ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ

ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«БАШКИРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

Отчет по лабораторной работе № 1

«Анализ выборки»

наименование лабораторной работы

Выполнил: студент гр. 1ФКЦ

Ахметов Руслан Олегович

Проверил: кандидат технических наук, доцент

Прудников Вадим Борисович

Уфа – 2023

1. Загрузка исходных данных проводится с помощью команды load(): load('var3.RData'). В результате в рабочее пространство загружен набор данных var3, имеющий 2000 наблюдений по 8 переменным. Рассмотрим набор данных командой View(var3) – рис. 1.

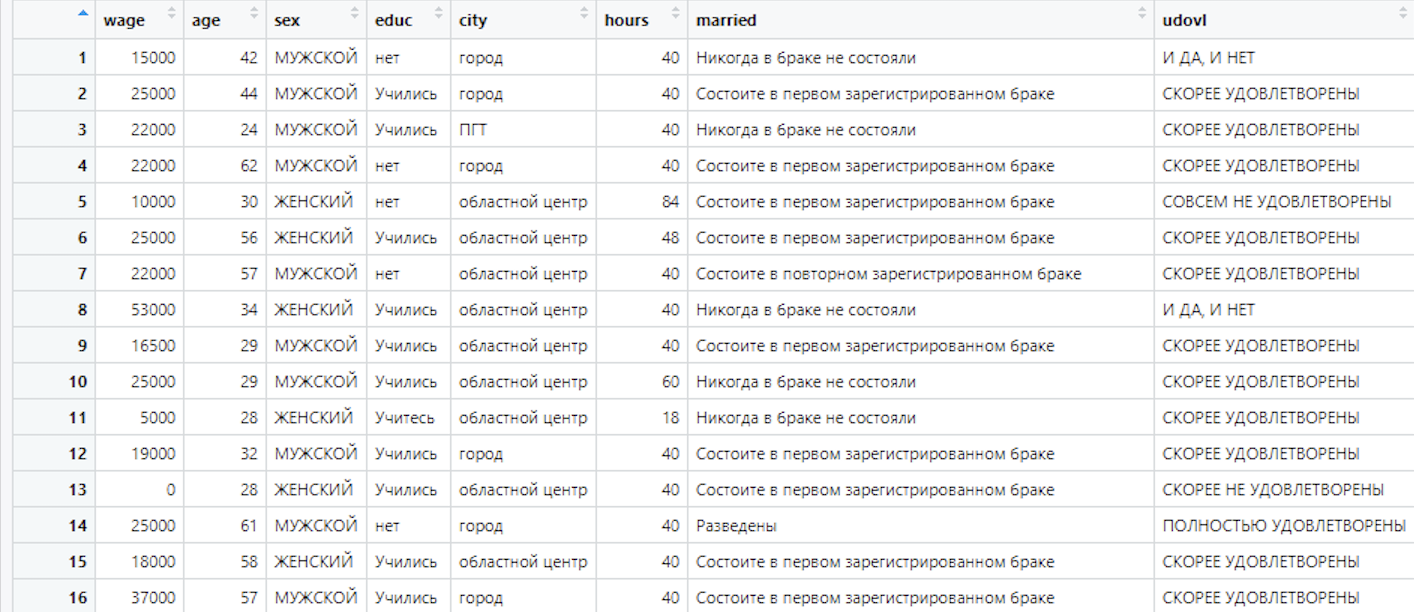


рисунок 1- Общий вид набора данных var3

Для определения типа данных каждой переменной используем функцию glimpse(var3).

2.Для проведения описательного анализа используем переменную wage.   
Выборочное среднее (mean):   
Выборочная дисперсия (var) :   
Выборочное среднее квадратическое (стандартное) отклонение (sd):   
Медиану и квартили выборки удобно определить с помощью функции summary(): summary(var3).   
  
Результат выполнения команд представлен на рис. 2.

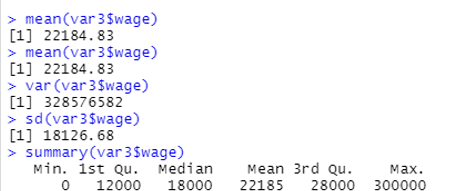


Рисунок 2 – Описательные статистики

Из рис. 2 можно заключить, что выборочная средняя составляет 22 185 руб., дисперсия –   
328 576 582 руб., а стандартное отклонение – 18 127 руб.

Выборочная медиана равна 18 000 руб. Это означает, что половина попавших в выборку респондентов имеют заработную плату, меньшую 18 000 руб., а четверть – меньше 12 000 руб. (первый квартиль), а 25% опрошенных имеют заработную плату, превышающую 28 000 руб. (третий квартиль). При этом максимальная заработная плата составляет 300 000 руб.

Для расчета показателей эксцесса (kurtosis) и скошенности (skewness) необходима загрузка и подключение пакета moments: install.packages('moments'); library(moments). Результаты расчета представлены на рис. 3.

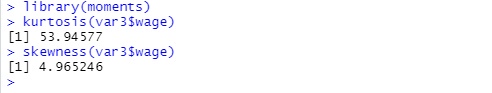


Рисунок 3 - Показатели выборочных эксцесса и скошенности

3.Построим гистограмму относительных частот с помощью функции hist. Для задания построения гистограммы именно относительных частот используется параметр функции freq (чистота), значение которого устанавливается равным F (False).

Функция hist() полезна для понимания распределения числовых переменных. Эта функция разбивает значения на числовые переменные. Его основная функциональность заключается в создании гистограммы заданного фрейма данных.

Используем команду:  
hist(var3$wage, breaks = 20, freq = FALSE, main = 'Гистограмма относительных частот',xlab = 'Среднемесячная зарплата, руб.', ylab = 'Отн. частота', col = 'blue').



Здесь параметр breaks задает число интервалов разбиения количественного признака (заработной платы), параметры main, xlab, ylab – задают соответствующие подписи, а параметр col – цвет заливки столбцов гистограммы.

Построенная гистограмма приведена на рис. 4.

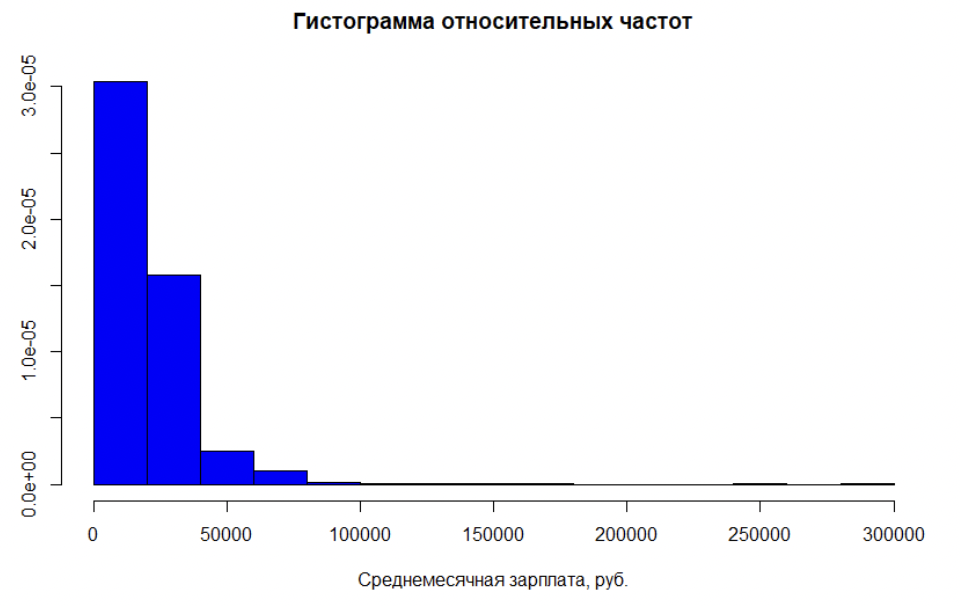


Рисунок 4 – Гистограмма относительных частот распределения заработной платы

4.Проверку на нормальность распределения проведем по трем предложенным методам.

Идея метода Q-Q графика заключается в сопоставлении квантилей выборочного (эмпирического) распределения с квантилями теоретического нормального распределения, математическое ожидание и дисперсия которого совпадают с выборочными характеристиками.

Для построения Q-Q графика теоретических и выборочных квантилей распределения используем функций qqnorm и qqline.

В qqnorm() функция создает нормальный QQ-график и qqline() добавляет линию, проходящую через первый и третий квартили.

На графике необходимо обращать внимание на то, что нанесенные точки лежат на прямой линии, проведенной из левого нижнего угла. Особенно важно соответствие значений между первым и третьим квартилями.

Проверим для нашей выборки var10$wage с помощью кода.  
qqnorm(var3$wage, pch=1, frame=FALSE, main='Q-Q график', xlab='Квантили теор. распределения', ylab='Квантили эмпир. распределения');   
qqline(var3$wage, col = "steelblue", lwd = 3).


Результат построения приведен на рис. 5.

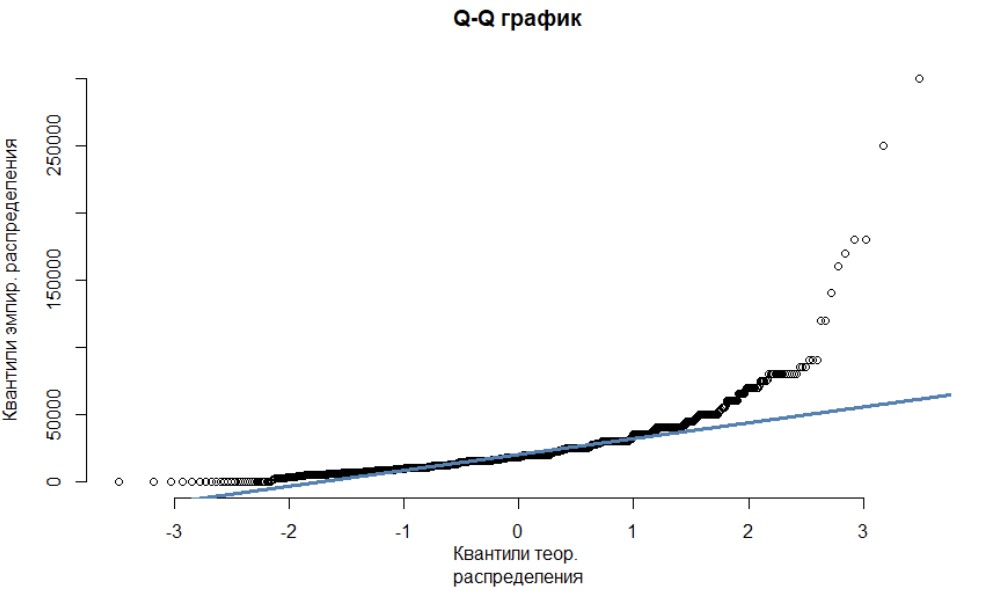


Рисунок 5 – Q-Q график

В данном случае из рис. 5 видно, что соответствия нормальному распределению нет, отклонения от прямой очень сильные, особенно в левом и правом хвостах распределения.

Необходимо отметить, что в общем случае вывод носит в достаточной степени субъективный характер, один исследователь может полагать, что экспериментальные точки значительно отклоняются от обозначенной диагонали, другой –нет.

На основании значений скошенности (асимметрии) и эксцесса можно сделать вывод о том, что выборочное распределение очень далеко от нормального.

Наконец, осуществим проверку нормальности с помощью формальных тестов Колмогорова- Смирнова и Шапиро-Уилки. Для этого в R предусмотрены функции ks.test и shapiro.test.

**Тест Колмогорова-Смирнова** - это тип непараметрического теста на равенство прерывного и непрерывного одномерного распределения вероятностей, который используется для сравнения выборки с эталонным вероятностным тестом (известным как одновыборочный тест KS) или между двумя выборками. (известный как двухвыборочный тест KS).

При проведении теста Колмогорова-Смирнова выдвигается нулевая гипотеза H0, состоящая в том, что выборочное распределение соответствует предполагаемому. Альтернативная гипотеза заключается в том, что такого соответствия нет.

Тестируемое распределение задается параметром функции (в нашем случае ‘pnorm’), также задаются математическое ожидание и стандартное отклонение, в качестве которых мы используем выборочные характеристики:

ks.test(var3$wage, 'pnorm', mean(var3$wage), sd(var3$wage)).

Результат проведения теста представлен на рис. 6.

Поскольку p-value значительно меньше любого разумного уровня значимости, H0 отвергается, соответствия нормальному распределению нет. На предупреждение (warning) не следует обращать внимание, оно носит информационный характер.

**Р-value** — это вероятность получить значение статистики критерия равное наблюдаемому или более нетипичное по сравнению с наблюдаемым при условии, что нулевая гипотеза верна. Поскольку p-value значительно меньше любого ра-зумного уровня значимости, H0 отвергается, соответствия нормальному распределению нет.

P-значение, или значение вероятности, это статистическая величина, выраженная числом от 0 до 1, которая используется для проверки гипотезы.

**Pnorm** - кумулятивная функция распределения. Кумулятивная функция распределения, или просто функция распределения, дает вероятность того, что случайная величина X меньше или равна определенному значению (x).

**Стандартное отклонение** (**sd**)- это наиболее распространенный показатель в теории вероятности и статистике, оценивающий среднеквадратичное отклонение случайной величины (x) относительно ее математического ожидания на основе несмещенной оценки ее дисперсии.

data.name символьная строка, дающая имя (имена) данных.

аlternative указывает на альтернативную гипотезу и должен быть одним из "two.sided".

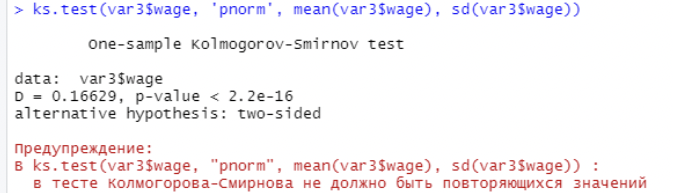


Рисунок 6 – Результаты теста Колмогорова-Смирнова

**Критерий Шапиро-Уилка** является специальным критерием нормальности и используется для проверки гипотезы о нормальном распределении.

Для проведения теста Шапиро-Уилка нулевая и альтернативная гипотезы формулируются в том же виде, команда: shapiro.test(var3$wage)

Результат проведения теста представлен на рис. 7. Поскольку p-value значительно меньше любого разумного уровня значимости, H0 отвергается, соответствия нормальному распределению нет, как и в предыдущем тесте.

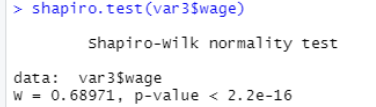
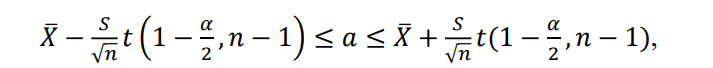


Рисунок 7 – Результаты теста Шапиро-Уилка

5.Построим доверительный 95%-й интервал для математического ожидания среднемесячной заработной платы при неизвестной дисперсии.

**Доверительный интервал** — это интервал, построенный с помощью случайной выборки из распределения с неизвестным параметром, такой, что он содержит данный параметр с заданной вероятностью.

Доверительный интервал для математического ожидания при неизвестной дисперсии имеет вид:

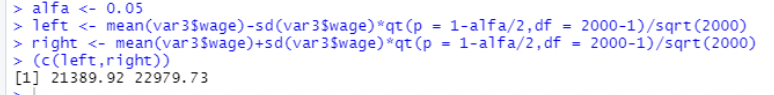


где S – выборочное стандартное отклонение, 𝑛 – объем выборки, t(1-a/2 ,n-1) – квантиль t-распределения Стьюдента.

Используем команды:

alfa <- 0.05  
left <- mean(var3$wage)-sd(var3$wage)\*qt(p = 1-alfa/2,df = 2000-1)/sqrt(2000)  
right <- mean(var3$wage)+sd(var3$wage)\*qt(p = 1-alfa/2,df = 2000-1)/sqrt(2000)  
(c(left,right))

В результате получаем, что с 95% вероятностью среднемесячная заработная плата находится в интервале от 21 389 руб. до 22 980 руб.



Вывод: в ходе лабораторной работы мы выяснили, что гипотеза H0, состоящая в том, что выборочное распределение соответствует предполагаемому, отвергается. Также мы выявили, что с 95 % вероятностью среднемесячная заработная плата находится в интервале от 21 389 р. до 22 980 р.