14. ОДНОТИПНЫЕ ВЫБОРОКИ

14.1. Однотипные выборки экспериментальных данных и задачи их обработки

По результатам экспериментов может накапливаться целый ряд выборок по однотипным средствам и комплексам. Однотипность не означает равноценности объектов по их показателям. Неоднородность означает, что выборки принадлежат различным законам распределения, которые различаются или только параметрами при одном и том же виде, или видом и параметрами распределения.

Задачи обработки однотипных выборок подразделяются на две группы.

К первой группе относятся задачи объединения выборок. Простое слияние однотипных, но неоднородных выборок для последующей оценки показателей по объединенной выборке, приводит к снижению качества оценок или даже к их полной непригодности. Необходимо применение специальных приемов объединения разнородных сведений в интересах использования всей содержащейся в выборках информации. Таким образом, при объединении выборок необходимо сначала проверить их однородность. Однородные выборки сливаются в одну общую выборку, которая обрабатывается с помощью обычных методов. Неоднородные выборки обрабатываются раздельно или объединяются с помощью специальных приемов.

Вторая группа задач связана с сопоставлением параметров распределения выборок, т.е. с определением существенных различий в значениях параметров однотипных выборок. Наиболее широкое распространение получил один из видов подобного рода задач, так называемый дисперсионный анализ. В дисперсионном анализе исследуются методы проверки гипотезы о равенстве математических ожиданий случайных величин, представленных выборками ограниченного объема. Непосредственное сравнение оценок математических ожиданий совокупности выборок оказывается менее эффективным, чем сопоставление оценок дисперсий, это обстоятельство и дало наименование методу.

Итак, пусть имеются выборки m (m≥2) однородных выборок, каждая выборка имеет свой объем n_i, . Априорных сведений об однородности или неоднородности различных выборок нет.

$$\begin{matrix} x_{11} & x_{12} & \dots & x_{1n} \\ y_{21} & y_{22} & \dots & y_{2n} \\ \dots & \dots & \dots & \dots \\ w_{m1} & w_{m2} & \dots & w_{mn} \end{matrix} \qquad (5.1)$$

Эта совокупность состоит из m слоев (строк). Каждая i-я строка (i=1,…,m) представляет собой однородную случайную выборку результатов наблюдений за значениями случайной величины $X, Y, \dots, W$ соответственно. Слой характеризуется своим, в общем случае векторным, параметром Θ_i распределения и может иметь свои статистики, т.е. свои функции от выборочных значений.

14.2. Объединение выборок

По возможности объединения информации из совокупности однотипных выборок можно выделить три типовые ситуации:

- различные слои представляют собой однородные выборки. В такой идеализированной ситуации выборки можно объединить и определять искомые параметры, используя традиционный аппарат математической статистики.
- совокупность слоев частично неоднородна. Тогда однородные слои, если таковые обнаружатся, целесообразно объединить, а оставшиеся неоднородные группы выборок обрабатывать раздельно.
- слои полностью или частично неоднородны. Но, в дополнение к результатам наблюдений, имеется априорная информация о взаимосвязи параметров Θ_i различных выборок. Чем выше уровень априорной информированности о взаимосвязях параметров, тем потенциально более высокой эффективности оценок показателей можно достичь.

Следовательно, объединение слоев всегда следует начинать с проверки однородности выборок.

14.2.1. Объединение однородных выборок

Постановка задачи проверки однородности выборок формулируется следующим образом.

Имеются результаты наблюдений в виде совокупности выборок типа (5.1), задан уровень значимости a для проверки статистической гипотезы об однородности выборок.

Необходимо проверить однородность слоев.

Допущение: законы распределения случайных величин для различных слоев неизвестны.

На практике используется последовательная процедура проверки и попарного объединения выборок. В качестве исходной выборки можно взять любую, например, наибольшую по количеству элементов. В качестве второй выбирается любая из оставшихся выборок. Эти две выборки проверяются на однородность. При ее наличии выборки объединяются в одну, а при ее отсутствии вторая выборка остается самостоятельной. Указанную проверку и объединение повторяют для всех слоев исходной выборки.

Определение однородности двух выборок проводится на основе проверки статистической гипотезы H_0 о том, что выборки и принадлежат одному, пусть и неизвестному, закону распределения. При этом может быть применен критерий Вилкоксона (Вилкоксона – Мана – Уитни).

Проверка однородности выборок по критерию Вилкоксона состоит в следующем. Пусть для случайной величины X имеется выборка объема n_x и для случайной величины Y выборка объема n_y. По этим выборкам необходимо с уровнем значимости a проверить гипотезу H_0 о том, что функция распределения F(x) случайной величин X равна функции распределения F(y) случайной

величины Y. Конкурирующая гипотеза – функции распределения случайных величин различны: F(x) < F(y) или F(x) > F(y), т.е. критическая область двусторонняя.

Сущность проверки основана на простой идее: если верна гипотеза H_0, то нельзя ожидать преобладания наблюдений одной из выборок на любом из концов вариационного ряда, иначе говоря, результаты наблюдений из каждого слоя должны быть рассеяны по всему вариационному ряду. Такая проверка осуществляется только по порядковым соотношениям x > y и x < y между элементами выборок.

Пусть $n_x > 3$, $n_y > 3$ и суммарный объем обеих выборок не превосходит 25. Проверка гипотезы осуществляется поэтапно:

- из выборок исключаются одинаковые элементы (вероятность совпадения элементов весьма невелика, поэтому число исключаемых членов выборок не будет большим);
- на основе элементов обеих выборок строится общий вариационный ряд, индексы и конкретные значения элементов можно опустить. В результате получится просто последовательность букв *y* и *x*, например *xxxyxyyxxxyyy*;
- подсчитывается сумма порядковых номеров *u* вариант первой (меньшей по объему) выборки. В приведенном примере $n_x > n_y$ ($n_x = 7$ и $n_y = 6$), поэтому первой будем считать выборку для величины *Y*. Буква *y* встречается на четвертом, шестом, седьмом, одиннадцатом, двенадцатом и тринадцатом местах, следовательно

$$u = 4 + 6 + 7 + 11 + 12 + 13 = 53.$$

Случайная величина *u* имеет распределение Вилкоксона. Для нее построена специальная таблица нижних критических точек распределения (Приложение)

- по таблице критических точек для $n_y = 6$, $n_x = 7$, заданного уровня значимости, например $a = 0{,}05$ (критическая область двусторонняя, следовательно, каждая сторона критической области соответствует уровню значимости $a/2 = 0{,}025$), определяется нижняя критическая точка ин. В данном случае $u_н = 27$;
- вычисляется верхняя критическая точка $u_в = (n_y + n_x + 1)n_y - u_н$. Для рассматриваемого примера

$$u_в = (6 + 7 + 1)\cdot 6 - 27 = 57;$$

- если $u < u_н$ или $u > u_в$, то нулевую гипотезу отвергают. В противном случае нет оснований для отклонения нулевой гипотезы. В приведенном примере нулевая гипотеза об однородности выборок принимается.

Сумма порядковых номеров вариант первой выборки с увеличением общего объема выборок стремится к нормальному распределению. *Нормальное распределение можно применять, если $n_x > 3$, $n_y > 3$ и объем хотя бы одной из выборок превосходит 25*. В таком случае значение нижней критической точки величины *u* при $n_x \cdot n_y$

$$u_n = \frac{(n_x + n_y + 1)n_y - 1}{2} - z_{(1-\alpha/2)}\sqrt{\frac{(n_x + n_y + 1)n_x n_y}{12}} \qquad (5.2)$$

где $z_{1-a/2}$ – квантиль уровня $1-a/2$ стандартизованной нормальной случайной величины.

Остальные этапы проверки ничем не отличаются от рассмотренных выше, применительно к малому объему слоев. В результате выполнения рассмотренных процедур однородные выборки будут объединены.

14.2.2. Объединение неоднородных выборок

Одним из простых и рациональных способов слияния является линейное объединение оценок показателей независимо от степени однородности имеющейся информации. При таком способе объединения неоднородной информации общая выборка рассматривается как смесь из *m* выборок однотипных наблюдений, каждая из которых имеет свои значения показателей. Подобное объединение возможно для несмещенных выборочных средних оценок (типа центральных моментов распределения, вероятностей свершения событий).

Пусть имеются выборочные средние оценки q_i отдельных слоев. Задача состоит в нахождении функции $\Theta = z(\theta_1, \ldots, \theta_m)$, которая была бы лучшей, в смысле принятого критерия, объединенной оценкой Θ^* параметра Θ. Типичным критерием оптимальности оценки является минимум дисперсии оценки. В качестве оценочной функции можно взять любую, но использование сложных функций вызывает трудно преодолимые препятствия по нахождению несмещенных и эффективных оценок. Лучше взять простую линейную комбинацию $\Theta = \sum_{i=1}^{m} v_i \Theta_i$ Коэффициенты выбирают из условия $\sum_{i=1}^{m} v_i = 1$ что обеспечивает получение несмещенной объединенной оценки. Значения коэффициентов v_i, обеспечивающие минимум дисперсии искомой оценки $m_2(\Theta) = \sum_{i=1}^{n} v_i^2 m_2(\Theta_i)$ равны

$$\upsilon_i = m_2^{-1}(\Theta) \Big/ \left(\sum_{i=1}^{n} \frac{1}{m_2(\Theta)} \right).$$

Применение рассмотренного подхода предполагает знание дисперсий оценок, которые, как правило, неизвестны. Замена дисперсии ее выборочной оценкой приводит к трудно оцениваемому смещению величины Θ^*. Преодоление данного недостатка возможно на основе объединения выборок с учетом доли каждой выборки в общем объеме имеющихся сведений, т.е. коэффициенты υ_i характеризуют относительный вклад каждого слоя в общую оценку. Значение коэффициента υ_i можно определить как отношение объема данной выборки к общему объему всех. Линейное объединение оценок приводит к их усреднению по всем выборкам. Иначе говоря, значение некоторого показателя в данном случае следует рассматривать как среднее значение случайной величины, принимающей значение Θ_i с вероятностью u_i.

Пример 6.1. По результатам наблюдения за пропускной способностью канала в различные дни испытаний сформированы упорядоченные выборки, табл. 5.1. При уровне значимости $a = 0{,}05$ необходимо проверить однородность выборок.

Решение. Возьмем в качестве исходной выборку *X*, соответствующую первому дню испытаний, и проверим ее на однородность с выборкой *Y*, составленной из результатов второго дня испытаний.

Перечислим последовательность элементов в общем вариационном ряду, составленном из элементов первой и второй выборки: *yyxyxxyxyxy*.

Таблица 5.1

День испытаний	Пропускная способность						
	1	2	3	4	5	6	7
1	259,14	260,06	260,97	262.43	267.83	273.14	
2	253.68	258.14	259.49	260.18	263.65	271.39	274.12
3	256.36	259.36	262.84	265.94	270.33	270.44	271.63

Сумма порядковых номеров вариант первого дня испытаний ($n_1<n_2$) составит

$$u=3+5+7+8+10+12 =45.$$

Количество элементов в обеих выборках меньше 25, поэтому следует воспользоваться распределением Вилкоксона для проверки гипотезы H_0 об однородности выборок. Значение нижней критической точки для двусторонней критической области при заданном уровне $a/2=0{,}025$, количестве наблюдений $n_1=6$, $n_2=7$ определим по табл. П.5. Оно составит

$$u_н = 27.$$

Значение верхней критической точки распределения равно

$$u_в = (n_1 + n_2 + 1)n_1 - u_н = (6 + 7 + 1)6 - 27 = 57.$$

Значение величины u превышает $u_н$ и меньше $u_в$, поэтому нет оснований отвергать нулевую гипотезу об однородности выборок. Обозначим объединенную выборку через *X*.

Проверим однородность объединенной выборки X и результатов третьего дня наблюдений *W*. Построим общий вариационный ряд из элементов выборки X и выборки *W*:

xwxxwxxxxxwxwxwwxwxx.

Сумма порядковых номеров вариант третьего дня испытаний (этих вариант меньше, чем в объединенном ряду *X*) составит

$$u=2+5+11+13+15+16+18=80.$$

Воспользуемся распределением Вилкоксона и определим при уровне значимости $a/2=0{,}025$, $n_1=7$, $n_2=13$ нижнюю критическую точку $u_н = 48$, табл. П.5. Верхняя критическая точка

$$u_в = (7+13+1)7 - 48=99.$$

В соответствии с выбранным критерием нет оснований отвергать нулевую гипотезу, следовательно, все три выборки однородны и их можно объединить в одну.