Lab 1

Yaroslav Ivchenkov
DCAM, MIPT

28 марта 2020 г.

Problem 2.3

Известно, что электричка "Вашингтон-Петушки"аварийно останавливается раз в несколько дней. Аналитики РЖД проанализировали, сколько дней электричка едет без поломок, и составили выборку:

$$\boldsymbol{x = (3, 22, 13, 6, 18, 5, 6, 10, 7, 15).}$$

РЖД хочет проверить гипотезу, что дисперсия распределения равна 9 против правосторонней альтернативы.

1. Предположим, что поломки происходят независимо друг от друга, для иного нам необходимы были бы дополнительные данные. Также предположим, что вероятность поломки не меняется с течением времени, то есть равна зафиксированному числу p. При сделанных предположениях мы приходим к выводу, что от выборки следует ожидать геометрического распределения(количество «неудач» до первого «успеха» в серии испытаний Бернулли):

$$\mathbb{P}(X = n) = q^n p\,, \quad q = 1 - p$$

2. Для данного вида распределения мы уже знаем математическое ожидание и дисперсию: $\mathbb{E}X = \frac{q}{p}$, $\mathbb{D}X = \frac{q}{p^2}$. Теперь мы можем сформулировать задачу:

$$\begin{aligned} \text{Выборка:} \quad & X \sim Geom(p) \\ \text{Нулевая гипотеза:} \quad & H_0 : \mathbb{D}X = 9 \\ \text{Альтернатива:} \quad & H_1 : \mathbb{D}X > 9 \end{aligned}$$

Однако, можно заметить, что дисперсия распределения однозначно определяет его параметр - функция $f(x) = \frac{1-x}{x^2}$ однозначно переводит отрезок $[0, 1]$ в $[0, +\infty]$. Тогда возникает возможность переформулирования задачи в более привычном виде:

$$\begin{aligned} \text{Выборка:} \quad & X \sim Geom(p) \\ \text{Нулевая гипотеза:} \quad & H_0 : p = \frac{-1 + \sqrt{37}}{18} \approxeq 0.282 \\ \text{Альтернатива:} \quad & H_1 : p < 0.282 \end{aligned}$$

3. Хотелось бы воспользоваться критерием отношения правдоподобий, но альтернатива в нем только двусторонняя, поэтому воспользуемся критерием меток и ещё раз перепишем нашу задачу:

$$\begin{aligned} \text{Выборка:} &\quad X \sim Geom(p) \\ \text{Нулевая гипотеза:} &\quad H_0 : p = 0.282 = p_0 \\ \text{Альтернатива:} &\quad H_1 : p < p_0 \\ \text{Статистика:} &\quad Z_s(X^n) = \frac{S(p_0)}{\sqrt{I(p_0)}} \end{aligned}$$

4. Выведем все вручную:

$$S(p) = \frac{\partial}{\partial p} \log L(X^n, p) = \dots$$

$$\log L(X^n, p) = \sum_{i=1}^{n} \log{(1-p)^{X_i} p} = \sum_{i=1}^{n} \log{(1-p)^{X_i}} + \sum_{i=1}^{n} \log p =$$

$$= \sum_{i=1}^{n} X_i \cdot \log{(1-p)} + n \log p$$

$$\dots = \frac{\partial}{\partial p} \left(\sum_{i=1}^{n} X_i \cdot \log{(1-p)} + n \log p \right) = \frac{n}{p} - \frac{\sum_{i=1}^{n} X_i}{1-p};$$

$$I(p) = -\mathbb{E}\left[\frac{\partial^2}{\partial^2 p} \log L\right] = -\mathbb{E}\left[\frac{\partial}{\partial p}\left(\frac{n}{p} - \frac{\sum_{i=1}^{n} X_i}{1-p}\right)\right] = -\mathbb{E}\left[-\frac{n}{p^2} - \frac{\sum_{i=1}^{n} X_i}{(1-p)^2}\right] =$$

$$= \frac{n}{p^2} + \frac{n\mathbb{E}X_i}{(1-p)^2} = \frac{n}{p^2} + \frac{n \cdot \frac{1-p}{p}}{(1-p)^2} = \frac{n}{p^2} + \frac{n}{p(1-p)} = \frac{n}{p^2(1-p)}$$

В нашей выборке: $n = 10$, $x = (3, 22, 13, 6, 18, 5, 6, 10, 7, 15)$, $\sum_{i=1}^{n} X_i = 105$. Тогда:

$$\begin{aligned} S(p_0) &\approxeq -110.78 \\ I(p_0) &\approxeq 175.14 \\ Z_s(X^n) &\approxeq -8.37 \end{aligned}$$

Т.к. $Z_s(X^n)\ \mathcal{N}(0,1)$, смотрим p-value и находим, что:

$$\text{p-value} < 10^{-15} \lll 0.05 \to \text{гипотезу отвергаем.}$$