Глава 1

Что такое не везёт и как это рассчитывать: p-values

Статистическая значимость

Словосочетание «статистическая значимость» (или его психологического доппельгангера, «достоверность»), наверное, слышали все. Медицина, генетика, опросы про зубную пасту и против зубной пасты - весь этот поток информации без статистики и без сравнения гипотез (вот для чего нужна значимость) превращается в хаос разнообразных наборов данных.

Слово «статистика» имеет несколько значений. К более техническому, важному для нас, мы вернёмся позже, а сейчас поговорим немного об общенаучном. Статистика и теория вероятности - это связанные способы исследования закономерностей мира. Теория вероятностей предсказывает события, исходя из моделей, и так пытается понять то, что мы видим вокруг. Монетка, игральная кость, нормальное распределение - всё это модели случайных переменных, а исходы этих переменных - это те события, которые мы можем увидеть - орёл, четыре, червяк длиной 4 сантиметра. Статистика же делает для теории вероятности черновую, обратную работу. По наблюдениям оцениваются параметры модели (об этом мы много говорили об этом в предыдущей главе), и проверяются гипотезы о пригодности модели для описания наблюдений.

Сравнение гипотез

Если мы одновременно думали о нескольких гипотезах (их может быть сколько угодно, но двух взаимоисключающих гипотез $B \equiv !A$ достаточно, чтоб понять, что происходит), и мы только что провели новое наблюдение (поставили опыт, взломали сайт, посчитали попугаев), то наше доверие к каждой из этих гипотез претерпело некоторые изменения от столкновения

с реальностью, представленной результатом наблюдения.

$$P(\mathrm{A}|\mathrm{obs}) = \frac{P(\mathrm{obs}|\mathrm{A})\,P(\mathrm{A})}{P(\mathrm{obs})} = \\ = \frac{P(\mathrm{obs}|\mathrm{A})\,P(\mathrm{A})}{P(\mathrm{obs}|\mathrm{A})\,P(\mathrm{A}) + P(\mathrm{obs}|\mathrm{B})\,P(\mathrm{B})} \tag{1.1}$$

$$P(\mathrm{B}|\mathrm{obs}) = \frac{P(\mathrm{obs}|\mathrm{B})\,P(\mathrm{B})}{P(\mathrm{obs})} = \\ = \frac{P(\mathrm{obs}|\mathrm{B})\,P(\mathrm{B})}{P(\mathrm{obs}|\mathrm{A})\,P(\mathrm{A}) + P(\mathrm{obs}|\mathrm{B})\,P(\mathrm{B})} \tag{1.2}$$

$P(\mathrm{A}|\mathrm{obs})$ – вероятность того, что A верна, при условии того (то есть после того), что получен результат наблюдения obs; $P(\mathrm{A})$ - это априорная (до наблюдения) вероятность того, что A верна, $P(\mathrm{obs}|\mathrm{A})$ - условная вероятность (правдоподобие) получить результат obs, если A верна.

Одинаковый знаменатель в формулах (1.1) и (1.2) – это вероятность наблюдения $P(\mathrm{obs})$ (в англоязычной литературе она называется evidence). Часто при сравнении гипотез её вообще опускают, и из (1.1) и (1.2) получается:

$$\frac{P(\mathrm{A}|\mathrm{obs})}{P(\mathrm{B}|\mathrm{obs})} = \frac{P(\mathrm{obs}|\mathrm{A})\,P(\mathrm{A})}{P(\mathrm{obs}|\mathrm{B})\,P(\mathrm{B})} \tag{1.3}$$

.

Или просто пишут

$$P(\mathrm{A}|\mathrm{obs}) \propto P(\mathrm{obs}|\mathrm{A})\,P(\mathrm{A}) \tag{1.4}$$

В каждом байесовском выражении у всех вероятностей справа после «|», мы должны были бы написать длинную цепочку утверждений, при условии которых это выражение имеет смысл. Тут будет такие предположения, как наблюдаемость мира, работоспособность приборов, правильность постановки эксперимента. Для формул (1.1) и (1.2) в этот набор входит это ещё и предположение, что верна одна и только одна из гипотез. Обычно мы все эти условия не пишем, потому что договорились считать это очевидным, но иногда, когда модели вложены или параметризованы, явная запись части условий необходима для того, чтобы формулы заработали.

Кстати, полезно иногда думать о том, что справа от вертикальной черты, как о контексте сообщения, а о том, что слева – как о сообщении.

Пусть у модели, описывающей реальность в предположении о верности гипотезы A (чтобы не плодить обозначения, будем называть модель и её гипотезу одной и той же буквой), есть параметр α. Считаем его апостериорное распределение после наблюдения obs. Это распределение, как и сам параметр α, имеет смысл, только если A верна,

$$P(\alpha|\mathrm{obs},\mathrm{A}) = \frac{P(\mathrm{obs}|\alpha,\mathrm{A})\,P(\alpha|\mathrm{A})}{P(\mathrm{obs}|\mathrm{A})} \tag{1.5}$$

$P(\text{obs}|\text{A})$ - это evidence в (1.5), он считается как статистическая сумма по всем возможным значениям α - и эта же условная вероятность - $P(\text{obs}|\text{A})$ - это значение правдоподобия гипотезы A в (1.1). Получается, что когда мы оцениваем апостериорные распределения параметра при условии, что сама гипотеза верна, evidence – это оценка адекватности гипотезы (модели) наблюдениям. Значение evidence позволяет сравнивать гипотезы (подробнее см. Skilling, 2006).

Нулевые и альтернативные гипотезы

Семейство моделей, о которых мы говорим (вернее, с пониманием молчим), когда заходит речь о статистической значимости или о p-value – это модели, соответствующие нулевой гипотезе (Null Hypothesis). Конкретное содержание нулевой гипотезы зависит от контекста наблюдения, но общий смысл всегда один и тот же - всё плохо. Эта оптимистичная мысль объединяет собой все нулевые гипотезы. Лекарство работает так же, как плацебо, преступность не отличается между двумя городами, ген одинаково экспрессируется в разных условия, носители разных аллелей одного локуса одинаково часто болеют чем попало – всё это примеры нулевых гипотез.

Если нулевая гипотеза верна, то в эксперименте, мы, конечно, всё равно не увидим идеального сходства условий, идеального совпадения экспрессии генов в разных группах и прочих идеальных вариантов выполнения предположения нулевой гипотезы – мы получим результат, порождённый шумом. Если же нулевая гипотеза не верна, то мы будем наблюдать некий содержательный сигнал, опять-таки искажённый шумом. Для того, чтобы на основании наблюдения (наблюдений), понять, насколько близка к истине нулевая гипотеза NULL по сравнению с альтернативной (ненулевой) !NULL, можно использовать формулы условной вероятности (1.1) - (1.2).

$$P(\text{NULL}|\text{obs}) = \frac{P(\text{obs}|\text{NULL})\,P(\text{NULL})}{P(\text{obs})} = \\ = \frac{P(\text{obs}|\text{NULL})\,P(\text{NULL})}{P(\text{obs}|\text{NULL})\,P(\text{NULL}) + P(\text{obs}|!\text{NULL})\,P(!\text{NULL})} \tag{1.6}$$

P-value

На самом деле, (1.6) использует редко: для этого нужно уметь оценивать распределение экспериментальных результатов не только для нулевой гипотезы, но и для альтернативной, а это требует, как минимум, построения модели содержательного сигнала. Для того, чтобы можно было сказать хоть что-нибудь о состоятельности нулевой гипотезы, не зная ничего про альтернативные, используют оценку, в чём-то родственную Байесовской (об этом ниже), но сильно упрощённую. Поскольку сравнивать нулевую гипотезу с

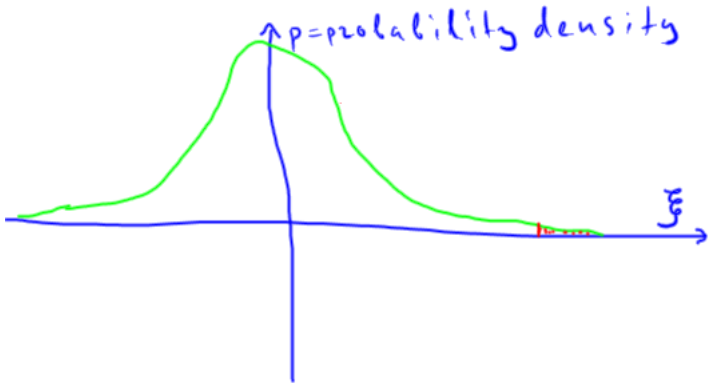

Рис. 1.1: Площадь красного сегмента графика плотности вероятности случайной величины ξ - это p-value, соответствующее значению ξ на границе сегмента

альтернативными мы не можем, то единственная мера адекватности нулевой гипотезы наблюдениям - это вероятность наблюдений при условии того, что нуль-гипотеза верна. Её априорную вероятность при этом вообще не учитывают, как в анекдоте о том, что вероятность того, что за углом стоит тигр, равна $1/2$ - действительно, он же там или стоит, или нет. Следующая трудность связана с тем, что вероятность одного значения непрерывной случайной величины - это бесконечно малая величина. Она имеет смысл либо если нам нужно отношение таких вероятностей при разных предположениях, либо если мы интегрируем её на каком-то интервале. Пока в баейсовской формуле мы считали отношения, всё было в порядке, теперь надо задуматься от интервалах.

Самый понятный интервал - это от столба и до обеда. Его и используют при оценке $p - value$: по определению, это вероятность наблюдать тот результат, который был получен, или более маргинальный результат, при условии того, что нулевая гипотеза верна. Очевидно, чтобы оценить p-value по наблюдению, надо знать распределение случайной величины, которая описывает результат нашего наблюдения при верной нулевой гипотезе. Но этого мало: надо ещё понимать, что значит «более маргинальный» применительно к нашим наблюдением.

Литература

Skilling, J. (2006). Nested sampling for general bayesian computation. *Bayesian Analysis*, 1(4):833–859. Publisher: International Society for Bayesian Analysis.

Куски про ошибки первого рода

Alexander Favorov, [16.04.21 00:14] вообще ошибка первого рода возникает в тот момент, когда мы её делаем

Alexander Favorov, [16.04.21 00:15] В смысле p-value это просто p-value

Alexander Favorov, [16.04.21 00:16] а дальше если нам пришло в голову почему-то (например, посмотрев на p-value или на звёзды) сказать, что этот эксперимент отверг нулевую гипотезу - вот тут уже мы что - то сделали, ошиблись ли? вероятность этого это вероятность ошибки первого рода

Alexander Favorov, [16.04.21 00:16] p-val это вообще не про ошибку

Alexander Favorov, [16.04.21 00:17] это вер-ть получить такое или ещё страннее если нуль-гипотеза верна

Alexander Favorov, [16.04.21 00:19] При этом не p-val это не p(obs | H0) а интеграл от obs до предельной кривизны наблюдения p(x | H0)

Alexander Favorov, [16.04.21 00:19] где x - воображаемое другое наблюдение

Sergey Isaev, [16.04.21 00:19] ага понял

Alexander Favorov, [16.04.21 00:19] Ну, переменная интегрирования она же

Sergey Isaev, [16.04.21 00:19] дада

Sergey Isaev, [16.04.21 00:20] Меня вот что смущает, что вот допустим у меня есть два эксперимента, да, и что я их обоих сравниваю с H0, да

Sergey Isaev, [16.04.21 00:20] И что в одном случае у меня p = 0.001, а в другом p = 0.0000000001

Sergey Isaev, [16.04.21 00:21] И что во втором случае как будто бы я менее вероятно проебался, чем в первом, когда отвергнул H0

Alexander Favorov, [16.04.21 00:24] приходит мужик и приносит тест на беременность с двумя полосками :):)

Sergey Isaev, [16.04.21 00:24] ну или он приносит 20 тестов на беременность с двумя полосками

Sergey Isaev, [16.04.21 00:25] ну типа во втором случае я как будто бы больше поверю, что он беременный

Alexander Favorov, [16.04.21 00:25] ну да

Alexander Favorov, [16.04.21 00:25] но это одна группа тестов примерно

Alexander Favorov, [16.04.21 00:25] но если ты его сравниваешь с барышней, у которое тоже две полоски, но в одном экземпляре и видно плохо, ты скорее поверишь что барышня

Alexander Favorov, [16.04.21 00:26] хотя у мужика 20 и видны за километр без очков