Опубликовано в «Эпсилоне»

p-value: то, что вы всегда хотели узнать, но боялись спросить

Мария Лысюк

Аннотация. Аннотация должна передавать краткое содержание работы. Она должна быть ясной, содержательной, релевантной и короткой (не более 150 слов). Аннотация должна содержать информацию, необходимую для поиска по базам научных работ. В аннотации не должно быть математических формул.

Ключевые слова: p-value, уровень значимости, гипотезы, интерпретация.

С завидным постоянством хотя бы раз в жизни студент, слушающий курс статистики, сталкивается с вопросом экзаменатора (который обычно ещё надеется, что вопрос очевиден и «вытягивает» с помощью него студента): «Мистер X, что показывает p-value?»

И тут для многих наступает этот неловкий момент, и лицо выглядит примерно вот так (Эдвард Мунк, видимо, тоже не знал):

Рис. 1. Лицо обычного человека, у которого спросили, что такое p-value

Для того чтобы осознать сие, безусловно, великое понятие, мы должны, как Будда, пройти 7 ступеней познания. Как водится, примеры красноречивее всего доносят нужную информацию до мозга, так что поговорим сегодня про машинки.

Вкратце *о ходе эксперимента*. Мы будем узнавать, существует ли какая-либо зависимость между штрафом за лихачество водителя и цветом его машины. Гипотеза $\mathcal{H}_0$ будет выглядеть следующим образом.

НИУ ВШЭ, Москва.

- $\mathcal{H}_0$: Выдача штрафа не зависит от цвета машины.
- $\mathcal{H}_1$: Водители с красными машинами чаще получают штрафы за превышение скорости по сравнению с синими машинами.

Итак, в добрый путь!

1 Семь ступеней познания p-value

Ступень 1. Выберите уровень значимости. Начнём со знакомого до боли. Строго говоря, уровень значимости — это мера, которая отражает наше предпочтение точности результатов: низкие уровни значимости говорят о маленькой вероятности того, что полученные экспериментальным путём результаты случайны, и наоборот. Согласно негласной конвенции, обычно используется 5%-й уровень значимости. Это означает, что вероятность того, что наши результаты случайны, равна 0,05, а вероятность того, что мы сами повлияли на результат, равна 0,95.

- *Пример.* Возьмём и мы уровень значимости в 5 %.

Ступень 2. Определите ожидаемые результаты эксперимента. Как правило, учёные, проводя эксперимент и наблюдая впоследствии результаты, имеют представление о том, какие результаты являются «типичными» до начала эксперимента. Это может быть основано на результатах из прошлых исследований, достоверных источников, научной литературы и т. д. Для вашего эксперимента определите ваши ожидаемые результаты любым из способов.

- *Пример.* Пусть предыдущие исследования показали, что штрафы за превышение скорости чаще получают водители красных машин по сравнению с синими. Также пусть результаты по всей стране показывают превышение красными в отношении 2 : 1 по сравнению с синими. Мы же хотим узнать, применимы ли результаты, характерные для всей страны, к нашему городу. Если мы возьмём случайную выборку из 150 машинок, которым выписали штрафы, мы будем ожидать, что 100 машин будут красными, а 50 — синими, *если наша полиция выписывает штрафы согласно национальной тенденции.*

Красная машинка	Синяя машинка
100	50

Рис. 2. Ожидаемые значения количества штрафов

Ступень 3. Определите наблюдаемые результаты эксперимента. После того как мы определили ожидаемые результаты, проводим реальный

эксперимент и получаем наблюдаемые результаты. Если мы каким-либо образом повлияли и наблюдаемые результаты отличаются от ожидаемых, возможна одна из двух ситуаций:

1. Это произошло случайно.
2. Те условия, в которых мы проводили эксперимент, *повлияли* на исход.

Как правило, цель нахождения p-value — определить, правда ли, что наблюдаемые результаты отличаются от ожидаемых настолько, что мы не можем отвергнуть нулевую гипотезу (гипотезу о том, что нет связи между переменными и наблюдаемым результатом).

Прим. ред. Что значит «определить наблюдаемые результаты»? Это как?

- *Пример.* Пусть в нашем городе мы произвольно выбрали 150 красных и синих машин нарушителей. Оказалось, что 90 штрафов выписали красным машинам, а 60 — синим. Это отличается от ожидаемых 100 и 50 соответственно. Правда ли, что те условия, в которых мы проводили эксперимент (в нашем случае смена источника данных с национальных на местные) послужила причиной изменения результатов, или действия городской полиции так же смещены, как и предсказывает национальная средняя оценка, и мы просто наблюдаем случайную вариацию? p-значение спешит на помощь!

Красная машинка	Синяя машинка
90	60

Рис. 3. Наблюдаемые количества штрафов

Ступень 4. Определите степени свободы в вашем эксперименте. Степени свободы отражают меру изменчивости, характерную для исследования, которая определяется количеством переменных, которые вы изучаете. Степени свободы определяются как $n-1$, где n — это количество переменных, используемых в эксперименте.

Прим. ред. Что это за misdirection? Степени свободы чего? В регрессии для Residual S. S. это, например, количество наблюдений минус количество параметров. Не лучше ли дать более общее и понятное определение?

- *Пример.* У нас есть две переменные: количество красных машин и количество синих машин. Поэтому степеней свободы всего $2-1=1$, т. е. одна.

Ступень 5. Сравните наблюдаемые результаты с ожидаемыми с помощью распределения χ^2. χ^2 — статистика, численно измеряющая

разницу между ожидаемыми и наблюдаемыми результатами. Уравнение:

$$\chi^2 = \sum_{i=0}^{n} \frac{(h_i - e_i)^2}{e_i},$$

где h — значение наблюдаемой переменной, а e — ожидаемой.

- *Пример.* Мы должны просуммировать значения для всех возможных переменных, то есть в нашем случае для синих и красных машинок:

$$\chi^2 = \sum_{i=0}^{1} \frac{(h_i - e_i)^2}{e_i} = \frac{(90-100)^2}{100} + \frac{(60-50)^2}{50} = \frac{(-10)^2}{100} + \frac{10^2}{50} = 1 + 2 = \boxed{3}.$$

Ступень 6. Используем таблицу χ2-распределения, чтобы аппроксимировать p-value. Скрестила пальцы: надеюсь, что все умеют пользоваться таблицами распределений.

- *Пример.* Наше значение статистики χ^2 равно 3. Далее пользуемся таблицей 1 для нахождения p-значения. У нас одна степень свободы (degree of freedom), поэтому берём первую строку и ищем там первое значение, превышающее значение нашего $\chi^2 = 3$. Оно равно 3,84. Соответствующее p-значение равно 0,05. Это означает, что наше p-value располагается между 0,05 и 0,1.

	p-value								
df	20%	10%	5%	2,5%	1%	0,5%	0,25%	0,1%	0,05%
1	1,64	2,71	3,84	5,02	6,63	7,88	9,14	10,83	12,12
2	3,22	4,61	5,99	7,38	9,21	10,60	11,98	13,82	15,20
3	4,64	6,25	7,81	9,35	11,34	12,84	14,32	16,27	17,73
4	5,99	7,78	9,49	11,14	13,28	14,86	16,42	18,47	20,00
5	7,29	9,24	11,07	12,83	15,09	16,75	18,39	20,52	22,11
10	13,44	15,99	18,31	20,48	23,21	25,19	27,11	29,59	31,42
20	25,04	28,41	31,41	34,17	37,57	40,00	42,34	45,31	47,50
30	36,25	40,26	43,77	46,98	50,89	53,67	56,33	59,70	62,16
40	47,27	51,81	55,76	59,34	63,69	66,77	69,70	73,40	76,09
50	58,16	63,17	67,50	71,42	76,15	79,49	82,66	86,66	89,56

Таблица 1. Критические статистики для распределения χ^2

Ступень 7. Вот мы и добрались до конца! Осталось решить, отвергается или нет нулевая гипотеза. Если p-value меньше, чем уровень значимости, то мои поздравления, можете отсылать вашу работу в топовые журналы! Вы доказали, что высока вероятность того, что есть значимая корреляция между переменными, которыми вы манипулируете, и наблюдаемыми результатами. Если ли же p-значение больше выбранного уровня значимости, вы не можете с точностью сказать, случайны ли полученные вами результаты, или они являются результатом ваших действий.

- *Пример.* Наше p-значение находится в границах от 0,05 до 0,1. Это определённо меньше, чем выбранный уровень значимости, равный 0,05, поэтому, к сожалению, мы не можем отвергнуть нулевую гипотезу. Другими словами, мы не достигли желаемого уровня в 95 %, чтобы с точностью сказать, что в нашем городе полиция выдаёт штрафы красным и синим машинам в пропорции, значительно отличающейся от национального уровня. Иначе говоря, есть вероятность 5–10 % того, что изменения в выдаче штрафов красным и синим машинам связаны не со сменой локации, а с чистой случайностью. Ввиду того что мы ищем вероятность, меньшую, чем 0,05, мы не можем быть *уверены*, что полиция нашего города более склонна выдавать штрафы красным машинам: есть маленькая, но статистически значимая вероятность того, что это не так.

А теперь, после того как мы проделали такой до-олгий путь к нирване, ввёдем, наконец, определение.

p-значение — это вероятность того, что случайная величина с данным распределением (распределением тестовой статистики при нулевой гипотезе) примет значение, не меньшее, чем фактическое значение тестовой статистики.

И напоследок. Господин Гудман (Goodman, 2008) написал чудную статью о недопонимании p-value и о тех ошибках в интерпретации, которые обычно допускают студенты. **Не делайте так! Опасно для жизни!** Помните:

- $p = 0{,}05$ не означает, что есть 5%-я вероятность того, что нулевая гипотеза верна.
- $p = 0{,}05$ не означает, что есть 5%-я вероятность ошибки первого рода.
- $p = 0{,}05$ не означает, что есть 95%-я вероятность того,что результаты будут такими же при повторении эксперимента.
- $p > 0{,}05$ не означает, что нет разницы между наблюдаемыми переменными.
- $p < 0{,}05$ не означает, что нулевая гипотеза не отвергается.

Список литературы

Goodman S. A Dirty Dozen: Twelve P-Value Misconceptions // Seminars in hematology. Т. 45. — Elsevier. 2008. — С. 135—140.

How to Calculate P Value. — 1 нояб. 2014. — URL: http://www.wikihow.com/Calculate-P-Value.

Statistics for Experimental Biologists. — 17 окт. 2014. — URL: http://labstats.net/articles/pvalue.html.