15. ДИСПЕРСИОННЫЙ АНАЛИЗ

15.1. Задачи дисперсионного анализа

При исследовании однотипных величин возникают задачи их сравнения. Сравнение случайных величин производится путем сопоставления законов распределения или их моментов.

Законы распределения можно сопоставить на основе критерия Вилкоксона при нулевой гипотезе H_0 о равенстве законов распределения двух случайных величин $F_x = F_y$ и конкурирующей гипотезе H_1 в виде: $F_x < F_y$ или $F_x > F_y$. В этих случаях критическая область является односторонней. Поэтому нижнюю критическую точку и квантиль распределения находят при уровне значимости a. Содержание остальных этапов проверки гипотез сохраняется. Следует отметить, принятие гипотезы H_1 о том, что

$$F_x < F_y, \text{ означает } X > Y.$$

Действительно, неравенство $F_x(x) < F_y(x)$ равносильно неравенству

$$P(X < x) < P(Y < x),$$

следовательно, $X > Y$.

Аналогично, если справедлива гипотеза $F_x > F_y$, то $X < Y$.

Вполне естественно сопоставление случайных величин на основе моментов проводить путем сравнения их математических ожиданий. *Однофакторный дисперсионный анализ* позволяет установить, оказывает ли существенное влияние некоторый фактор Ф, который имеет несколько уровней, на исследуемую случайную величину.

Задача сравнения выборок случайных величин формулируется следующим образом.

Имеются результаты наблюдений в виде совокупности слоев типа (6.1), задан уровень значимости a для проверки статистической гипотезы. В данном случае отдельные слои трактуются как выборки одной и той же случайной величины, полученные по результатам наблюдения за одним объектом при различных значениях фактора Ф (количество уровней фактора равно m).

Требуется проверить нулевую гипотезу H_0 о равенстве математических ожиданий случайных величин всех выборок.

Допущения: генеральные совокупности, соответствующие каждому слою, распределены *нормально*; *дисперсии* слоев *одинаковы*,

Основная идея дисперсионного анализа состоит не в сопоставлении математических ожиданий случайных величин, а в сравнении оценки "*факторной дисперсии*", порождаемой воздействием фактора, и оценки "*остаточной дисперсии*", обусловленной случайными причинами. Если различие между этими оценками значимо, то фактор оказывает существенное влияние на случайную величину, в противном случае влияние фактора несущественно.

Дисперсионный анализ выполняется поэтапно. Такими этапами являются следующие:

- проверка выборок на принадлежность к нормальному закону распределения. Этап необходим, когда нет априорной информации о законах распределения слоев. Если принадлежность нормальному закону не подтвердится, то аппарат дисперсионного анализа, вообще говоря, применять нельзя. Некоторые исследователи допускают его применение при больших объемах выборок (объем каждой выборки должен быть не менее 30) независимо от вида закона распределения;
- проверка равенства оценок дисперсий во всех слоях выборки (проверка однородности дисперсий). Если однородность не подтвердится, то методы дисперсионного анализа не применимы;
- вычисление оценки факторной и остаточной дисперсии;
- сравнение средних значений величин методом дисперсионного анализа и формирование выводов по результатам сравнения.

15.2. Проверка однородности совокупности дисперсий

Для каждого слоя вычисляется несмещенная оценка дисперсии, обозначим эти оценки через $S_0^2(x)$, S_0^2 (y), …, S_0^2 (w) соответственно. Числа степеней свободы этих оценок

$$k_1 = n_1 - 1, \quad k_2 = n_2 - 1, \ldots, \quad k_m = n_w - 1.$$

Гипотеза H_0 состоит в том, что выборки, по которым определены оценки дисперсии, получены из генеральных совокупностей, обладающих одинаковыми дисперсиями

$$S_0^2(x) = S_0^2(y) = \ldots = S_0^2(w) = S_0^2,$$

при этом величина дисперсии S_0^2 остается неизвестной. Следует выяснить, являются ли величины $S_0^2(x)$, S_0^2 (y), …, S_0^2 (w) оценками одной и той же генеральной дисперсии μ_2.

Рассмотрим сначала случай, когда объем выборок по слоям хотя бы частично различается. В такой ситуации применяется критерий однородности Бартлетта. Проверка однородности реализуется в несколько шагов.

Вычисляется усредненная оценка несмещенной дисперсии по всем слоям

$$S_0^2 = \sum_{i=1}^{m} k_i S_0^2(i)/k, \qquad k = \sum_{i=1}^{m} k_i, \tag{6.1}$$

где S_0^2(i) – несмещенная оценка дисперсии для слоя i.

Рассчитывается значение критерия

$$B = \frac{2.303\left[k \lg \mu_2 - \sum_{i=1}^{m} k_i \lg \mu_2(i)\right]}{1 + \frac{1}{3(m-1)}\left[\sum_{i=1}^{m} \frac{1}{k_i} - \frac{1}{k}\right]} \qquad (6.2)$$

Бартлетт установил, что случайная величина B при условии справедливости нулевой гипотезы распределена приближенно как хи-квадрат с $m-1$ степенями свободы, если все n_i больше трех. По заданному уровню значимости a, числу степеней свободы $m-1$ для правосторонней критической области определяется критическое значение $c^2_{кр}(m-1;a)$. Если соблюдается условие

$$B < c^2_{кр}(m-1;a),$$

то нет оснований отвергнуть нулевую гипотезу. Если $B > c^2_{кр}(m-1;a)$, то нулевая гипотеза отвергается. Критерий Бартлетта чувствителен к отклонениям распределения от нормального, поэтому к результатам сравнения следует относится осторожно, а при одинаковом объеме всех слоев вместо критерия Бартлетта лучше применять критерий Кочрена (Кохрена).
Итак, если $k_1 = k_2 = \ldots = k_m$, то применяется *критерий Кочрена*

$$G = S^2_{0\max} \Big/ \sum_{i=1}^{m} S_0^2(i), \qquad (6.3)$$

где S_0^2 – максимальная оценка дисперсии по всем слоям.

Критическая область для критерия Кочрена правосторонняя. Критическую точку $G_{кр}(k_1,m;a)$ находят по таблице распределения Кочрена, (Приложение). Критическая область определяется неравенством $G > G_{кр}(k_1,m;a)$.

15.3. Сравнение факторной и остаточной дисперсий

Пусть все выборки (6.1) характеризуют одну случайную величину X при различных значениях фактора Ф, т.е. каждый слой соответствует одному количественному или качественному значению фактора. Сравнение дисперсий производится в следующем порядке:

- рассчитывается среднее значение (оценка математического ожидания) по всей совокупности наблюдений

$$m^* = \frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij},$$

где $n = n_1 + n_2 + \ldots + n_m$, а x_{ij} – j-й *элемент* i-го слоя;

- вычисляются средние значения для всех слоев (групп)

$$m_{гpi} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}, \quad i = 1,\ldots,m;$$

- определяется общая сумма квадратов отклонений наблюдаемых значений от оценки математического ожидания

$$S_{общ} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(x_{ij} - m^*)^2; \qquad (6.4)$$

- определяется факторная сумма квадратов отклонений средних по слоям от оценки математического ожидания (характеризует рассеяние между слоями)

$$S_{факт} = \sum_{i=1}^{m} n_i(\mu_{гpi} - m^*)^2; \qquad (6.5)$$

- определяется остаточная сумма квадратов отклонений наблюдаемых значений внутри слоя от своей средней

$$S_{ост} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(m^*_{гpi} - x_{ij})^2. \qquad (6.6)$$

Величина $S_{факт}$ характеризует влияние фактора Ф. Это положение можно пояснить следующим образом. Пусть фактор оказывает существенное влияние на величину X. Тогда результаты наблюдения для одного слоя, вообще говоря, отличаются от результатов, представленных в других слоях. Следовательно, различаются и средние значения по слоям, причем они тем больше отличаются от оценки математического ожидания по всей выборке, чем больше проявляется влияние фактора. Таким образом, сумма квадратов отклонений средних по слоям от общей средней и характеризует влияние фактора (возведение отклонений во вторую степень исключает взаимную компенсацию положительных и отрицательных отклонений).

Наблюдения внутри одного слоя различаются из-за воздействия случайных причин. Именно сумма квадратов отклонений наблюдаемых значений в каждом слое от среднего значения в слое и характеризует воздействие этих причин, т.е. величина $S_{ост}$ отражает суммарное влияние случайных причин на значение величины X.

Величина $S_{общ}$, как сумма квадратов отклонений конкретных значений от среднего значения, характеризует суммарное влияние фактора и случайных причин. Можно показать, что

$$S_{общ} = S_{ост} + S_{факт},$$

тогда для вычисления остаточной суммы квадратов можно воспользоваться более простым соотношением

$$S_{ост} = S_{общ} - S_{факт}.$$

Разделив суммы квадратов отклонений на соответствующее число степеней свободы, получим оценки общей, факторной и остаточной дисперсий:

$$S^2_{0,общ} = \frac{S_{общ}}{n-1}; \quad S^2_{0,фак} = \frac{S_{факт}}{m-1}; \quad S^2_{0,ост} = \frac{S_{ост}}{n-m}. \tag{6.7}$$

Если средние значения случайной величины, вычисленные по отдельным выборкам одинаковы, то оценки факторной и остаточной дисперсий являются несмещенными оценками генеральной дисперсии и различаются несущественно. Тогда сопоставление оценок этих дисперсий по критерию Р. Фишера

$$F = S_0{}^2_{факт} / S_0{}^2_{ост}$$

должно показать, что нулевую гипотезу о равенстве факторной и остаточной дисперсий отвергнуть нет оснований. Если $\mu_{2факт} < \mu_{2ост}$, то нет необходимости прибегать к вычислению критерия Р. Фишера – из неравенства сразу следует вывод о выполнении нулевой гипотезы. Итак, из справедливости гипотезы о равенстве средних величин по группам следует соблюдение гипотезы о равенстве факторной и остаточной дисперсий.

Если нулевая гипотеза о равенстве средних величин по слоям является ложной, то с увеличением расхождения между слоями возрастает оценка факторной дисперсии, а вместе с ней и величина критерия $F = \mu_{2факт} / \mu_{2ост}$. В результате значение F превысит критическое значение, и гипотеза о равенстве дисперсий будет отвергнута.

Рассуждая от противного, можно доказать справедливость утверждений: из справедливости (ложности) гипотезы о дисперсиях следует истинность (ложность) гипотезы о математических ожиданиях. Таким образом, вместо проверки нулевой гипотезы H_0 о равенстве средних значений для совокупности выборок следует проверить гипотезу о равенстве факторной и остаточной дисперсий.

Пример 6.2. Методом дисперсионного анализа при уровне значимости 0,05 проверить нулевую гипотезу о равенстве средних значений по слоям, применительно к результатам наблюдений, табл. 6.1. Предполагается, что выборки принадлежат нормальному распределению, а каждый слой соответствует некоторому значению фактора Ф.

Решение. Необходимо проверить однородность дисперсий, а затем непосредственно провести дисперсионный анализ. Проверим гипотезу об однородности дисперсий. Для этого вычислим:

- оценки математического ожидания по слоям (групповые средние)

$$m_{гр\ 1} = 263{,}93; \quad m_{гр\ 2} = 262{,}95; \quad m_{гр\ 3} = 265{,}32;$$

- несмещенные оценки дисперсии по слоям

$$m^*_2\ (1) = 29{,}79; \ m^*_2\ (2) = 54{,}20; \ m^*_2\ (3) = 34{,}61;$$

- усредненную оценку несмещенной дисперсии по всем слоям

$$m^*_2 = (29{,}79 \times 5 + 54{,}20 \times 6 + 34{,}61 \times 6)/17 = 40{,}11.$$

- значение критерия Бартлетта

$$B = a/c = 0{,}56/1{,}08 = 0{,}52,$$

где $a = 2{,}303 \cdot (17 \cdot \lg 40{,}11 - (5 \cdot \lg 29{,}79 + 6 \cdot \lg 54{,}20 + 6 \cdot \lg 34{,}61)) = 0{,}56;$

$c = 1 + (1/5 + 1/6 + 1/6 - 1/17) / [3(3 - 1)] = 1{,}08.$

Критическое значение хи-квадрат для правосторонней области

$$c^2_{кр}(2;\ 0{,}05) = 6{,}0.$$

Поскольку величина B меньше $c^2_{кр}(2;\ 0{,}05)$, отвергнуть нулевую гипотезу об однородности дисперсий нет оснований.

Дисперсионный анализ предусматривает вычисление:

- суммы квадратов

$$S_{общ} = 701{,}65; \quad S_{факт} = 19{,}81; \quad S_{ост} = 681{,}84;$$

- оценок дисперсий

$$\mu^*_{2\ общ} = 701{,}65/19 = 36{,}93; \quad \mu^*_{2\ факт} = 19{,}81/2 = 9{,}91; \quad \mu^*_{2\ ост} = 681{,}84/17 = 40{,}10.$$

Оценка факторной дисперсии меньше оценки остаточной дисперсии, поэтому можно сразу утверждать справедливость нулевой гипотезы о равенстве математических ожиданий по слоям выборки, Иначе говоря, в данном примере фактор Ф не оказывает существенного влияния на случайную величину.