Лекция 11. Идентификация случайных закономерностей.

Исходным материалом для идентификации законов распределения являются экспериментальные данные, полученные в результате большого числа наблюдений и образующие выборку заданного объема.

Идентификация числовых характеристик выборки

Среднее арифметическое, или просто среднее, одно из основных характеристик выборки, с помощью которого описывается (или идентифицируется) математическое ожидание. Среднее, как и другие числовые характеристики выборки, может вычисляться как по необработанным первичным данным, так и по результатам группировки этих данных. Точность вычисления по необработанным данным всегда выше, но процесс вычисления оказывается трудоемким при большом объеме выборки.

Для несгруппированных данных среднее арифметическое определяется по следующей формуле:

$$\bar{x} = (1/n) \sum_{j=1}^{n} x_i, \qquad (64)$$

где n – объем выборки, x_i – элементы выборки.

Если данные сгруппированы, то

$$\bar{x} = (1/n) \sum_{i=1}^{k} n_i x_j, \qquad (65)$$

где n_i – частоты интервалов, k – число интервалов группировки, x_i – срединные значения интервалов.

Среднее арифметическое, вычисленное по формуле (65), называют также взвешенным средним, так как x_i суммируются с весами, равными частотам попадания в интервалы группировки.

Для идентификации дисперсии используется выборочная дисперсия S^2, которая вычисляется по следующим формулам:

$$S^2 = \frac{1}{n-1} \sum_{j=1}^{n} (x_i - \bar{x})^2 \qquad (66)$$

для несгруппированных данных;

$$S^2 = \frac{1}{n-1} \sum_{j=1}^{n} n_i (x_i - \bar{x})^2 \qquad (67)$$

для сгруппированных данных.

Формулы (66) и (67) неудобны с вычислительной точки зрения, поэтому на практике используются другие расчетные формулы, более удобные как для ручных расчетов, так и для вычислений на компьютере.

Для несгруппированных данных:

$$S^2 = \frac{1}{n-1} \left(\sum_{j=1}^{n} x_i^2 - n\bar{x}^2 \right) \qquad (68)$$

или

$$S^2 = \frac{1}{n-1} \left[\sum_{j=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2 / n \right] \qquad (69)$$

Если данные сгруппированы, то

$$S^2 = \frac{1}{n-1}\left(\sum_{j=1}^{n} n_i x_i^2 - n\bar{x}^2 \right) \tag{70}$$

или

$$S^2 = \frac{1}{n-1}\left[\sum_{j=1}^{n} n_i x_i^2 - \left(\sum_{i=1}^{n} n_i x_i \right)^2 / n \right] \tag{71}$$

Формулы (68) и (70) применяются для определения дисперсии, если $\bar{x}$ уже вычислено. Формулы (69) и (71) используются в тех случаях, когда $\bar{x}$ и S^2 вычисляются одновременно.

Для идентификации коэффициентов корреляции между парой случайных величин используется выборочный коэффициент корреляции Бравайса-Пирсона

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\cdot \left[\sum_{i=1}^{n} (x_i - \bar{x})^2 \right] \cdot \left[\sum_{i=1}^{n} (y_i - \bar{y})^2 \right]}}.$$

Для практических расчетов более удобной является формула

$$r = \frac{n \sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i \right) \cdot \left(\sum_{i=1}^{n} y_i \right)}{\sqrt{\cdot \left[n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2 \right] \cdot \left[n \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i \right)^2 \right]}}.$$

Значения полученного по этой формуле коэффициента корреляции находятся в пределах

$-1 \le r \le +1$.

Идентификация функций плотностей непрерывных случайных величин

Процесс идентификации функций плотности условно можно разбить на шесть этапов.

На первом из них осуществляется группировка элементов заданной выборки объемом *n*, т.е. вся область значений элементов выборки подразделяется на некоторое число непересекающихся интервалов. Число интервалов *k* группировки зависит от объема выборки *n*.

Приведем некоторые рекомендации по выбору числа *k*.

При выборе *k* следует учитывать, что при слишком большом его значении картина распределения искажается случайными зигзагами частот, слишком малочисленными при узких интервалах. При слишком малом числе интервалов сглаживаются и затушевываются характерные особенности распределения. Поэтому желательны многовариантные расчеты с разным числом интервалов.

Для определения числа *k* можно использовать, например, следующие выражения:

1) $k = 1 + 3{,}32 \lg(n)$ (формула Стерджеса);

2) $k = 5 \lg(n)$ и $6 \le k \le 20$;

3) n/k = 50/8, 100/10, 500/13, 1000/15, 10000/20;

4) $k = \min\left(\sqrt{n},\ 30\right)$.

Следующим этапом является определение длины и границ интервалов группировки. Если интервалы группировки берутся одинаковыми, то их длина вычисляется по формуле

$$D = \frac{1{,}02\left(X_{\max} - X_{\min}\right)}{k},$$

где X_{max}, X_{min} – максимальные и минимальные значения выборки.

Границы отдельных интервалов определяются следующим образом:

$$\left[X_{\min} + (j-1)\alpha - \Delta, X_{\min} + j\alpha - \Delta\right], \quad j = \overline{1,k},$$

где j – номер интервала, а $\Delta = 0{,}01\alpha$.

Длины интервалов группировки можно выбрать и по другому принципу из условия равновероятности попадания реализаций случайной величины во все интервалы. При постороении схемы разбиения, основанной на некотором ожидаемом законе распредения, необходимо рассматривать не область значений, заданных выборкой, а теоретическую область возможных значений случайной величины с ожидаемым законом распределения.

На третьем этапе идентификации плотности распределения подсчитываются относительные частоты $\nu_i = \frac{n_i}{n}$, $i = \overline{1,k}$, наблюдений, попавших во все интервалы, и строится гистограмма данных (распределение относительных частот).

Четвертый этап является самым ответственным. Здесь по виду полученной гистограммы подбираются подходящие к данному случаю теоретические распределения, которых может быть и несколько.

На пятом этапе путем сравнения выборочных числовых характеристик (среднее арифметическое, выборочная дисперсия) с их теоретическими значениями выбирается наиболее подходящее теоретическое распределение.

Однако окончательный результат идентификации закона распределения можно получить только на шестом этапе по итогам проверки выбранного распределения с помощью одного из критериев согласия.

Идентификация законов распределения дискретных случайных величин

Схема идентификации законов распределения дискретных случайных величин не имеет принципиальных отличий от схемы для непрерывных величин. К непринципиальным различиям этих схем можно отнести следующие:

1. Число группировок k здесь совпадает с числом всех возможных значений дискретной случайной величины. Поэтому схема идентификации начинается с третьего этапа.

2. При невозможности подбора подходящего теоретического закона распределения можно ограничиться табличной формой задания закона распределения дискретной случайной величины. В качестве оценок вероятностей используются относительные частоты появления каждого из возможных значений дискретной случайной величины.

Оценка результатов идентификации

Для статистической оценки гипотезы о том, что совокупность эмпирических данных незначительно отличается от той, которую можно ожидать при выбранном теоретическом законе распределения, чаще всего используются критерии согласия Пирсона, Колмогорова-Смирнова, Мизеса. Каждый из этих критериев имеет свои сильные и слабые стороны и относительно выбора между ними можно дать лишь самые общие рекомендации.

Критерий Пирсона эффективен при больших объемах выборки (n >100). Критерий Колмогорова-Смирнова дает хорошие результаты при $10 \le n \le 100$. При объеме выборки меньше 10 удовлетворительные результаты можно получить, пожалуй, только с помощью критерия Мизеса. При использовании критериев Пирсона и Колмогорова-Смирнова необходимо задать число интервалов группировки. В случае применения критерия Пирсона это число определяется из условия, чтобы в каждый интервал попало не менее 5 экспериментальных точек. В то же время в случае использования критерия Колмогорова-Смирнова данные можно как группировать, так и относить каждое наблюдение к отдельной группе. Это условие открывает возможность эффективного анализа при малых выборках.

Учитывая, что наибольшее распространение среди критериев согласия получил критерий Пирсона (критерий χ^2), рассмотрим его более подробно.

Пусть имеется выборка из n реализаций x_i, $i = \overline{1,n}$ случайной величины η с искомым законом распределения. По результатам группировки элементов заданной выборки уже

вычислены все относительные частоты ν_j, $j=\overline{1,k}$ и соответствующие им вероятности p_j, $j=\overline{1,k}$ подобранного теоретического (гипотетического) распределения.

Выдвигается гипотеза о том, что случайная величина η действительно имеет подобранный в результате идентификации закон распределения. Чтобы проверить эту гипотезу, надо выбрать некоторую меру расхождения статистического распределения с гипотическим.

В качестве такой меры R при использовании критерия Пирсона берется сумма квадратов отклонений ($\nu_i - p_j$) статистических частот от гипотетических вероятностей p_j, взятых с некоторыми весами c_j:

$$R = \sum_{j=1}^{k} c_j (\nu_j - P_j)^2 . \qquad (72)$$

Коэффициенты c_j вводятся потому, что отклонения, относящиеся к разным значениям p_j, нельзя считать равнозначными, так как одно и то же по абсолютной величине отклонение ($\nu_i - p_j$) может быть малозначительным, если сама вероятность p_j велика, и очень заметным, если она мала. Поэтому веса c_j необходимо взять обратно пропорциональными вероятностям p_j. Пирсон доказал, что если положить

$$c_j = \frac{n}{p_j},\ j=\overline{1,k},$$

то при большом числе опытов n закон распределения величины R обладает следующими свойствами: он практически не зависит от закона распределения случайной величины η и мало зависит от числа опытов n, а зависит только от числа k (число интервалов группировки) и при увеличении n приближается к распределению χ^2 с функцией плотности

$$f_k(r) = \left[2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right) \right]^{-1} r^{\frac{k}{2}-1} e^{\frac{r}{2}}.$$

При таком выборе коэффициентов c_j мера расхождения R обычно обозначается через χ^2

$$R = \chi^2 = \frac{n \sum_{j=1}^{k} \left(\nu_j - P_j\right)^2}{P_j},$$

или, учитывая, что $\nu_j = n_j/n$, где n_j – число элементов выборки в j-м интервале, получим

$$R = \chi^2 = \frac{n \sum_{j=1}^{k} \left(n_j - nP_j\right)^2}{nP_j}. \qquad (73)$$

Распределение χ^2, как известно, зависит от параметра l, называемого "числом степеней свободы". При использовании критерия Пирсона число степеней свободы полагается равным числу интервалов k минус число независимых условий (связей), наложенных на частоту ν_j. Примерами таких условий могут быть $\sum_{j=1}^{k} x_j \nu_j = m_x$ если требуется совпадение статистического среднего с гипотетическим математическим

ожиданием, или $\sum_{j=1}^{k} \nu_j = 1$ если сумма частот была равна 1 (это требование накладывается во всех случаях) и т.д.

Принцип оценки результата индентификации закона распределения с помощью критерия $\aleph^2$ основывается на теореме 8 (Пирсона К.). Каковы бы ни были исходная случайная величина η и выбранное число интервалов k (такое, что $p_j > 0$, $j = \overline{1,k}$), при каждом $r > 0$ имеет место

$$\lim_{n \to \infty} P\{\aleph^2 \geq r\} = \int_r^{\infty} f_{k-1}(u)du .$$

Пусть α – значение такого уровня значимости, что событие с вероятностью α считается уже практически невозможным. Тогда, решая уравнение

$$\int_r^{\infty} f_{k-1}(u)du = \alpha , \qquad (74)$$

найдем значение $r = \aleph_\alpha^2$, отвечающее фиксированному нами уровню значимости α. Тогда при выполнении условия $R = \aleph^2 < \aleph_\alpha^2$ результат осуществленной идентификации закона распределения случайной величины η принимается, а если

$$R = \aleph^2 \geq \aleph_\alpha^2 , \qquad (75)$$

то этот результат отвергается.

Указанный вывод, очевидно, зависит от выбранного уровня значимости и поэтому не имеет абсолютного характера. Чаще других в качестве α принимают значения 0,1; 0,05; 0,01; 0,005.

На практике для определения значений $\aleph_\alpha^2$ используются таблицы, в которых приведены решения интегрального уравнения (74) для всех возможных значений α и (k-1). Эти таблицы приведены во всех учебниках по теории вероятностей и математической статистике.

Приведенная схема оценки результатов идентификации закона распределения по критерию Пирсона представим в виде следующего алгоритма:

Шаг 1. Вычисление вероятностей p_i, i = 0, 1, ..., k по гипотетическому закону распределения.

Шаг 2. Вычисление меры $R = \aleph^2$.

Шаг 3. Выбор величины уровня значимости $\alpha_m = \max\{\alpha_1, \alpha_2, \ldots, \alpha_l\}$.

Шаг 4. Проверка условия $\alpha_m > 0$. При нарушении этого условия переход на шаг 9.

Шаг 5. Определить $\aleph_\alpha^2 = R(l, \alpha)$.

Шаг 6. Проверить условие $\aleph^2 < \aleph_\alpha^2$. При его выполнении переход на шаг 8.

Шаг 7. Положить $\alpha_m = 0$. Возврат на шаг 3.

Шаг 8. Вывод текста "Выбранный закон распределения отвергается".

Шаг 10. Конец.

Контрольные вопросы:

1. Какие критерии применяются для оценки результатов идентификации?

2. Из каких этапов состоит процесс идентификации законов распределения непрерывных случайных величин?

3. Приведите блок-схему алгоритма оценки результатов идентификации закона распределения по критерию Пирсона.