Ник Сингх

Автор бестселлера Amazon, автор интервью Ace the Data Science и создатель курса Ace the Data Job Hunt.

Основатель DataLemur и ранее инженер-программист в Facebook и Google.
Присоединяйтесь к моему бесплатному 9-дневному ускоренному курсу Data Interview!

40 вопросов для интервью о вероятности и статистике по науке о данных, заданных FAANG и Уолл-Стрит

КОНСУЛЬТАЦИИ ПО КАРЬЕРЕ В ОБЛАСТИ РАЗРАБОТКИ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ

Мы не можем лгать — интервью по науке о данных — это **СЛОЖНО** . Особенно сложно — вопросы о вероятности и статистике, которые задают ведущие технологические компании и хедж-фонды в процессе собеседования по науке о данных, аналитику данных и Quant Trading.

Вот почему мы собрали **40 реальных вопросов для собеседования по науке о вероятностях и статистических данных** , которые задают такие компании, как Facebook, Amazon, Two Sigma и Bloomberg. У нас есть решения для всех 40 задач и еще 161 задачи интервью с данными по SQL, машинному обучению и продукту/бизнесу в нашей книге Ace The Data Science Interview . Вы также можете попрактиковаться с некоторыми из этих точных вопросов в разделе вопросов статистического интервью DataLemur .

Итак, без лишних слов, вот:

- концепции вероятности и статистики, которые необходимо рассмотреть перед собеседованием с DS
- 20 вопросов о вероятностях, которые задают ведущие технологические компании и Уолл-стрит
- 20 вопросов по статистике, заданных FANG & Hedge Funds
- ответы на 5 вероятностных вопросов
- ответы на 5 вопросов статистики
- ссылки на дополнительные ресурсы для интервью по науке о данных

Концепции вероятности и статистики, которые следует изучить перед

Ник Сингх

Автор бестселлера Amazon, автор интервью Ace the Data Science и создатель курса Ace the Data Job Hunt.

Основатель DataLemur и ранее инженер-программист в Facebook и Google.
Присоединяйтесь к моему бесплатному 9-дневному ускоренному курсу Data Interview!

собеседованием по науке о данных

Поскольку вероятность и статистика лежат в основе науки о данных, перед собеседованием вы должны просмотреть:

- Основы вероятности и случайные величины
- Распределения вероятностей
- Проверка гипотезы
- Регрессивный анализ

Если эти статистические концепции кажутся вам чуждыми, ознакомьтесь с некоторыми из наших любимых книг по статистике для аналитиков данных , чтобы освежить их в памяти.

Основы вероятности и случайные величины

Истоки вероятности начинаются с размышлений о выборочных пространствах, основных принципах подсчета и комбинаторики. Хотя не обязательно знать все тонкости комбинаторики, полезно понимать основы упрощения задач. Одним из классических примеров здесь является метод подсчета «звезды и полосы».

Другой ключевой темой для изучения являются случайные величины. Знание понятий, связанных с ожиданием, дисперсией, ковариацией, наряду с основными распределениями вероятностей, имеет решающее значение.

Распределения вероятностей

Для моделирования случайных величин важно знать основы различных вероятностных распределений. Понимание как дискретных, так и непрерывных примеров в сочетании с ожиданиями и отклонениями имеет решающее значение. Наиболее распространенными распределениями, обсуждаемыми в интервью, являются Равномерное и Нормальное, но есть множество других хорошо известных распределений для конкретных случаев использования (Пуассона, Биномиальное, Геометрическое).

В большинстве случаев достаточно знать основы и их применение. Например, под каким распределением будет подбрасываться монета? Как насчет ожидания события? Никогда не помешает иметь возможность делать выводы для ожидания, дисперсии или других более высоких моментов.

Проверка гипотезы

Проверка гипотез является основой статистического вывода и может быть разбита на несколько тем. Первая — это Центральная предельная теорема, которая играет важную роль при изучении больших выборок данных. Другие основные элементы проверки гипотез: распределения выборки, p-значения, доверительные интервалы, ошибки I и II типов. Наконец, стоит взглянуть на различные тесты, включающие пропорции и другие тесты гипотез.

Большинство этих концепций играют решающую роль в A/B-тестировании, которое часто задают во время интервью в таких компаниях, занимающихся потребительскими технологиями, как Facebook, Amazon и Uber. Полезно не только понимать технические детали, но и концептуально, как работает A/B-тестирование, каковы предположения, возможные ловушки и приложения к реальным продуктам.

Ник Сингх

Автор бестселлера Amazon, автор интервью Ace the Data Science и создатель курса Ace the Data Job Hunt.

Основатель DataLemur и ранее инженер-программист в Facebook и Google.
Присоединяйтесь к моему бесплатному 9-дневному ускоренному курсу Data Interview!

БЛОГ

ИНТЕРВЬЮ С ЭЙСОМ О НАУКЕ О ДАННЫХ

ПОДГОТОВКА К СОБЕСЕДОВАНИЮ

DATALEMUR ПО SQL

14 КНИГ, КОТОРЫЕ ИЗМЕНИЛИ МОЮ ЖИЗНЬ

ОБО МНЕ

Моделирование

Моделирование опирается на глубокое понимание распределения вероятностей и проверки гипотез. Поскольку это широкий термин, мы будем называть моделирование областями, которые имеют сильное статистическое пересечение с машинным обучением. Сюда входят такие темы, как линейная регрессия, оценка максимального правдоподобия и байесовская статистика. Для интервью, посвященных моделированию и машинному обучению, знание этих тем необходимо.

20 проблем на вероятностном собеседовании, заданных ведущими технологическими компаниями и Уолл-Стрит

1. **[Facebook - Легко]** [Тест монеты на честность на DataLemur] Есть честная монета (с одной стороны решка, с одной стороны решка) и нечестная монета (с обеих сторон решка). Вы выбираете одну из них наугад, переворачиваете ее 5 раз и наблюдаете, что все пять раз она выпадает решкой. Какова вероятность того, что вы подбрасываете нечестную монету?
2. **[Lyft - Easy]** Вы и ваш друг играете в игру. Вы двое будете продолжать подбрасывать монету, пока не появится последовательность HH или TH. Если HH появится первым, вы выиграете. Если TH появится первым, ваш друг выиграет. Какова вероятность того, что вы выиграете?
3. **[Google – Easy]** Какова вероятность того, что серия из семи игр будет состоять из 7 игр?
4. **[Facebook – Легко]** В Facebook есть команда по контенту, которая помечает фрагменты контента на платформе как спам или не спам. 90% из них являются прилежными оценщиками и помечают 20% контента как спам и 80% как не спам. Оставшиеся 10 % не являются прилежными оценщиками и помечают 0 % контента как спам, а 100 % — как не спам. Предположим, что части контента помечены независимо друг от друга для каждого оценщика. Учитывая, что оценщик отметил 4 части контента как хорошие, какова вероятность того, что они являются прилежными оценщиками?
5. **[Bloomberg - Easy]** Допустим, вы рисуете круг и выбираете два случайных аккорда. Какова вероятность того, что эти хорды пересекутся?
6. **[Amazon - Easy]** 1/1000 человек имеют определенное заболевание, и есть тест, который дает 98% точных результатов, если у вас есть это заболевание. Если у вас нет болезни, вероятность ошибки составляет 1%. Если у кого-то положительный результат, каковы шансы, что у него есть болезнь?
7. **[Facebook - Easy]** В наборе 50 карт 5 разных цветов. Карты каждого цвета пронумерованы от 1 до 10. Вы выбираете 2 карты наугад. Какова вероятность того, что они не одного цвета и не одного номера?
8. **[Tesla - Easy]** Честная шестигранная кость подбрасывается дважды. Какова вероятность того, что при первом броске выпало 1, а при втором броске не 6?
9. **[Facebook – Easy]** Какое ожидаемое количество бросков нужно, чтобы увидеть все 6 сторон игральной кости?
10. **[Microsoft - Легко]** Трое друзей в Сиэтле сказали вам, что идет дождь, и вероятность того, что каждый из них солжет, равна 1/3. Какова вероятность того, что в Сиэтле дождливо? Предположим, что вероятность дождя в любой день в Сиэтле равна 0,25.
11. **[Убер - Легко]** Допустим, вы бросаете три кубика один за другим. Какова вероятность того, что вы получите 3 числа в строго возрастающем порядке?
12. **[Bloomberg - Medium]** Три муравья сидят в углах равностороннего треугольника. Каждый муравей случайным

Ник Сингх

Автор бестселлера Amazon, автор интервью Ace the Data Science и создатель курса Ace the Data Job Hunt.

Основатель DataLemur и ранее инженер-программист в Facebook и Google.
Присоединяйтесь к моему бесплатному 9-дневному ускоренному курсу Data Interview!

БЛОГ

ИНТЕРВЬЮ С ЭЙСОМ О НАУКЕ О ДАННЫХ

ПОДГОТОВКА К СОБЕСЕДОВАНИЮ

DATALEMUR ПО SQL

14 КНИГ, КОТОРЫЕ ИЗМЕНИЛИ МОЮ ЖИЗНЬ

ОБО МНЕ

образом выбирает направление и начинает двигаться по краю треугольника. Какова вероятность того, что ни один из муравьев не столкнется? А что, если это k муравьев во всех k углах равностороннего многоугольника?

13. **[Две сигмы - средняя]** Какое ожидаемое количество подбрасываний монеты необходимо, чтобы выпали два последовательных орла?
14. **[Amazon – Medium]** Сколько карт вы ожидаете вытянуть из стандартной колоды, прежде чем увидите первого туза?
15. **[Robinhood - Medium]** A и B играют в игру, в которой у A есть n+1 монета, у B есть n монет, и каждый из них подбрасывает все свои монеты. Какова вероятность того, что у A будет больше орлов, чем у B?
16. **[Airbnb — Medium]** Допустим, вам дали нечестную монету с неизвестным уклоном в сторону орла или решки. Как вы можете получить справедливые шансы, используя эту монету?
17. **[Quora - Medium]** Допустим, у вас есть N iid рисунков нормального распределения с параметрами μ и σ. Какова вероятность того, что k таких розыгрышей больше некоторого значения Y?
18. **[Spotify - Hard]** Правильная игральная кость бросается n раз. Какова вероятность того, что наибольшее выпавшее число равно r для каждого r из 1..6?
19. **[Snapchat - Hard]** Есть две группы из n пользователей, A и B, и каждый пользователь в A дружит с пользователями в B, и наоборот. Каждый пользователь в A будет случайным образом выбирать пользователя в B в качестве своего лучшего друга, и каждый пользователь в B будет случайным образом выбирать пользователя в A в качестве своего лучшего друга. Если два человека выбрали друг друга, они взаимные лучшие друзья. Какова вероятность того, что взаимных лучших дружеских отношений не будет?
20. **[Tesla - Hard]** Предположим, предстоит запуск нового автомобиля. Исходные данные предполагают, что в любой данный день в какой-либо части транспортного средства может произойти авария или авария с вероятностью p, которая затем потребует замены. Кроме того, каждое транспортное средство, которое существует n дней, должно быть заменено. Какова долгосрочная частота замены транспортных средств?

20 статистических задач, которые задают FAANG и хедж-фонды

1. **[Facebook – Легко]** Как бы вы объяснили доверительный интервал нетехнической аудитории?
2. **[Две сигмы — легко]** Допустим, вы используете множественную линейную регрессию и считаете, что есть несколько коррелирующих предикторов. Как повлияют результаты регрессии, если они действительно коррелированы? Как бы вы справились с этой проблемой?
3. **[Uber - Easy]** Опишите p-значения простыми словами.
4. **[Facebook – Легко]** Как бы вы построили и протестировали метрику для сравнения двух ранжированных списков предпочтений пользователей в отношении фильмов/телешоу?
5. **[Microsoft - Easy]** Объясните статистическую подоплеку власти.
6. **[Twitter – Easy]** Опишите A/B-тестирование. Каковы некоторые распространенные ловушки?
7. **[Google — Medium]** Как бы вы получили доверительный интервал на основе серии подбрасываний монеты?
8. **[Полоса — средний]** Допустим, вы моделируете время жизни для набора клиентов, используя экспоненциальное распределение с параметром λ, и у вас есть история жизни (в месяцах) n клиентов. Каково ваше лучшее предположение для λ?
9. **[Lyft - Medium]** Получите среднее значение и дисперсию равномерного распределения U(a, b).
10. **[Google - Medium]** Скажем, у нас есть X ~ Uniform(0, 1) и Y ~ Uniform(0, 1). Каково ожидаемое значение минимума X и Y?

Ник Сингх

Автор бестселлера Amazon, автор интервью Ace the Data Science и создатель курса Ace the Data Job Hunt.

Основатель DataLemur и ранее инженер-программист в Facebook и Google.
Присоединяйтесь к моему бесплатному 9-дневному ускоренному курсу Data Interview!

БЛОГ

ИНТЕРВЬЮ С ЭЙСОМ О НАУКЕ О ДАННЫХ

ПОДГОТОВКА К СОБЕСЕДОВАНИЮ

DATALEMUR ПО SQL

14 КНИГ, КОТОРЫЕ ИЗМЕНИЛИ МОЮ ЖИЗНЬ

ОБО МНЕ

11. **[Spotify — Medium]** Вы выбираете из равномерного распределения [0, d] n раз. Какова ваша наилучшая оценка d?
12. **[Quora - Medium]** Вы рисуете из нормально распределенной случайной величины X ~ N(0, 1) один раз в день. Каково примерное ожидаемое количество дней, прежде чем вы получите значение больше 2?
13. **[Facebook — Medium]** Получите математическое ожидание для случайной величины с геометрическим распределением.
14. **[Google - Medium]** Монету подбрасывали 1000 раз, и 550 раз выпадал орел. Считаете ли вы монету необъективной? Почему или почему нет?
15. **[Robinhood - Medium]** Допустим, у вас есть n целых чисел 1... n, и выберите случайную перестановку. Для любых целых чисел i, j пусть обмен определяется как когда целое число i находится в j-й позиции, и наоборот. Какова ожидаемая стоимость общего количества свопов?
16. **[Uber - Hard]** В чем разница между MLE и MAP? Опишите его математически.
17. **[Google – сложно]** Допустим, у вас есть два подмножества набора данных, для которых вы знаете их средние значения и стандартные отклонения. Как рассчитать смешанное среднее и стандартное отклонение всего набора данных? Можете ли вы расширить его до K подмножеств?
18. **[Lyft - Hard]** Как вы случайным образом равномерно выбираете точку из круга с радиусом 1?
19. **[Две сигмы — сложно]** Допустим, вы непрерывно выбираете случайные величины из некоторого iid с равномерным распределением (0, 1) до тех пор, пока сумма переменных не превысит 1. Сколько раз вы планируете производить выборку?
20. **[Uber — Hard]** Как при наличии генератора случайных проб Бернулли вернуть значение, выбранное из нормального распределения?

Решения вопросов вероятностного интервью

Проблема №1 Решение:

Здесь мы можем использовать теорему Байеса. Пусть U обозначает случай, когда мы подбрасываем нечестную монету, а F обозначает случай, когда мы подбрасываем честную монету. Поскольку монета выбирается случайно, мы знаем, что P(U) = P(F) = 0,5. Пусть 5T обозначает событие, когда мы подбрасываем 5 орлов подряд. Затем нас интересует P(U|5T), т. е. вероятность того, что мы подбрасываем нечестную монету, учитывая, что мы увидели 5 решек подряд.

Мы знаем, что P(5T|U) = 1, так как по определению нечестная монета всегда будет выпадать решкой. Кроме того, мы знаем, что P(5T|F) = 1/2^5 = 1/32 по определению честной монеты. По теореме Байеса имеем:

$$P(U|5T) = \frac{P(5T|U) * P(U)}{P(5T|U) * P(U) + P(5T|F) * P(F)} = \frac{0.5}{0.5 + 0.5 * 1/32} = 0.97$$

Следовательно, вероятность того, что мы выбрали нечестную монету, составляет около 97%.

Проблема №5 Решение:

По определению, хорда — это отрезок, две конечные точки которого лежат на окружности. Поэтому две произвольные хорды всегда можно представить любыми четырьмя точками, выбранными на окружности. Если вы решите представить первую хорду двумя из четырех точек, то у вас есть:

Ник Сингх

Автор бестселлера Amazon, автор интервью Ace the Data Science и создатель курса Ace the Data Job Hunt.

Основатель DataLemur и ранее инженер-программист в Facebook и Google.
Присоединяйтесь к моему бесплатному 9-дневному ускоренному курсу Data Interview!

БЛОГ

ИНТЕРВЬЮ С ЭЙСОМ О НАУКЕ О ДАННЫХ

ПОДГОТОВКА К СОБЕСЕДОВАНИЮ DATALEMUR ПО SQL

14 КНИГ, КОТОРЫЕ ИЗМЕНИЛИ МОЮ ЖИЗНЬ

ОБО МНЕ

$$\binom{4}{2} = 6$$

варианты выбора двух точек для представления хорды 1 (и, следовательно, две другие точки будут представлять хорду 2). Однако обратите внимание, что в этом подсчете мы дублируем счет каждой хорды дважды, поскольку хорда с концами p1 и p2 такая же, как хорда с концами p2 и p1. Следовательно, правильное количество допустимых аккордов:

$$\frac{1}{2}\binom{4}{2} = 3$$

Среди этих трех конфигураций только одна из хорд будет пересекаться, поэтому искомая вероятность равна:

$$p = \frac{1}{3}$$

Задача №13 Решение:

Пусть X будет количеством подбрасываний монеты, необходимых для выпадения орла. Затем мы хотим найти E[X]. Пусть H обозначает флип, в результате которого выпал орел, а T обозначает флип, в результате которого выпала решка. Обратите внимание, что E[X] может быть записано в терминах E[X|H] и E[X|T], т. е. ожидаемого количества необходимых подбрасываний при условии, что подбрасывание выпадет либо орлом, либо решкой соответственно.

При условии первого флипа имеем:

$$E[X] = \frac{1}{2}(1 + E[X|H]) + \frac{1}{2}(1 + E[X|T])$$

Обратите внимание, что E[X|T] = E[X], так как если решка перевернута, нам нужно начать сначала, чтобы получить две решки подряд.

Чтобы найти E[X|H], мы можем дополнительно обусловить его следующим исходом: либо орлом (HH), либо решкой (HT).

Таким образом, мы имеем:

$$E[X|H] = \frac{1}{2}(1 + E[X|HH]) + \frac{1}{2}(1 + E[X|HT])$$

Обратите внимание, что если результатом является HH, то E[X|HH] = 0, так как результат был достигнут, и что E[X|HT] = E[X], поскольку решка была перевернута, нам нужно начать сначала, поэтому :

$$E[X|H] = \frac{1}{2}(1 + 0) + \frac{1}{2}(1 + E[X]) = 1 + \frac{1}{2}E[X]$$

Подставив это в исходное уравнение, мы получим E[X] = 6 подбрасываний монеты.

Задача №15 Решение:

Рассмотрим первые n монет, которые подбрасывает A, и n монет, которые подбрасывает B.

Возможны три сценария:

1. У A больше голов, чем у B
2. A и B имеют одинаковое количество голов
3. У A меньше голов, чем у B

Обратите внимание, что в сценарии 1 игрок A всегда будет выигрывать (независимо от монеты n+1), а в сценарии 3 игрок A

Ник Сингх

Автор бестселлера Amazon, автор интервью Ace the Data Science и создатель курса Ace the Data Job Hunt.

Основатель DataLemur и ранее инженер-программист в Facebook и Google.
Присоединяйтесь к моему бесплатному 9-дневному ускоренному курсу Data Interview!

всегда будет проигрывать (независимо от монеты n+1). По симметрии эти два сценария имеют равную вероятность возникновения.

Обозначим вероятность любого сценария как x, а вероятность сценария 2 — как y.

Мы знаем, что 2x + y = 1, поскольку эти 3 сценария являются единственными возможными исходами. Теперь рассмотрим монету n+1. Если при подбрасывании выпадет орёл с вероятностью 0,5, то A выиграет после сценария 2 (что происходит с вероятностью y). Следовательно, общие шансы игрока A на победу в игре увеличиваются на 0,5y.

Таким образом, вероятность того, что A выиграет игру, равна:

$$x + \frac{1}{2}y = x + \frac{1}{2}(1 - 2x) = \frac{1}{2}$$

Задача №18 Решение:

Пусть B будет событием, когда все n бросков имеют значение, меньшее или равное r. Тогда у нас есть:

$$P(B_r) = \frac{r^n}{6^n}$$

поскольку все n бросков должны иметь значение меньше или равное r. Пусть A — событие, состоящее в том, что наибольшее число равно r. У нас есть:

$$B_r = B_{r-1} \cup A_r$$

а поскольку два события в правой части не пересекаются, имеем:

$$P(B_r) = P(B_{r-1}) + P(A_r)$$

Следовательно, вероятность A определяется как:

$$P(A_r) = P(B_r) - P(B_{r-1}) = \frac{r^n}{6^n} - \frac{(r-1)^n}{6^n}$$

Решения для статистических вопросов интервью

Проблема №2 Решение:

Основных проблем будет две. Во-первых, оценки и знаки коэффициентов будут сильно различаться в зависимости от того, какие именно переменные вы включаете в модель. В частности, некоторые коэффициенты могут даже иметь доверительные интервалы, включающие 0 (это означает, что трудно сказать, связано ли увеличение этого значения X с увеличением или уменьшением Y). Во-вторых, результирующие значения p будут вводить в заблуждение: важная переменная может иметь высокое значение p и считаться незначительной, хотя на самом деле она важна.

Вы можете решить эту проблему, либо удалив, либо объединив коррелированные предикторы. При удалении предикторов лучше всего понять причины корреляции (т. е. включили ли вы посторонние предикторы или такие, как X и 2X). Для объединения предикторов можно включить условия взаимодействия (произведение двух). Наконец, вы также должны 1) центрировать данные и 2) попытаться получить больший размер выборки (что приведет к более узким доверительным интервалам).

Проблема №9 Решение:

Ник Сингх

Автор бестселлера Amazon, автор интервью Ace the Data Science и создатель курса Ace the Data Job Hunt.

Основатель DataLemur и ранее инженер-программист в Facebook и Google.
Присоединяйтесь к моему бесплатному 9-дневному ускоренному курсу Data Interview!

БЛОГ

ИНТЕРВЬЮ С ЭЙСОМ О НАУКЕ О ДАННЫХ

ПОДГОТОВКА К СОБЕСЕДОВАНИЮ

DATALEMUR ПО SQL

14 КНИГ, КОТОРЫЕ ИЗМЕНИЛИ МОЮ ЖИЗНЬ

ОБО МНЕ

Для X~U(a, b) имеем следующее:

$$f_X(x) = \frac{1}{b-a}$$

Следовательно, мы можем рассчитать среднее значение как:

$$E[X] = \int_a^b x f_X(x) dx = \int_a^b \frac{x}{b-a} dx = \frac{x^2}{2(b-a)}\Big|_a^b = \frac{a+b}{2}$$

Точно так же для дисперсии мы хотим:

$$Var(X) = E[X^2] - E[X]^2$$

И у нас есть:

$$E[X^2] = \int_a^b x^2 f_X(x) dx = \int_a^b \frac{x^2}{b-a} dx = \frac{x^3}{3(b-a)}\Big|_a^b = \frac{a^2+ab+b^2}{3}$$

Поэтому:

$$Var(X) = \frac{a^2+ab+b^2}{3} - (\frac{a+b}{2})^2 = \frac{(b-a)^2}{12}$$

Задача №12 Решение:

Поскольку X нормально распределяется, мы можем посмотреть на кумулятивную функцию распределения (CDF) нормального распределения:

$$\Phi(x) = P(X \leq x)$$

Чтобы проверить, что вероятность X не меньше 2, мы можем проверить (зная, что X распределено как стандартное нормальное):

$$\Phi(2) = P(X \leq 2) = P(X \leq \mu + 2\sigma) = 0.977$$

Поэтому P(X > 2) = 1 - 0,977 = 0,023 для любого данного дня. Поскольку розыгрыши каждый день независимы, то ожидаемое время до розыгрыша X > 2 подчиняется геометрическому распределению с p = 0,023. Пусть T — случайная величина, обозначающая количество дней, тогда имеем:

$$E[T] = \frac{1}{p} = \frac{1}{.024} \approx 43 \text{ days}$$

Задача №14 Решение:

Поскольку размер выборки флипов велик (1000), мы можем применить центральную предельную теорему. Поскольку каждый отдельный бросок является случайной величиной Бернулли, мы можем предположить, что вероятность выпадения орла равна p. Затем мы хотим проверить, равно ли p 0,5 (т. е. справедливо ли оно). Центральная предельная теорема позволяет нам аппроксимировать общее количество орлов, рассматриваемых как нормально распределенные.

В частности, количество видимых орлов должно следовать биномиальному распределению, поскольку оно является суммой случайных величин Бернулли. Если монета не смещена (p = 0,5), то по ожидаемому количеству орла имеем следующее:

$$\mu = np = 1000 * 0.5 = 500$$

а дисперсия определяется как:

$$\sigma^2 = np(1-p) = 1000 * 0.5 * 0.5 = 250, \sigma = \sqrt{250} \approx 16$$

Ник Сингх

Автор бестселлера Amazon, автор интервью Ace the Data Science и создатель курса Ace the Data Job Hunt.

Основатель DataLemur и ранее инженер-программист в Facebook и Google.
Присоединяйтесь к моему бесплатному 9-дневному ускоренному курсу Data Interview!

Поскольку это среднее значение и стандартное отклонение определяют нормальное распределение, мы можем рассчитать соответствующий z-показатель для 550 голов:

$$z = \frac{550 - 500}{16} > 3$$

Это означает, что, если бы монета была честной, событие появления 550 орлов должно произойти с вероятностью < 1% при предположениях о нормальности. Следовательно, монета, вероятно, необъективна.

Задача №20 Решение:

Предположим, у нас есть n испытаний Бернулли, каждое с вероятностью успеха p:

$$x_1, x_2, \ldots x_n,\ x_i \sim Ber(p)$$

Предполагая iid испытаний, мы можем вычислить выборочное среднее для p из большого количества испытаний:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Мы знаем, что математическое ожидание этого выборочного среднего равно:

$$E[\hat{\mu}] = \frac{np}{n} = p$$

Кроме того, мы можем вычислить дисперсию среднего значения этой выборки:

$$Var(\hat{\mu}) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$$

Предположим, мы делаем выборку из большого числа n. В соответствии с центральной предельной теоремой наше выборочное среднее будет нормально распределено:

$$\hat{\mu} \sim N(p, \frac{p(1-p)}{n})$$

Следовательно, мы можем взять z-оценку нашего среднего значения по выборке как:

$$z(\hat{\mu}) = \frac{\hat{\mu} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

Затем этот z-показатель будет смоделированным значением стандартного нормального распределения.

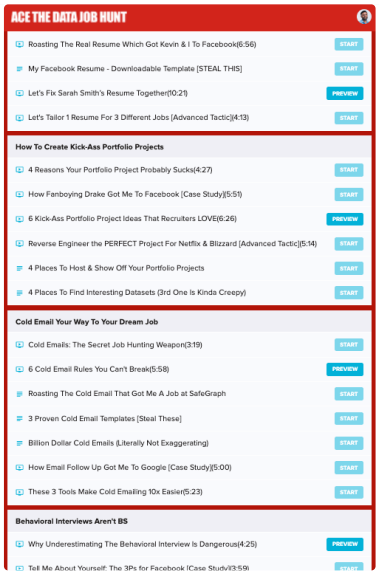
Как получить дополнительные ресурсы для подготовки к собеседованию по науке о данных

Обязательно купите полную 301-страничную книгу на Amazon, а также запишитесь на видеокурс Ace the Data Job Hunt , который охватывает резюме, проект портфолио, холодную электронную почту и поведенческие аспекты интервью для получения работы вашей мечты в данных!

Ник Сингх

Автор бестселлера Amazon, автор интервью Ace the Data Science и создатель курса Ace the Data Job Hunt.

Основатель DataLemur и ранее инженер-программист в Facebook и Google.
Присоединяйтесь к моему бесплатному 9-дневному ускоренному курсу Data Interview!

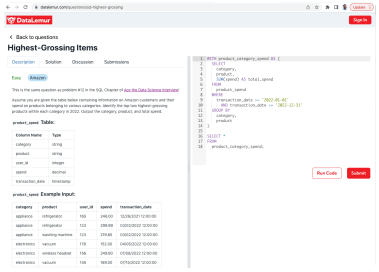


Более 25 видеомодулей , которые помогут вам провести больше интервью с данными!

Вам, вероятно, также понравятся 30 вопросов по SQL и базам данных, которые мы собрали вместе. Хотя это не так сложно, как вопросы о статистике/проблеме здесь, хорошее понимание SQL и дизайна базы данных имеет решающее значение для любого практикующего специалиста по данным или аналитика данных. А если вам нужно меньше практических вопросов и более общие советы о том, как подготовиться к собеседованию по SQL, ознакомьтесь с моим руководством по подготовке к собеседованию по SQL из 5000 слов .

30 вопросов на собеседовании по SQL и БД

И поскольку практика делает совершенным, решайте настоящие вопросы интервью FAANG SQL на DataLemur . Например, ниже приведен снимок экрана с вопросом для собеседования по Amazon SQL на платформе!



Ознакомьтесь с полным списком вопросов для собеседования по SQL !

Ник Сингх

Автор бестселлера Amazon, автор интервью Ace the Data Science и создатель курса Ace the Data Job Hunt.

Основатель DataLemur и ранее инженер-программист в Facebook и Google.
Присоединяйтесь к моему бесплатному 9-дневному ускоренному курсу Data Interview!

БЛОГ
ИНТЕРВЬЮ С ЭЙСОМ О НАУКЕ О ДАННЫХ
ПОДГОТОВКА К СОБЕСЕДОВАНИЮ
DATALEMUR ПО SQL
14 КНИГ, КОТОРЫЕ ИЗМЕНИЛИ МОЮ ЖИЗНЬ
ОБО МНЕ