

Национальный исследовательский университет  
"Высшая школа экономики"

Факультет компьютерных наук  
Департамент анализа данных и искусственного интеллекта

Домашнее задание  
по анализу и разработке данных

Выполнили студенты БПМИ133:

**Стеценко Макар**

**Корытова Александра**

**Милеев Алексей**

Москва 2015

## Домашнее задание №1

1. В настоящих данных приводится статистика по NEA (Near Earth Objects) и кометам, обнаруженным исследовательской миссией NEOWISE под руководством NASA. Near-Earth Objects - это кометы и астероиды, которые были притянуты гравитационным полем ближайших планет, в следствии чего они смогли сблизиться с Землей.

Каждый объект описывается следующим набором признаков:

- Discovery Date [Дата открытия]

Качественный признак в формате YYYY-MM-DD

- H (mag) [Магнитуда]

Количественный признак, абсолютная величина

С помощью абсолютной магнитуды вычисляется диаметр астероида, чем ниже значение H, тем больше размер объекта.

- MOID (AU) - Minimum Orbit Distance [Минимальная дистанция орбиты]

Количественный признак, измеряемый относительно AU (The astronomical unit).

//AU - астрономическая единица измерения длины, приблизительно показывающая расстояние между Землей и Солнцем. Равна 149597870700 метров (примерно 150 млн км).

Minimum orbit intersection distance (MOID) - мера, используемая в астрономии для оценки потенциальных сближений и рисков столкновений между астрономическими объектами.

- q (AU) perihelion distance

Количественный признак

Perihelion - точка на орбите планеты, кометы или другого объекта, расстояние от которой до Солнца минимально.

- Q (AU) aphelion distance

Количественный признак

Aphelion - точка, в которой небесное тело максимально удалено от Солнца.

- period (yr) [Период]

Количественный признак, показывающий период обращения объекта вокруг Солнца, измеряется в годах.

- PHA (Potentially Hazardous Asteroids)

Признак, показывающий принадлежит ли астероид к классу PHA. Принимает два значения (Y/N), для удобства можно считать количественным.

- Orbit Class [Класс орбиты]

Качественный признак, множество принимаемых значений: {Apollo, Aten, Amor, Atiras}.

## 2. Предметная область

Научный интерес к таким объектам проявлен во многом из-за их происхождения. Так, например, астероиды по сути являются уцелевшими осколками после формирования нашей солнечной системы. Поскольку эти объекты могут столкнуться, они оказывали и будут оказывать влияние на биосферу Земли. Так же астероиды являются богатым источником ресурсов. Выяснилось, что минеральных запасов в астероидном поясе Марса и Юпитера хватит, чтобы каждому человеку на Земле дать 100 миллиардов долларов.

По имеющимся данным можно пробовать строить модели для определения принадлежности небесного объекта к классу РНА.

Источник: <http://neo.jpl.nasa.gov/stats/wise/>.

### Домашнее задание №2

1. Был выбран количественный признак  $H$  (mag) [Магнитуда]. Поскольку этот признак позволяет определить размер исследуемого объекта, то его подробное изучение позволит лучше понять, каких размеров достигают наиболее встречаемые астероиды. В используемом наборе данных  $H$  принимает следующие значения:

15.6 16.2 17.0 17.5 18.3 18.3 18.7 18.7 18.9 19.0 19.1 19.2 19.2 19.3 19.3 19.3 19.4 19.4 19.4 19.5 19.5  
19.5 19.6 19.6 19.7 19.7 19.7 19.7 19.8 19.9 19.9 19.9 20.0 20.1 20.1 20.2 20.2 20.3 20.3 20.4 20.6 20.7  
20.7 20.7 20.7 20.8 20.8 20.8 20.9 20.9 20.9 21.0 21.0 21.1 21.3 21.4 21.4 21.5 21.5 21.6 21.8 21.8 22.0  
22.1 22.3 22.5 22.6 22.6 23.2 24.1

Построим гистограмму:

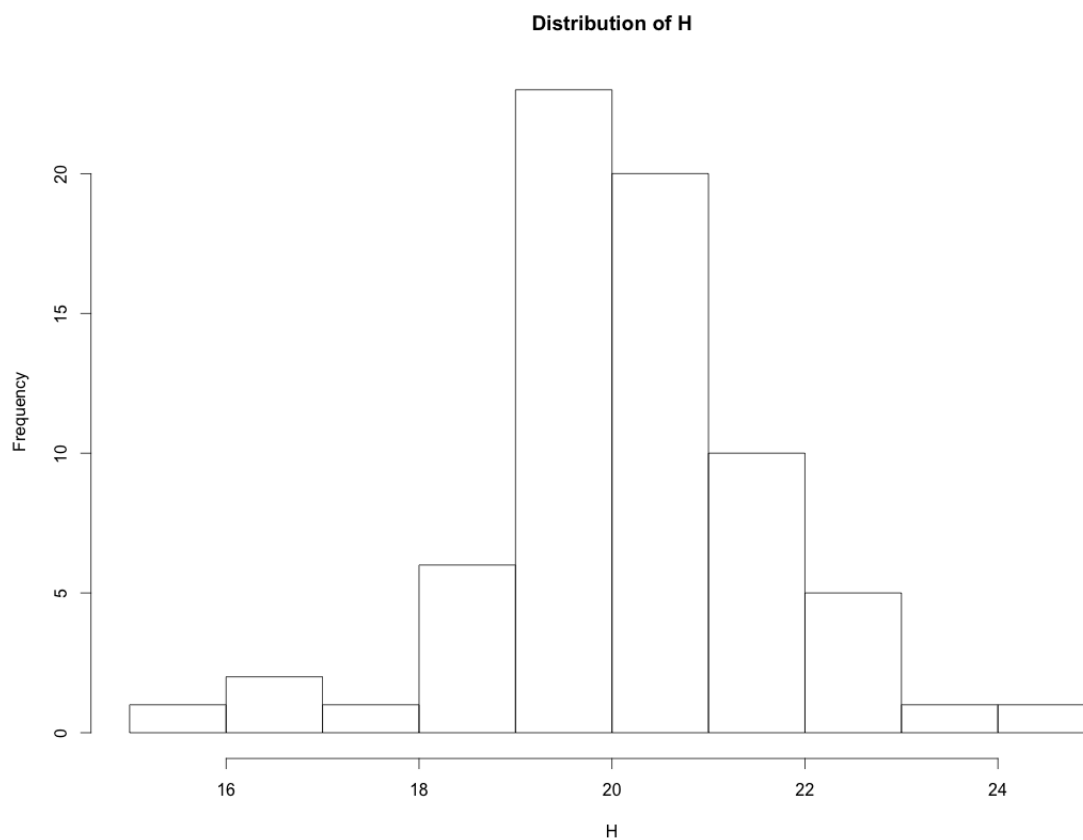


Рис. 1: Гистограмма для признака H (mag)

Полученная гистограмма позволяет нам предположить, что распределение признака H похоже на нормальное. А также понять, в каком диапазоне лежат наиболее встречаемые значения H (примерно от 19 до 21). Этот факт подтвердится, когда мы найдем моду. Построим бокс-плот:

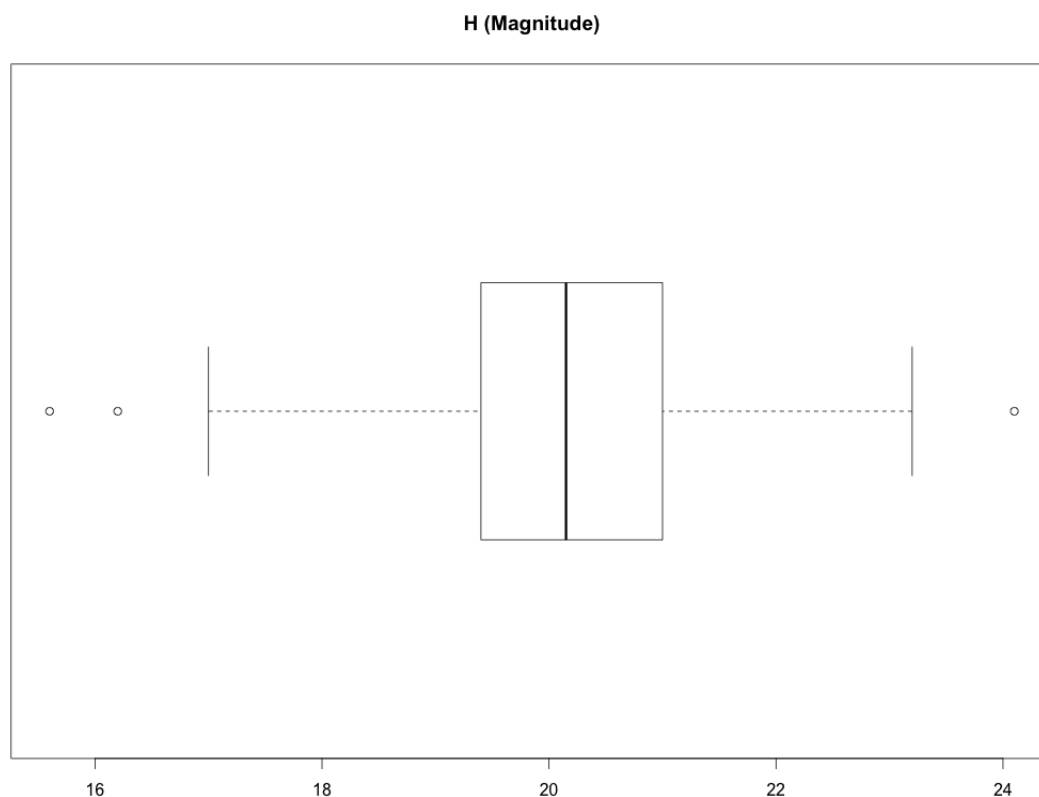


Рис. 2: Бокс-плот для H (mag)

Видно, что у нас есть 3 выброса, а именно: [16.2, 24.1, 15.6]. Так же видно значение медианы. Найдем среднее значение, моду и медиану

Среднее	Медиана	Мода
20.21	20.15	19.7 и 20.7

Как видно, найденные значения не равны, это свидетельствует о том, что величина H не подчиняется нормальному распределению, а немного отклоняется от него. Однако, если убрать из расчетов найденные выбросы [16.2, 24.1, 15.6] и пересчитать, то получим равные между собой значения:

Среднее	Медиана	Мода
20.2	20.2	$(19.7 + 20.7) / 2 = 20.2$

2. Теперь построим доверительные интервалы тремя методами:

1. Статистический
2. Опорный бутстрэп
3. Безопорный бутстрэп

Так как наше распределение похоже на нормальное, то

$$CI = \left( mean - 1.965 \frac{std}{\sqrt{n}}; mean + 1.965 \frac{std}{\sqrt{n}} \right)$$

$$CI = (19.859740; 20.560260)$$

Для 5000 средних значений от случайных выборок с повторениями построим гистограмму:

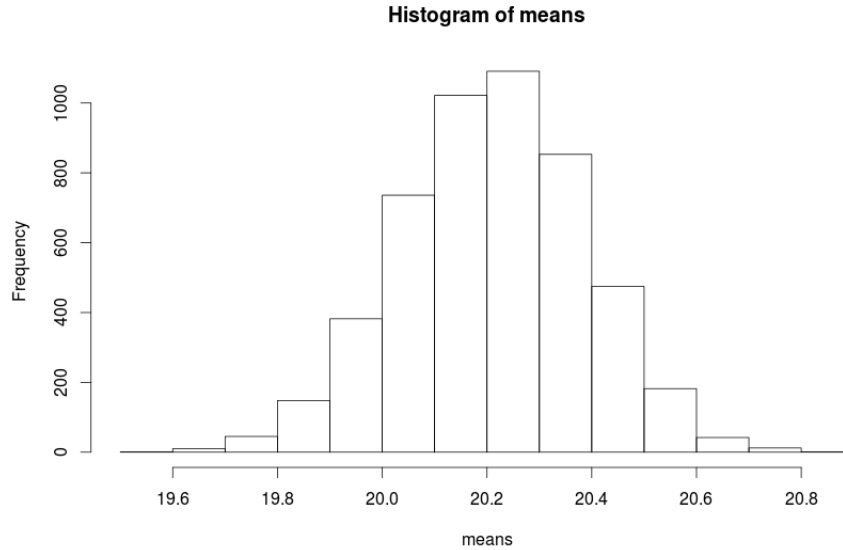


Рис. 3: Гистограмма для признака средних значений

Видим, что гистограмма похожа на нормальное распределение, а значит применяем метод опорного бутстрэпа и получаем интервал:

$$PCI = (20.207460; 20.217272)$$

И теперь безопорный бутстрэп:

$$NPCI = (19.864286; 20.542857)$$

3. Покажем, что для моды и медианы нельзя использовать технику опорного бутстрэпа. Для это построим гистограммы аналогично случаю со средним.

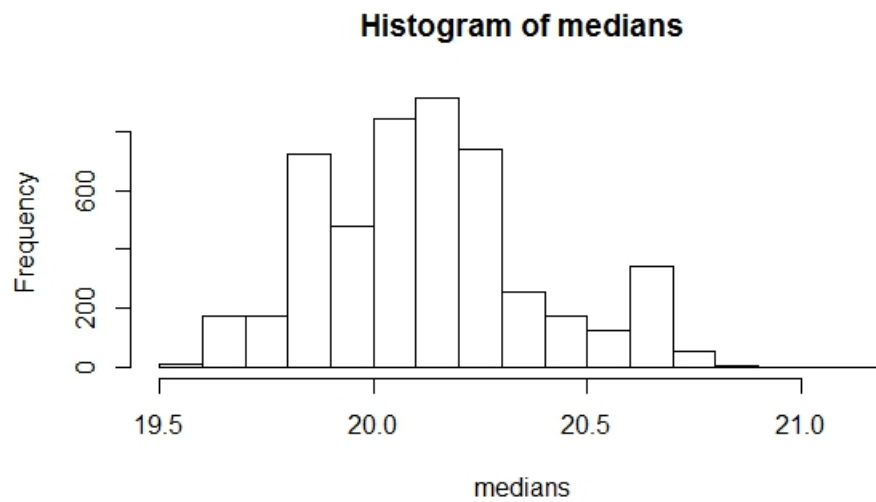


Рис. 4: Гистограмма для медиан

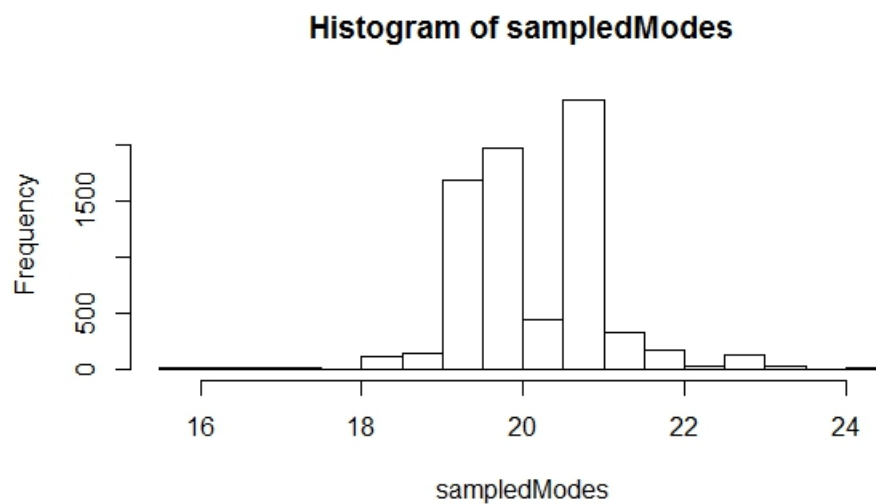


Рис. 5: Гистограмма для мод

Видим, что распределения совсем не напоминают Гауссовские. Значит, мы можем использовать только безопорный бутстрэп.

Доверительный интервал для медианы:

$$NPCI = (19.700000; 20.700000)$$

Доверительный интервал для моды:

$$NPCI = (18.300000; 20.700000)$$