Математические методы первичной обработки статистических данных, определяющих состояние факторов риска

Бердышев О.В., кандидат педагогических наук, доцент, Пермский национальный исследовательский политехнический университет, Пермь, Россия
Долинов А.Л., кандидат технических наук, доцент, Пермский национальный исследовательский политехнический университет, Пермь, Россия

Аннотация. Представленная статья посвящена вопросам подготовки собранных значений исследуемых факторов риска к дальнейшей их оценке и управлению. Включает в себя процедуры проверки выбранного закона распределения значений на адекватность, оценку наличия среди собранных значений так называемых «ложных» и оценку случайности собранных данных.

Ключевые слова: значения факторов риска, закон распределения, ложность, случайность.

Mathematical methods of primary treatment of the statistical data defining a condition of risk factors

Berdyshev O.V., candidate of Pedagogical Sciences, Associate Professor, Perm national research polytechnic university, Perm, Russia
Dolinov A.L., candidate of Technical Sciences, Associate Professor, Perm national research polytechnic university, Perm, Russia

Annotation. Work is devoted to questions of preparation of collected values of the studied risk factors for further assessment and managements. Includes assessment of the choice of value distribution on adequacy, existence assessment among collected values so-called «chance» and assessment of accident of collected data.

Keywords: values of risk factors, distribution law, falsity, randomness.

Введение

Исследование проблем, связанных с возникающими рисковыми ситуациями, актуально всегда и везде пока есть в практической деятельности человека хоть какая-то неопределенность. Сами рисковые ситуации корректно анализируемы лишь тогда, когда их можно измерить. Причем измерение должно быть объективным. Фактически результаты измерений факторов риска представляют собой совокупности статистических данных. И на основе анализа этих данных риски оценивают, ими пытаются управлять. Однако, прежде чем работать с данными необходимо выяснить, а действительно ли они могут быть обрабатываемы имеющейся совокупностью методов.

Вопросам проверки собранных данных на возможность их статистической обработки и вопросам подготовки этих данных к обработке, если в этом есть необходимость посвящена настоящая работа.

В данной статье приводятся процедуры, обеспечивающих достоверную количественную оценку значений факторов риска и корректность методов управления ими.

К этим процедурам относятся следующие:

1) проверка соответствия закона распределения значений факторов риска заявленному;

2) выявление и исключение значений факторов риска резко отличающихся от основной массы;

3) проверка значений факторов риска на случайность их формирования в исследуемой совокупности.

Результаты исследования

1. Проверка соответствия закона распределения значений факторов риска заявленному

Еще на этапе идентификации рисков необходимо осуществить так называемую первичную (предварительную) обработку данных.

Это утверждение является следствием того, что исходные числовые данные по рискам являются данными статистическими и к ним для корректного

использования методов статистической обработки предъявляются определенные требования. Одним из наиболее распространенных требований является условие нормальности распределения значений факторов риска.

Подавляющая совокупность критериев, используемых при исследовании значений параметров факторов риска, ориентированы именно на этот закон распределения. Однако однозначно утверждать, что значения распределены нормально, разумеется, нельзя.

Предположения относительно закона распределения вырабатываются, обосновываются и формулируются изначально экспертно на основе теоретических предположений о распределении значений факторов риска, однако существует математическая методика объективной оценки рассматриваемого закона распределения. Такой методикой является критерий Пирсона (χ^2).

Выбор закона распределения следует начать с разбиения диапазона значений на частичные интервалы.

Для этого сначала находят минимальное (x_{min}) и максимальное (x_{max}) из значений факторов риска, а затем весь промежуток исходных данных [x_{min}, x_{max}] разбивают на частичные интервалы. Длина частичного интервала вычисляется по следующей формуле:

$$\Delta x = \frac{x_{\max} - x_{\min}}{1 + 3{,}3 \lg n}.$$

Число частичных интервалов (r) в дальнейшем может быть уменьшено, за счет их объединения. Объединение осуществляется посредством присоединения к интервалам, в которые попало относительно большое количество значений факторов риска интервалов смежных с ними. Смежные интервалы присоединяются если в них не попало ни одного значения, или попало значений относительно мало.

Количество значений факторов риска, попавших в интервал номер i обозначим за n_i. На основе величин n_i (абсолютных частот попадания значений факторов риска в интервал номер i) можно построить таблицу абсолютных

частот или гистограмму. Визуальное представление распределения абсолютных частот существенно упрощает процесс выбора соответствующего закона распределения значений факторов риска.

Сигналом нормальности распределения является гистограмма и полигон частот, которые должны иметь вид схожий с так называемой «кривой Гаусса».

В ней наиболее наполненными являются центральные интервалы разбиения, в то время как крайние интервалы содержат преимущественно единичные попадания. Кроме того, кривая Гаусса симметрична.

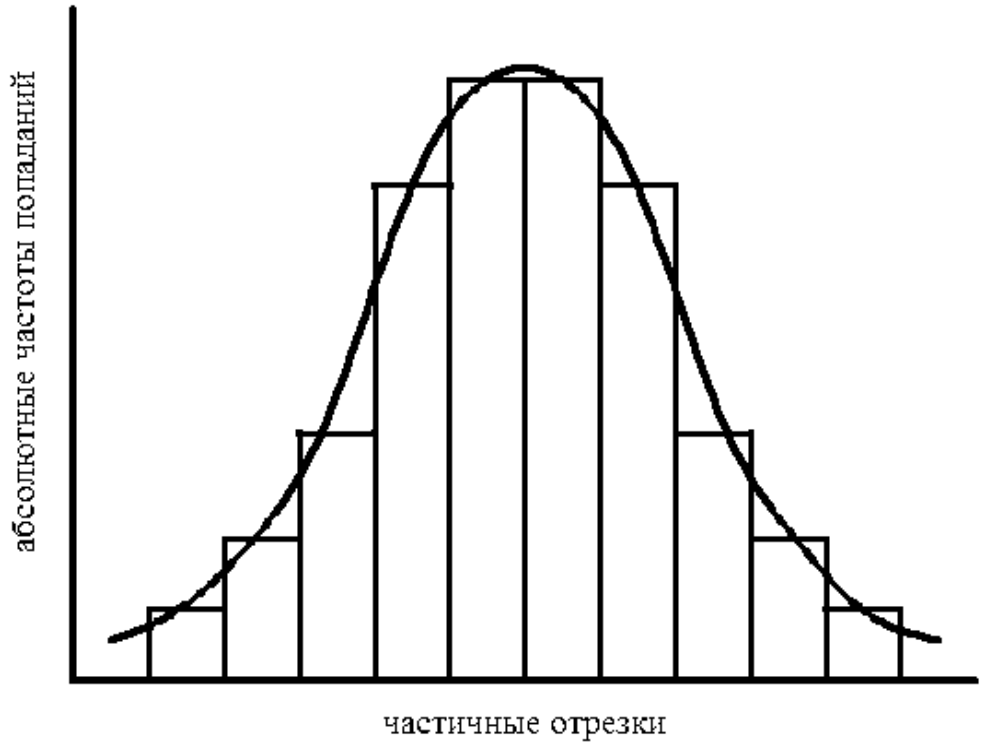

Рис. 1 – Кривая Гаусса

Для проверки согласованности фактического распределения с выбранным задается уровень значимости α – вероятность ошибки.

Обычно α принимается из отрезка [0,01; 0,05].

Сам критерий Пирсона основывается на проверке выполнения следующего неравенства:

$$\chi^2_{расч} < \chi^2_{кр}.$$

Если неравенство выполняется, то с вероятностью 1-α можно утверждать, что выбранный закон распределения значений факторов риска адекватен.

В противном случае следует отвергнуть гипотезу об адекватности выбранного закона распределения.

Расчетное значение критерия ($\chi^2_{расч}$) вычисляется по следующей формуле:

$$\chi^2_{расч} = \sum_{i=1}^{r} \frac{(n_i - n_i')^2}{n_i'}.$$

Здесь n_i' – теоретическое число данных, попавших в i-ый интервал.

Значения n_i' вычисляются по формуле:

$$n_i' = nP(x_i < x < x_{i+1}),$$

где n – объем исходной выборочной совокупности (общее число рассматриваемых значений факторов риска).

Согласно свойствам функции распределения случайной величины, теоретическое число данных можно определить следующим образом:

$$n_i' = nP(x_i < x < x_{i+1}) = n\big(F(x_{i+1}) - F(x_i)\big),$$

где $F(x_{i+1})$ и $F(x_i)$ – значения функции распределения на правой и левой границе частичного интервала номер i.

Выбор конкретного закона распределения значений факторов риска осуществляется именно сейчас. Выбрав закон распределения, исследователь тем самым выбирает соответствующую функцию распределения. После чего вычисляет ее значения в конкретных точках (границах частичных отрезков).

Этот этап для исследователя является особо ответственным, по следующим причинам:

1) экспертный выбор закона распределения требует от лица, принимающего решение достаточно значительного опыта;

2) ошибка на этом этапе неизбежно повлечет за собой ошибки как в оценке рисков, так и в управлении ими.

В частности, если предполагается нормальность распределения, теоретическое значение n_i' вычисляется по следующей формуле:

$$n_i' = n\left[\Phi\left(\frac{x_{i+1} - a}{\sigma}\right) - \Phi\left(\frac{x_i - a}{\sigma}\right)\right].$$

Здесь a – математическое ожидание, σ – среднеквадратическое отклонение значений факторов риска, а $\Phi(x)$ – стандартная функция Лапласа, значения которой можно найти в соответствующих статистических таблицах.

Менее распространенным, чем нормальное, но, тем не менее, тоже часто встречающимся на практике распределением значений факторов риска является экспоненциальное распределение. Оно особенно востребовано в исследовании отказов функционирования технических систем.

Функция распределения значений факторов риска при экспоненциальном законе распределения имеет следующий вид:

$$F(x) = 1 - e^{-\lambda x}.$$

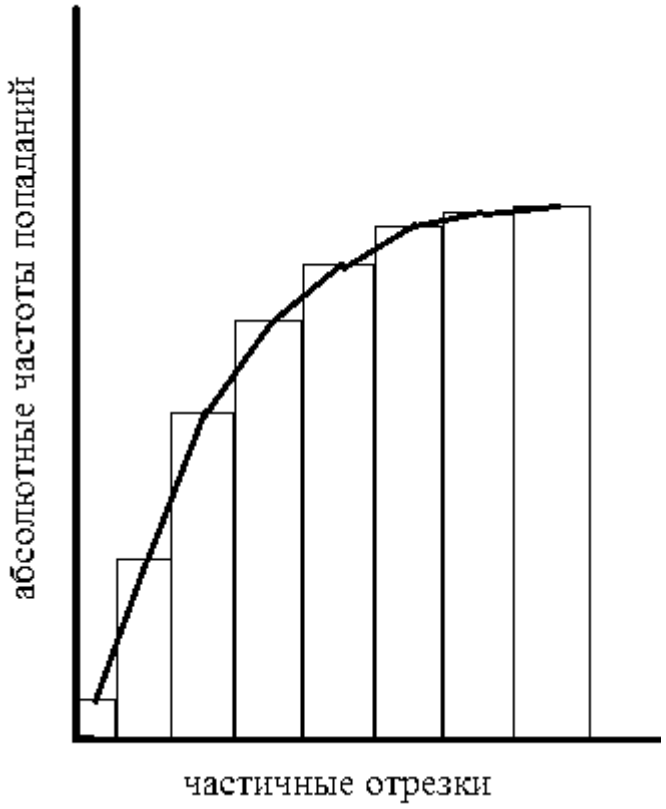

Рис. 2 – Экспоненциальное распределение

Здесь λ – интенсивность появления значений факторов риска. Вычисляется по следующей формуле:

$$\lambda = \frac{1}{a}.$$

При экспоненциальном распределении теоретические значения опытных частот будут вычисляться следующим образом:

$$n_i' = n\left[1 - e^{-\lambda x_{i+1}} - 1 + e^{-\lambda x_i}\right] = n\left[e^{-\lambda x_i} - e^{-\lambda x_{i+1}}\right].$$

Следует отметить, что, кроме выше приведенных законов распределения, достаточной практически ориентированной популярностью пользуется биномиальный закон распределения.

Это распределение предполагает реализацию схемы повторяющихся событий.

Например, заключается N однотипных страховых договоров с принципиально равной вероятностью наступления страхового события в каждой из рассматриваемых рисковых ситуациях, определяемых условиями этих договоров.

Или на конвейер поступает N одинаковых деталей (изделий) с принципиально равной вероятностью брака для каждой конкретной детали (изделия).

Тогда в случае гипотезы о биномиальном распределении значений факторов риска, вероятность того, что среди N случайных событий, рисковое событие произойдет ровно K раз, вычисляется по следующей формуле:

$$P_N(K) = C_N^K p^K q^{N-K}.$$

Здесь p – вероятность возникновения рискового события в каждом конкретном случае, q – вероятность не возникновения рискового события в этом случае, C_N^K – число сочетаний из N по K, вычисляется по следующей формуле:

$$C_N^K = \frac{N!}{(N-K)!K!}.$$

В этом сценарии теоретические абсолютные частоты вычисляются по следующей формуле:

$$n_k' = NC_N^K p^K q^{N-K}.$$

Отметим, что гистограмма и полигон абсолютных частот не будут иметь какого-то особенного, принципиально характерного именно для этого закона распределения, вида.

После вычисления расчетного значения критерия Пирсона, определяется значение критическое ($\chi^2_{кр}$). Критическое значение выбирается из соответствующих статистических таблиц в зависимости от установленного в задаче уровня значимости α и числа k степеней свободы.

Число степеней свободы вычисляется по следующей формуле:

$$k = r - 1 - m,$$

где m – это число параметров предполагаемого распределения.

В частности, у нормального распределения значений факторов риска два параметра (a, σ), у экспоненциального распределения один параметр (λ), у биномиального распределения два параметра (N, p).

Сделаем ряд практических замечаний, востребованных при реализации критерия Пирсона.

1. Для эффективного использования критерия объём выборки значений факторов риска должен быть достаточно большим, так как теоретически критерий справедлив при бесконечно большом объеме совокупности статистических данных. Обычно требуют, чтобы исследуемое множество содержало не менее 150-200 значений.

2. Достаточно большим должен быть не только общий объем n выборочной совокупности, но и величины абсолютных частот n_i попаданий в каждый из частичных интервалов разбиения. В противном случае, как уже было отмечено раннее, интервалы разбиения следует объединять.

2. Выявление и исключение значений факторов риска резко отличающихся от основной массы

После проверки адекватности выбранного закона распределения осуществляется процедура выявления и исключения значений факторов риска резко отличающихся от основной массы.

Следует отметить, представленный ниже математический аппарат ориентирован именно на нормальный закон распределения и при выборе иного распределения использоваться не может.

Итак, в процессе формирования совокупности значений факторов риска возможны ошибки.

Ошибки бывают связаны с неточностью наблюдения за исследуемым процессом, а также с воздействием внешних и внутренних случайных факторов. Ошибки могут быть систематическими и случайными. Систематические ошибки могут возникать из-за сбоя в процедурах измерения или по техническим причинам. Случайные ошибки часто возникают в результате действия человеческого фактора, например, стресса или усталости оператора.

Кроме того, нельзя исключать случай, когда ошибки появляются в результате намеренной фальсификации данных.

И, наконец, ошибки как таковой может и не быть, а вот вследствие вполне объективных, оправданных воздействий на контролируемую систему появляются данные сильно отличающиеся от общей совокупности.

Например, торговый центр объявил какую-то особенную распродажу в некоторый день и, как следствие, именно в этот день количество посетителей оказалось значительно (возможно в разы) больше обычного.

Или, например, в техническом комплексе осуществлялся натурный эксперимент по оценке длительности незащищенности системы. В результате чего именно в этот день количество тепла генерируемого системой вышло за допустимое, что повлекло за собой аварию.

Инсайдерская на фондовом рынке также причина возникновения «ложных» данных.

Все это приводит к нарушению массовости значений факторов риска и, как следствие, к снижению качества оценки и эффективности управления.

Сам процесс выявления ложных данных состоит в проверке однородности собранной совокупности.

Те элементы совокупности, которые заметно выбиваются из общей массы, скорее всего и являются ложными.

Разумно предположить, что если такие значений есть, то они должны быть среди наименьших или наибольших.

Настоящая процедура, также как и предыдущая, выполняется при заданном уровне значимости α.

Определим минимальное (x_{min}) и максимальное (x_{max}) из значений факторов риска.

Вычислим следующие расчетные характеристики совокупности опытных данных:

$$\tau_{расч}^{\min} = \frac{|x_{\min} - a|}{\sigma}, \ \tau_{расч}^{\max} = \frac{|x_{\max} - a|}{\sigma} .$$

Настоящие расчетные величины показывают, на сколько сильно отличаются крайние значения факторов риска от ожидаемых с учетом имеющегося в совокупности разброса.

Расчетные значения следует сопоставить с критическим. Для этого проверяют выполнение следующего неравенства:

$$\tau_{расч} < \tau_{кр}.$$

Если это неравенство выполняется, можно утверждать, что с вероятностью 1-α, соответствующее крайнее (максимальное или минимальное) значение ложным не является. В противном случае, это значение признается ложным и исключается из дальнейшего рассмотрения.

Критическая величина определяется в соответствии с количеством значений исследуемого фактора риска.

При небольших выборках объемом до 25 значений, в качестве критического значения рекомендуется использовать таблицу квантилей распределения максимального относительного отклонения $\tau_{кр}=\tau(\alpha,n)$.

Таблица некоторых значений квантилей максимального относительного отклонения для двух различных значений уровня значимости α приведена ниже.

Таблица 1

Квантили максимального относительного отклонения

α	Объем выборочной совокупности													
	12	13	14	15	16	17	18	19	20	21	22	23	24	25
0,01	2.66	2.71	2.76	2.8	2.84	2.87	2.9	2.93	2.96	2.98	3.01	3.03	3.05	3.07
0,05	2.39	2.43	2.46	2.49	2.52	2.55	2.58	2.60	2.62	2.64	2.66	2.68	2.70	2.72

Более полные таблицы рассматриваемой статистики содержатся в справочниках и учебниках по математической статистике.

При большой совокупности значений, следует пользоваться статистиками распределения Стьюдента.

В этом случае критическое значение вычисляется по следующей формуле:

$$\tau_{кр} = \frac{t(\alpha, n-2)\sqrt{n-1}}{\sqrt{(n-2)+t^2(\alpha, n-2)}}.$$

Здесь $t(\alpha,n-2)$ – это критические точки распределения Стьюдента.

Таблица некоторых значений критических точек критерия Стьюдента для двух различных значений уровня значимости α приведена ниже.

Таблица 2

Критические точки распределения Стьюдента

α	Объем выборочной совокупности										
	26	27	28	29	30	32	34	36	40	44	50
0.01	2,4786	2,4727	2,4621	2,4620	2,4573	2,4487	2,4411	2,4620	2.4233	2.4141	2.4033
0.05	1,7056	1,7033	1,7011	1,6991	1,6973	1,6939	1,6909	1,6883	1,6839	1,6802	1,6759

Более полные таблицы рассматриваемой статистики также как и в случае квантилей максимального относительного отклонения содержатся в соответствующих литературных источниках.

Если в результате выше описанной процедуры были выявлены и исключены некоторые из значений факторов риска, эту процедуру следует провести повторно, так как теперь крайними данными стали другие, и они также могут оказаться ложными. Процесс выявления и исключения ошибочных данных следует проводить до тех пор, пока среди исследуемых значений таких не останется.

3. Проверка значений факторов риска на случайность их формирования в исследуемой совокупности

При осуществлении сбора данных значений факторов риска возможны случаи нарушения еще одного обязательного условия, которое при этом не относится к наличию или отсутствию в выборочной совокупности ложных данных.

Этим обязательным условием, регламентирующим саму возможность применения к исследуемым данным аппарата статистической обработки, является случайность этих данных.

Например, оценка наблюдений стоимости актива осуществляется по мере поступления регистрируемых значений. Однако, периодичность (цикличность) проявления ряда факторов (объемы энергопотребления зависят от времени суток) влияет на значение стоимости. И это влияние не случайно. Это влияние определено.

Или, например, аварийная разгерметизация сосудов, работающих под давлением – событие в общем случае случайное. Однако если плановые испытания (проверки, диагностики) осуществляются с нарушением норм промышленной безопасности, они могут спровоцировать внештатную разгерметизацию. И периодичность таких испытаний лишает процесс разгерметизации сосудов случайности.

Существуют разные критерии выявления закономерностей среди исследуемых значений факторов риска.

В настоящей работе будет рассмотрен так называемый критерий «медианы выборки».

Эта процедура дает возможность выявить своеобразное монотонное смещение среднего выборочного значения относительно медианы в процессе наблюдения. Такое смещение может явиться следствием того, что фактические значения факторов риска не являются случайными и это смещение является следствием какой-то скрытой закономерности.

Для реализации критерия «медианы выборки» следует вычислить медиану выборочной совокупности по упорядоченным значениям факторов риска по следующей формуле:

$$x_{мед}(n) = \begin{cases} x_{\lfloor \frac{n}{2} \rfloor}, \text{ если } n \text{ не четно}, \\ \dfrac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}, \text{ если } n \text{ четно}. \end{cases}$$

Здесь через $\lfloor f \rfloor$ обозначена целая часть числа f, то есть ближайшее к f целое число, не превосходящее f.

Каждое значение анализируемого фактора риска необходимо сравнить с вычисленной медианой.

Если фактическое значение больше медианы, ему в соответствие следует поставить «+», если меньше – «-», если они равны « » (пробел, ничего не ставим).

В итоге получится совокупность последовательностей (называемых сериями), состоящих из плюсов и минусов.

Обозначим за ν число таких последовательностей, а за τ длину самой длинной последовательности.

Совокупность значений факторов риска с вероятностью 0,95 признается случайной если выполняется следующая система неравенств:

$$\begin{cases} \nu > \left\lfloor \dfrac{1}{2}(n+1) - 1{,}96\sqrt{n-1} \right\rfloor, \\ \tau < \lfloor 3{,}3\ln(n) + 1 \rfloor. \end{cases}$$

Согласно [1], при анализе рисков необходимо рассматривать достоверность в определении уровня риска и его чувствительность к предварительным условиям и допущениям. Кроме того, эффективность управления рисками на объекте любого рода во многом зависит не только от выбранных методов анализа, подробно представленных в [2], но и от качества исходной информации, представляющей исследуемые факторы риска. Как следствие первичная обработка их значений является востребованной и ценной.

Заключение

В представленной работе были рассмотрены три базовые процедуры обработки статистических данных, подготавливающих факторы риска к дальнейшему анализу. Эти процедуры предназначены для повышения качества исходных данных и, как следствие, для повышения эффективности оценивания рисков и управления ими.

Библиографический список

1. ГОСТ Р ИСО 31000-2010 Менеджмент риска. Принципы и руководство.

2. ГОСТ Р ИСО 31010-2011 Менеджмент риска. Методы оценки риска.

3. Профессиональный риск. Оценка и определение. Практическое пособие. – М.: Изд-во «Альфа-Пресс», 2010. 336 с.

4. Финансовые риски: учебное пособие / М.Л. Кричевский. – 2-е изд., стер. – М.: КНОРУС, 2013. 248 с.

5. Валентинов В.А. Эконометрика: Учебник / В.А. Валентинов. – 2-е изд. – М.: Издательско-торговая корпорация «Дашков и К^0», 2012. 448 с.

References

1. GOST R ISO 31000-2010 Risk management. Principles and guidelines.

2. GOST R ISO 31010-2011 Risk management. Risk assessment methods.

3. Occupational risk. Evaluation and determination. Practical guide. – M.: Publisher «Al'fa-Press», 2010. 336 p.

4. Financial risks: a tutorial / M.L. Krichevskij. – 2-e izd., ster. – M.: KNORUS, 2013. 248 p.

5. Valentinov V.A. Econometrics: Tutorial / V.A. Valentinov. – 2-e izd. – M.: Publishing and Trading Corporation «Dashkov i K0», 2012. 448 p.