Вероятность и статистика

Почему условие меняет вероятность, а одинаковое среднее не означает одинаковые данные?

Слово «приз» встретилось и в спаме, и в обычных сообщениях. Можно ли по нему судить о классе? Для ответа нужно уточнить, среди каких сообщений мы считаем долю. Этот вопрос приведёт нас к вероятности и статистике.

1. Событие и частота

Бросим обычный шестигранный кубик. Результат — одно из чисел от одного до шести. **Событие** объединяет результаты, которые нас интересуют. Например, «выпало чётное число» включает два, четыре и шесть. Если все шесть исходов равновероятны, вероятность события равна трём шестым, то есть одной второй.

Вероятность описывает модель случайного опыта. **Частота** описывает уже наблюдавшиеся результаты: сколько раз событие произошло, делённое на число попыток. Если в десяти бросках чётное выпало семь раз, частота равна 0,7. Это не противоречит вероятности 0,5: короткая серия может заметно отклониться от неё.

У вероятности значения от нуля до единицы. Сумма вероятностей всех взаимоисключающих исходов равна единице. Вероятность дополнения события равна единице минус вероятность события. Если шанс дождя по модели 0,3, шанс отсутствия дождя в той же постановке равен 0,7.

Но складывать вероятности произвольных событий без поправки нельзя. На кубике «чётное» и «больше трёх» пересекаются в четырёх и шести. Простая сумма посчитала бы эти исходы дважды. Сначала перечислите исходы объединения: два, четыре, пять и шесть. Его вероятность равна четырём шестым.

> В 20 попытках событие произошло 6 раз. Какова наблюдаемая частота? Ответьте числом от 0 до 1.

Пример

Если события «выпало один» и «выпало два» относятся к одному броску, они несовместны: не произойдут одновременно. Это не означает независимость: знание первого сразу исключает второе.

2. Условие меняет знаменатель

Рассмотрим учебную группу из ста сообщений. Двадцать из них спам. Слово «приз» встретилось в восьми спам-сообщениях и двух обычных. Остальные двенадцать спам-сообщений и семьдесят восемь обычных этого слова не содержат. Числа придуманы для ручного расчёта.

	Есть «приз»	Нет «приз»	Всего
Спам	8	12	20
Обычные	2	78	80

	Есть «приз»	Нет «приз»	Всего
Всего	10	90	100

Вопрос «какова доля спама среди сообщений со словом „приз“?» ограничивает группу десятью сообщениями. Восемь из них спам, поэтому получаем 80 процентов. Это пример **условной вероятности**, оценённой по частотам. Условие «слово присутствует» выбирает новый знаменатель.

Обратный вопрос даёт другое число: «какова доля сообщений со словом „приз“ среди спама?» Теперь группа состоит из двадцати спам-сообщений; восемь содержат слово. Получается 40 процентов. Совпадает числитель, но знаменатель различается. Поменять местами условие и событие без пересчёта нельзя.

В опыте меняется число обычных сообщений с этим словом. Число спам-сообщений со словом остаётся восемью. Когда обычных становится восемь, доля спама среди сообщений со словом падает до половины. Само слово не даёт гарантии класса: его полезность зависит от того, насколько часто оно встречается в обеих группах.

Обозначения

Запись $P(A \mid B)$ читается «вероятность A при условии B». Для частот выбираем только наблюдения, удовлетворяющие B, затем считаем среди них долю A.

Проверьте свой прогноз



Среди сообщений со словом «приз»: спам — 8, обычные — 2. Доля спама: 8 / 10 = 80,0 %. Меняется знаменатель.

Восемь спам-сообщений со словом сохраняются. Меняйте число обычных со словом и следите за долей спама.

Интерактивный эксперимент — в электронной версии урока.

> Из 15 сообщений со словом «скидка» 9 оказались спамом. Какова доля спама среди сообщений с этим словом?

Профи. Формула Байеса: как пересчитать условие в обратную сторону

Условная вероятность по частотам — отношение двух долей: $P(A \mid B) = P(A \cap B) / P(B)$. В таблице из ста сообщений «спам и „приз“» встречается в 8 случаях из 100, «приз» — в 10 из 100,

значит $P(\text{спам} \mid \text{приз}) = 0{,}08/0{,}10 = 0{,}8$. Та же дробь, записанная через другое условие, даёт **формулу Байеса**:

$$P(\text{спам} \mid \text{приз}) = \frac{P(\text{приз} \mid \text{спам})\,P(\text{спам})}{P(\text{приз})}.$$

Проверим: $0{,}4 \cdot 0{,}2/0{,}1 = 0{,}8$. Формула нужна, когда известна «прямая» доля $P(\text{приз} \mid \text{спам})$, а требуется обратная.

Ответ зависит от доли спама в группе. Пусть в другой группе спама 5 %, а слово по-прежнему встречается в 40 % спама и в 2,5 % обычных сообщений. Знаменатель собираем из двух частей: $P(\text{приз}) = 0{,}4 \cdot 0{,}05 + 0{,}025 \cdot 0{,}95 = 0{,}02 + 0{,}02375 = 0{,}04375$. Тогда $P(\text{спам} \mid \text{приз}) = 0{,}02/0{,}04375 \approx 0{,}46$. То же слово, те же прямые доли — а шанс спама упал с 80 % до 46 %.

Типичная ошибка — взять $P(\text{приз} \mid \text{спам}) = 0{,}4$ за ответ на вопрос «спам ли сообщение со словом». Ловится она просто: проверьте, что в знаменателе стоит именно группа, заданная условием.

> В группе из 200 сообщений 36 спам, остальные обычные. Слово «бонус» встречается в половине спам-сообщений и в 5 % обычных. Какова доля спама среди сообщений со словом «бонус»? Ответ с точностью до сотых.

Эпик. **Слово помогает в каждом канале и мешает в сумме**

Фильтр работает в двух каналах: SMS и почта. Подсчёт по каналам:

Канал	Со словом: спам / всего	Без слова: спам / всего
SMS	9 / 10	70 / 100
Почта	10 / 100	1 / 20

В SMS доля спама среди сообщений со словом $0{,}9$ против $0{,}7$ без слова; в почте $0{,}1$ против $0{,}05$. В обоих каналах слово повышает шанс спама. Теперь сложите каналы и посчитайте те же две доли для всех 230 сообщений вместе. Без слова получится $71/120 \approx 0{,}59$, а со словом — заметно меньше. Знак эффекта перевернулся: это **парадокс Симпсона**.

Причина не в арифметике, а в скрытом условии. Слово почти всегда приходит из почты, где спама мало, поэтому «сообщение со словом» в объединённой группе — почти синоним «сообщение из почты». Складывая каналы, мы незаметно сравнили разные группы. Правило «условие выбирает знаменатель» ломается, когда знаменатели составлены из неравных частей.

Перенос: та же ловушка возникает при сравнении двух школ по среднему баллу, если в одной больше старших классов, или двух больниц по смертности, если в одну везут тяжёлых пациентов. Прежде чем складывать, спросите, одинаков ли состав групп.

> В SMS слово встретилось в 10 сообщениях, из них 9 спам. В почте слово встретилось в 100 сообщениях, из них 10 спам. Какова доля спама среди всех 110 сообщений со словом по обоим каналам вместе? Ответ с точностью до сотых.

3. Независимость и ожидание

События независимы, если знание одного не меняет вероятность другого. Для двух отдельных бросков честного кубика знание результата первого не изменяет вероятность чётного во втором. Тогда вероятность совместного события можно получить произведением вероятностей: одна вторая умножить на одну вторую равно одной четверти.

Несовместность устроена иначе: если произошло одно событие, другое в том же опыте невозможно. Два несовместных события с положительными вероятностями не независимы, потому что знание первого уменьшает условную вероятность второго до нуля. Эти слова описывают разные отношения.

Случайная величина присваивает исходу число. Допустим, игра приносит ноль баллов с вероятностью 0,75 и четыре балла с вероятностью 0,25. Её **математическое ожидание** равно сумме значений, умноженных на вероятности: `0 * 0.75 + 4 * 0.25 = 1` . Это среднее по распределению.

Ожидание не обязано быть возможным результатом одной попытки. В этой игре нельзя получить ровно один балл, но при большом числе независимых одинаково устроенных попыток средний выигрыш обычно приближается к единице. Ожидание не обещает выигрыш в конкретной серии и не описывает величину колебаний вокруг среднего.

Заметка

Условную вероятность по частотам нельзя считать, если условная группа пуста. Деление на ноль означает, что для этого вопроса нет наблюдений; это не вероятность ноль.

> Игра даёт 0 баллов с вероятностью 0,5, 2 балла с вероятностью 0,25 и 6 баллов с вероятностью 0,25. Найдите ожидание.

4. Среднее, медиана и разброс

Для наблюдений 1, 1, 3, 3 среднее равно двум. Отклонения от среднего равны −1, −1, 1, 1. Если их просто сложить, положительные и отрицательные части взаимно уничтожатся. Поэтому для измерения разброса удобно возвести отклонения в квадрат и усреднить.

Так получается **описательная дисперсия**: среднее квадратов отклонений от среднего. В нашем наборе каждый квадрат равен единице, значит дисперсия равна единице. **Стандартное отклонение** — квадратный корень из дисперсии; оно возвращает

исходные единицы измерения. Если числа были минутами, дисперсия измеряется в квадратных минутах, стандартное отклонение в минутах.

Набор 2, 2, 2, 2 имеет такое же среднее два, но нулевую дисперсию. Набор 0, 0, 4, 4 тоже имеет среднее два, однако дисперсия равна четырём. Одно среднее не различает эти три ситуации. Для содержательного отчёта полезно рассмотреть и середину, и разброс.

Здесь делим сумму квадратов на число наблюдений. При оценке неизвестной дисперсии по случайной выборке часто используют деление на число наблюдений минус один. Это другая задача и другая формула. У библиотечных функций проверьте, какая нормировка выбрана по умолчанию: одинаковое название не гарантирует одинаковый делитель.

Ссылки
Python: statistics различает `pvariance` с делением на число значений и `variance` с делением на число значений минус один.

Профи. Вторая формула дисперсии и когда ей нельзя верить

Раскроем квадрат отклонения: $(x - m)^2 = x^2 - 2mx + m^2$, где m — среднее. Усредним по всем наблюдениям: среднее x^2 минус $2m \cdot m$ плюс m^2. Получается

$$\mathrm{Var} = \overline{x^2} - m^2,$$

дисперсия равна среднему квадратов минус квадрат среднего. Для 1, 1, 3, 3 среднее квадратов $(1 + 1 + 9 + 9)/4 = 5$, квадрат среднего 4, разность 1 — совпало с расчётом через отклонения. Для 0, 0, 4, 4: $8 - 4 = 4$.

Формула удобна: не нужно сначала находить среднее, а потом второй раз проходить по данным. Но у неё есть ловушка. Если наблюдения — большие числа с маленьким разбросом, например 1000001 и 1000003, оба слагаемых около 10^{12}, а их разность — единица. В арифметике с ограниченной точностью такое вычитание теряет верные цифры, и дисперсия может выйти даже отрицательной. Поймать ошибку просто: дисперсия никогда не меньше нуля, а результаты двух формул на одних данных обязаны совпасть.

Две формулы — один ответ

Цель: Вычислите дисперсию через среднее квадратов и сравните с расчётом через отклонения.

```
values = [2, 4, 4, 4, 5, 5, 7, 9]
mean = sum(values) / len(values)
variance1 = sum((x - mean) ** 2 for x in values) / len(values)
mean_sq = sum(values) / len(values)  # TODO: среднее квадратов
variance2 = mean_sq - mean  # TODO: квадрат среднего
print(f"Через отклонения: {variance1}; через квадраты: {variance2}")
```

Эпик. Что гарантирует дисперсия: оценка худшего случая

Среднее и дисперсия не говорят, как именно устроены данные, но дают гарантию. Пусть у набора среднее m и дисперсия s^2. Какая доля наблюдений может отстоять от среднего на t и больше?

Каждое наблюдение с $|x - m| \ge t$ вносит в среднее квадратов не меньше t^2, остальные — не меньше нуля. Если таких наблюдений доля q, то $s^2 \ge q\,t^2$, откуда

$$q \le \frac{s^2}{t^2}.$$

Это **неравенство Чебышёва** в описательной форме. Оно выполняется для любого набора — распределение знать не нужно.

Пример: страница загружается в среднем за 50 мс при стандартном отклонении 5 мс. Доля загрузок, отклоняющихся на 20 мс и больше, не превосходит $25/400 = 0{,}0625$. Это оценка сверху, а не точное значение: для колоколообразных данных реальная доля много меньше.

В худшем случае оценка точна. Возьмите десять значений: восемь нулей, -2 и 2. Среднее 0, дисперсия $8/10 = 0{,}8$. При $t = 2$ граница даёт $0{,}8/4 = 0{,}2$ — ровно два наблюдения из десяти отстоят на 2. Улучшить неравенство без дополнительных сведений нельзя.

Перенос: детектор аномалий «дальше трёх стандартных отклонений» отмечает не больше $1/9$ наблюдений при любом распределении — и именно поэтому его граница бывает слишком грубой.

> Среднее время ответа сервера 120 мс, стандартное отклонение 8 мс. Какую верхнюю границу доли запросов, отклоняющихся от среднего на 20 мс и больше, даёт неравенство Чебышёва? Ответ долей от 0 до 1 с точностью до сотых.

5. Выборка ограничивает вывод

Представьте опрос о времени в телефоне, проведённый только в чате любителей мобильных игр. Участников может быть много, но способ отбора выделяет определённую группу. Увеличение числа ответов из того же чата само по себе не превращает результат в описание всех школьников.

Случайные колебания и смещение отбора требуют разных решений. Больше независимых наблюдений может уменьшить случайную неопределённость среднего, но не исправляет систематическое исключение части людей. В отчёте укажите, кого и как отбирали, кто мог не ответить и на какую группу переносите вывод.

В лаборатории посчитайте описательную дисперсию. Среднее уже вычислено. Исправьте сумму отклонений так, чтобы каждое отклонение сначала возводилось в квадрат. Затем проверьте постоянный набор: у него должно получиться ноль. Пустой набор потребовал бы отдельной обработки, потому что среднее и деление на количество там не определены.

Посчитать разброс

Цель: Вычислите среднее квадратов отклонений от среднего.

```
values = [1, 1, 3, 3]
mean = sum(values) / len(values)
variance = sum(x - mean for x in values) / len(values)  # TODO: квадраты
print(f"Среднее: {mean}; дисперсия: {variance}")
```

Практическая проверка вывода начинается не с формулы, а с вопроса о группе. К кому относится знаменатель? Какие исходы возможны? Что произошло с отсутствующими ответами? После этого вероятность, ожидание и дисперсия становятся точными инструментами описания, а не числами без ясного смысла.

Где это встретится дальше

Сначала определите исходы и группу наблюдений. Вероятность, среднее и разброс отвечают на разные вопросы об этой группе.

Вычислить частоту и условную вероятность по числам наблюдений.

Найти ожидание и описательную дисперсию.

Объяснить, почему способ отбора ограничивает вывод.

Домашняя работа

1. Найдите ожидание: 1 с вероятностью 0,6 и 6 с вероятностью 0,4.
2. Найдите описательную дисперсию чисел 2 и 6.
3. ★ Можно ли по ста ответам друзей оценить мнение всего города? Назовите ограничение.

Источники

- OpenStax: вводная статистика (https://openstax.org/details/books/introductory-statistics-2e).
- NIST: EDA (https://www.itl.nist.gov/div898/handbook/eda/eda.htm).
- Python: statistics (https://docs.python.org/3/library/statistics.html).

История

Вероятностные методы развивались вместе с задачами игр и наблюдений; современная статистика различает модель случайности и способ получения выборки. OpenStax: вводная статистика соединяет эти темы с примерами.

Подумай

Если опросили всех участников одного игрового чата, ошибки выборки относительно этого чата нет. Но перенос результата на всех школьников всё ещё требует обоснования.