## 1.9. Упражнения

### Идентификация наборов данных

1.1. *[3]* Укажите, где находятся в вебе интересные наборы данных, подхо-
щие для следующих областей.
(a) Книги.
(b) Скачки.
(c) Курсы акций.
(d) Риски болезней.
(e) Колледжи и университеты.
(f) Индексы преступности.
(g) Наблюдение за птицами.
Для каждого из этих источников данных объясните, что следует сдела
чтобы превратить данные в формат, пригодный для использования на в
шем компьютере для анализа.

1.2. *[3]* Предложите подходящие источники данных для следующих конкур
сов по прогнозированию для передачи *The Quant Shop*. Укажите, каки
источники данных, вы уверены, должны *у кого-то* быть, и какие данны
вполне определенно доступны для вас.
(a) *Miss Universe*.
(b) *Movie gross*.
(c) *Baby weight*.
(d) *Art auction price*.
(e) *White Christmas*.
(f) *Football champions*.
(g) *Ghoul pool*.
(h) *Gold/oil prices*.

1.3. *[3]* Посетите сайт `http://data.gov` и выявите пять наборов данных,
которые покажутся вам интересными. Для каждого сделайте краткое опи
сание и предложите три интересные вещи, которые вы могли бы сделать
с ними.

### Задайте вопросы

1.4. *[3]* Для каждого из следующих источников данных предложите три ин
тересных вопроса, на которые вы сможете ответить, проанализировав их
(a) Данные расчетов по кредитным карточкам.
(b) Данные о щелчках на `http://www.amazon.com`.
(c) Справочники адресов и телефонов.



1.5. *[5]* Посетите портал Национального центра биотехнологической инфор
мации (NCBI). Посмотрите, какие источники данных доступны, особен
но ресурсы Pubmed и Genome. Предложите три интересных проекта ис
следования каждого из них.

1.6. *[5]* Вы хотели бы провести эксперимент и установить, предпочитают ли
ваши друзья вкус обычной кока-колы или диетической. Коротко набро
сайте проект такого исследования.

1.7. *[5]* Вы хотели бы провести эксперимент и установить, учатся ли студенты
лучше, если они делают это без музыки, с инструментальной музыкой или
с лирическими песнями. Коротко набросайте проект такого исследования.

1.8. *[5]* Традиционные опросы, в частности Gallup, используют такую проце
дуру, как звонок по случайному номеру, который состоит из строки слу
чайных цифр, а не выбирается из телефонной книги. Укажите, почему
такие опросы проводятся с использованием случайных наборов цифр в
номере.

### Реализация проектов

1.9. *[5]* Напишите программу для очистки списка книжных бестселлеров на
Amazon.com. Используйте ее для составления графика рангов всех книг
Скиены за все время. Какая из этих книг должна быть следующей вашей
покупкой? Если у вас есть друзья, которым вы желаете добра, то не хоти
те ли вы сделать им ценный подарок?

1.10. *[5]* Создайте для вашего любимого вида спорта (бейсбол, американский
футбол, баскетбол, крикет или футбол) набор данных с историческими
статистическими записями обо всех главных участниках. Разработайте и
реализуйте систему рангов, чтобы выявить наилучшего игрока в каждой
позиции.

### Вопросы на интервью

1.11. *[3]* Для каждого из следующих вопросов: (1) сделайте быстрое предполо
жение на основании только вашего понимания мира, а затем (2) исполь
зуйте Google для поиска приемлемых чисел, чтобы получить более прин
ципиальную оценку. Насколько отличаются ваши две оценки?
(a) Сколько настройщиков фортепьяно есть во всем мире?
(b) Сколько весит лед на хоккейном поле?
(c) Сколько бензоколонок в Соединенных Штатах?
(d) Сколько людей входит и выходит из аэропорта Ла Гуардия каждый день?
(e) Сколько галлонов (3,78 литра) мороженого продается в Соединенных
Штатах каждый год?

(f) Сколько баскетболистов покупает Национальная баскетбольная ассоциация (NBA) каждый год?
(g) Сколько рыб плавает во всех океанах в мире?
(h) Сколько людей во всем мире летит в воздухе прямо сейчас?
(i) Сколько теннисных мячей может вместить большой коммерческий самолет?
(j) Сколько миль (1609,34 метра) мощеных дорог в вашей любимой стране?
(k) Сколько долларов находится в бумажниках всех людей в университете Стоуни-Брук?
(l) Сколько галлонов бензина продает в день типичная бензоколонка?
(m) Сколько слов в этой книге?
(n) Сколько котов живет в Нью-Йорке?
(o) Сколько стоило бы заполнить бензобак типичного автомобиля кофе из Starbuck?
(p) Сколько чая в Китае?
(q) Сколько текущих счетов в Соединенных Штатах?

1.12. *[3]* В чем разница между регрессией и классификацией?
1.13. *[8]* Как вы построили бы управляемую данными систему рекомендаций? Каковы ограничения этого подхода?
1.14. *[3]* Как вы заинтересовались наукой о данных?
1.15. *[3]* Как, по-вашему, наука о данных — это искусство или наука?

### Конкурсы Kaggle

1.16. Кто пережил кораблекрушение "Титаника"?
`https://www.kaggle.com/c/titanic`
1.17. Куда какое такси едет?
`https://www.kaggle.com/c/pkdd-15-predict-taxi-service-trajectory-i`
1.18. Сколько времени займет конкретная поездка на такси?
`https://www.kaggle.com/c/pkdd-15-taxi-trip-time-prediction-ii`

# Глава 2

# Математические основы

> Аналитик данных — это тот, кто знает о статистике больше, чем программист, и больше об информатике, чем статистик.
>
> — Джош Блюменсток (Josh Blumenstock)

Вы должны научиться ходить, прежде чем сможете бегать. Аналогично существует определенный уровень математической зрелости, достичь которого необходимо прежде, чем вам станут доверять выполнение чего-либо значимого с числовыми данными.

Создавая эту книгу, я предполагал, что у читателей есть некоторый уровень знаний в областях вероятности и статистики, линейной алгебры и непрерывной математики. Я также предполагал, что они, вероятно, забыли большинство из этого или, возможно, не всегда видят лес (почему вещи важны, и как их использовать) за деревьями (все подробности определений, доказательства и операции).

Эта глава будет пытаться обновить ваше понимание определенных элементарных математических концепций. Следуйте за мной и вытаскивайте ваши старые учебники на случай, если понадобится справка. Более сложные концепции будут представлены в книге по мере необходимости.

## 2.1. Вероятность

Теория вероятности служит формальной основой для рассуждения о вероятности событий. Поскольку это формальная дисциплина, существует множество взаимосвязанных определений, позволяющих точно определить, о чем мы рассуждаем.

- *Эксперимент* (experiment) — это процедура, приводящая к одному из нескольких возможных результатов. В качестве примера будем рассматривать эксперимент, в ходе которого бросают две шестигранные игральные кости (кубика), одна красного, другая синего цвета, на каждой грани которых расположены целые числа $\{1, \ldots, 6\}$.