Введение

Социологические исследования имеют крайне широкое распространение, их целью является сбор информации о различных социальных процессах и явлениях, а также связях между ними.

Целью работы является изучение методов обработки и анализа данных социологических исследований, создание на их основе умного помощника для специалистов, проводящих социологические исследования, и его практическое применение к эмпирическим данным, полученным в результате социологического опроса трудоустроенных представителей молодёжи Крыма.

Задачи работы:

1) Изучение правил проведения и методов выборки в социологических исследованиях.
2) Изучение методов обработки и анализа эмпирических данных социологических исследований.
3) Создание умного помощника для специалистов, проводящих социологические исследования.
4) Сбор данных среди трудоустроенных представителей молодёжи Крыма для социологического исследования.
5) Обработка и анализ эмпирических данных социологического исследования с помощью умного помощника, в ходе которого будет получено представление об участии граждан в общественной жизни их региона.

Объектом исследования являются методы обработки и анализа эмпирических данных, полученных в ходе проведения социологического исследования.

Предметом исследования является реализация методов обработки социологических данных с помощью средств языка программирования высокого уровня Python.

Для проведения социологического исследования необходимо выполнить следующие этапы: провести статистическое наблюдение, обработать полученные данные, провести их анализ, проверки статистических гипотез.

Работа состоит из двух разделов. Первый раздел называется «Статистические методы обработки эмпирических данных» и

рассматривает правила и методы проведения выборки, методы обработки и анализа данных социологических исследований.

Второй раздел «Создание умного помощника для анализа данных с помощью средств языка Python» - включает в себя практическую часть, в которой применяются методы, рассмотренные в первом разделе при создании умного помощника для специалистов, проводящих социологические исследования, и их применение к эмпирическим данным, полученным в ходе проведения социологического исследования «Изучение участия трудоустроенных представителей молодёжи Крыма в общественной жизни их региона».

РАЗДЕЛ 1. СТАТИСТИЧЕСКИЕ МЕТОДЫ ОБРАБОТКИ ЭМПИРИЧЕСКИХ ДАННЫХ

1.1. Основные понятия

Социологическое исследование – системный процесс изучения различных процессов и явлений в обществе в целях получения новой научной информации об общественных явлениях и процессах в обществе. Эта информация может позволить предугадывать возможные итоги происходящих в обществе процессов, находить их причины, и показывать, как можно повлиять на эти процессы и явления [1].

Методы обработки и анализа эмпирических данных в социологии могут давать прогнозы только с той или иной степенью вероятности.

Для осуществления любого социологического исследования необходимо выполнить следующие этапы:

1) Статистическое наблюдение, которое можно представить в виде совокупности случайных событий и случайных величин.
2) Обработка данных, полученных в ходе статистического наблюдения, и их представление в удобном для последующего анализа данных виде.
3) Анализ обработанных данных, работа с целью проверки статистических гипотез, изучения результатов исследования и создания выводов на основе проведённого анализа данных. Данные выводы позволят далее оказывать влияние на происходящие явления и процессы.

В процессе социологического исследования производится измерение, то есть придание, согласно правилам, зависящим от постановки задачи, некоторых числовых значений признакам объектов и объектам. Вышеуказанные действия необходимы для получения требуемых результатов социологического исследования в виде математической модели исследования.

В ходе социологического исследования используются разнообразные измерительные шкалы. В процессе социологического исследования шкалу измерений называют основным инструментом измерения. Она служит эталоном для определённого набора значений, которые рассматриваются в

процессе исследования. В результате своих действий с её помощью исследователь может привести к количественным показателям, которые будут сопоставимыми, значения, которые ранее не могли сопоставляться, так как были качественно различными. Характер признаков, которые планируется измерять, и постановка задачи определяют то, какой тип шкалы будет применён. Существуют номинальные, ранговые (порядковые) и метрические шкалы [2].

Метрические шкалы делятся на два подтипа: шкала отношений и интервальная шкала. Для отражения отношений пропорции используется шкала отношений. Во время анализа силы проявления силы свойств объектов, выражаемых некоторыми величинами, которые разделены на интервалы равной длины, используется интервальная шкала. При этом нуль на данном типе шкалы является условным [2].

Ранговая (порядковая) шкала, как правило, используется при сравнении силы проявления признаков по убыванию и возрастанию. При использовании порядковой шкалы каждому рангу присваивается некоторое число. Вместе с этим, если числа на шкале заменить на иные, то порядок расположения рангов на шкале сохраняется, не изменяя уже имеющегося порядка [2].

Номинальная шкала, как правило, используется при классификации объектов и их характеристик. Она используется для определения различных групп объектов, когда значения на шкале не поддаются сравнению между собой, то есть каждой группе присваивается определённая позиция на шкале [2].

Для сбора информации в социологических исследованиях применяют различные методы. Основные из них: анкетирование, интервью, наблюдение, эксперимент, анализ документов. Каждый из этих методов имеет свои преимущества и недостатки. В данной работе применяется только анкетирование.

В социологическом исследовании для построения графиков и различных расчётов можно не использовать программное обеспечение, однако его использование позволяет существенно ускорить ход вычислений, увеличить точность вычислений. Для подобных целей может применяться множество прикладных программ, например, Libre Office, Microsoft Office, STATISTICA, SPSS и другие, так же возможно применение языков программирования со специальными библиотеками, как Python, R и другие.

1.2. Язык программирования для анализа данных

Существует множество языков программирования, которые имеют готовые библиотеки для специализированных методов обработки и анализа данных, а также их отображения, которые применяются при проведении социологического исследования.

Одними из наиболее часто применяемых языков программирования при проведении социологических исследований используют языки программирования Python и R. Оба языка имеют широкий выбор уже существующих библиотек, которые позволяют упростить написание кода, который будет использоваться при проведении социологического исследования.

Язык программирования R был специально создан для математических расчётов, статистического анализа данных и машинного обучения.

Язык программирования Python является универсальным языком программирования, имея множество сфер применения, и при этом не уступает языку R в функциональности при использовании в математических расчётах, статистическом анализе и машинном обучении.

В данной работе был выбран язык программирования Python, так как он ни в чём значительно не уступает другим языкам программирования при использовании в социологическом исследовании, и именно в Python имелся наибольший опыт написания кода.

При создании умного помощника использовались библиотеки Python: numpy, pandas, matplotlib, math, scipy, warnings, copy. В программе это реализовано следующим образом:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from math import *
import scipy as sci
import warnings
import copy
```

Листинг 1.2.1. Подключение библиотек.

1.3. Правила и методы сбора эмпирических данных

Генеральная совокупность в социологии является совокупностью всех людей с некоторой определенной характеристикой. Опросить всех представителей генеральной совокупности часто физически или по разумным причинам невозможно. Тогда применяются методы выборочного опроса, он заключается в отборе небольшого числа людей из общей совокупности по определенным правилам в качестве социальной модели, воспроизводящей структуру объектов исследования. Процесс отбора группы таких людей, а также сама группа людей называется выборкой [2].

В то же время результаты исследования зависят от правильного проведения выборки, поскольку выборка представляет собой упрощенную форму общей совокупности и не в полной мере отражает ее разнообразие.

Поэтому необходимо соблюдать правила составления выборочной совокупности:

1) В качественном исследовании выборка должна быть не однородной (геторогенной) [2].
2) В количественном исследовании выборка должна быть однородной (гомогенной) [2].
3) Выборка должна быть репрезентативной [2].
4) Каждый элемент генеральной совокупности должен иметь одинаковую вероятность попадания в выборку [2].

В социологии под репрезентативностью понимается свойства выборки, которые позволяет ей выступать в качестве модели генеральной совокупности на момент переписи. Репрезентативными выборками считаются выборки, основные характеристики которых соответствуют аналогичным характеристикам генеральной совокупности [2].

Однородные группы людей должны совпадать по основным характеристикам, а разнородные группы должны отличаться по основным характеристикам.

Метод выборки является способом создания выборки. Каждый тип выборки имеет свои собственные математические инструменты и процедуры [2].

Методы выборки делятся на вероятностные (статистические) и целевые (не статистические) [2].

Вероятностные методы включают в себя:

1) Простой случайный отбор.
2) Стратифицированный отбор.
3) Кластерный (гнездовой) случайный отбор.
4) Систематический отбор [2].

Простой случайный отбор из генеральной совокупности заключается в том, что генеральная совокупность однородна, и что все ее элементы также могут быть использованы в исследованиях в равной степени, имеется список всей совокупности элементов, составляющих генеральную совокупность. Для получения полного списка используются процедуры случайного отбора (в частности, применение генератора случайных чисел) [2].

Стратифицированная выборка делит объём выборки между всеми стратами пропорционально их численности, и извлекает простые случайные выборки из каждой страты. Этот метод обеспечивает равномерно распределённое представительство всевозможных групп и (или) типов населения в выборочной совокупности [2].

Кластерная выборка (гнездовая) является типом выборки, при котором выбранными объектами образуется кластер (гнездо, группа) из меньших единиц. Группы выбираются случайным образом (в некоторых случаях с вероятностью, пропорциональной их количеству), и подвергаются изучению полностью или выборочно [2, 4].

Системный отбор заключается в выборе из списка представителей генеральной совокупности людей с помощью определённого количества номеров [2].

Целевые методы включают в себя:

1) Квотную выборку.
2) Метод «снежного кома».
3) Метод типичных представителей.
4) Метод стихийного отбора на основе принципа удобства.
5) Метод на основе суждений [2].

Квотная выборка является уменьшенной моделью объекта социологического исследования. Она устанавливается на основе статистических данных (квотных параметров) о социальном и демографическом характере элементов генеральной совокупности. Методология основан на намеренном установлении структуры выборочной совокупности. Например, в ходе исследования была предпринята попытка опросить некоторое количество людей определенных возрастов, пола,

уровней образования и профессий. Удельные квоты в выборочной совокупности обязаны быть приведены в соответствие к её удельному весу в генеральной совокупности. Обычно квотная выборка используется при последних стадиях проведения отбора [2].

Метод "снежного кома" применим тогда, когда существует предположение о том, что отбор дополнительных (последующих) респондентов производится после ссылки на них ранее отобранных. Этот метод применяется для изучения особенных, редких и неслучайных совокупностей [3].

Метод типичных представителей лучше всего применим на последних стадиях отбора, в случае, когда нужно использовать небольшую численность объектов. Только тогда, когда существует обоснование выбора объектов, данный метод может в значительной мере обеспечить репрезентативность выборки. Для этого нужно собрать дополнительную информацию о тех признаках, которые позволяют рассматривать их в качестве контрольных [3].

Метод стихийного отбора аналогичен случайному отбору, но в случае стихийного отбора необходимо опрашивать тех, кто внешне похож на представителей генеральной совокупности [3].

Смысл метода отбора на основе принципа удобства сводится к тому, чтобы создать экземпляр наиболее удобным способом с точки зрения исследователя, например, с точки зрения доступности респондентов и (или) с меньшими затратами времени и усилий.

Создание выборки на основе суждений основывается на учёте суждений квалифицированных экспертов о составе выборки. С использованием этого подхода часто создаются элементы фокус-групп [4].

На практике часто используется многоступенчатая выборка, при которой набор объектов, выбранных на предыдущем этапе, используется в качестве исходного объекта на следующем этапе. Объекты самого низкого этапа, полученные из непосредственно собранных данных, называются единицами наблюдения. Многоуровневая выборка используется, когда генеральная совокупность велика и разнообразна, а рандомизация, достигаемая другими методами, приводит к чрезмерной дисперсии выборки.

После сбора эмпирических данных каким-либо из выше представленных методов перед их использованием с помощью умного блокнота для помещения их в форму, которую сможет обработать

программа умного помощника, может применяться множество прикладных программ, например, Libre Office, Microsoft Office, в которых необходимо создать таблицу с собранными данными в формате .csv.

Необходимо поместить данные о каждом критерии об исследуемом объекте в отдельный столбец с соответствующем названием. Столбцы должны располагаться вплотную друг к другу в левой верхней части таблицы. Используется кодировка UTF-8.

1.4. Методы обработки социологических данных

Характеристика, рассматриваемая в социологическом исследовании, принимается за X. Значения X (т. е. случайной величины), получаемые в процессе сбора эмпирических данных, обозначаются через $x^{(1)}, x^{(2)}, \dots, x^{(n)}$ и называются вариантами – значениями признака. Количество объектов, имеющих данный признак, называется частотой варианты. Множество всех возможных вариант называется генеральной совокупностью. Любое конечное подмножество из генеральной совокупности называется выборкой [4].

$\{x^{(1)}, x^{(2)}, \dots, x^{(n)}\} = X$, где n – объём выборки. Значения $x^{(i)}$ располагают в порядке возрастания:

$$x_1, x_2, \dots, x_n \ (x_1 < x_2 < \dots < x_n).$$

Некоторые варианты x_i могут встречаться в выборке несколько раз.

Предположим, было опрошено некоторое количество респондентов с помощью анкеты.

1. Пол:

1) Мужской
2) Женский

2. Сколько вам полных лет?

3. Как часто вы посещаете мероприятия культурной направленности в вашем регионе?

1) Постоянно
2) Часто
3) Редко
4) Никогда

…

Здесь представлены признаки, которые можно распределить на шкалах трёх типов: результаты первого вопроса можно изобразить на номинальной шкале, второй – на метрической, третий – на порядковой. Необходимо перенести данные из анкет в специальную табл. 1.4.1.

Таблица 1.4.1

Условный пример: данные социологического опроса 30 респондентов

Номер анкеты			
	Пол	Возраст	Частота просмотра новостей культурной направленности о регионе респондента
1	1	1	3
2	2	4	1
…	…		…
30	1	6	4

Такие таблицы называют матрицами данных.

В программе входные данные представляются в виде объекта типа DataFrame библиотеки Pandas:

```
df_start = pd.read_csv(path_empiric_data)
```

Листинг 1.4.1. Загрузка данных из таблицы.

Тут df_start – переменная типа pandas.DataFrame для хранения загруженной таблицы, path_empiric_data – переменная типа string для хранения пути к файлу на диске.

Вариационным рядом называют ряд, разделённый в порядке возрастания или убывания вариант с соответствующими им весами. Различают непрерывные и дискретные вариационные ряды.

В табл. 1.4.1 представлены дискретные вариационные ряды различных признаков.

Так же можно составить табл. 1.4.2, которая называется дискретным вариационным рядом выборки. В ней содержатся частоты, частности и проценты.

Таблица 1.4.2

Дискретный вариационный ряд выборки

Варианты, x_i	x_1	x_2	…	x_k
Частоты, n_i	n	n_2	…	n_k
Относительные частоты, w_i	w_1	w_2	…	w_k

Для отображения, какое количество раз в данных выборки встречается варианта x_i используется эмпирическая частота выборки n_j. Вместе с этим

$$\sum_{i=1}^{k} n_i = n \tag{1.4.1}$$

Отношение

$$w_i = \frac{n_i}{n} \tag{1.4.2}$$

называется относительной частотой варианты x_i. Тут n_i – частота появления варианты x_i, а $\boldsymbol{n}$ – объём выборки [5].

Пусть имеется дискретный вариационный ряд выборки, тогда ломаная линия с вершинами (x_k, n_k) называется полигоном частот. А ломаная линия с вершинами (x_k, n_k) – полигоном относительных частот [6].

Чаще всего при $\boldsymbol{n}$ > 30 данные помещают в интервальный вариационный ряд.

Для создания вариационного ряда вычисляется диапазон изменения R признака X, как разность между значениями признака - наибольшим x_{max} и наименьшим x_{min}:

$$R = x_{max} - x_{min} \tag{1.4.3}$$

R разбивается на k равных интервалов. Как правило, пользуются одним из правил для выбора числа k, хотя теоретически можно взять любое k:

1) $6 \le k \le 20$
2) $k \approx \sqrt{n}$ (1.4.4)
3) $k \approx 1 + \log_2 n \approx 1 + 3{,}221 \lg n = 1 + 1.44 \cdot \ln n$

k приравнивается от 6 до 10, а каждому интервалу присваивается ширина h=R/k в том случае, когда объём выборки не большой. Как правило, h округляется до некого значения d [6].

Составляется таблица 1.4.3, называемая интервальным вариационным рядом выборки.

Таблица 1.4.3

Интервальный вариационный ряд выборки

Интервалы	$[x_0; x_1)$	$[x_1; x_2)$	…	$[x_{k-1}; x_k)$
Частоты, n_j	n_1	n_2	…	n_k
Относительные частоты, w_i	w_1	w_2	…	w_k

Можно составить по полученным данным гистограмму выборки, или гистограмму относительных частот выборки.

$F_n^*(x)$ является эмпирической функцией распределения выборки $\{x_k\}_1^n$, и определяется равенством:

$$F_n^*(x) = \frac{n_x}{n}, \quad (1.4.5)$$

где n – объём выборки, n_x – число вариант выборки, меньших, чем x.

Теоретическая функция определяет вероятность события $X < x$. Эмпирическая функции различаются тем, что теоретическая функция определяет относительную частоту этого события.

Эмпирическая функция всегда принимает значения в диапазоне значений [0; 1] (рис. 1.4.1).

Выброс (экстремальное значение, аномальное значение) – редкое значение характеристики, значительно отличающееся от других значений.

Выбросы могут быть вызваны как ошибкой статистического сбора данных (например, неточность приборов измерения), так и значительно отличающимися редкими крайне значениями, которые реально встречаются на практике.

Для того, чтобы решить, какие данные необходимо считать выбросами, определяются границы выбросов. Они могут устанавливаться по разным правилам с учётом исследуемых данных и задач исследования.

Удаляются выбросы потому, что оказывают значительное воздействие на различные метрики и результаты анализа данных, которое не пропорционально относительной частоте их наблюдения.

Распространённой практикой считаются границы выбросов, равные не попадают в интервал

$$[Q1 - 1{,}5 * IQR; Q3 + 1{,}5 * IQR], \quad (1.4.6)$$

где Q1 и Q3 – первый и третий квартили, IQR – расстояние между первым и третьим квартилем.

Все значения, не попадающие в этот интервал, считаются выбросами.

В программе нахождение границ выбросов реализовано следующим образом:

```
# Первый и третий квартили:
q1 = table[s].quantile(0.25)
q3 = table[s].quantile(0.75)
# Межквартальный размах:
iqr = q3 - q1
# Границы выбросов:
lower_bound = q1 - 1.5 * iqr
upper_bound = q3 + 1.5 * iqr
```

Листинг 1.4.2. Вычисление границ выбросов.

Тут table – переменная типа pandas.DataFrame для хранения таблицы, s – переменная типа str для хранения строки с названием признака - столбца таблицы, q1, q3 - переменные типа numpy.float64 для хранения значений первого и третьего квартилей, iqr – переменная типа numpy.float64 для хранения значения межквартильного интервала, lowerbound и upper_bound – переменные типа numpy.float64 для хранения значений играниц выбросов.

Существует три метода обработки выбросов: удаление, трансформация, вменение в вину.

Самым простым методом можно считать удаление. Этот метод обычно применяется тогда, когда значения уверенно можно считать ошибочными, или же можно пренебречь крайне редкими и значительно отличающимися значениями. В этом случае удаляются данные, содержащие выбросы для того, чтобы избежать их какого бы то ни было воздействия на дальнейшие вычисления [9].

В случае, когда необходимо учесть значения выбросов, и вместе с этим снизить их воздействие на всю выборку, может применяться трансформация. Этот метод может подразумевать различные преобразования, в частности, нахождения значений натурального или десятичного логарифмов, квадратных корней, обратных величин [9].

В случае, когда необходимо сохранить значения, содержащие выбросы, учесть тенденции, на которые могут указывать выбросы, и вместе с этим снизить из воздействие на всю выборку в ещё большей степени, чем в случае трансформации, применяется вменение в вину [9].

Для дальнейшего анализа данных необходимо сделать сопоставимыми эмпирические значения признаков. В этих целях данные можно подвергнуть нормализации или стандартизации.

Нормализация – способ преобразования эмпирических данных путём изменения масштаба значений, в ходе которого значения x_i признака X располагаются в интервале [0; 1].

Нормализованные данные рассчитываются по формуле:

$$x_i' = \frac{(x_i - \bar{x})}{x_{max} - x_{min}}, \quad (1.4.7)$$

где x_i' нормализованное значение, x_{max} и x_{min} – максимальные значения признака X.

В программе нормализация реализована следующим образом:

```
table[s] = (table[s] - x_min) / (x_max - x_min)
```

Листинг 1.4.3. Нормализация данных.

Тут table – переменная типа pandas.DataFrame для хранения таблицы со значениями, s – переменная типа string для хранения имени столбца таблицы, x_min и x_max – переменные типа numpy.int64 или numpy.float64 для хранения максимального и минимального значений в столбце таблицы.

Стандартизация – способ преобразования эмпирических данных путём изменения масштаба значений, в ходе которого значения x_i признака X располагаются таким образом, что их среднее арифметическое приравнивается 0, а среднее квадратическое отклонение приравнивается единице.

Стандартизованные данные рассчитываются по формуле:

$$X_i' = \frac{(x_i - \bar{x})}{S}, \quad (1.4.8)$$

где x_i' нормализованное значение, S-среднее квадратическое отклонение признака X.

В программе стандартизация реализована следующим образом:

```
table[s] = (table[s] - mean) / std
```

Листинг 1.4.4. Стандартизация данных.

Тут table – переменная типа pandas.DataFrame для хранения таблицы со значениями, s переменная типа string для хранения имени столбца

таблицы, mean и std – переменные типа numpy.float64 для хранения среднего арифметического и среднего квадратического отклонения.

1.5. Расчёт описательных статистик распределения и их реализация в умном помощнике

Для характеристики распределения значений случайной величины применяют описательные статистики, которые используют различные вычисляемые показатели. Это мода, медиана, математическое ожидание, дисперсия и среднее отклонение, максимум, минимум и размах, асимметрия A_s и эксцесс E_x.

Выборочными показателями называют различные средние показатели, используемые в качестве характеристики свойств статистического распределения.

Пусть будет выборка объёма n со значениями x_1, x_2,…, x_n признака X.

Мода выборки – это то значение варианты выборки, которое имеет наибольшую частоту.

В программе мода вычисляется следующим образом:

```
mode = sci.stats.mode(np_data)[0]
```

Листинг 1.5.1. Вычисление моды.

Тут mode - переменная типа numpy.float64 для хранения значения моды, np_data – переменная типа nupy.ndarray для хранения массива numpy со значениями некоторой характеристики.

Медиана выборки – серединное значение вариационного ряда значений случайной величины.

Медиана обозначается как *Me.*

Если n=2k+1, то:

$$Me = x_{k+1} \tag{1.5.1}$$

Если n=2k, то:

$$Me=\frac{x_k+x_{k+1}}{2} \tag{1.5.2}$$

В программе медиана вычисляется следующим образом:

```
median = np.median(np_data)
```

Листинг 1.5.2. Вычисление медианы.

Тут median – переменная типа numpy.float64 для хранения значения медианы, np_data – переменная типа numpy.ndarray для хранения массива numpy со значениями некоторой характеристики.

В случае, если признак X имеет значения x_i, которые не сгруппированы в вариационные ряды (табл. 1.4.2, 1.4.3), а объём выборки n не большой, то математическое ожидание a_n^* находится по формуле:

$$a_n^* = \frac{1}{n}\sum_{i=1}^{n} x_i \quad (1.5.3)$$

В программе математическое ожидание вычисляется следующим образом:

```
mean = round(np_data.mean(), 3)
```

Листинг 1.5.3. Вычисление математического ожидания.

Тут mean – переменная типа numpy.float64 для хранения значения математического ожидания, np_data – переменная типа numpy.ndarray для хранения массива numpy со значениями некоторой характеристики.

Дисперсия $(\delta_n^*)^2$ находится по формуле:

$$(\delta_n^*)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - a_n^*)^2 \quad (1.5.4)$$

Однако, если был образован дискретный вариационный ряд (табл. 1.4.2), то формулы имеют следующий вид:

$$a_n^* = \frac{1}{n}\sum_{k=1}^{n} x_k,\ n = \sum_{k=1}^{k} n_k \quad (1.5.5)$$

$$(\delta_n^*)^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - a_n^*)^2 n_k \quad (1.5.6)$$

В программе дисперсия вычисляется следующим образом:

```
variance = round(np.var(np_data, ddof = 1), 3)
```

Листинг 1.5.4. Вычисление дисперсии.

Тут variance – переменная типа numpy.float64 для хранения значения дисперсии, np_data – переменная типа numpy.ndarray для хранения массива numpy со значениями некоторой характеристики.

Очевидно, что формулы (1.5.1) и (1.5.3) дают такие же результаты, как (1.5.2) и (1.5.4) для a_n^* и $(\delta_n^*)^2$ соответственно.

Стандартная ошибка – характеристика точности величины, для которой она вычисляется. Ошибка для математического ожидания считается по формуле:

$$m_{a_n^*} = \frac{a_n^*}{\sqrt{n}} \quad (1.5.7)$$

Формулу (1.5.4) используют, когда объём выборки $n \leq 50$, в противном случае используют исправленную дисперсию для простой выборки:

$$(S_n^*)^{\frac{2}{2}} \frac{1}{n-1} \sum_{i=1}^{n} (x_i - a_n^*)^2 \quad (1.5.8)$$

или для взвешенной выборки:

$$\left(\hat{S}_n^*\right)^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - a_n^*)^2 n_i \quad (1.5.9)$$

Чем слабее разброс значений признака относительно своего математического ожидания, тем слабее варьируются результаты респондентов в данной группе.

Среднее квадратичное отклонение показывает, насколько отдельные члены ряда отклоняются от среднего значения при различных объёмах выборки:

$$S = \sqrt{(S_n^*)^2} \text{ или } \hat{s} = \sqrt{(\hat{s}_n^*)^2} \quad (1.5.10)$$

В программе среднее квадратическое вычисляется следующим образом:

```
std = round(np.std(np_data), 3)
```

Листинг 1.5.5. Вычисление дисперсии.

Тут std – переменная типа numpy.float64 для хранения значения среднего квадратического, np_data – переменная типа numpy.ndarray для хранения массива numpy со значениями некоторой характеристики.

Ошибка среднего квадратичного отклонения вычисляется по формуле:

$$m_s = \frac{s}{\sqrt{2n}} \quad (1.5.11)$$

Коэффициент вариации – это отношение среднего квадратичного отклонения к средней арифметической, которое выражено в процентах:

$$c_v = \frac{S}{a_n^*} \cdot 100\% \quad (1.5.12)$$

Ошибки коэффициента вариации:

$$m_v = \frac{c_v}{\sqrt{2n}} \quad (1.5.13)$$

Для характеристики форм распределения используется асимметрия. Она показывает, в какую сторону относительно среднего сдвинуто большинство значений случайной величины. Симметричность эмпирического распределения относительно среднего значения

характеризуется нулевым значением. На сдвиг распределения в сторону больших значений указывает $A_s < 0$, в сторону меньших $A_s > 0$. В большинстве случаев за нормальное распределение принимается распределение с асимметрией $A_s \in [-1; +1]$.

Формула асимметрии: [7]

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s}\right)^3 \quad (1.5.14)$$

В программе асимметрия вычисляется следующим образом:

```
A = round(np.mean(((np_data - mean) ** 3) / variance ** (3 / 2)), 3)
```

Листинг 1.5.6. Вычисление асимметрии.

Тут A – переменная типа numpy.float64 для хранения значения асимметрии, mean – переменная типа numpy.float64 для хранения значения среднего арифметического, variance – переменная типа numpy.float64 для хранения значения дисперсии, np_data – переменная типа numpy.ndarray для хранения массива numpy со значениями некоторой характеристики.

Для характеристики формы распределения применяют такую характеристику как эксцесс. Если эксцесса близко к нулю, то форма распределения близка к теоретическому виду. Если $E_S > 0$, следовательно, распределение имеет плоскую вершину. Если $E_S \gg 5.0$, то распределение имеет острую вершину и его график вытянут по вертикальной оси. Если $E_s \in [-1; +1]$, -, то распределение соответствует нормальному виду. Формулы, по которым происходит вычисление данных характеристик формы распределения описаны в п. 1.4 (1.5.12 и 1.5.13).

Формула эксцесса [7]:

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s}\right)^4, \quad (1.5.15)$$

В программе эксцесс вычисляется следующим образом:

```
E = round(np.mean(((np_data - mean) ** 4) / (variance ** 2)), 3)
```

Листинг 1.5.6. Вычисление эксцесса.

Тут E – переменная типа numpy.float64 для хранения значения эксцесса, mean – переменная типа numpy.float64 для хранения значения среднего арифметического, variance – переменная типа numpy.float64 для

хранения значения дисперсии, np_data – переменная типа numpy.ndarray для хранения массива numpy со значениями некоторой характеристики.

1.5. Способы визуализации данных

Для визуализации собранных статистических данных может использоваться множество способов. Наиболее распространёнными способами отображения данных являются гистограммы, коробчатые диаграммы (диаграммы размаха), точечные диаграммы.

Гистограмма является способом отображения данных, в ходе которого отображается координатная плоскость с прямоугольниками, характеризующими количество нахождений значений x_i.

Обычно горизонтальная ось (ось абсцисс) отображает значения признака X, а вертикальная ось (ось ординат) отображает количество нахождений значения x_i, представленного на оси абсцисс. Для отображения на гистограмме значения признака X делятся на равные интервалы. На координатной плоскости изображаются прямоугольники, чья длина характеризует количество значений x_i признака X в указанном интервале.

Диаграмма размаха является способом отображения данных, который позволяет изобразить основные данные о распределении признака X с помощью построения на координатной плоскости прямоугольника с двумя отрезками, исходящими из противоположных сторон прямоугольника. На оси, вдоль которой расположен прямоугольник и отрезки, изображены значения x_i признака X.

Стороны прямоугольника, из которых исходят отрезки, расположены на уровне значений, указывающих на границы между первым и вторым квартилями и третьим и четвёртым квартилями признака X. Черта внутри прямоугольника указывает на расположение медианы признака X. Конец линий, выходящих из прямоугольника влево и вправо, указывают на наблюдаемый максимум и минимум значений признака X, без учёта выбросов. Возможные кружки за пределами отрезков указывают на выбросы.

Точечная диаграмма является способом отображения данных, в ходе которого можно наглядно наблюдать зависимость значений одного признака от значений другого признака.

На координатной плоскости на горизонтальной оси (оси абсцисс) и вертикальной оси (оси ординат) отображаются значения двух некоторых признаков X и Y, а точки на плоскости указывают на наблюдения, имеющие значения x_i и y_i, соответствующие шкалам на координатных осях.

1.6. Проверка статистических гипотез

Нулевая гипотеза - гипотеза об отсутствии различий, её обозначают через H_0. Нулевая гипотеза – это предположение, которое мы хотим опровергнуть. Нулевую гипотезу можно опровергнуть только с определённой вероятностью, зависящей от расчётов [10].

Альтернативная гипотеза – это гипотеза о значимости различий, её обозначают через H_1. Альтернативная гипотеза или экспериментальная гипотеза – это предположение, которое мы хотим доказать [10].

Зачастую производится проверка близости эмпирического распределения и нормального распределения, как наиболее часто встречающегося распределения в природе.

Функция нормального распределения имеет вид:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}}, \qquad (1.7.1)$$

где a – математическое ожидание, σ – среднее квадратическое отклонение.

Для проверки гипотез могут применяться множество критериев, например, критерий Эппса-Палли, критерий χ^2 Пирсона и др.

Преимущество критерия Эппса-Палли заключается в том, что он имеет статус одного из наиболее мощных критериев для проверки распределения на нормальность [11].

Критерий Эппса-Палли применим при объёме выборки n $n \geq 8$. Выборки при $n < 8$ при обнаружении отклонений от нормального распределения не дают достоверных результатов [8].

В качестве результата применения критерия Эппса-Палли к эмпирическим данным возможны следующие гипотезы:

1) H_0: полученное эмпирическое распределение признака не отличается от теоретического распределения.
2) H_1: полученное эмпирическое распределение признака отличается от теоретического распределения.

Статистику критерия T_{EP}Эппса-Палли вычисляют по формуле

$$T_{EP} = 1 + \frac{n}{\sqrt{3}} + \frac{2}{n} \sum_{k=2}^{n} \sum_{j=1}^{k-1} exp\left\{\frac{-(x_j - x_k)^2}{2m_2}\right\} - \sqrt{2} \sum_{j=1}^{n} \left\{\frac{-(x_j - \bar{x})^2}{4m_2}\right\}, \quad (1.7.2)$$

где m_2 – выборочный центральный момент второго порядка, $\boldsymbol{n}$ объём выборки, $\bar{x}$ – среднее арифметическое [8].

Выборочный центральный момент второго порядка m_2 рассчитывается по формуле:

$$m_2 = \frac{1}{n} \sum_{j=1}^{n} \left(x_j - \bar{x}\right)^2 \quad (1.7.3)$$

В программе критерий T_{EP}Эппса-Палли вычисляется следующим образом:

```
    m_2 = np.var(np_data, ddof = 0)
    A = sqrt(2) * np.sum([exp(-(np_data[i] - mean)**2 / (4*m_2))
    for i in range(n)])
    B = 2/n * np.sum([
        np.sum([
            exp(-(np_data[j] - np_data[k])**2 / (2*m_2)) for j in
range(0, k)
            ]) for k in range(1, n)])
    T_EP_empiric = round(1 + n / sqrt(3) + B - A, 3)
```

Листинг 1.7.1. Вычисление значения критерия Эппса-Палли.

Тут np_data – переменная типа numpy.ndarray, хранящая массив значений признака, n – переменная типа int, хранящая длину выборки, mean – переменная типа numpy.float64, хранящая среднее арифметическое.

Нулевую гипотезу отклоняют, если вычисленное значение статистики T_{EP} превышает p-квантиль при данных уровне значимости α и объёме выборки $\boldsymbol{n}$. p -Квантили статистики критерия T_{EP} при $p = 1 - \alpha =$ 0,90; 0,95; 0,975 и 0,99 приведены в приложении А [8].

Нередко в статистике необходимо установить корреляцию двух переменных в эмпирическом распределении.

Корреляция – это взаимосвязь между значениями двух и более переменных.

Для проверки гипотез может применяться множество методов, например, использование коэффициента корреляции Пирсона.

Коэффициент корреляции Пирсона может принимать значения от 1 до -1:

1) Если коэффициент корреляции меньше 0, то корреляция является отрицательной, т. е. при увеличении одной переменной другая уменьшается [13].
2) Если коэффициент корреляции больше 0, то корреляция является положительной, т. е. при увеличении одной переменной увеличивается и другая [13].
3) Чем ближе коэффициент корреляции к 0, тем меньше выражена зависимость одной переменной от другой [13].

Коэффициент корреляции Пирсона вычисляется по формуле:

$$r_{x,y} = \frac{cov(x,y)}{\sqrt{s_x^2 s_y^2}}, \quad (1.7.3)$$

где $cov(x,y)$ –ковариация, s_x^2 и s_y^2 -квадраты средних квадратических отклонений.

Ковариация находится по формуле:

$$cov(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}), \quad (1.7.4)$$

$\bar{x}$ и $\bar{y}$ - средние арифметические признаков X и Y, x_i и y_i – значения признаков, n – объём выборки.

В программе коэффициент корреляции Пирсона вычисляется следующим образом:

```
r = round(float(sci.stats.pearsonr(data_list_1, data_list_2)[0]), 3)
```

Листинг 1.7.2. Вычисление коэффициента корреляции Пирсона.

Тут переменные data_list_1 и data_list_2 типа list содержат в себе значения признаков X и Y, переменная r типа float хранит значение коэффициента корреляции Пирсона.

Для проверки значимости можно использовать t-распределение Стьюдента. Значение t-распределения вычисляется по формуле:

$$t = \frac{r_{xy}\sqrt{n-2}}{1-r_{xy}^2}, \quad (1.7.5)$$

Где r_{xy} – значение коэффициента корреляции Пирсона, n – длина выборки.

Так как распределение Стьюдента симметрично, и рассматривается возможность отклонения вычисленного значения $r_{x,y}$ от истинного как в

большую, так и в меньшую стороны, то двустороннее t-распределение будет зависеть от $\alpha/2$ [12]:

$$t \in \left[t_{\alpha/2}, t_{1-\alpha/2}\right] \quad (1.7.6)$$

В программе значение t-распределения вычисляется следующим образом:

```
t = r * np.sqrt((df) / (1 - r**2))
```

Листинг 1.7.3. Вычисление значения t-распределения.

Тут df – переменная типа int, хранящая количество степеней свободы, r – переменная типа float для хранения значения корреляции Пирсона, t – переменная типа numpy.float64 для хранения значения статистики t-распределения.

Полученное значение t-распределения сравнивается со значением в таблице значений t-распределения Стьюдента. Таблица приведена в приложении Б.

В программе нахождение табличного значения t-распределения реализовано следующим образом:

```
t_critical = sci.stats.t.ppf(1 - alpha/2, df)
```

Листинг 1.7.4. Вычисление табличного значения t-распределения.

Тут переменная t_critical типа numpy.float64 хранит табличное значение t-критерия, переменная alpha типа float содержит значение α, переменная df типа int содержит количество степеней свободы, равное $n - 2$.

Если табличное значение меньше модуля вычисленного значения t-статистики, то коэффициент корреляции считается значимым.

Для классификации значений коэффициента корреляции часто используется шкала Чеддока [15].

Таблица 1.6.1

Шкала Чеддока

Значение коэффициента корреляции	Качественная характеристика силы связи
0.1-0.3	Слабая
0.3-0.5	Умеренная
0.5-0.7	Заметная

0.7-0.9	Высокая
0.9-0.99	Весьма высокая

Следует помнить, что это деление весьма условно, и может давать результаты, не корректные по отношению к генеральной совокупности. Его границы могут пересматриваться в зависимости от объекта и задач исследования [15].

РАЗДЕЛ 2. СОЗДАНИЕ УМНОГО ПОМОЩНИКА ДЛЯ АНАЛИЗА ДАННЫХ С ПОМОЩЬЮ СРЕДСТВ ЯЗЫКА PYTHON

2.1. Сбор данных

Сбор информации производился добровольным и анонимным анкетированием среди трудоустроенных представителей молодёжи Крыма методом простого случайного отбора в интернете с помощью сервиса https://www.google.com/intl/ru/forms/about. Ссылка на анкету, используемую при опросе: https://forms.gle/gksFRTBbWiy9xseB6.

Для создания анкеты применялись вопросы закрытого типа. После этого собранные эмпирические данные были закодированы с помощью порядковой кодировки. В ходе использования анкеты (приложение В) были представлены признаки, порождающие порядковые, метрические и номинальные шкалы [1].

Объектом исследования являлись представители трудоустроенной молодёжи Крыма, то есть граждане возрастом от 18 до 35 лет [14].

Предметом исследования являлось участие представителей трудоустроенной молодёжи Крыма в общественной жизни региона.

Собранные эмпирические данные представлены в приложении Г.

Были собраны эмпирические данные по следующим признакам:

1) X_1 - «1. Пол».
2) X_2 - «2. Возраст».
3) X_3 - «3. Образование».
4) X_4 - «4. Наличие брака».
5) X_5 - «5. Средний доход в месяц».
6) X_6 - «6. Членство в молодёжной организации».
7) X_7 - «7. Посещение мероприятий культурной направленности».
8) X_8 - «8. Посещение мероприятий политической направленности».
9) X_9 - «9. Посещение мероприятий развлекательной направленности».

В ходе анкетирования была получена выборка объёмом n = 160.

Для значений признаков X_1, X_4, X_6 использовались номинальные измерительные шкалы.

Для значений признаков X_2, X_3, X_7, X_8, X_9 использовались ранговые измерительные шкалы.

Для значений признака X_5 использовалась метрическая измерительная шкала.

Данные были собраны в файле с расширением .csv в виде таблицы с кодировкой UTF-8 с запятой в качестве разделителя значений.

2.2. Обработка данных умным помощником

Собранные эмпирические данные были загружены в программу (приложение Д), и далее обрабатывались и анализировались в ней.

Значения признаков, для которых использовались номинальные и ранговые шкалы, в целях дальнейшей обработки и анализа были зашифрованы. Для признаков с различными данными и типами шкал использовались разные шифры. Для значений, использующих метрические шкалы – значений признака X_5 - шифр не использовался.

Данные признака X_1 были зашифрованы следующим образом:

Таблица 2.2.1

Шифр для признака X_1

Не зашифрованное значение	Зашифрованное значение
Мужской	1
Женский	2

Данные признака X_2 были зашифрованы следующим образом:

Таблица 2.2.2

Шифр для признака X_2

Не зашифрованное значение	Зашифрованное значение
18-20	19
21-23	22
24-26	25
27-29	28
30-32	31
33-35	34

Данные признака X_3 были зашифрованы следующим образом:

Таблица 2.2.3

Шифр для признака X_3

Не зашифрованное значение	Зашифрованное значение
Общее	1
Среднее профессиональное	2
Высшее	3

Данные признаков X_4 и X_6 были зашифрованы следующим образом:

Таблица 2.2.4

Шифр для признаков X_4 и X_6

Не зашифрованное значение	Зашифрованное значение
Да	1
Нет	0

Данные признаков X_7, X_8 и X_9 были зашифрованы следующим образом:

Таблица 2.2.5

Шифр для признаков X_7, X_8 и X_9

Не зашифрованное значение	Зашифрованное значение
Постоянно	4
Часто	3
Редко	2
Никогда	1

Пример процесса создания шифров:

Тут переменные dict_1 и dict_2 типа dict содержат словари с парами «значение – шифр для значения».

Пример процесса применения шифров:

Тут переменная df_data типа pandas.DataFrame хранит таблицу со значениями, dict_1 и dict_2 – словари, содержащие пары «значение – шифр для значения», question_4_6 – переменная типа list для хранения списка названий столбцов таблицы, для которых применяется шифр.

После применения шифров можно изобразить полученные данные на гистограммах и диаграммах размаха:

Для построения гистограммы применялась следующая функция:

Тут функция принимает в качестве аргументов переменную table типа pandas.DataFrame для хранения таблицы со значениями, переменную

series типа list для хранения списка с заголовками столбцов таблицы, для которых строится гистограмма, переменную типа path_to_save типа string для хранения имени подпапки для сохранения изображений, или пустого значения в случае, если сохранять изображения не требуется.

Для построения диаграммы размаха применялась следующая функция:

Тут функция принимает в качестве аргументов переменную table типа pandas.DataFrame для хранения таблицы со значениями, переменную series типа list для хранения списка с заголовками столбцов таблицы, для которых строится диаграмма размаха, переменную типа path_to_save типа string для хранения имени подпапки для сохранения изображений, или пустого значения в случае, если сохранять изображения не требуется.

Для построения точечной диаграммы применялась следующая функция:

Собранные данные были проверены на наличие пустых значений. После применения шифра среди значений признака X_5 присутствовали строчные значения «NaN» и числовые значения. Значения «NaN» были получены из незашифрованных значений «Нет», и считались пустыми значениями. Строки, содержащие эти значения, были удалены из таблицы.

Для обработки пустых значений применялась следующая функция:

Тут функция принимает в качестве аргументов переменную table типа pandas.DataFrame для хранения таблицы со значениями, переменную series типа list для хранения списка с заголовками столбцов таблицы, в которых обрабатываются пустые значения, переменную mod типа string, определяющую то, как будут обрабатываться пустые значения – удаляться, или заменяться на медиану или среднее арифметическое.

Собранные данные были проверены на наличие выбросов. Границы выбросов рассчитывались по формуле (1.4.6). Среди значений признака X_5 были обнаружены выбросы, которые были заменены на среднее арифметическое, так как они могли указывать на некоторую тенденцию значений признака в выборке.

После удаления пустых значений и выбросов мощность выборки стала равна n = 106.

Для удаления выбросов применялась следующая функция:

Тут функция принимает в качестве аргументов переменную table типа pandas.DataFrame для хранения таблицы со значениями, переменную series типа list для хранения списка с заголовками столбцов таблицы, в

которых обрабатываются выбросы, переменную mod типа string, определяющую то, как будут обрабатываться выбросы – заменяются на среднее арифметическое, удаляются, или подвергаются преобразованию квадратного корня.

Для полученных значений можно сформировать интервальные и дискретные вариационные ряды.

Для вычисления частоты и относительной частоты использовались формулы (1.4.1) и (1.4.2).

Полученные значения для возможности сравнивать их между собой и дальнейшего анализа данных необходимо нормализовать. Для нормализации использовалась формула (1.4.7).

После нормализации данные готовы для их анализа.

Ниже представлены вариационные ряды, гистограммы и диаграммы размаха для обработанных значений признаков.

Таблица2.2.1

Дискретные вариационные ряда для признаков X_1 и X_2.

Показатель	Признак								Всего
	X_1		X_2						
	Номер варианты ответа								
	1	2	19	22	25	28	31	34	
Частоты, n_i	57	49	4	25	21	15	15	26	106
Относительные частоты, w_i	0,54	0,46	0,04	0,24	0,2	0,14	0,14	0,25	1
Процент, %	54%	46%	4%	24%	20%	14%	14%	25%	100%

Таблица 2.2.2

Дискретные вариационные ряды для признаков X_3, X_4, X_6.

Показатель	Признак							Всего
	X_3			X_4		X_6		
	Номер варианты ответа							
	1	2	3	0	1	0	1	
Частоты, n_i	22	45	39	52	54	72	34	106
Относительные частоты, w_i	0,21	0,42	0,37	0,49	0,51	0,68	0,32	1
Процент, %	21%	42%	37%	49%	51%	68%	32%	100%

Таблица 2.2.3

Дискретные вариационные ряды для признаков X_7 и X_8.

Показатель	Признак								Всего
	X_7				X_8				
	Номер варианты ответа								
	1	2	3	4	1	2	3	4	
Частоты, n_i	20	34	36	16	25	31	31	19	106
Относительные частоты, w_i	0,19	0,32	0,34	0,15	0,24	0,29	0,29	0,18	1
Процент, %	19%	32%	34%	15%	24%	29%	29%	18%	100%

Таблица 2.2.4

Дискретный вариационный ряд для признака X_9.

Показатель	Признак				Всего
	X_9				
	Номер варианты ответа				
	1	2	3	4	
Частоты, n_i	14	41	33	18	106
Относительные частоты, w_i	0,13	0,39	0,31	0,17	1
Процент, %	13%	39%	31%	17%	100%

Таблица 2.2.5

Интервальный вариационный ряд для признака X_5.

Показатель	Признак					Всего
	X_5					
	Интервал					
	15000-20000	20000-25000	25000-30000	30000-35000	35000-40000	
Частоты, n_i	16	30	37	15	8	106
Относительные частоты, w_i	0,15	0,28	0,35	0,14	0,07	1
Процент, %	15%	28%	35%	14%	7%	100%

Для получения вариационных рядов применялась следующая функция:

Тут функция принимает в качестве аргументов переменную table типа pandas.DataFrame для хранения таблицы со значениями, переменную series типа list для хранения списка с заголовками столбцов таблицы, для которых строятся вариационные ряды, переменную mod типа string, определяющую, будет строиться дискретный или интервальный вариационный ряд, переменную path_to_save типа string для хранения имени подпапки для сохранения таблиц, или пустого значения в случае, если сохранять изображения не требуется.

Для нормализации применялась следующая функция:

Тут функция принимает в качестве аргументов переменную table типа pandas.DataFrame для хранения таблицы со значениями, переменную series типа list для хранения списка с заголовками столбцов таблицы, данные которых подвергаются нормализации.

Для стандартизации использовалась следующая функция:

Тут функция принимает в качестве аргументов переменную table типа pandas.DataFrame для хранения таблицы со значениями, переменную series типа list для хранения списка с заголовками столбцов таблицы, данные которых подвергаются стандартизации.

После нормализации описательные метрики различных признаков можно сравнивать между собой.

Вычисление описательных метрик эмпирических данных.

Медиана вычисляется согласно формулам (1.5.1, 1.5.2).

Среднее арифметическое вычисляется согласно формуле (1.5.3).

Дисперсия вычисляется согласно формулам (1.5.4, 1.5.8).

Среднее квадратическое отклонение вычисляется согласно формуле (1.5.10).

Асимметрия вычисляется согласно формуле (1.5.14).

Эксцесс вычисляется согласно формуле (1.5.15).

Табл. 2.2.6

Описательные метрики признаков(начало).

Признак	n	Мода	Медиана	Среднее арифметическое
X_1	106	0	0	0.462
X_2	106	1	0.6	0.570
X_3	106	0.5	0.5	0.580
X_4	106	1	1	0.509

X_5	106	0.4	0.4	0.418
X_6	106	0	0	0.321
X_7	106	0.667	0.333	0.484
X_8	106	0.333	0.333	0.472
X_9	106	0.333	0.333	0.506

Табл. 2.2.7

Описательные метрики признаков (окончание).

Признак	Дисперсия	Среднее квадратическое отклонение	Асимметрия	Эксцесс
X_1	0.251	0.499	0.149	1.004
X_2	0.104	0.321	0.041	1.597
X_3	0.139	0.371	-0.262	1.816
X_4	0.252	0.500	-0.037	0.983
X_5	0.052	0.227	0.343	2.728
X_6	0.220	0.467	0.757	1.560
X_7	0.104	0.321	0.007	1.997
X_8	0.120	0.345	0.074	1.807
X_9	0.096	0.308	0.087	2.108

Для вычисления описательных метрик применялась следующая функция:

Тут функция принимает в качестве аргументов переменную table типа pandas.DataFrame для хранения таблицы со значениями, переменную series типа list для хранения списка с заголовками столбцов таблицы, для которых вычисляются метрики, переменную path_to_save типа string для хранения имени подпапки для сохранения таблицы, или пустого значения в случае, если сохранять изображения не требуется.

2.3. Анализ данных умным помощником

После обработки данные можно использовать для их анализа и проверки статистических гипотез.

Пусть будет проведена проверка гипотезы для всех полученных признаков о том, что между хотя бы одной парой признаков есть

статистически значимая корреляция, являющаяся хотя бы заметной согласно шкале Чеддока, т. е. $|r| \geq 0{,}5$. Тогда альтернативной гипотезой будет отсутствие статистически значимой корреляции, которая будет хотя бы заметной согласно шкале Чеддока.

Выдвинем 2 гипотезы:

H_0: между хотя бы одной парой признаков из полученных X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9 есть статистически значимая корреляция со значением коэффициента корреляции $|r| \geq 0{,}5$.

H_1: между хотя бы одной парой признаков из полученных X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9 нет статистически значимой корреляции со значением коэффициента корреляции $|r| \geq 0{,}5$.

Найдём коэффициент корреляции Пирсона для всех пар признаков, проверим их статистическую значимость с помощью t-распределения, и выведем информацию только о значимых коэффициентах корреляции.

В рамках этой задачи пусть будет α = 0.05.

Выведем таблицу с парами признаков и соответствующими им значениями коэффициентов корреляции, в которые по модулю будут больше 0.5, а также некоторые соответствующие им точечные диаграммы:

Таблица 2.3.1

Вычисление коэффициентов корреляции.

X_i	X_j	n	t-статистика (t)	Табличное критическое значение t-статистики (t крит.)	|t| > t крит.	r
X_2	X_3	106	8.276	1.983	ИСТИНА	0.63

Из результатов работы программы можно видеть, что в выборочной совокупности наблюдается заметная положительная корреляция между признаками X_2 и X_3.

Указанная пара признаков имеет между собой коэффициенты корреляции, удовлетворяющие условию $|r| \geq 0{,}5$.

Вывод: гипотеза H_0 принимается, отклоняется гипотеза H_1.

Для вычисления коэффициентов корреляции применяется следующая функция:

Тут функция принимает в качестве аргументов переменную table типа pandas.DataFrame для хранения таблицы со значениями, переменную series типа list для хранения списка с заголовками столбцов таблицы, для

которых вычисляются коэффициенты корреляции, переменную alpha типа float для хранения значения α, переменную only_meaning типа bool для хранения информации о необходимости вывода строк с только статистически значимыми коэффициентами корреляции, переменную critical типа float для хранения значения, при модуле коэффициента корреляции ниже которого строка не будет выводиться, переменную scatter_diagram_plotting типа bool для хранения информации о необходимости построения точечных диаграмм, переменную path_to_save типа string для хранения имени подпапки для сохранения таблицы, или пустого значения в случае, если сохранять изображения не требуется.

Пусть будет проведена проверка гипотезы о соответствия эмпирического распределения нормальному распределению для признака X_5.

Выдвинем 2 гипотезы:

H_0: распределение признака X_5 является нормальным.

H_1: распределение признака X_5 не является нормальным.

Вычислим значение критерия Эппса-Палли для признака X_5, сравним его с табличным значением, и выведем результаты проверки распределения на нормальность.

В рамках этой задачи пусть будет α = 0.05.

Выведем таблицу с результатами работы программы, а также гистограмму значений признака X_5.

Таблица 2.3.2

Проверка распределения на нормальность.

X_i	n	p	Значение критерия (T_EP)	Табличное значение критерия (T_EP крит.)	T_EP < T_EP крит.	Нормальное распределение
X_5	106	0.95	0.182	0.376	ИСТИНА	Выполняется

Из результатов работы программы можно видеть, что распределение признака X_5 является нормальным.

Вывод: гипотеза H_0 принимается, отклоняется гипотеза H_1.

Для проверки распределения на нормальность применяется следующая функция:

Тут функция принимает в качестве аргументов переменную table типа pandas.DataFrame для хранения таблицы со значениями, переменную series типа list для хранения списка с заголовками столбцов таблицы, для которых вычисляются коэффициенты корреляции, переменную p_level типа float для хранения значения p, переменную path_to_save типа string для хранения имени подпапки для сохранения таблицы, или пустого значения в случае, если сохранять изображения не требуется.

ЗАКЛЮЧЕНИЕ

Социологические исследования имеют широкую сферу применения, их проводят с целью сбора информации о различных социальных процессах и явлениях и связях между ними, а также для получения возможных выводов о возможности действий с целью влияния на социальные процессы и явления.

В любом социологическом исследовании проводятся следующие этапы: проведение статистического наблюдения, обработка полученных данных, анализ обработанных данных, проверка статистических гипотез.

Первый раздел работы содержит описание некоторых правил и методов проведения выборки в социологических исследованиях, методов обработки и анализа эмпирических данных социологических исследований, а также описание коэффициента корреляции Пирсона и критерия Эппса-Палли, используемых для проверки статистических гипотез.

Второй раздел включает в себя обработку и анализ эмпирических данных социологического исследования, проведённого среди представителей трудоустроенной молодёжи Крыма. В исследовании участвовало 160 человек.

Сбор данных производился методом анкетирования, выборка среди респондентов проводилась методом простого случайного отбора.

Были изучены методы обработки и анализа данных социологических исследований, в ходе работы был разработан умный помощник на языке программирования высокого уровня Python, и практически применён к эмпирическим данным, полученным в результате социологического опроса трудоустроенных представителей молодёжи Крыма.

В ходе работы был разработан умный помощник на языке программирования высокого уровня Python. Эмпирические данные, полученные в результате анкетирования, были обработаны умным помощником. В результате обработки и анализа ответов респондентов программой стало известно, что:

1) Среди респондентов находилось 54% мужчин и 46% женщин.
2) 4% респондентов возрастом 18-20 лет, 24% респондентов возрастом 21-23 года, 20% респондентов возрастом 24-26 лет, 14% респондентов возрастом 27-29 лет, 14% респондентов возрастом 30-32 года, 25% респондентов возрастом 33-35 лет.

3) 21% респондентов имеет среднее общее образование, 42% респондентов имеет среднее профессиональное образование, 37% респондентов имеет высшее образование.
4) 51% респондент находится в официально зарегистрированном или официально не зарегистрированном браке, а 49% респондентов не находятся в официально зарегистрированном или официально не зарегистрированном браке.
5) 32% респондентов состоят в какой-либо молодёжной организации, а 68% респондентов не состоят ни в какой молодёжной организации.
6) 15% респондентов постоянно посещает общественные мероприятия культурной направленности в регионе, 34% респондентов делает это часто, 34% респондентов делает это редко, а 19% респондентов никогда не посещает общественные мероприятия культурной направленности.
7) 18% респондентов постоянно посещает общественные мероприятия политической направленности в регионе, 29% делают это часто, 29% делают это редко, а 24% респондентов никогда не посещает общественные мероприятия политической направленности в регионе.
8) 17% респондентов постоянно посещает общественные мероприятия политической культурной направленности в регионе, 31% делает это часто, 39% делает это редко, а 13% никогда не постоянно посещает общественные мероприятия политической культурной направленности в регионе.

Коэффициент корреляции Пирсона показал, что между уровнями образования и возрастом респондентов присутствует заметная положительная корреляция со значением $r = 0{,}63$.

Критерий Эппса-Палли показал, что уровень среднего дохода респондентов имеет нормальное распределение.

В разработанном умном помощнике поэтапно реализованы различные методы обработки и анализа данных, каждый из которых сопровождается подробными пояснениями и комментариями в коде.

Основными преимуществами по сравнению с существующими программными продуктами являются встроенное поэтапное руководство для пользователя по использованию умного помощника вместе с подробным комментированием кода, наличие программной реализации

критерия Эппса-Палли, которая отсутствует в широко известных библиотеках языка Python, открытый код, который свободен для модификации пользователем.

В дальнейшем результаты данного социологического исследования могут применяться в других социологических исследованиях, для разработки программных продуктов, или для других целей.