2. Метод Уорнера

Метод Уорнера проведения рандомизированных опросов позволяет повысить доверие респондентов к социологу. В отличие от метода Хетманспергера, теперь предлагается ответить на один и тот же вопрос, но с рандомизированным условием, что “скрывает” ответ каждого из респондентов.

Предположим, что все люди разделены по некоему принципу на две группы A или B (состав групп социологу не известен). В опросе Хетманспергера группы A или B определялись отношением к курению марихуаны.

Задачей социолога является оценка количества людей из группы A. Для этого социолог случайным образом выбирает n человек и каждому из них предлагает ответить на один вопрос, но перед этим предлагает каждому респонденту запустить юлу с указателем, который после остановки может указывать на одно из двух полей: либо на поле с буквой A (вероятность этого события равна p), либо на поле с буквой B (вероятность этого события равна $1-p$). Поле, на которое указывает юла, социологу также не известно. Пусть G — это группа, к которой относится респондент, а L — поле, на которое указывает юла. Вопрос, на который респонденту необходимо дать ответ, звучит так

`одинаковы ли` G `и` L`?`

Таким образом, каждый из респондентов должен сказать “*да*” или “*нет*” в зависимости от того совпадают или нет его G и L. Социологу не сообщаются ни поле, на которое указывает юла, ни группа каждого респондента. Предположим, что респонденты уверены в анонимности, то есть их ответы являются правдивыми.

Рассмотрим такие случайные величины

$$X_i = \begin{cases} 1, & \text{если } i\text{-ый опрашиваемый отвечает “}да\text{”}, \\ 0, & \text{если } i\text{-ый опрашиваемый отвечает “}нет\text{”}. \end{cases}$$

Обозначим через π вероятность того, что случайно выбранный человек принадлежит к группе A (именно эту вероятность социологу предстоит оценить). Согласно формуле полной вероятности

$$\begin{aligned} &\mathsf{P}(X_i = 1) = \pi p + (1-\pi)(1-p), \\ &P(X_i = 0) = (1-\pi)p + \pi(1-p). \end{aligned} \tag{2}$$

Понятно, что

$$\pi p + (1-\pi)(1-p) + (1-\pi)p + \pi(1-p) = 1, \tag{3}$$

то есть X_i является случайной величиной Бернулли с вероятностью "*успеха*" $\pi p + (1-\pi)(1-p)$.

Доказательство равенств (2). Доказательство каждого из равенств (2) основано на формуле полной вероятности (1) и естественном предположении о том, что результат запуска волчка и принадлежность к группе являются независимыми. Для краткости будем писать $i \in A$, если индивидуум i принадлежит к группе A. Аналогично определяется сокращение $i \in B$. Будем также писать $\to A$ вместо фразы "*указатель волчка указывает на поле* A". Мы используем аналогичное сокращение $\to B$ и для поля B. Поскольку $\mathsf{P}(i \in A) = \pi$ и $\mathsf{P}(i \in B) = 1 - \pi$, то согласно (1)

$$\begin{aligned} \mathsf{P}(X_i = 1) &= \mathsf{P}(i \in A)\mathsf{P}\left(X_i = 1 \,/ i \in A\right) \\ &\qquad + \mathsf{P}(i \in B)\mathsf{P}\left(X_i = 1 \,/ i \in B\right) \\ &= \pi \cdot \mathsf{P}\left(X_i = 1 \,/ i \in A\right) \\ &\quad + (1-\pi) \cdot \mathsf{P}\left(X_i = 1 \,/ i \in B\right). \end{aligned}$$

Ясно, что

$$\begin{aligned} &\mathsf{P}\left(X_i = 1 \,/ i \in A\right) \\ &\quad = \mathsf{P}\left(X_i = 1, \to A \,/ i \in A\right) + \mathsf{P}\left(X_i = 1, \to B \,/ i \in A\right) \\ &\quad = \mathsf{P}\left(X_i = 1, \to A \,/ i \in A\right) \\ &\quad = \mathsf{P}\left(\to A \,/ i \in A\right) = \mathsf{P}(\to A) = p, \end{aligned}$$

поскольку события $\{\to A\}$ и $\{i \in A\}$ независимы. Таким же образом доказываем, что $\mathsf{P}\,(X_i = 1\,/i \in B\,) = 1 - p$. Отсюда и вытекает первое равенство в (2). Второе равенство в (2) доказывается аналогично. $\square$

2.1. Функция правдоподобия. *Функцией правдоподобия* для *выборки* $X_1, \ldots, X_n$ называется

$$L(x_1, \ldots, x_n) = \mathsf{P}(X_1 = x_1, \ldots, X_n = x_n).$$

Таким образом, функция правдоподобия для выборки размера n является функцией n аргументов $x_1, \ldots, x_n$.

Поскольку респонденты дают ответы независимо друг от друга, то случайные величины $X_1, \ldots, X_n$ (ответы респондентов) независимы в совокупности и поэтому

$$L \stackrel{\text{def}}{=} L(x_1, \ldots, x_n) = \mathsf{P}(X_1 = x_1) \ldots \mathsf{P}(X_n = x_n).$$

Следовательно, если хотя-бы один из аргументов $x_1, \ldots, x_n$ отличен и от 0, и от 1, то $L = 0$. В дальнейшем мы рассматриваем функцию правдоподобия только на множестве

$$\{0,1\}^n = \underbrace{\{0,1\} \times \cdots \times \{0,1\}}_{n \text{ раз}},$$

которое состоит из векторов длины n, каждая координата которых равна либо 0, либо 1.

Если количество ответов “*да*” обозначить через n_1, то количество ответов “*нет*” равно $n - n_1$, а функция правдоподобия сводится к

$$L = [\pi p + (1-\pi)(1-p)]^{n_1}[(1-\pi)p + \pi(1-p)]^{n-n_1}. \tag{4}$$

Поэтому логарифм функции правдоподобия равен

$$\begin{aligned} \log L = {} & n_1 \log[\pi p + (1-\pi)(1-p)] \\ & + (n - n_1) \log[(1-\pi)p + \pi(1-p)]. \end{aligned} \tag{5}$$

2.2. Максимум функции правдоподобия. Возвращаясь к поставленной задаче оценить количество людей, принадлежащих группе A, заметим, что она эквивалентна оценке вероятности π.

Одним из лучших статистических методов оценки параметров по выборке является *метод максимального правдоподобия*. Этот метод заключается в нахождении максимума функции правдоподобия по отношению к изучаемому параметру и выборе в качестве оценки (которая называется *оценкой метода максимального правдоподобия*) такого параметра, при котором L достигает максимума.

В нашем случае оцениваемым параметром является вероятность π. Если $p \neq \frac{1}{2}$, то максимум логарифма функции правдоподобия (5) достигается для того значения π, для которого

$$\frac{\partial \log L}{\partial \pi} = \frac{n_1(2p-1)}{\pi p + (1-\pi)(1-p)} + \frac{(n-n_1)(1-2p)}{(1-\pi)p + \pi(1-p)} = 0$$

или

$$\pi p + (1-\pi)(1-p) = \frac{n_1}{n}. \tag{6}$$

2.3. Оценка метода максимального правдоподобия. Решение уравнения (6) является оценкой максимального правдоподобия для π. Если $p \neq \frac{1}{2}$, то решение уравнения (6) равно

$$\hat{\pi} = \frac{p-1}{2p-1} + \frac{n_1}{(2p-1)n}. \tag{7}$$

Замечание 1. При определенных соотношениях между p, n и n_1 может случиться, что $\hat{\pi} \notin (0,1)$. Однако оцениваемый параметр π является вероятностью, то есть $\pi \in (0,1)$.

В таких случаях оценка максимального правдоподобия становится бессмысленной. Например, если $n_1 = n$, то $\hat{\pi} < 0$, если $p < \frac{1}{2}$, и $\hat{\pi} > 1$, если $\frac{1}{2} < p < 1$. Это означает, что при $n_1 = n$ оценка максимального правдоподобия (7) имеет смысл только при $p = 0$ или $p = 1$. В дальнейшем мы считаем, что p, n и n_1 таковы, что $\hat{\pi} \in (0, 1)$ (см. упражнение 3).

2.4. Моменты оценки максимального правдоподобия. Поскольку $n_1 = \sum_{i=1}^{n} X_i$, то математическое ожидание оценки $\hat{\pi}$ равно

$$\begin{aligned} \mathsf{E}[\hat{\pi}] &= \frac{1}{2p-1}\left[p - 1 + \frac{1}{n}\sum_{i=1}^{n}\mathsf{E}[X_i]\right] \\ &= \frac{1}{2p-1}\left[p - 1 + \pi p + (1-\pi)(1-p)\right] \\ &= \pi. \end{aligned} \tag{8}$$

Свойство (8) оценки, а именно $\mathsf{E}[\hat{\pi}] = \pi$, считается желательным при статистическом анализе, а такие оценки называются *несмещенными*.

Если $p \neq \frac{1}{2}$, то в силу (3) дисперсия величины $\hat{\pi}$ равна

$$\begin{aligned} \mathsf{var}[\hat{\pi}] &= \frac{n\,\mathsf{var}[X_1]}{(2p-1)^2 n^2} \\ &= \frac{[\pi p + (1-\pi)(1-p)]\,[(1-\pi)p + \pi(1-p)]}{(2p-1)^2 n}. \end{aligned} \tag{9}$$

Из равенства (9) мы получаем $\mathsf{var}[\hat{\pi}] \to 0$, $n \to \infty$. Это свойство также является весьма желательным в статистическом анализе и называется *состоятельностью оценки*.

2.5. Доверительный интервал. *Доверительным интервалом* уровня доверия α для параметра π называется числовой интервал $I \stackrel{\text{def}}{=} [\pi_1, \pi_2)$, для которого

$$\mathsf{P}(\pi \in I) = \mathsf{P}(\pi_1 \leq \pi < \pi_2) = \alpha. \tag{10}$$