**Перечень заданий по микрофайлам домохозяйств**

Имеются следующие 3 выборки данных по выборочному опросу домохозяйств по уровню жизни[[1]](#footnote-1):

1) demo2017.sav – данные по членам домашних хозяйств, включающие данные по возрасту и зарплатам отдельных людей (принадлежность к конкретному домохозяйству определяется специальным кодом); перечень переменных – подмножество из «Описание микрофайла по составу домашних хозяйств.pdf»;

2) health2017.sav – дополнительные данные по членам домашних хозяйств, включающие данные по состоянию здоровья отдельных людей; перечень переменных – подмножество из «Описание микрофайла по составу домашних хозяйств.pdf»;

3) income2017.sav – показатели по домохозяйствам в целом, включая различные статьи доходов и расходов.

При выполнении заданий указанные исходные файлы .sav, а также промежуточные рабочие файлы (.csv) следует добавить в список исключений, чтобы запретить git их отслеживание. Это поможет избежать многократного закачивания файлов на сервер GitHub. С использованием кода всегда можно повторить все вычисления, а для этого достаточно иметь файлы с данными исключительно локально на компьютере. Это поможет избежать распространения файлов данных в Интернете.

**1. Задания по предварительному анализу**

1) подготовить данные для своего варианта заданий:

а) объединить выборки в одну, используя коды домохозяйств и их отдельных членов (функция merge в pandas); при объединении контролировать, чтобы в итоговой выборке не было дубликатов переменных;

б) выбрать с помощью случайного выбора (без повторений) подвыборку в 2000 человек; задавать начальное значение датчика случайных чисел как число в формате «ГГГГММДД» (использовать день рождения); дальше работать с данной выборкой для выполнения личных заданий;

в) для выборки из *б* провести анализ типов каждой переменной; проверить корректность типов, значений и меток категориальных переменных;

г) подготовить «взвешенный» вариант своей выборки на основе переменной, обозначающей статистический вес каждого респондента (члена домашнего хозяйства): в выборке должны быть продублированы значения каждого наблюдения ровно столько раз, сколько указано в весовой переменной (данный подход позволяет «приблизить» выборку к генеральной совокупности); веса округлить до целого, отбросив дробную часть.

Теоретически, пункт г можно не делать, но в последствии использовать веса для расчета взвешенного среднего зарплаты и других показателей, найти способы использования весов в методах и алгоритмах как это выполняется в SPSS Statistics. В результате механического повторения наблюдений в выборке, ее объем может сильно увеличиться.

д) каждую выборку сохранять в папке проекта SDAE, не закачивать их на сервер GitHub – для этого, после создания выборок и сохранения в файлы (.csv или другие форматы) поставить ограничения в .gitignore отдельно для этих файлов, чтобы прекратить отслеживание.

e) в дальнейшем, в анализе можно использовать выборку б, однако для некоторых задач (где будет указано) использовать взвешенную выборку.

2) провести анализ распределений переменных:

а) проводить анализ следующих количественных переменных: все доходы, все расходы, inc\_n и exp\_n (n – номер варианта, будет присвоен на занятии путем случайного выбора), зарплата, возраст, масса тела и рост; графики – гистограмма, статистики – среднее, медиана, мода, станд. отклон., ассиметрия, эксцесс; при анализе некоторых переменных следить за пропущенными либо релевантными для анализа значениями (исключать их предварительно);

б) вычислить производные количественные переменные: Индекс массы тела (вес тела в кг разделить на квадрат роста в метрах), сбережения (разность между всеми доходами и расходами); проанализировать распределения переменных;

в) провести частотный анализ дискретных переменных (категориальных и номинальных): пол, регион, тип домохозяйства, количество человек в семье, распределение по уровню образования, по оценке своего общего уровня здоровья; провести анализ пропущенных значений; построить графики (столбчатые диаграммы или круговые);

г) на взвешенной выборке провести анализ распределения зарплаты, сбережений и сравнить их с распределениями на невзвешенной выборке;

3) провести анализ преобразованных значений количественных переменных из 2а:

а) как изменяется распределение и статистики переменной при ее преобразовании с помощью натурального логарифма;

б) тот же вопрос из 3а, если применить стандартизирование: становится ли распределение некоторых переменных после преобразования похожим на нормальное?

Таблица 1 – Распределение номеров вариантов относительно номеров студентов в списке (случайный выбор, нач. знач. датчика СЧ – текущая дата)

|  |  |
| --- | --- |
| **Номер в списке (по алфавиту)** | **Вариант inc\*, exp** |
| 1 | 8 |
| 2 | 3 |
| 3 | 12 |
| 4 | 2 |
| 5 | 11 |
| 6 | 9 |
| 7 | 4 |
| 8 | 7 |
| 9 | 1 |
| 10 | 5 |
| 11 | 10 |
| 12 | 6 |

inc12=Privlg

**2. Задания по исследование однородности данных домохозяйств методами проверки гипотез, дисперсионного анализа и таблиц сопряженности**

1) проверить, равна ли средняя зарплата по выборке (использовать как невзвешенное, так и взвешенное среднее при расчете) статистически равна средней номинальной начисленной зарплате за 2017 г. (815.25), можно (при необходимости, использовать одновыборочный t-тест);

2) проверить гипотезу об равенстве средних значений (двухвыборочный t-test) и дисперсий (критерий Ливиня) количественных в двух группах:

а) зарплату для мужчин и женщин;

б) inc\_n, exp\_n (согласно своему варианту) для домохозяйств из любых двух регионов (выбрать самостоятельно, можно сравнить свой родной регион с Минском, если Минск не является местом рождения);

3) создать на основе количественных переменных категориальные:

а) для возраста: выделить следующие категории: 18-24, 25-34, 35-44, 45-54, 55-64, остальным наблюдениям, не для которых возраст не входит в указанные категории, присвоить для данной переменной пропущенные значения;

б) для зарплаты (нижняя граница включается, верхняя нет): от 0 до 400, от 400 до 500, от 500 до 700, от 700 до 1000, от 1000; значение 0 считать при расчетах пропущенным;

4) построить таблицы сопряженности для следующих пар переменных (их категориальных версий): пол и интервалы зарплат, возраст и интервалы зарплат, уровень образования и интервалы зарплат, уровень образования и частота занятий спортом (только на выборке понятных значений); двухмерную таблицу для пола (sex) и бинарной переменной, характеризующей наличие привычки курения (smoker);

5) протестировать с помощью критерия Хи-квадрат равенство распределения частот по уровню зарплат (для категориальной переменной ) для мужчин и женщин, по образованию для мужчин и женщин; для двухмерной таблицы сопряженности sex-smoker протестировать наличие значимой взаимосвязи между полом и наличием привычки курения (на всей выборке, а также отдельно в разрезе уровней образования);

6) выполнить однофакторный дисперсионный анализ для зависимой количественной переменной зарплаты относительно следующих вариантов дискретной переменной-фактора (провести отдельный анализ по порядку): региона (включая и исключая Минск), общей оценки состояния здоровья;

7) с помощью многомерной T-статистики Хоттелинга проверить, имеют ли домохозяйства из двух произвольных регионов одновременной одинаковые средние значений переменных cashing, InKind, Privlg; можно попытаться найти такую пару регионов, для которых это выполняется.

**3. Задания по корреляционному и регрессионному анализу данных домохозяйств**

1) построить корреляционную матрицу для показателей inc\_n, cashing, InKind, Privlg, totalinc, exp\_n, totalexp:

а) использовать коэффициенты корреляции Пирсона;

б) вывести p-values;

в) различаются ли корреляционные связи по регионам – выбрать два любых региона для сравнения;

2) вычислить частные выборочные коэффициенты корреляции для каждой пары переменной в каждой тройке переменных, которые имеют достаточно сильную корреляцию друг с другом;

3) построить множественную регрессию для переменной exp\_n относительно:

а) других переменных из 1 пункта, которые имееют наименьшую корреляцию друг с другом (для исключения эффекта коллинеарности);

б) тоже самое для переменной totalexp относительно других переменных;

в) попробовать найти связи для зависимых переменных в a и b c переменной зарплаты, рассчитанной как сумма по всему домохозяйству;

г) вывести t-статистики для каждого коэффициента регрессии, а также их p-значения; вывести значение F-критерия и его p-value; какие коэффициенты модели значимы на 5% уровне и значима ли на этом уровне модель в целом?

4) применить многофакторный дисперсионный анализ для количественной переменной зарплаты относительно следующих категориальных переменных: уровень образования, пол и возраст (категориальное представление);

а) использовать обобщенную линейную статистическую модель;

б) определить значимые факторы и их комбинации, а также F-статистику для модели в целом;

**4. Дискриминантный и кластерный анализ данных домохозяйств**

1)

1. https://www.belstat.gov.by/informatsiya-dlya-respondenta/vyborochnye-obsledovaniya/vyborochnoe-obsledovanie-domashnih-hozyaistv-po-urovnyu-zhizni/opisanie-mikrofailov-po-vyborochnomu-obsledovaniyu-domashnih-hozyaistv-po-urovnyu-zhizni/ [↑](#footnote-ref-1)