

# Глава 1

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

### Статистическая значимость

Словосочетание «статистическая значимость» (или его психологического допсельгангера, «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.

$$\begin{aligned} P(A|\text{obs}) &= \frac{P(\text{obs}|A) P(A)}{P(\text{obs})} = \\ &= \frac{P(\text{obs}|A) P(A)}{P(\text{obs}|A) P(A) + P(\text{obs}|B) P(B)} \end{aligned} \quad (1.1)$$

$$\begin{aligned} P(B|\text{obs}) &= \frac{P(\text{obs}|B) P(B)}{P(\text{obs})} = \\ &= \frac{P(\text{obs}|B) P(B)}{P(\text{obs}|A) P(A) + P(\text{obs}|B) P(B)} \end{aligned} \quad (1.2)$$

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

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

$$\frac{P(A|\text{obs})}{P(B|\text{obs})} = \frac{P(\text{obs}|A) P(A)}{P(\text{obs}|B) P(B)} \quad (1.3)$$

Или просто пишут

$$P(A|\text{obs}) \propto P(\text{obs}|A) P(A) \quad (1.4)$$

Значение *evidence* несёт важную информацию о наборе моделей (гипотез), или иными словами, о предположениях, в которых мы их сравниваем (подробнее см. Skilling, 2006). В каждом байесовском выражении у всех вероятностей справа после  $|$  мы можем написать длинную цепочку условий, при которых это выражение имеет смысл. Тут будут такие предположения, как наблюдаемость мира, работоспособность приборов, правильность постановки эксперимента. Для (1.1) – (1.2) это ещё и предположение, что верна одна и только одна из гипотез. Обычно мы все эти условия не пишем, как очевидные, но иногда, когда модели вложены или параметризованы, явная запись части условий необходима для того, чтобы формулы заработали.

Пусть у модели  $A$  есть параметр  $\alpha$ . Считаем его апостериорное распределение после наблюдения  $\text{obs}$ . Запись этого распределение имеет смысл, только если  $A$  верна,

$$P(\alpha|\text{obs}, A) = \frac{P(\text{obs}|\alpha, A) P(\alpha|A)}{P(\text{obs}|A)} \quad (1.5)$$

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

## 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 условиях, 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).

$$\begin{aligned} P(\text{NULL}|\text{obs}) &= \frac{P(\text{obs}|\text{NULL}) P(\text{NULL})}{P(\text{obs})} = \\ &= \frac{P(\text{obs}|\text{NULL}) P(\text{NULL})}{P(\text{obs}|\text{NULL}) P(\text{NULL}) + P(\text{obs}|\text{!NULL}) P(\text{!NULL})} \end{aligned} \quad (1.6)$$

## 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 альтернативными мы не можем, то единственная мера адекватности нулевой гипотезы наблюдениям - это вероятность наблюдений при условии то-

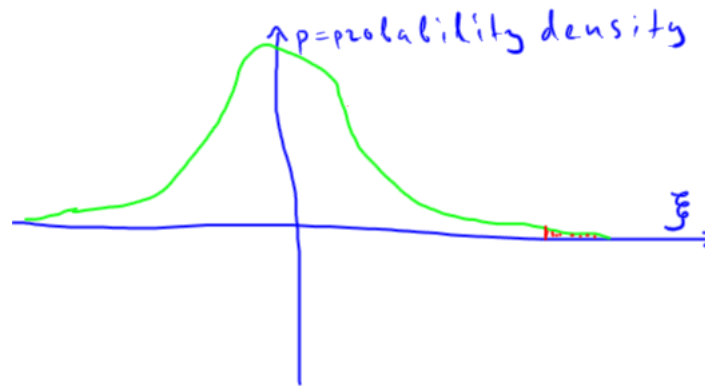


Рис. 1.1: Площадь красного сегмента графика плотности вероятности случайной величины  $\xi$  - это p-value, соответствующее значению  $\xi$  на границе сегмента

го, 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 байесовской 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. 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 наблюдениям.

# Литература

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