Действительно ли Д. Мартин жесток к Старкам? Проверка статистических гипотез

Болычев Антон, ЛМШ 54

22 июня 2019 г.

По старой математической традиции сначала сформулируем условие. Скорее всего, формулировка вам покажется дикой и непонятной, но не стоит беспокоится: я постараюсь тщательно разжевать ниже написанное, чтобы к концу занятия вы были способны полностью решить данную задачу.

1 Постановка задачи

Говорят, Джордж Р.Р. Мартин, автор цикла "Песнь Льда и Огня", истребляет Старков: чаще "убивает" персонажей, относящихся к этому дому, чем персонажей других домов. В таблице 1 приведено количество персонажей, относящихся к тому или иному дому, упомянутых за первые 4 книги, а так же количество погибших персонажей. Предлагается протестировать отличие уровня смертности дома Старков от уровня смертности каждого из других домов на 5% уровне значимости. Необходимо привести значения оценок вероятностей смертельных исходов для всех домов, найти p-value для каждого их трех тестов, а также проделать данные эксперименты с использованием метода Бонферрони.

Дом	Упомянутые персонажи	Погибшие персонажи
House Stark	72	18
House Lannister	49	11
House Greyjoy	41	12
Night's Watch	105	41

Таблица 1: Данные взяты из датасета `https://www.kaggle.com/mylesoneill/game-of-thrones`

2 Зачем и почему?

Конечно, можно просто взглянуть на табличку, на пальцах сравнив доли погибших персонажей внутри каждого дома и выдать ответ. Однако в реальной жизни так не работает. Если мы занимаемся чем-то более серьёзным, например, оцениваем насколько эффективен некий фармацевтический аппарат, то «статистика на пальцах» неприемлима! Нужны конкретные нормализованные математические методы, помогающие ответить на поставленные вопросы. Данные ответы предоставляет такой раздел математической статистики, как **проверка статистических гипотез**.

3 Немного тервера

Давайте мы будем воспринимать каждого персонажа как *случайную величину*, которая случайным образом принимает ровно два значения:

1. Герой выжил
2. Герой умер

Мы математики и любим работать с цифрами, поэтому предлагается «закодировать» данные два пункта нулём и единицей. Таким образом, все герои — это случайные величины, которые могут принимать только два значения: *ноль*, что соответствует тому, что данный герой выжил, и *единица*, что соответствует тому, что данный герой умер.

Замечание 1. *В таком прекрасном разделе математики как теория вероятностей случайные величины, которые принимают всего два значения* 0 *и* 1*, называют* ***бернуллиевскими*** *случайными величинами. При этом считается, что единица принимается с конкретной вероятностью* p*, а ноль, соответственно, с вероятностью* $1-p$*. Число* p *называют параметром бернуллиевской случайной величины.*

Спойлер 1. *Мы далее для примера предположим, что все персонажи Старков — это бернуллиевские случайные величины с вероятностью смерти* p_S*, а все персонажи Ланнистеров — бернуллиевские случайные величины с вероятностью смерти* p_L*. Потом попытаемся статистически ответить на вопрос: «Можно ли нам по данным таблицы 1 сказать, что* $p_S = p_L$*, или нельзя?». Но это позже, а сейчас нам нужно сделать ряд приготовления для формулировки одной важной теоремы, которой мы воспользуемся при решении задачи.*

Задача 1. *Нарисуйте график функции* $\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}$.

Задача 2. *Пусть* $F(x)$ *— это площадь под графиком функции* $\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}$ *от* $-\infty$ *до* x.

1. *Как будет выглядеть график функции* $F(x)$*?*

   P.S. Без доказательства считаем, что $F(0) = 0.5$ *и* $F(x) \to 1$ *при* $x \to +\infty$*, или, по-другому, что прямая* $y = 1$ *— это горизонтальная асимптота* $F(x)$*.*

2. *А как будет выглядеть график функции* $F^{-1}(\alpha)$*, обратной к* $F(x)$*?*

Задача 3. *Пусть* $X_1, X_2, X_3, \ldots, X_N$ *— бернуллиевские случайные величины с параметром* p*.*

1. *Какие значения может принимать* $X_1 + X_2 + \cdots + X_N$*? С какими вероятностями?*
2. *А какие значения может принимать* $\frac{X_1+X_2+\cdots+X_N}{N}$*? И с какими вероятностями?*
3. *Нарисуйте график функции* $f(t)$*, где* $f(t) = \mathsf{P}(\frac{X_1+X_2+\cdots+X_N}{N} = t)$ *— вероятность того, что* $\frac{X_1+X_2+\cdots+X_N}{N} = t$*. Если* $\frac{X_1+X_2+\cdots+X_N}{N}$ *не может принимать конкретное значение* t*, то тогда считаем, что* $f(t) = 0$*. Для простоты положим* $N = 5$ *и* $p = \frac{1}{2}$*.*

Можно заметить, что график функции $f(t)$ очень похож по форме на график функции $\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}$, и это неспроста. Оказывается, что есть определённая очень важная связь между данными функциями, которую мы сформулируем ниже. Это своего рода фундаментальная теорема теории вероятностей. Итак,

Теорема 1 (Центральная предельная теорема для бернуллиевских случайных величин)**.** *Пусть $X_1, X_2, \ldots, X_N$ — бернуллиевские случайные величины с параметром p. Обозначим*

$$\hat{p} = \frac{X_1 + X_2 + \cdots + X_N}{N}$$

тогда

$$\mathsf{P}\left(\frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}} \leq F^{-1}(\alpha)\right) \approx \alpha$$

при достаточно больших N. Как правило, данная теорема хорошо работает при $N \geqslant 30$.

Но нам на самом деле понадобиться другая формулировка.

Теорема 2 (Другая центральная предельная теорема для бернуллиевских случайных величин)**.** *Пусть $X_1, \ldots, X_N$ — бернуллиевские случайные величины с параметром p_X, а $Y_1, \ldots Y_M$ — бернуллиевские случайные величины с параметром p_Y. Обозначим*

$$\hat{p}_X = \frac{X_1 + \cdots + X_N}{N}, \qquad \hat{p}_Y = \frac{Y_1 + \cdots + Y_M}{M}$$

тогда

$$\mathsf{P}\left(\frac{(\hat{p}_X - \hat{p}_Y) - (p_X - p_Y)}{\sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{N} + \frac{\hat{p}_Y(1-\hat{p}_Y)}{M}}} \leq F^{-1}(\alpha)\right) \approx \alpha$$

Замечание 2. *ЦПТ формулируется чуть-чуть по-другому, однако в приведенных теоремах я постарался максимально хорошо передать её суть и привёл наиболее удобные формулировки в рамках нашей задачи.*

4 Проверка статистических гипотез

Итак, настало время матстата. Когда мы говорим про случайные величины, то мы отнюдь не ограничиваемся лишь бернуллиевскими случайными величинами. Есть много других видов. У каждого вида есть свои параметры, аналогичные параметру p. Некое предположение о данных параметрах конкретных видов случайных величин называется статистической гипотезой. Этому предположению, называемому гипотезой H_0, противопоставляется т.н. альтернативная гипотеза H_1. Наша задача ставится следующим образом: готовы ли мы принять гипотезу H_0 или же нам придётся её отвергнуть, склонив наш выбор в сторону H_1? Чтобы было понятней, приведём конкретный пример. Пусть

$$X_1, X_2, \ldots, X_N$$

бернуллиевские случайные величины aka персонажи Старков. Будем считать, что каждый Старк умирает с вероятностью p_S, то есть для любого i

$$\mathsf{P}(X_i = 1) = p_S, \qquad \mathsf{P}(X_i = 0) = 1 - p_S$$

Аналогично, пусть

$$Y_1, Y_2, \ldots, Y_M$$

бернуллиевские случайные величины aka персонажи Ланнистеров и для любого i

$$\mathsf{P}(Y_i = 1) = p_L, \qquad \mathsf{P}(Y_i = 0) = 1 - p_L$$

Задача 4. *Пусть*

$$\hat{p}_S = \frac{X_1 + \cdots + X_N}{N}, \qquad \hat{p}_L = \frac{Y_1 + \ldots Y_M}{M}$$

По таблице 1 определите, чему равны $\hat{p}_S$ *и* $\hat{p}_L$*.*

Итак, можно ли утверждать, что Старки умирают так же, как Ланнистеры? Сформулируем данный вопрос на математическом языке:

$$H_0 : p_S = p_L, \qquad H_1 : p_S \neq p_L \tag{1}$$

В данном аккуратном вопросе, начинающемся с «можно ли», скрыто ещё одно важное обстоятельство: нам нужен конкретный критерий, согласно которому мы считаем, когда *можно*, а когда *нельзя*. Данный критерий носит название *уровня значимости*. В нашей задаче он по условию равен 5%. Я сейчас попробую расшифровать, что это такое.

Обозначим $\alpha = 0.05$ и ответим на вопрос: что делают математики при проверке статистических гипотез? Они буквально говорят: «Построим множество, в которое при верности H_0 некий крокодил попадёт с вероятностью не менее $1 - \alpha$». Данное множество называется *критическим*. Далее математики действуют очень просто: они смотрят на крокодила, попутно проверяя, принадлежит ли он критическому множеству или нет. Если принадлежит, то принимают H_0, иначе — отвергают H_0 в пользу H_1. Всё действительно просто. В качестве так называемого крокодила предлагается взять разность $p_S - p_L$, которая должна быть равна 0 в случае верности H_0. Итак, если H_0 верна, то, согласно теореме 2, получаем[1]

$$\mathsf{P}\left(\frac{(\hat{p}_S - \hat{p}_L) - 0}{\sqrt{\frac{\hat{p}_S(1-\hat{p}_S)}{N} + \frac{\hat{p}_L(1-\hat{p}_L)}{M}}} \leq F^{-1}(1-\alpha)\right) \approx 1 - \alpha \iff$$

$$\iff \mathsf{P}\left(0 \geq (\hat{p}_S - \hat{p}_L) - F^{-1}(1-\alpha)\sqrt{\frac{\hat{p}_S(1-\hat{p}_S)}{N} + \frac{\hat{p}_L(1-\hat{p}_L)}{M}}\right) \approx 1 - \alpha$$

$$\iff \mathsf{P}\left(0 \in \left[(\hat{p}_S - \hat{p}_L) - F^{-1}(1-\alpha)\sqrt{\frac{\hat{p}_S(1-\hat{p}_S)}{N} + \frac{\hat{p}_L(1-\hat{p}_L)}{M}}, +\infty\right)\right) \approx 1 - \alpha$$

Кажется, мы получили то, что хотели. Если наш крокодил 0 принадлежит критическому множеству:

$$\left[(\hat{p}_S - \hat{p}_L) - F^{-1}(1-\alpha)\sqrt{\frac{\hat{p}_S(1-\hat{p}_S)}{N} + \frac{\hat{p}_L(1-\hat{p}_L)}{M}}, +\infty\right)$$

то мы принимаем H_0, иначе — отвергаем.

Задача 5. *Проверьте* (1) *на* 5% *уровне значимости, если* $F^{-1}(0.95) \approx 1.645$*.*

Задача 6. *Проверьте на* 5% *уровне значимости, действительно ли Старки умирают чаще*

(a) Грейджоев?

(b) Ночного Дозора?

[1] Отметим, что я просто в формулировку подставил $p_S - p_L = 0$

Мы уже находимся на финишной прямой. Мне хотелось бы рассказать про ещё одну очень популярную и нужную штуковину. Она называется p-value. Но прежде давайте вспомним про уровень значимости. Отметим, что чем он больше, тем меньше критическое множество, и, соответсвенно, тем вероятнее мы отвергнем H_0. Так вот, p-value — это минимальный уровень значимости, при котором мы отвергаем H_0. Давайте поясним на примере нашей задачи. Итак, мы отвергаем H_0, когда

$$0 < (\hat{p}_S - \hat{p}_L) - F^{-1}(1-\alpha)\sqrt{\frac{\hat{p}_S(1-\hat{p}_S)}{N} + \frac{\hat{p}_L(1-\hat{p}_L)}{M}}$$

И нам нужно найти минимальное α, при котором данное неравенство будет выполнятся. Это и будет искомое p-value. Строго говоря, минимум не может быть достигнут в принципе в силу строгого неравенства. По факту это то же самое, что и потребовать найти минимальное действительное x, при котором $x > 2$. На самом деле, в правильном определении используется не минимум, а так называемый инфимум, но об этом вам расскажут позже на мехмате[2], поэтому мы (пока) закроем глаза на математическую строгость высказывания и положим, что p-value — это решение уравнения

$$0 = (\hat{p}_S - \hat{p}_L) - F^{-1}(1-\text{p-value})\sqrt{\frac{\hat{p}_S(1-\hat{p}_S)}{N} + \frac{\hat{p}_L(1-\hat{p}_L)}{M}}$$

По факту это будет «пограничным значением» уровня значимости: если p-value меньше заданного в условии уровня значимости α, то мы отвергаем H_0, иначе — принимаем. Посчитаем p-value для каждого из случаев

	Старки vs Ланнистеры	Старки vs Грейджои	Старки vs Ночной Дозор
p-value	0.37255026511742795	0.6871878448525566	0.9779318062843195

Теперь мы обладаем достаточной математической грамотностью, чтобы ответить на финальный вопрос: действительно ли Старки умирают чаще? Мы по отдельности сравнили Старков с Ланнистерами, Грейджоями и Ночным Дозором. По факту у нас есть 3 отдельные гипотезы с тремя отдельными p-value, однако мы хотим оценить ситуацию в общем! Воспользуемся методом Бонферонни: если все найденные p-value меньше чем $\alpha/3$ (тройка в знаменателе — количество гипотез), то тогда признаем, что Старки не умирают чаще остальных персонажей, иначе — сделаем вывод, что Мартин чрезмерно жесток к Старкам.

Задача 7. *Действительно ли Джордж Мартин жесток к Старкам?*

[2]Если вы, конечно, поступите, что вряд ли