Лесных Кирилл, БСЦ-153

# Отчет о выполнении домашнего задания №1

## Краткое описание данных

Приведенная база данных представляет собой результаты панельного исследования National Longitudinal Survey of Youth 1979–1994 (NLSY79), целью которого являлось установление взаимосвязи между уровнем дохода и рядом социально-демографических характеристик, таких, как пол, возраст, расовая принадлежность и, что более важно, уровень образования и уровень дохода. Выборку составляют молодые люди в количестве 540 людей от 14 до 21 года.

## Распределение переменных

Было проанализировано распределение следующих переменных: **ASVABC**, **HEIGHT**, **WEIGHT85**, **WEIGHT02**, **W\_EXP**, **EARNINGS**, **S**.

Перечисленные переменные измеряют следующие характеристики:

**ASVABC** – Суммарная переменная по знаниям арифметики, знанию слов и пониманию текстового материала.

**HEIGHT** – Рост, в дюймах, в 1985 году.

**WEIGHT85** – Вес, в фунтах, в 1985 году.

**WEIGHT02** – Вес, в фунтах, в 2002 году.

**EARNINGS** – Текущий часовой заработок в $ в 2002 году.

**S** – Количество лет обучения, по состоянию на 2002 году.

**W\_EXP** – Полное количество лет опыта работы вне школы в 2002 году.

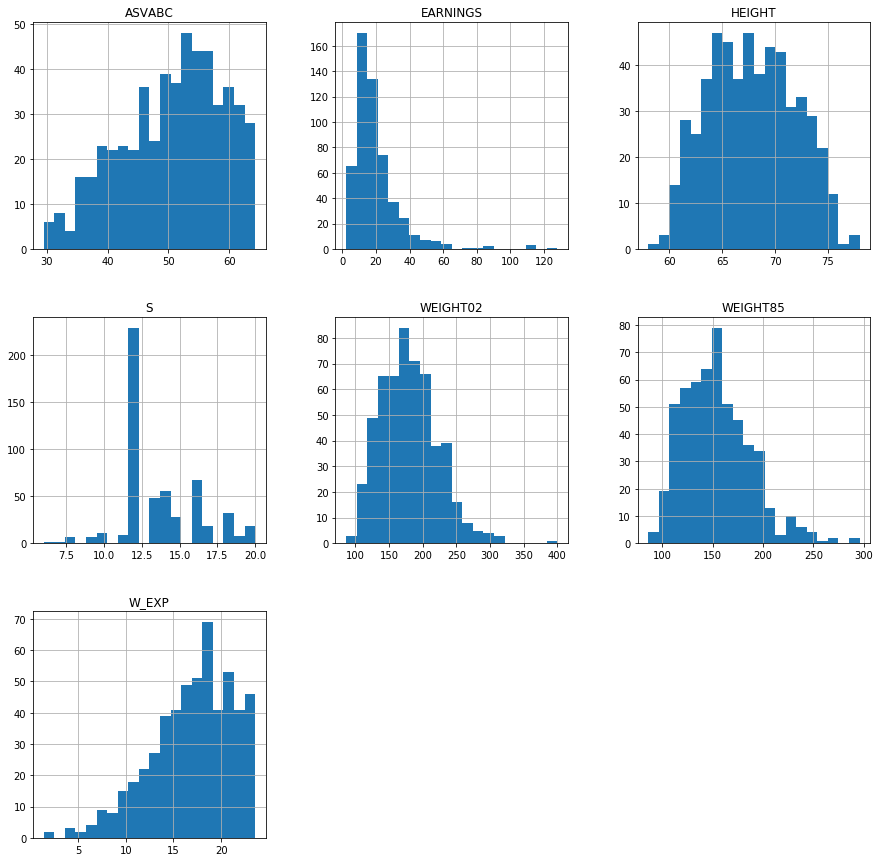
Среди вышеперечисленных переменных не все имеют визуальное отделимое нормальное распределение. Таковым обладают переменные **HEIGHT**, **WEIGHT85**, **WEIGHT02**. Данное распределение можно объяснить тем, что физические характеристики людей обладают некоторым средним значением, некоторые значения роста и веса гораздо часто встречаются по сравнению с другими среди всех людей, поэтому распределяются нормально (*График 1*).

Менее визуально выраженным нормальным распределением, но имеющим смещение в ту или иную сторону, обладают **EARNINGS**, **ASVABC** и **W\_EXP**. Распределение уровня заработка говорит о том, что большая часть выборки обладает схожим размером дохода, но в выборке присутствует доля людей, чей заработок значительно превосходит средний уровень (может превосходить в 5 или 6 раз). Уровень арифметики, набора слов и восприятия текста говорит о вербальных способностях людей, этот показатель говорит о том, что большая доля выборки обладает значением этого показателя на уровне «выше среднего». Распределение количества лет опыта работы говорит о том, значительная часть выборки представляет собой работающих людей, поэтому наиболее часто встречающееся значение этого показателя смещено на графике в правую сторону (*График 1*).

Что касается распределения для **S**, его тоже можно в определенной степени считать нормальным, но в нем наблюдается четкий перевес одного центрального значения по сравнению с другими. Это можно объяснить тем, что для большего количества людей характерно одно количество лет обучения, которое складывается из фиксированного количества лет обучения в школе и фиксированного количества лет в университете (*График 1*).

Выводы, сделанные на основе данных гистограмм, также подтверждаются QQ-графиками (коды их построения приведены в файле *153\_Lesnykh\_HW\_1.ipynb*). Нормальное распределение характерно для всех переменных, однако в одних оно является более выраженным, в других в меньшей степени.

Иные графики, описывающие распределение и построенные в ходе выполнения задания, (*гистограммы в виде кривых*, «*ящики с усами»*) подтверждают вышесказанные выводы.



*График 1.* Столбчатые гистограммы

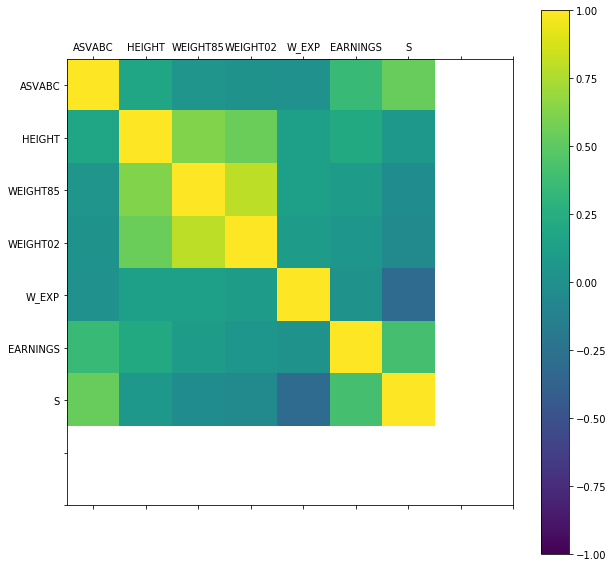
Таким образом, для перечисленных переменных из приведенной части базы данных характерно нормальное распределение. Но для ряда переменных оно является более визуально выраженным, для других в меньшей степени, что можно объяснить смещением в ту или иную сторону.

## Зависимости между переменными

Зависимости между переменными в данной работе можно проанализировать с помощью таблиц корреляций и диаграмм рассеяния.

Прежде чем приступать к анализу связей, стоит отметить, что в данной работе выводы о зависимостях между переменными строятся исключительно только на визуальном анализе построенных графиков и диаграмм, в данной работе не проверяются статистические значимости различий и взаимосвязей, поэтому выводы в данной работе являются лишь предположениями и приблизительными.

Таблица корреляций по всем переменным представлена далее (*График 2*).



*График 2.* Корреляционная матрица

Согласно данной матрице, наибольшая положительная корреляционная связь[[1]](#footnote-1)\* характерна для пар переменных **S – ASVABC**, **S – EARNINGS** и **ASVABC – EARNINGS**, отрицательная корреляционная связь характерна для пары **S – W\_EXP**.

Положительную связь между **S – ASVABC** можно объяснить тем, что люди, которые больше лет потратили на учебу, имеют больший уровень общих вербальных навыков решения математических примеров, словарного запаса и текстового анализа.

Положительную связь между **S – EARNINGS** можно объяснить тем, что люди, которые больше лет потратили на учебу, являются более образованными и, следовательно, являются более способными занимать высокооплачиваемые должности.

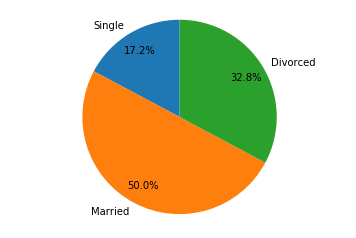
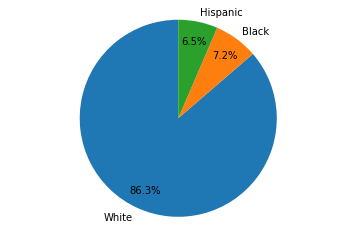
Положительную связь между **ASVABC – EARNINGS** можно объяснить тем, что люди, которые обладают более высоким уровнем общих вербальных навыков решения математических примеров, словарного запаса и текстового анализа, являются более способными занимать высокооплачиваемые должности.

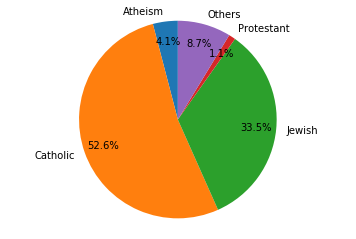
Отрицательную связь между **S – W\_EXP** можно объяснить тем, что у людей, которые больше лет потратили на учебу, оставалось меньше времени на работу. Это утверждение работает и в обратную сторону.

Диаграммы рассеяния по приведенным выше парам также подтверждают сделанные выводы о характере связи (диаграммы приведены в файле с кодом *153\_Lesnykh\_HW\_1.ipynb*).

## Распределение выборки по разным социально-демографическим характеристикам

Также, в работе был проведен анализ распределения выборки по разным социально-демографическим признакам: *этническая принадлежность*, *семейное положение*, *религия*. Диаграммы по данному распределению представлены далее.





*График 3.* Распределение выборки по признакам   
этнической принадлежности, семейного положения и религии

На основе визуального анализа данных круговых диаграмм можно сделать следующие выводы о выборке:

* Большую часть выборки составляет население с белым цветом кожи (**86,3%**).
* Люди в браке составляют около половины выборки, доля разведенных составляет около **30%**, а одиночки – всего **17,2%.**
* Около половины выборки составляют католики, **33,5 %** составляет доля евреев, **4,1%** – атеисты, оставшаяся доля – иные религии.

Таким образом, именно такие вышеперечисленные выводы можно сделать на основе проведенного анализа. Стоит отметить, что все выводы являются предположениями, выдвинутыми только на основе визуального анализа графиков, статистическая значимость ни одного из утверждений не была проверена.

1. \* положительная корреляционная связь между **HEIGHT** и **WEIGHT** не рассматривается, она является вполне логичной и понятной, люди с большим физическим ростом будут обладать большим физическим весом, и наоборот [↑](#footnote-ref-1)