

1



Рис. 5.6. Коэффициент корреляции между площадью дома и ценой продажи составляет 0,62 (определяется степенью близости точек данных к линии тренда)

Корреляция

Корреляция — это мера связанности двух переменных. Наиболее распространенный коэффициент корреляции в сфере бизнеса — коэффициент корреляции Пирсона. Он принимает значения в диапазоне от -1 до 1 и измеряет степень линейной зависимости (простая прямая линия) между парами чисел, отображаемыми на диаграмме рассеяния. Корреляция может быть положительной, когда увеличение одной переменной сопровождается увеличением другой: большие дома продаются за большие деньги. Корреляция также может быть отрицательной: более тяжелые автомобили менее экономичны в плане расхода топлива. Коэффициент корреляции между размером дома и ценой продажи составляет 0,62 (рис. 5.6). Чем ближе точки к линии тренда, тем выше степень корреляции⁴¹.

2

ОБНАРУЖИЛИ ЛИ ВЫ НОВЫЕ ВОЗМОЖНОСТИ В ДАННЫХ?

Разведочный анализ данных — это не просто процесс, позволяющий лучше разобраться в данных и наметить путь решения стоящих перед нами проблем. Это еще и шанс найти дополнительные возможности в этих данных, которые могут оказаться ценными для вашей организации. Дата-сайентист может обнаружить что-то интересное или странное в наборе данных и сформулировать проблему.

⁴⁵ Fisher, R. A. (1958). Cancer and smoking. *Nature*, 182 (4635), 596.



Чтобы стать главным по данным, вам необходимо постоянно заниматься разведочным анализом данных. Это позволит вам:

3

- Наметить более четкий путь решения проблемы.
- Уточнить исходную бизнес-задачу с учетом выявленных в данных ограничений.
- Сформулировать новые проблемы, которые можно решить с помощью этих данных.
- Отменить проект. Хотя это не приносит удовлетворения, EDA считается успешным, если он предотвращает трату времени и денег на решение тупиковой проблемы.



Вероятность наступления множества событий

4 При моделировании вероятности наступления множества событий нотация и правила зависят от того, происходят ли они одновременно (наводнение и отключение электричества) или происходит только одно из них (или наводнение, или отключение электричества).

Одновременное наступление двух событий

Сначала поговорим о двух событиях, наступающих одновременно.

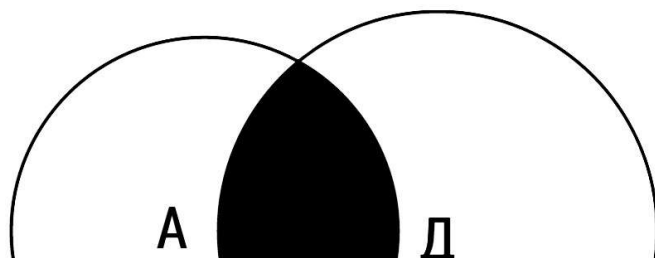
$P(\text{выпадения орла при подбрасывании монеты}) = P(O) = 1/2.$

$P(\text{выбора карты пиковой масти из колоды карт}) = P(\Pi) = 13/52 = 1/4.$

Наступление одного или другого события

Что, если наступает одно или другое событие? Статистика и теория вероятностей учит нас тому, что все зависит от обстоятельств. Начните с предположения и корректируйте его, опираясь на имеющуюся информацию.

Когда два события не могут произойти одновременно, все сводится к простому сложению вероятностей. При бросании кубика не может одновременно выпасть 1 и 2, поэтому вероятность выпадения 1 или 2 равна $P(K = 1 \text{ или } K = 2) = P(K = 1) + P(K = 2) = 1/6 + 1/6 = 2/6 = 1/3$.

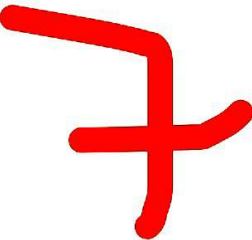




БУДЬТЕ ОСТОРОЖНЫ, ДЕЛАЯ ПРЕДПОЛОЖЕНИЯ О НЕЗАВИСИМОСТИ СОБЫТИЙ

Если события не зависят друг от друга, вы можете перемножить вероятности их наступления. Например, вероятность выпадения двух орлов подряд при подбрасывании честной монеты составляет $P(O) \times P(O) = 1/2 \times 1/2 = 1/4$. Однако не все события являются независимыми, поэтому с осторожностью делайте соответствующее предположение при вычислении или анализе вероятностей.

Не допускайте ошибку игрока




С другой стороны, некоторые события являются независимыми, но не воспринимаются таковыми. Это порождает другой вид риска, благодаря которому процветают казино. В данном случае люди переоценивают вероятность наступления того или иного события, основываясь на предшествующих событиях.

Если при подбрасывании честной монеты 10 раз подряд выпадет орел, то вероятность выпадения орла в результате следующего броска все равно будет составлять $P(O) = 50\%$. В случае с независимыми событиями вероятность наступления одного из них не увеличивается и не уменьшается в зависимости от предыдущих результатов. Однако игроки ошибочно полагают, что величина вероятности меняется — отсюда и название «ошибка игрока»⁵¹.



ВСЕ ВЕРОЯТНОСТИ ЯВЛЯЮТСЯ УСЛОВНЫМИ



Все вероятности в некотором смысле условны. Вероятность выпадения орла при подбрасывании монеты $P(O)$ равна 50% при условии, что монета является честной. То же самое касается вероятности выпадения единицы при бросании кубика: $P(K == 1) = 1/6$. Вероятность успеха проекта по работе с данными зависит от коллективного разума группы аналитиков, правильности данных, сложности проблемы, отсутствия вирусов на компьютерах, риска закрытия компании из-за пандемии и так далее.



Теорема Байеса

Теорема Байеса, сформулированная в XVIII веке, — это способ работы с условными вероятностями, который применяется повсюду, начиная с планирования сражений и управления финансами и заканчивая расшифровкой ДНК⁵⁵. Для двух событий A и B теорема Байеса утверждает следующее:

$$P(A | B) \times P(B) = P(B | A) \times P(A)$$

Пусть вас не пугает эта формула. Самое важное — не запомнить ту или иную формулу, а понять, что она делает и почему о ней стоит знать.

Теорема Байеса позволяет связать условную вероятность двух событий. Вероятность наступления события A при условии наступления события B связана с вероятностью наступления события B при условии наступления события A . Они не равны, но связаны приведенным выше уравнением.

Это может пригодиться, когда вам известна одна из условных вероятностей и вы хотите определить другую. Например:

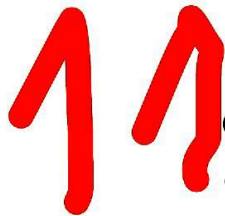
- Медицинские исследователи хотят знать вероятность того, что у человека будет положительный результат скринингового теста на рак при условии, что этот человек болен раком, $P(+ | P)$. Тогда они смогут создать более точные тесты, позволяющие немедленно приступить к лечению. Разработчики политики хотят знать обратное — вероятность того, что человек болен раком при условии положительного результата скринингового теста, $P(P | +)$, потому что они не хотят подвергать людей ненужному лечению на основании ложноположительного результата (когда тест показывает наличие болезни при ее отсутствии).
- Прокуроры хотят знать вероятность того, что подсудимый виновен при условии наличия доказательств, $P(B | Д)$. Это зависит от вероятности обнаружения доказательств при условии, что человек виновен, $P(Д | В)$.

Калибровка

Определяемые вероятности должны иметь смысл.

Например, при условии равных затрат и выгод проект с вероятностью успеха 60% сопряжен с большим риском, чем проект с вероятностью успеха 75%.

Мы знаем, что это кажется очевидным, однако люди часто оценивают события с вероятностью 60% или 75% как весьма вероятные, потому что их вероятность превышает 50%. Но если бы это было так, вероятности не имели бы никакого значения и сводились бы к бинарным решениям типа «да/нет», при которых полностью утрачивается смысл статистического мышления и работы с неопределенностью.



Плохая калибровка делает невозможной точную оценку риска. Если вы самоуверенный юрист, который думает, что выиграет дело с вероятностью 90%, хотя до этого выигрывал только в 60% случаев, вы переоцениваете свои шансы на успех. Это пример плохой калибровки.

Итак, вероятности должны иметь смысл. Помните о том, что редкие события не являются невозможными, а высоковероятные события не обязательно наступают.

Редкие события могут случаться и случаются

12

Редкое событие может не произойти с вами или с кем-либо из ваших знакомых, но это не значит, что оно не произойдет вообще. Тем не менее у нас часто возникают сложности с пониманием редких событий.

Это правда: вы вряд ли сорвете джекпот в лотерею, однако некоторые люди в нее все-таки выигрывают. Если учесть количество лотерей, проводимых по всему миру каждый день, вероятность того, что столь редкое событие произойдет с кем-то из жителей планеты, пусть даже не с вами, оказывается не такой уж и низкой.

13

Не перемножайте вероятности без необходимости

Не перемножайте вероятности прошлых событий без особой необходимости. В противном случае вы можете сделать то или иное событие практически невероятным.

Давайте прикинем вероятность того, что вы читаете именно эту строку на этой странице этой самой книги. Помимо данной строки на этой странице еще примерно 35 строк ($1/35$), в книге — еще 300 страниц ($1/300$), а в мире — миллионы книг. Если вы перемножите эти вероятности, то получите бесконечно малое число. Очевидно, мы были созданы друг для друга!

14 Итак, какова вероятность того, что игрок, реализующий 50% своих бросков, промахнется 10 раз подряд?

Отталкиваясь от базовой вероятности, вы выполняете некоторые расчеты. Вероятность того, что он промахнется один раз, составляет 50%. Вероятность двух промахов подряд составляет $50\% \times 50\% = 25\%$ (при условии, что результаты бросков не зависят друг от друга, как говорилось в предыдущей главе). Продолжая эту логику, вы умножаете показатель 50% сам на себя 10 раз: $0,5^{10} = 0,00098$, то есть 0,1%, или примерно 1 из 1000.

15

Таким образом, вероятность данного конкретного результата, то есть 10 промахов подряд, при условии, что стажер, по его словам, способен реализовать 50% бросков, составляет 1 из 1000.

Эта вероятность, равная 1 из 1000 или 0,001, называется p -значением (p означает probability — «вероятность»). Теперь вы должны решить, был ли у стажера просто неудачный день или ваша нулевая гипотеза, согласно которой процент реализации бросков стажера составляет 50%, ошибочна?

Десять пропущенных бросков лишь подрывают доверие. Однако то, что вероятность неудачного дня составляет 1 из 1000, довольно убедительно доказывает то, что первоначальное утверждение стажера вряд ли было истинным. Скорее всего, вы отвергли нулевую гипотезу на более ранних этапах игры в пользу альтернативной гипотезы, H_a : процент реализации бросков стажера $< 50\%$.

16

Поскольку Вселенная полна вариаций, вы должны смириться с некоторым уровнем случайности (и количеством промахов). Иногда человек может плохо играть без всяких причин. Таким образом, уровень значимости — это некий условный установленный вами предел, до которого вы можете мириться со случайностью и необъяснимыми вариациями, продолжая считать нулевую гипотезу верной. Если p -значение меньше уровня значимости, вы отбрасываете нулевую гипотезу и говорите, что результат статистически значим.

Урок: проверка того, что p -значение не превышает уровня значимости, с целью отбрасывания нулевой гипотезы — ключевая часть процесса построения статистического вывода. Разумеется, наличие вариаций и произвольный выбор уровня значимости чреваты ошибками при принятии решений.

17

ПРОЦЕСС ПОСТРОЕНИЯ СТАТИСТИЧЕСКОГО ВЫВОДА

В предыдущих пяти кратких уроках мы рассмотрели несколько компонентов процесса статистического вывода. Пришло время понять, как эти компоненты сочетаются друг с другом. Давайте попробуем обобщить их, чтобы вы как главный по данным могли понять и четко объяснить весь процесс построения статистического вывода.

Если вкратце, то в ходе этого процесса вы должны выполнить следующие действия:

1. Задайте осмысленный вопрос.
2. Сформулируйте гипотезы для проверки, используя статус-кво в качестве нулевой гипотезы, а свое предположение — в качестве альтернативной.
3. Задайте уровень значимости. (Чаще всего используется произвольное значение в 5% или 0,05.)
4. Вычислите p -значение на основе результата статистического теста.
5. Вычислите соответствующие доверительные интервалы.
6. Отклоните нулевую гипотезу в пользу альтернативной, если p -значение оказалось меньше уровня значимости; в противном случае не отклоняйте нулевую гипотезу.