УДК 51(071)

Об одной непоследовательности при использовании критерия χ^2

Ехилевский С.Г., Вилкова И.В.

Донецкий национальный технический университет

В роботі обгрунтовано доцільність врахування варіантів із нульовим емпіричними частотами при перевірці статистичних гіпотез за методом Пірсона. Запропонований спосіб є строгим та конструктивним, до того ж він практично не потребує додаткових обчислень.

Согласно имеющимся в литературе рекомендациям [1], критерий Пирсона может применяться при любом (достаточно большом) числе вариантов признака, которые для удобства могут быть объединены в произвольное (достаточно большое) количество интервалов (не обязательно равных). Это означает, что в принципе, часть статических данных может быть проигнорирована при составлении расчетного значения критерия $\chi^2_{рас}$. Конечно, поступать так неразумно, ибо при прежнем уровне значимости это в среднем уменьшит разницу между расчетным и критическим значением критерия $\chi^2_{кр}$. Так в таблице №8 работы [2], на которую мы будем дальше ссылаться, представлены 9 – ть интервалов. На каждый из них приходится в среднем по 1,017 в $\chi^2_{рас}$ = 9,152. Именно на столько в среднем уменьшится $\chi^2_{рас}$ в результате случайного отбрасывания одного из интервалов. При этом уменьшение $\chi^2_{кр}$, связанное с изменением на единицу числа степеней свободы, для использованного в [2] α = 0,05 составит 12,6-11,1 = 1,5>1,017. т.е., двигаясь по этому пути, можно так «обкорнать» выборку, что вывод об истинности проверяемой гипотезы сменится на противоположный.

В этой связи, безусловно, оправдано использование при составлении $\chi^2_{рас}$ всех имеющихся статистических данных. Однако весьма распространенно заблуждение, что для этого достаточно учесть все интервалы (варианты) с отличными от нуля эмпирическими частотами. Покажем, что это не так.

Очевидно, что пустые по данным выборки интервалы дают вклад в критерий, если в них отличны от нуля теоретические частоты. Иными словами, равенство нулю эмпирических частот – тоже статистические данные, которые не сле-

дует игнорировать. Тем более, что их учет не составляет ни какого труда. Согласно основной расчетной формуле (см. [1,2]) вклад от пустых интервалов равен сумме приходящихся на них теоретических частот.

Ошибочно считать такую поправку пренебрежимо малой. Так, в рассмотренной в [2] задаче, на неучтенные пустые интервалы $(-\infty, 3{,}93)$ и $(4.56, \infty)$ при объеме выборки 100 и нормальном распределении признака с σ_B = 0,12 и a = 4,21[1] приходится Ф(-2,3331) – Ф($-\infty$) + Ф(∞) – Ф(2,917) = 0,0117 вероятности или 1,17 теоретической частоты. Это не просто 10% $\chi^2_{рас}$, но больше среднего вклада, приходящегося на один учитываемый интервал. Т.е. поправка, о которой идет речь, имеет не меньше оснований учитываться, чем любой из рассмотренных интервалов.

Кроме того, при ее учете разница между $\chi^2_{рас}$ и $\chi^2_{кр}$ увеличивается, т.к. возрастает число степеней свободы κ. В нашем случае $\Delta\kappa$ = 2 и $\Delta\chi^2_{кр}$ = =15,5 – 12,6 = 2,9. Т.е. разница между $\chi^2_{рас}$ и $\chi^2_{кр}$ увеличилась на 2,9 – 1,17 = = 1,73, что по сравнению с исходным значением $\chi^2_{рас}$ составляет почти 20%. Эти 20% могут уверенно склонить чашу весов в тех случаях, когда проверяемая гипотеза не четко просматривается в данных выборки.

Основная погрешность при практическом применении метода Пирсона возникает при определении теоретических частот с помощью таблиц, содержащих лишь дискретные значения аргумента. Так сумма учитываемых в [2] теоретических частот равна 98,96. Т.е. не учтены 100 – 98,96 = 1,04, что не равно 1,17, найденным ранее. Однако неточность 1,17 – 1,04 = 0,13 на порядок меньше предлагаемой поправки, что делает ее корректной и с этой точки зрения.

Таким образом, общее правило заключается в следующем. Для полного использования статистической информации нужно к значению $\chi^2_{рас}$, полученному с помощью интервального ряда, прибавить разницу между объемом выборки и суммой учтенных теоретических частот. При этом число степеней свободы нужно увеличить на число добавленных пустых интервалов.

---

[1] Все данные взяты из рассмотренной в [2] задачи, для которой и построена упоминавшаяся таблица 8.

В заключение заметим, что попытки учесть «края» теоретического распределения имеются. В частности интервалы, содержащие малые эмпирические частоты, рекомендуют сливать с соседними. С полубесконечными интервалами делать этого ни в коем случае нельзя, ибо смещается оценка матожидания, для получения которой берут серединные значения рассматриваемых интервалов.

Литература

1. Логинов Э.А. Математическая статистика. М. Изд – во МИСИ, 1977г.,93с.
2. Методические указания к выполнению семестрового индивидуального задания по математической статистике / Сост.: Ю.Ф. Косолапов, Н.Г. Плаксина. – Донецк: ДПИ, 1989. – 48 с.