1], [1, 2], [2, 2], [2, 3]])
y = np.dot(X, np.array([1, 2])) + 3
reg = LinearRegression().fit(X, y)
print(reg.score(X, y))
print(reg.coef_)
print(reg.intercept_)
print(reg.predict(np.array([[3, 5]])))
```

Листинг 6. Код для расчёта параметров парной регрессии и построения прогноза

Для оценки качества описания связи между выборками с помощью найденной регрессии можно использовать значение коэффициента детерминации R^2, отражающего, насколько лучше регрессия с объясняющей переменной x описывает значения элементов выборки y по сравнению с выборочной оценкой математического ожидания My. В языке программирования Python значение коэффициента детерминации может быть рассчитано с помощью команды r2_score библиотеки sklearn.metrics.

Ход работы

Даны две выборки:
X = {1,2,3,4,5,6,7,8,9,10}
Y = {10, 34, 46, 55, 61, 76, 84, 99, 103, 110}
Построить линейную регрессию Y=f(X) для приведенных выборок, нарисовать её график и привести формулу. Посчитать коэффициент детерминации R^2.

Вопросы

1. Линейная регрессия. Формула парной регрессии.
2. Выборка остатков. Коэффициент детерминации.

Литература

1. Гмурман, В. Е. Теория вероятностей и математическая статистика — 12-е изд. — Москва : Издательство Юрайт, 2014. — 479 с.
2. Hastie, Trevor, Tibshirani, Robert, Friedman, Jerome. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition. Springer, 2009.
3. Крамер Г. Математические методы статистики. М.: Мир, 1975. - 648 с

Лабораторная работа № 6
Анализ влияния факторов

Цель работы

Целью лабораторной работы является получение студентом навыков оценки влияния фактора, задаваемого категориальной переменной, на значения интервальной переменной на основе использования критериев Манна-Уитни и Краскела-Уоллиса.

Краткие теоретические сведения

Пусть изучается влияние фактора A на значения некоторой переменной S (например, влияние диеты на вес пациентов, влияние терапии на артериальное давление, влияние широты города на дневную температуру). Пусть влияние фактора отражается некоторым конечным числом уровней, тогда A представляется категориальной переменной. Задача оценки влияния фактора может быть решена с помощью подтверждения или опровержения гипотезы о неизменности генерального среднего M(S) при изменении значения A. В ситуации, когда о законе распределении выборочных данных ничего не известно заранее, для решения указанной задачи могут применяться непараметрические методы.

Случай 1. Фактор A имеет два уровня, A_1 и A_2. Предположим, что значения переменной S при уровнях фактора A_1 и A_2 задаются выборками x и y соответственно. Тогда анализ влияния фактора A на переменную S сводится к проверке гипотезы H_0, заключающейся в том, что M(x)=M(y), т.е. математическое ожидание переменной S не зависит от уровня фактора. Данная гипотеза может быть проверена с помощью критерия Манна-Уитни. Процедура применения критерия совпадает с описанной в лабораторной работе 2 и заключается в следующем:

- Фиксируется уровень значимости критерия α.
- В результате применения критерия Манна-Уитни к выборкам x и y (в языке программирования Python может быть применена команда mannwhitneyu библиотеки scipy.stats) получается p-значение критерия.
- По результатам сравнения p-значения с выбранным уровнем значимости α делается вывод о принятии или отклонении гипотезы H_0.

Случай 2. Фактор A имеет несколько уровней, A_1, A_2, …, A_k. Предположим, что значения переменной S при уровнях фактора A_i задаются выборками x_i, i=1,2,…,k. Анализ влияния фактора A на переменную S сводится к проверке гипотезы H_0, заключающейся в равенстве математических ожиданий $M(x_1)$, $M(x_2)$, …, $M(x_k)$. Для проверки данной гипотезы может быть использован

критерий Краскела-Уоллиса. Последовательность шагов для принятия или отклонения гипотезы аналогична случаю двух уровней факторов. Применение критерия Краскела-Уоллиса к проверяемым выборкам x_i на языке программирования Python осуществляется командой kruskal библиотеки scipy.stats.

Несложно заметить, что критерий Краскела-Уоллиса может быть применен для двух уровней факторов, тем самым дополняя критерий Манна-Уитни.

Ход работы

Заданы выборки X, Y, Z цен на определенное лекарство в аптеках г. Москвы, Санкт-Петербурга и Омска:

X = {100, 129, 205, 134, 0, 130, 156, 130, 141},

Y= {98, 110, 102, 96, 97, 93, 101, 90, -110, 91, 94, 105, 90},

Z = {56, 78, 96, 76, 69, 89, 61, 63, 60, 71, 68}.

1. Проанализировать данные на наличие грубых ошибок, выдвинуть и обосновать предположения относительно элементов выборок, которые следует исправить или удалить.
2. С помощью применения статистических критериев ответить на вопрос: «В каком из трёх городов наиболее выгодно покупать данное лекарство?» Обосновать выбор использованных критериев.

Вопросы

1. Оценка влияния факторов. Непараметрические методы.
2. Метод Манна-Уитни для двух уровней факторов. Метод Краскела-Уоллиса для нескольких уровней факторов.

Литература

1. Гмурман, В. Е. Теория вероятностей и математическая статистика — 12-е изд. — Москва : Издательство Юрайт, 2014. — 479 с.
2. Hastie, Trevor, Tibshirani, Robert, Friedman, Jerome. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition. Springer, 2009.
3. Кобзарь А. И. Прикладная математическая статистика. — М.: Физматлит, 2006. — 466—468 с.

ПРИЛОЖЕНИЕ 1

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
САНКТ–ПЕТЕРБУРГСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

Отчет

о выполнении лабораторной работы № X
«Тема работы»

Работу выполнил:
ст. группы <Номер группы> Фамилия И.О.
Работу принял:
Леоненко В.Н.

Санкт–Петербург
202_

Цель работы

Какая цель преследуется при выполнении лабораторной работы (2–3 строки).

Постановка задачи

Задача, которая решается при выполнении этой лабораторной работы (1 абзац на 0.2 – 0.3 стр.).

Краткая теоретическая часть

Краткие сведения о теме дисциплины, по которой выполняется лабораторная работа. Сведения об используемых методах, методиках, алгоритмах: свойства, достоинства, недостатки (не более 1 стр.).

Результаты

Представление результатов (промежуточные и итоговые изображения). Краткое обсуждение результатов (что означают конкретные значения) – 1–3 стр.

Заключение

Что сделано. Какие навыки и умения приобретены. Прогноз возможностей применения навыков и умений, а также полученных результатов (5–10 строк).

Леоненко Василий Николаевич

ВЕРОЯТНОСТНЫЕ МЕТОДЫ АНАЛИЗА ДАННЫХ

Учебно–методическое пособие

В авторской редакции
Редакционно–издательский отдел Университета ИТМО
Зав. РИО Н.Ф. Гусарова
Подписано к печати
Заказ №
Тираж
Отпечатано на ризографе

Редакционно–издательский отдел
Университета ИТМО
197101, Санкт–Петербург, Кронверкский пр., 49