ОП «Политология», 2019-20
Математика и статистика, часть 2
Описание выборок – часть 1. (10.04.2020)

А. А. Макаров, А. А. Тамбовцева, Н. А. Василёнок

Задача 1.

(a) Дана выборка:

$$0 \quad 0 \quad 2 \quad 2 \quad 2 \quad 0 \quad 0 \quad 2$$

Может ли эта выборка быть правдоподобной (репрезентативной) выборкой из биномиального распределения с параметрами $n = 2$, $p = 0.5$? Обоснуйте свой ответ.

Решение.

Биномиальная величина $\mathrm{X} \sim \mathrm{Binom}(n = 2;\ p = 0.5)$ описывает число успехов в серии из двух испытаний Бернулли. В данном случае эта случайная величина описывает количество выпавших орлов в двукратном подбрасывании правильной монетки. Запишем закон распределения случайной величины X:

X	0	1	2
p	0.25	0.5	0.25

Исходы двукратного подбрасывания монетки: ОО, ОР, РО, РР. Вероятности соответствующие.

Вероятности отдельных значений биномиальной случайной величины можно было бы рассчитать и по формуле:

$$\mathrm{P}(\mathrm{X} = k) = C_n^k \times \mathrm{p}^k \times (1 - \mathrm{p})^{(n-k)}.$$

Получаем:

$$\mathrm{P}(\mathrm{X} = 0) = C_2^0 \times 0.5^0 \times 0.5^2 = 0.25;$$
$$\mathrm{P}(\mathrm{X} = 1) = C_2^1 \times 0.5^1 \times 0.5^1 = 0.5;$$
$$\mathrm{P}(\mathrm{X} = 2) = C_2^0 \times 0.5^2 \times 0.5^0 = 0.25.$$

Доля каждого из возможных значений случайной величины в нашей выборке должна быть близка к теоретической вероятности соответствующих значений. Примерно 50% выборки должны составлять значения 1, и по 25% – 0 и 2 (допустимы небольшие отклонения, особенно на маленькой выборке).

Однако мы видим, что в выборке нет ни одного значения 1, а значения 0 и 2 распределились примерно поровну. Скорее всего, такая выборка не является правдоподобной выборкой из заданного распределения.

(b) Дана выборка:

$$-20 \quad 25 \quad 5 \quad 78 \quad 27 \quad 16 \quad -11 \quad 10 \quad 18$$

Может ли предложенная выборка быть правдоподобной (репрезентативной) выборкой из нормального распределения $N(2, \sigma^2 = 16)$? Обоснуйте свой ответ.

Решение.

Вспомним, что: а) нормальное распределение симметрично относительно среднего; б) для него работает «правило трех сигм». Давайте рассчитаем, чему равны границы одного, двух и трех стандартных отклонений от среднего для данного распределения.

В пределах 1 стандартного отклонения от среднего находится примерно 68% выборки:

$$[2 - \sqrt{16};\ 2 + \sqrt{16}] = [-2;\ 6].$$

В пределах 2 стандартных отклонений от среднего находится примерно 95% выборки:

$$[2 - 2 \times \sqrt{16};\ 2 + 2 \times \sqrt{16}] = [-6;\ 10].$$

Следовательно, на промежутки $[-6; -2) \cup (6; 10]$ приходится $95 - 68 = 27\%$ выборки. Наконец, в пределах 3 стандартных отклонений должна находиться почти вся выборка (99.8%):

$$[2 - 3 \times \sqrt{16};\ 2 + 3 \times \sqrt{16}] = [-10;\ 14].$$

Но в данной нам выборке **только 2** значения (выделены красным) укладываются в границы 3 стандартных отклонений! Следовательно, данная выборка не является правдоподобной для рассматриваемого распределения.

Можно также было говорить про симметричность относительно среднего, но одного этого свойства все-таки недостаточно.

Задача 2. Дана выборка:

$$8 \quad 5 \quad 7 \quad 10 \quad 12 \quad 30 \quad 4 \quad 2 \quad 8$$

(a) Найдите медиану выборки.

Упорядочим выборку по возрастанию:

$$2 \quad 4 \quad 5 \quad 7 \quad 8 \quad 8 \quad 10 \quad 12 \quad 30$$

Медиана выделена синим:

$$\mathrm{Me} = x_{0.5} = 8.$$

(b) Найдите нижний и верхний квартили выборки.

«Меньшая» подвыборка, нижняя половина выборки:

$$2 \quad 4 \quad 5 \quad 7 \quad 8.$$

Нижний квартиль – медиана «меньшей» подвыборки – выделен синим:

$$\mathrm{Q}_1 = x_{0.25} = 5.$$

«Большая» подвыборка, верхняя половина выборки:

$$8 \quad 8 \quad 10 \quad 12 \quad 30.$$

Верхний квартиль – медиана «большей» подвыборки – выделен синим:

$$\mathrm{Q}_3 = x_{0.75} = 10.$$

(c) Проверьте, есть ли в выборке нетипичные наблюдения (выбросы). Если есть, укажите их.

Рассчитаем границы характерных значений:

$$[x_{0.25} - 1.5 \times (x_{0.75} - x_{0.25});\; x_{0.75} + 1.5 \times (x_{0.75} - x_{0.25})]$$

$$[5 - 1.5 \times (10 - 5);\; 10 + 1.5 \times (10 - 5)]$$

$$[-2.5;\; 17.5]$$

Одно наблюдение в нашей выборке – 30 – оказалось выходящим за границы типичных значений. Оно и будет выбросом.