ЗАЩИТА ИНФОРМАЦИИ И ИНФОРМАЦИОННАЯ БЕЗОПАСНОСТЬ

А.С. Крюковский[1]
Т.В. Лебедева[2]

МЕТОДИКА ОЦЕНКИ РИСКОВ УТЕРИ КОНФИДЕНЦИАЛЬНОЙ ИНФОРМАЦИИ В КОМПАНИИ

Данная работа посвящена описанию методики, позволяющей оценить риск потери конфиденциальной информации, а также помогающей определить меры для снижения вероятности ее потери.

Ключевые слова: *экспертный анализ, информационные риски, оценка рисков, оценка стоимости информации.*

A.S. Kryukovsky
T.V. Lebedeva

A TECHNIQUE OF AN ESTIMATION OF RISKS OF A LOSS OF THE CONFIDENTIAL INFORMATION IN THE COMPANY

This work describes methodology to assess the risk of loss of confidential information, as well as helping to identify measures to reduce the likelihood of loss.

Keywords: *expert analysis, information risk, risk assessment, valuation information.*

Бытует мнение, что управление рисками – удел топ-менеджеров либо сотрудников специализированных подразделений, риск-менеджеров, аналитиков. Однако на деле все обстоит поиному. Регулирование, а тем более оценка рисков являются прикладными задачами. И сфера информационной безопасности (ИБ) – не исключение. Специалисты в области ИБ должны скрупулезно отслеживать возникающие угрозы, анализировать связанные с ними риски и представлять руководству уже готовый отчет-план, какими средствами бороться за сохранность корпоративных данных.

Анализ рисков в области ИБ может быть качественным и количественным. Количественный анализ позволяет получить конкретные значения рисков, но он отнимает заметно больше времени, что не всегда оправданно. Чаще всего бывает достаточно быстрого качественного анализа, задача которого – распределение факторов риска по группам. Шкала качественного анализа может различаться в разных методах оценки, но всё сводится к тому, чтобы выявить самые серьезные угрозы [1]. В настоящей работе мы рассмотрим методику, использующую качественный анализ.

Методика оценки рисков состоит из следующих этапов: обработка выборки экспертов, анализ оценок, полученных в ходе опроса, и, на основе полученных результатов, формирование рекомендаций для лица, принимающего решения.

Вопросы, задаваемые экспертам относительно ряда категорий конфиденциальных сведений, делятся на две группы. Ответы на вопросы первой группы даются в виде оценок в баллах от 0 до 10, которым могут быть сопоставлены вероятностные характеристики. Ответы на вопросы второй группы даются в виде денежных оценок

Суть математической обработки данных, полученных в ходе ответов экспертов на вопросы первой группы, сводится к определению, насколько эксперты едины в своих вероятностных оценках относительно каждого вопроса по каждой категории конфиденциальных сведений.

Рассмотрим подробнее процедуру математической обработки данных для первой группы вопросов.

[1] Доктор физико-математических наук, профессор, декан факультета информационных систем и компьютерных технологий НОУ ВПО «Российский новый университет».

[2] Аспирантка НОУ ВПО «Российский новый университет».

В исследуемой группе присутствуют следующие вопросы.

1. Степень важности неразглашения конфиденциальных сведений.

2. Вероятность использования конфиденциальной информации злоумышленником (утрата конфиденциальности).

3. Вероятность появления кризисной ситуации в случае использования конфиденциальной информации злоумышленником.

4. Вероятность появления кризисной ситуации в случае утраты конфиденциальных сведений.

5. Вероятность модификации и искажения конфиденциальных сведений.

6. Вероятность появления кризисной ситуации в случае модификации и искажения конфиденциальных сведений злоумышленником.

7. Вероятность нарушения процесса, связанного с данным видом конфиденциальных сведений.

8. Вероятность быстрого восстановления процесса после потенциальной кризисной ситуации.

9. Актуальность обеспечения защиты данных конфиденциальных сведений через три года.

Допустим, что мы имеем k ответов n экспертов на некоторый вопрос из перечисленных выше для некоторой категории конфиденциальных сведений.

Обозначим ответы экспертов как Y_k, где k меняется от 1 до n.

Оценим математическое ожидание ответа эксперта на вопрос – среднее арифметическое по всем ответам Y_k. Суммируем все Y_k и делим на n:

$$\overline{Y} = \frac{1}{n}\sum_{k=1}^{n} Y_k. \qquad (1)$$

Далее находим центрированную оценку для каждого ответа. Для этого вычитаем оценку «математическое ожидание» из каждого Y_k:

$$y_k = Y_k - \overline{Y}. \qquad (2)$$

Просуммировав полученные центрированные оценки каждого ответа, возведенные в квадрат, и поделив эту сумму на $(n - 1)$, получаем оценку несмещенной дисперсии:

$$\hat{D} = \frac{1}{n-1}\sum_{k=1}^{n} y_k^2. \qquad (3)$$

Обозначим буквой σ квадратный корень из дисперсии:

$$\hat{\sigma} = \sqrt{\hat{D}}. \qquad (4)$$

Величина $\hat{\sigma}$ – несмещенная оценка среднеквадратичного отклонения – показывает, насколько совпадают ответы экспертов на вопрос. Чем $\hat{\sigma}$ меньше, тем большее единодушие демонстрируют эксперты при даче вероятностной оценки по данному вопросу. И наоборот, чем больше $\hat{\sigma}$, тем больше разброс в оценках экспертов по данному вопросу.

Оценка математического ожидания $\overline{Y} \pm \hat{\sigma}$ дает нам интервал средней вероятностной оценки экспертов для данного вопроса.

Для каждого математического ожидания может быть построен доверительный интервал:

$$\left(\overline{Y} - \frac{\hat{\sigma}}{\sqrt{n}} t_c,\ \overline{Y} + \frac{\hat{\sigma}}{\sqrt{n}} t_c\right) \qquad (5)$$

по уровню значимости α. Стандартным значением параметра α является 0,05, что соответствует мере надёжности 0,95. Это означает, что с вероятностью 95% математическое ожидание окажется в доверительном интервале. Параметр t_c определяется как двусторонняя квантиль:

$$t_c = t(\alpha, n-1) \qquad (6)$$

из распределения Стьюдента по таблице 1.

Таблица 1

Число степеней свободы j	$t(0{,}05; j)$
4	2,78
5	2,57
6	2,45
7	2,36
8	2,31
9	2,26
10	2,23
11	2,20
12	2,18
13	2,16
14	2,14
15	2,13

Вместо таблицы можно воспользоваться эмпирической формулой:

$$t_c = 3{,}93689 - 0{,}419522\, j + $$

$$+ 0{,}0351859\, j^2 - 0{,}00102305\, j^3.$$

Можно проверить гипотезу о значимости выборочной средней. Для этого рассчитаем контрольный параметр H:

$$H = \overline{Y}\frac{\sqrt{n}}{\sigma}. \qquad (7)$$

Если параметр H по модулю больше t_c:

$$|H| > t_c, \qquad (8)$$

то на поставленный вопрос следует дать положительный ответ, в противном случае:

$$|H| < t_c \qquad (9)$$

– отрицательный.

Основываясь на результатах обработки данных, лицо, принимающее решение, может сделать вывод по каждому из перечисленных вопросов.

Пример. Пусть имеется группа экспертов, состоящая из 15 респондентов. Результатом опроса этой группы являются 15 ответов на первый вопрос одной из категорий конфиденциальной информации.

Обозначим ответы экспертов как Y_k, где k меняется от 1 до 15 (таблица 2).

Таблица 2

Номер ответа	Ответ
Y_1	8
Y_2	0
Y_3	1
Y_4	0
Y_5	0
Y_6	2
Y_7	1
Y_8	10
Y_9	9
Y_{10}	8
Y_{11}	10
Y_{12}	2
Y_{13}	1
Y_{14}	0
Y_{15}	0

Находим оценку математического ожидания ответа эксперта на вопрос, то есть среднее арифметическое по всем ответам Y_k. Суммируем все Y_k и делим на 15:

$$\overline{Y} = \frac{1}{n}\sum_{k=1}^{n} Y_k = \frac{1}{15}\sum_{k=1}^{15} Y_k = \frac{52}{15} = 3,467.$$

Далее находим центрированную оценку для каждого ответа. Для этого вычитаем оценку математического ожидания из каждого Y_k (таблица 3).

Таблица 3

Номер ответа	Ответ	y_k
Y_1	8	4,533
Y_2	0	–3,467
Y_3	1	–2,467
Y_4	0	–3,467
Y_5	0	–3,467
Y_6	2	–1,467
Y_7	1	–2,467
Y_8	10	6,533
Y_9	9	5,533
Y_{10}	8	4,533
Y_{11}	10	6,533
Y_{12}	2	–1,467
Y_{13}	1	–2,467
Y_{14}	0	–3,467
Y_{15}	0	–3,467

Просуммировав полученные центрированные оценки каждого ответа, возведенные в квадрат, и поделив эту сумму на (n – 1), получаем оценку несмещенной дисперсии:

$$\hat{D} = \frac{1}{n-1}\sum_{k=1}^{n} y_k^2 = \frac{1}{15-1}\sum_{k=1}^{15} y_k^2 = \frac{239,733}{14} = 17,123.$$

Обозначим буквой $\hat{\sigma}$ квадратный корень из оценки дисперсии:

$$\hat{\sigma} = \sqrt{\hat{D}} = \sqrt{17,123} = 4,138.$$

Поскольку значение $\hat{\sigma}$ достаточно велико, то можно сделать вывод, что оценки экспертов по данному вопросу неоднозначны, так как интервал $\overline{Y} \pm \hat{\sigma}$ имеет вид $\left(\overline{Y} - \frac{\hat{\sigma}}{\sqrt{n}} t_c,\ \overline{Y} + \frac{\hat{\sigma}}{\sqrt{n}} t_c\right)$.

Для каждого математического ожидания построим доверительный интервал:

$$\left(\overline{Y} - \frac{\hat{\sigma}}{\sqrt{n}} t_c,\ \overline{Y} + \frac{\hat{\sigma}}{\sqrt{n}} t_c\right) =$$

$$= \left(3,467 - \frac{4,138}{\sqrt{15}} t_c,\ 3,467 + \frac{4,138}{\sqrt{15}} t_c\right)$$

по уровню значимости α. Стандартным значением параметра α является 0,05, что соответствует мере надежности 0,95. Это означает, что с вероятностью 95% математическое ожидание окажется в доверительном интервале. Параметр t_c определяется как двусторонняя квантиль:

$$t_c = t(\alpha, n-1) = t(0,05; 14)$$

из распределения Стьюдента, в описываемом случае оно равно 2,14. Таким образом, для найденного ранее математического ожидания доверительный интервал будет следующим: (1,18; 5,75).

Далее проверим гипотезу о значимости выборочной средней. Для этого рассчитаем контрольный параметр H:

$$H = \overline{Y}\frac{\sqrt{n}}{\sigma} = 3,245.$$

Так как параметр H по модулю больше t_c, то на поставленный вопрос следует дать положительный ответ, т.е. математическое ожидание отлично от нуля.

Найденная несмещенная оценка среднеквадратичного отклонения описывает меру рассеяния. Если она велика, то необходимо провести кластерный анализ. В случае если мера рассеяния несущественна, то вся выборка будет составлять одну группу.

Проблема кластерного анализа – обоснование и построение наилучшего «по сходству» (в том или ином смысле) разбиения исходного множества объектов. При этом к каждому кластеру будут отнесены объекты, имеющие характерную

общность, сами же кластеры имеют существенные различия. Понятие о наилучшем разбиении уточняется в каждой конкретной задаче путем выбора критерия оптимальности разбиения (построения кластеров), отражающего «сходство» элементов, отнесенных к данному кластеру.

Существует множество методов кластерного анализа. Рассмотрим наиболее распространенный – метод древовидной кластеризации.

Метод древовидной классификации – это пошаговый метод разбиения выборки на отдельные группы. Его принцип достаточно прост.

Шаг 1. Каждый человек признается единственным представителем своего кластера (типа). Количество типов равно объему выборки.

Шаг 2. Находится несколько человек, которые наиболее похожи на первого. Теперь эти люди составляют один кластер. Количество кластеров уменьшается.

Шаг 3. Продолжаем искать кластеры, наиболее похожие друг на друга, и объединять их. Теперь вся выборка разделена на некоторое количество групп, внутри которых люди очень схожи по своим характеристикам. Это продолжается, пока объединение не закончится и наступит последний шаг.

Шаг 4. Вся выборка объединяется в один кластер. Этот шаг не является информативным, так же как и первый шаг, но неизбежен в связи с процедурой. [2]

Пример. Рассмотрим группу экспертов, состоящую из 15 экспертов, результаты опроса которых указаны в таблице 2. Проведем кластерный анализ. В качестве математического критерия кластеризации возьмем меру расстояния между ответами экспертов.

Шаг 1. У нас имеется 15 кластеров.

Шаг 2. Кластер 1: Y_1, Y_{10}.

Шаг 3. Кластер 2: Y_2, Y_4, Y_5, Y_{14}, Y_{15}.
Кластер 3: Y_3, Y_7, Y_{13}.
Кластер 4: Y_6, Y_{12}.
Кластер 5: Y_8, Y_{11}.
Кластер 6: Y_9.

Наиболее близкие по значениям кластеры 1, 5, 6 и кластеры 2, 3, 4, поэтому объединяем их.

Кластер 1: Y_1, Y_{10}, Y_8, Y_{11}, Y_9.
Кластер 2: Y_2, Y_4, Y_5, Y_{14}, Y_{15}, Y_3, Y_7, Y_{13}, Y_6, Y_{12}.

Шаг 4. Кластер 1: Y_1, Y_{10}, Y_8, Y_{11}, Y_9, Y_2, Y_4, Y_5, Y_{14}, Y_{15}, Y_3, Y_7, Y_{13}, Y_6, Y_{12}.

Итогом кластеризации будет «дерево», изображенное на рисунке 1.

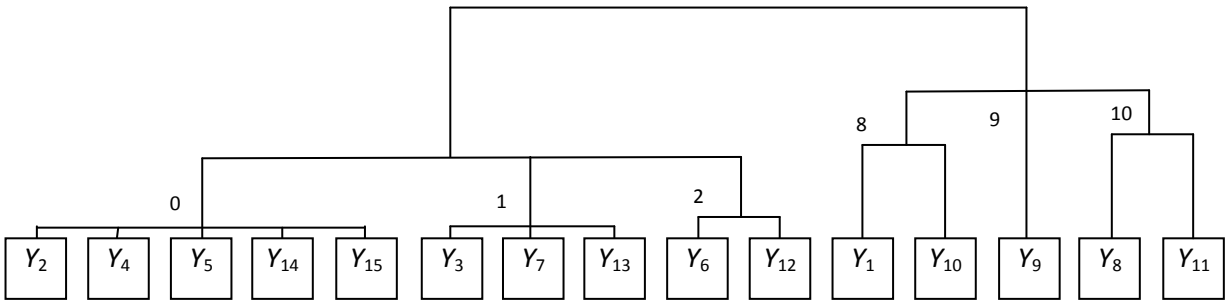

Рис. 1. Итог кластеризации

Таким образом, все респонденты поделились на две группы.

Так же можно было бы воспользоваться другим распространенным методом – методом *k*-средних. В отличие от древовидной классификации, метод *k*-средних разбивает всю выборку по заданным признакам на указанное количество кластеров. Таким образом, чтобы использовать этот метод, нужно знать или предполагать, сколько кластеров мы хотим иметь [2].

Следующий этап методики – обработка ответов второй группы вопросов анкеты.

Ответы на вопросы второй группы даются в виде денежных оценок. Указывается максимальная и минимальная сумма по каждому моменту времени в настоящем и будущем, либо при выполнении некоторого условия.

Математическая обработка данных, полученных в ходе ответов экспертов на вопросы второй группы, дает информацию о возможном факте наличия корреляции между стоимостью конфиденциальных сведений и потенциальным ущербом от их утраты, хищения, искажения, либо модификации.

Вопросы рассматриваемой группы для некоторой категории конфиденциальных сведений – это парные денежные оценки максимальной и минимальной стоимости информации и величины максимального и минимального ущерба от ее утраты (либо повреждения) на определенный момент времени в настоящем или будущем, либо при выполнении некоторого условия. Предполагается, что на вопросы ответили *n* экспертов.

Сначала вычислим средние минимальные и максимальные значения стоимости (X) и ущерба (Z) по каждому вопросу:

$$\overline{X}_{\min} = \frac{1}{n}\sum_{k=1}^{n} X_{k\min}, \quad \overline{Z}_{\min} = \frac{1}{n}\sum_{k=1}^{n} Z_{k\min} \qquad (10)$$

$$\overline{X}_{\max} = \frac{1}{n}\sum_{k=1}^{n} X_{k\max}, \quad \overline{Z}_{\max} = \frac{1}{n}\sum_{k=1}^{n} Z_{k\max} \qquad (11)$$

Знание этих величин позволяет оценить количественно стоимость информации, а также возможную величину ущерба как по категориям, так и в совокупности.

Обозначим X_k среднее значение между максимальной и минимальной денежной оценкой k-го эксперта стоимости конфиденциальных сведений в определенный момент времени

$$X_k = \frac{X_{k\min} + X_{k\max}}{2} \qquad (12)$$

и Z_k – среднее значение между максимальной и минимальной денежной оценкой k-го эксперта ущерба от утраты (либо повреждения) конфиденциальных сведений в тот же момент времени

$$Z_k = \frac{Z_{k\min} + Z_{k\max}}{2}, \qquad (13)$$

индекс k меняется от 1 до n.

Аналогично предыдущей процедуре обработки ответов Y_k, находим для X_k и Z_k оценки математических ожиданий и центрированные оценки:

$$\overline{X} = \frac{1}{n}\sum_{k=1}^{n} X_k, \qquad \overline{Z} = \frac{1}{n}\sum_{k=1}^{n} Z_k \; ; \qquad (14)$$

$$x_k = X_k - \overline{X}, \qquad z_k = Z_k - \overline{Z}. \qquad (15)$$

Вычисляем оценки дисперсии для X_k и Z_k: находим произведения сумм по k от 1 до n квадратов центрированных оценок X_k и Z_k и делим на $(n - 1)$:

$$\hat{D}_x = \frac{1}{n-1}\sum_{k=1}^{n} x_k^2, \quad \hat{D}_z = \frac{1}{n-1}\sum_{k=1}^{n} z_k^2. \qquad (16)$$

Находим несмещенные оценки среднеквадратичных отклонений:

$$\hat{\sigma}_x = \sqrt{\hat{D}_x}, \quad \hat{\sigma}_z = \sqrt{\hat{D}_z}. \qquad (17)$$

Как и в случае первой группы вопросов, для каждого математического ожидания могут быть построены доверительные интервалы:

$$\left(\overline{X} - \frac{\hat{\sigma}_x}{\sqrt{n}} t_c, \; \overline{X} + \frac{\hat{\sigma}_x}{\sqrt{n}} t_c\right), \qquad \left(\overline{Z} - \frac{\hat{\sigma}_z}{\sqrt{n}} t_c, \; \overline{Z} + \frac{\hat{\sigma}_z}{\sqrt{n}} t_c\right), \qquad (18)$$

рассчитаны контрольные параметры H:

$$H_x = \overline{X}\frac{\sqrt{n}}{\hat{\sigma}_x}, \qquad H_z = \overline{Z}\frac{\sqrt{n}}{\hat{\sigma}_z} \qquad (19)$$

и проверены гипотезы о значимости выборочной средней.

Таким образом, могут быть исключены из дальнейшего рассмотрения (полностью или частично) конфиденциальные сведения, стоимости которых (или возможные ущербы от утери которых) незначительны.

Возможна связь между стоимостью конфиденциальной информации и величиной ущерба от ее утраты. Для выяснения этого определяем оценку ковариации для X_k и Z_k. Суммируем по k от 1 до n произведения центрированных оценок X_k и Z_k и делим на $(n - 1)$:

$$\hat{\mu} = \frac{1}{n-1}\sum_{k=1}^{n} x_k \, z_k. \qquad (20)$$

Делением полученной оценки ковариации на среднеквадратичные отклонения получаем оценку значения коэффициента корреляции $\hat{\rho}$ для X_k и Z_k:

$$\hat{\rho} = \frac{\hat{\mu}}{\sigma_x \sigma_z}. \qquad (21)$$

Необходимо оценить значимость коэффициента корреляции $\hat{\rho}$. Для этого используем T-критерий. Вычислим контрольный параметр T по формуле:

$$T = t_c \sqrt{\frac{1-\hat{\rho}^2}{n-2}}. \qquad (22)$$

Если оценка коэффициента корреляции, взятая по модулю, меньше, чем вычисленный контрольный параметр $|\hat{\rho}| < T$, то корреляция отсутствует. Если больше $|\hat{\rho}| > T$, то корреляция есть. Здесь:

$$t_c = t(\alpha, n-2). \qquad (23)$$

Таким образом, можно сделать вывод о связи стоимости конфиденциальной информации и ущербе при ее утрате. Если коэффициент корреляции значим, то такая связь имеется.

Пример. Предполагается, что на вопросы ответили 15 экспертов (таблица 4).

Вычисляем средние минимальные и максимальные значения стоимости и ущерба:

$$\overline{X}_{\min}(1) = 41,2; \; \overline{X}_{\min}(2) = 40,933; \; \overline{X}_{\min}(3) = 42,6;$$

$$\overline{X}_{\max}(1) = 144,067; \; \overline{X}_{\max}(2) = 116,667; \; \overline{X}_{\max}(1) = 144,067;$$

$$\overline{Z}_{\min}(1) = 53,867; \; \overline{Z}_{\min}(2) = 34,667; \; \overline{Z}_{\min}(3) = 59,6;$$

$$\overline{Z}_{\max}(1) = 127,067; \; \overline{Z}_{\max}(2) = 109,667; \; \overline{Z}_{\max}(3) = 133,467.$$

Находим для X_k и Z_k оценки математических ожиданий:

$$\overline{X}(1) = 75,833; \; \overline{X}(2) = 78,8; \; \overline{X}(3) = 78,733;$$

$$\overline{Z}(1) = 90,467; \; \overline{Z}(2) = 72,167; \; \overline{Z}(3) = 96,533$$

и центрированную оценку (таблица 5).

Таблица 4

Номер эксперта	Оценка (в тыс. руб.) возможной стоимости информации (минимальной и максимальной)						Оценка (в тыс. руб.) возможной суммы ущерба (минимальной и максимальной)					
	в настоящее время (1)		через 3 года (2)		в случае реализации (3)		в настоящее время (1)		через 3 года (2)		в случае реализации (3)	
	min	max	min	max	min	max	min	max	min	max	min	max
1	30	128	30	132	42	93	43	141	23	125	59	110
2	34	80	59	128	34	96	47	93	52	121	51	113
3	40	92	37	92	47	104	53	105	30	85	64	121
4	53	121	43	97	43	135	66	134	36	90	60	152
5	49	97	53	108	42	97	62	110	46	101	59	114
6	33	83	27	85	25	119	46	96	20	78	42	136
7	48	96	50	123	37	108	61	109	43	116	54	125
8	47	148	43	138	47	115	60	161	36	131	64	132
9	47	139	32	100	35	102	60	152	25	93	52	119
10	40	135	24	122	53	149	53	148	17	115	70	166
11	37	111	28	89	50	80	50	124	21	82	67	97
12	33	142	50	145	35	107	46	155	43	138	52	124
13	31	98	59	120	52	146	44	111	52	113	69	163
14	45	136	55	148	46	126	53	149	48	141	63	143
15	51	105	24	123	51	146	64	118	28	116	68	187

Таблица 5

Номер эксперта	Оценка (в тыс. руб.) возможной стоимости информации (минимальной и максимальной)			Оценка (в тыс. руб.) возможной суммы ущерба		
	в настоящее время (1)	через 3 года (2)	в случае реализации (3)	в настоящее время (1)	через 3 года (2)	в случае реализации (3)
1	3,167	2,2	–11,233	1,533	1,833	–12,033
2	–18,833	14,7	–13,733	–20,467	14,333	–14,533
3	–9,833	–14,3	–3,233	–11,467	–14,667	–4,033
4	11,167	–8,8	10,267	9,533	–9,167	9,467
5	–2,833	1,7	–9,233	–4,467	1,333	–10,033
6	–17,833	–22,8	–6,733	–19,467	–23,167	–7,533
7	–3,833	7,7	–6,233	–5,467	7,333	–7,033
8	21,667	11,7	2,267	20,033	11,333	1,467
9	17,167	–12,8	–10,233	15,533	–13,167	–11,033
10	11,667	–5,8	22,267	10,033	–6,167	21,467
11	–1,833	–20,3	–13,733	–3,467	–20,667	–14,533
12	11,667	18,7	–7,733	10,033	18,333	–8,533
13	–11,333	10,7	20,267	–12,967	10,333	19,467
14	14,667	22,7	7,267	10,533	22,333	6,467
15	–24,833	–5,3	19,767	0,533	–0,167	30,967

Вычисляем несмещенные оценки среднеквадратичных отклонений X_k и Z_k:

$$\sigma_x(1) = 209,06,\ \sigma_x(2) = 202,779,\ \sigma_x(3) = 164,888,$$

$$\sigma_z(1) = 157,695,\ \sigma_z(2) = 200,631, \sigma_z(3) = 208,374.$$

Строим доверительные интервалы:

$$\left(\overline{X} - \frac{\sigma_x}{\sqrt{n}} t_c,\ \ \overline{X} + \frac{\sigma_x}{\sqrt{n}} t_c\right)_{(1)} = (-39,682;\ 191,348),$$

$$\left(\overline{X} - \frac{\sigma_x}{\sqrt{n}} t_c,\ \ \overline{X} + \frac{\sigma_x}{\sqrt{n}} t_c\right)_{(2)} = (-39,244;\ 190,844),$$

$$\left(\overline{X} - \frac{\sigma_x}{\sqrt{n}} t_c,\ \ \overline{X} + \frac{\sigma_x}{\sqrt{n}} t_c\right)_{(3)} = (-12,375;\ 169,842),$$

$$\left(\overline{Z} - \frac{\sigma_z}{\sqrt{n}} t_c,\ \ \overline{Z} + \frac{\sigma_z}{\sqrt{n}} t_c\right)_{(1)} = (3,333;\ 177,6),$$

$$\left(\overline{Z} - \frac{\sigma_z}{\sqrt{n}} t_c,\ \ \overline{Z} + \frac{\sigma_z}{\sqrt{n}} t_c\right)_{(2)} = (-38,691;\ 183,024),$$

$$\left(\overline{Z} - \frac{\sigma_z}{\sqrt{n}} t_c,\ \ \overline{Z} + \frac{\sigma_z}{\sqrt{n}} t_c\right)_{(3)} = (-18,603;\ 211,669).$$

Рассчитываем контрольные параметры H:

$$H_x(1)=20,313,\ \ H_x(2)=21,432,\ \ H_x(3)=23,747,$$

$$H_z(1)=27,901,\ \ H_z(2)=19,733,\ \ H_z(3)=25,9.$$

Поскольку все контрольные параметры больше $t_c = 2,14$, то делаем вывод, что мы не можем исключить из дальнейшего рассмотрения ни один вид конфиденциальных сведений, так как их стоимости (или возможные ущербы от утери) являются существенными.

Ищем оценки значений коэффициентов корреляции $\hat{\rho}$ для X_k и Z_k:

$$\hat{\rho}(1)=0,873,\quad \hat{\rho}(2)=0,995,\quad \hat{\rho}(3)=0,981.$$

Вычисляем контрольные параметры T:

$$T(1)=0,289,\quad T(2)=0,059,\quad T(3)=0,115.$$

Поскольку $|\hat{\rho}|>T$, то можно сделать вывод о связи стоимости конфиденциальной информации и ущербе при ее утрате. Коэффициент корреляции значим, следовательно, связь имеется.

Следующим этапом исследования является анализ оценок в динамике. Линейная форма связи между случайными переменными (линейная регрессия) занимает особое место в теории корреляции – многие практические задачи хорошо описываются линейной моделью.

Итак, если корреляционная связь между признаками установлена, то, в общем виде, *регрессионная модель* может быть представлена в виде

$$y=\varphi(x)+\varepsilon, \qquad (24)$$

где *возмущение* ε – случайная переменная, характеризующая отклонение от модельной функции регрессии (расхождения между эмпирическими и теоретическими значениями признака Y). При учете возмущения любой индивидуальный признак Y имеет возможность не попасть на линию регрессии. Основными причинами наличия возмущения, как правило, являются следующие:

• на вариабельность признака Y влияют помимо признака X и другие факторы;

• рассматриваемая экономическая система помимо общего влияния всех имеющих отношение к данному явлению факторов испытывает воздействие основного и непредсказуемого элемента случайности;

• значения переменной Y могут содержать ошибки измерения.

Линейный регрессионный анализ рассматривает функцию $\varphi(x)$, *линейную относительно оцениваемых параметров*, а не относительно переменной X, например следующие зависимости линейны относительно параметров α_i:

$$M_x(Y)=\alpha_0+\alpha_1 x;\ M_x(Y)=\alpha_0+\alpha_1\tfrac{1}{x};$$
$$M_x(Y)=\alpha_0+\alpha_1 x+\alpha_2 x^2. \qquad (25)$$

Если для оценки параметров модельной функции регрессии из двумерной генеральной совокупности взята выборка объема n, то регрессионная модель имеет следующий вид:

$$y_i=\varphi(x_i)+\varepsilon_i, \qquad (26)$$

где $(x_i, y_i)(i=\overline{1,n})$ – результат i-го наблюдения.

Введя в модель *возмущение* (слагаемое, характеризующее случайную ошибку) ε_i, определим характеристики распределения вероятностей этой величины. [3]

Простейшая модель регрессионного анализа *линейна* и по параметрам, и по переменным x_i, то есть имеет вид

$$y_i=\alpha_0+\alpha_1 x_i+\varepsilon_i. \qquad (27)$$

С помощью параметров α_0 и α_1 учитывается влияние на зависимую переменную Y объясняющей (предсказывающей) переменной X. Воздействие неучтенных факторов и случайных ошибок наблюдений определяется с помощью *остаточной дисперсии* σ_ε^2. Итак, в уравнении (27) величины α_0, α_1, ε неизвестны, причем величину возмущения будет трудно исследовать, поскольку она меняется от наблюдения к наблюдению. Но α_0 и α_1 остаются постоянными, и, даже не умея находить их значения точно без изучения *всех* возможных сочетаний Y и X, можно использовать информацию имеющейся выборки для получения *оценок* a_0, a_1 параметров α_0 и α_1. Оценкой линейной модели (27) по выборке является уравнение регрессии:

$$y_x=a_0+a_1 x \ \text{ или } \ y_x-\overline{y}=r_{xy}\cdot\frac{\sigma_y}{\sigma_x}(x-\overline{x}). \qquad (28)$$

Аналогично, линейное приближение x_y от y дается формулой линейной регрессии

$$x_y=b_0+b_1 y \ \text{ или } \ x_y-\overline{x}=r_{xy}\cdot\frac{\sigma_x}{\sigma_y}(y-\overline{y}). \qquad (29)$$

Параметры a_0, a_1, b_0, b_1 определяются на основе метода наименьших квадратов (МНК). Так как это основная процедура оценивания регрессионных характеристик, остановимся на ней подробнее.

Аналитическая процедура МНК заключается в следующем: рассмотрим уравнение регрессии

$$y_i=\alpha_0+\alpha_1 x_i+\varepsilon_i\ \ (i=\overline{1,n}). \qquad (30)$$

Сумма квадратов отклонений от «истинной» линии равна

$$S=\sum_{i=1}^{n}\varepsilon_i^2=\sum_{i=1}^{n}(y_i-\alpha_0-\alpha_1 x_i)^2. \qquad (31)$$

Подберем значения оценок a_0 и a_1 так, чтобы их подстановка вместо α_0 и α_1 в уравнение регрессии давала наименьшее возможное (минимальное) значение функции S. Для это-

го находим ее частные производные по параметрам:

$$\begin{cases} \dfrac{\partial S}{\partial \alpha_0} = -2\sum_{i=1}^{n}(y_i - \alpha_0 - \alpha_1 x_i) \\ \dfrac{\partial S}{\partial \alpha_1} = -2\sum_{i=1}^{n} x_i(y_i - \alpha_0 - \alpha_1 x_i), \end{cases} \quad (32)$$

минимальное значение получим при подстановке оценок (a_0, a_1) вместо (α_0, α_1) и приравнивая выражения частных производных к нулю, т.е.

$$\begin{cases} \sum_{i=1}^{n}(y_i - a_0 - a_1 x_i) = 0 \\ \sum_{i=1}^{n} x_i(y_i - a_0 - a_1 x_i) = 0. \end{cases} \quad (33)$$

После перегруппировки и приведения подобных получим *систему нормальных уравнений*:

$$\begin{cases} a_0 n + a_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ a_0 \sum_{i=1}^{n} x_i + a_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i. \end{cases} \quad (34)$$

Решив эту систему уравнений любым известным способом, получим значение оценок коэффициентов регрессии. Можно выразить оценки (a_0, a_1) явным образом. Решение системы нормальных уравнений относительно угла наклона прямой дает уравнение

$$a_1 = \frac{\sum x_i y_i - \left[\left(\sum x_i\right)\left(\sum y_i\right)\right]/n}{\sum x_i^2 - \left(\sum x_i\right)^2/n} = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2}. \quad (35)$$

Решение системы нормальных уравнений относительно свободного члена дает такой результат:

$$a_0 = \overline{y} - a_1 \overline{x}. \quad (36)$$

Итак, предсказывающее (подобранное) уравнение регрессии

$$y_x = a_0 + a_1 x. \quad (37)$$

Подставив в него полученное выражение для свободного члена a_0, получим оцениваемое уравнение регрессии:

$$y_x = \overline{y} + a_1(x - \overline{x}). \quad (38)$$

Из последнего уравнения видно, что если положить $x = \overline{x}$, то $y_x = \overline{y}$, т.е. *точка* $(\overline{x}, \overline{y})$ *лежит на подобранной линии регрессии.*

Коэффициент регрессии $a_1(b_1)$, характеризующий *линейную* связь, позволяет рассчитать, *на сколько в среднем изменится признак при изменении на единицу меры другого, связанного с ним, признака*.

По уравнению регрессии можно составить таблицу предсказанных значений y_{x_i} для каждого значения x_i из выборки, а значит, найти *остатки* $y_i - y_{x_i}$ – разности между тем, что наблюдалось, и тем, что предсказывается с помощью регрессионного уравнения.

В теории математической статистики выведена формула *выборочной оценки* s_e^2 *остаточной дисперсии* σ_e^2; определены формулы нахождения *доверительного интервала для прогнозов* значений (определения неизвестных возможных значений) зависимой переменной по уравнению регрессии; дан метод *оценки уравнения регрессии* – возможность установить, соответствует ли математическая модель, выражающая зависимость между переменными, экспериментальным данным и достаточно ли включенных в уравнение объясняющих переменных для описания зависимой переменной (дисперсионный анализ) [3].

На практике для проверки согласия построенной линии регрессии с результатами эксперимента можно воспользоваться идеей любой регрессии: часть изменений измеряемой величины Y связать с изменением внешних переменных (в двумерном случае – с X). Предполагая, что Y не зависит от X, за меру разброса результатов эксперимента принимаем сумму квадратов отклонений от среднего арифметического, т.е. величину

$$e_1 = \sum_{i=1}^{n}(y_i - \overline{y})^2, \text{ где } \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i. \quad (39)$$

Далее предполагаем линейную зависимость между признаками, то есть считаем, что построена регрессия $y_x = a_0 + a_1 x$. Теперь за меру разброса принимаем сумму квадратов отклонений от линии регрессии:

$$e_2 = \sum_{i=1}^{n}[y_i - (a_0 + a_1 x_i)]^2. \quad (40)$$

Если $e_1 \approx e_2$, то аппроксимирующая функция выбрана неудачно и подходящую линию регрессии стоит искать не среди прямых, а среди парабол, гипербол и т.п., то есть кривых другого вида [4].

Пример. Предположим, что зависимость между стоимостью информации и возможным ущербом (оценивается ситуация в настоящее время) описываются уравнением (35).

Найдем сумму квадратов отклонений от «истинной» линии:

$$a_1(1) = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2} = 0,758.$$

Решение системы нормальных уравнений относительно свободного члена дает такой результат:

$$a_0 = \overline{y} - a_1 \overline{x} = 32,956.$$

Тогда предсказывающее (подобранное) уравнение регрессии:

$$y_x = 32,956 + 0,758x.$$

Проверим согласие построенной линии регрессии с результатами эксперимента. За меру разброса результатов эксперимента принимаем сумму квадратов отклонений от среднего арифметического:

$$e_1 = \sum_{i=1}^{n} (y_i - \overline{y})^2 = 157,695.$$

Далее предполагаем линейную зависимость между признаками, то есть считаем, что построена регрессия $y_x = a_0 + a_1 x$. Теперь за меру разброса принимаем сумму квадратов отклонений от линии регрессии:

$$e_2 = \sum_{i=1}^{n} [y_i - (a_0 + a_1 x_i)]^2 = 524,372.$$

Поскольку $e_1 < e_2$, то делаем вывод, что аппроксимирующая функция выбрана верно.

Построим тренд по данному уравнению (рис. 2).

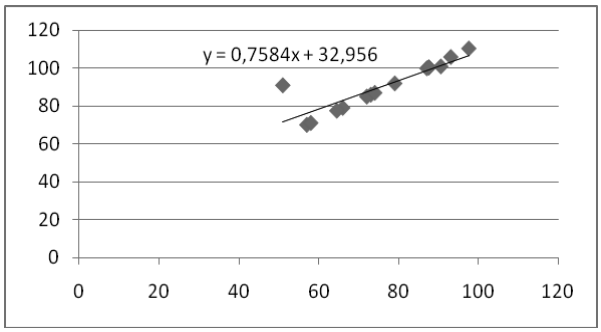

Рис. 2

С помощью найденного уравнения регрессии мы можем сделать прогноз возможного ущерба при заданной стоимости информации в данных условиях безопасности.

Следующим этапом исследования является анализ оценок средств защиты, который проводится на основе ответов экспертов на вопросы первой и второй групп.

Поскольку возможно множество вариантов, то в целях структуризации определим контрольные точки для первой группы вопросов – это 4 и 7. Тогда процесс обработки вопросов и формулировки выводов будет выполнен по следующему алгоритму.

Находим по формуле (1) оценку математических ожиданий ответов экспертов на вопрос $\left(\overline{Y}\right)$.

1. Если $\overline{Y} < 4$, находим математическое ожидание ответов экспертов на вопросы второй группы. Если оценка математического ожидания оценки величины ущерба меньше оценки математического ожидания стоимости информации более чем в 1,5 раза, то предлагаем службе безопасности уменьшить затраты на четверть на политику безопасности по данному критерию и провести опрос заново. Если же математическое ожидание оценки величины ущерба больше или равно оценке математического ожидания стоимости информации либо меньше менее чем в 1,5 раза, то советуем службе безопасности оставить политику неизменной, но через некоторое время провести повторное исследование.

2. Если $4 \le \overline{Y} < 7$, находим оценки математического ожидания ответов экспертов на вопросы второй группы, и здесь возможно несколько случаев:

a) оценка математического ожидания оценки величины ущерба меньше оценки математического ожидания стоимости информации более чем в 1,5 раза. Результатом исследования будет совет службе безопасности оставить политику неизменной, но через некоторое время провести повторное исследование;

b) оценка математического ожидания оценки величины ущерба равна оценке математического ожидания стоимости информации, в этом случае предлагаем службе безопасности оставить политику неизменной;

c) если же оценка математического ожидания оценки величины ущерба больше стоимости информации, то советуем ужесточить политику безопасности.

3. Если $\overline{Y} \ge 7$, находим оценки математического ожидания ответов экспертов на вопросы второй группы и тогда, вне зависимости от размеров величины ущерба и стоимости безопасности, необходимо ужесточить меры по обеспечению информационной безопасности.

Таким образом, описанная нами методика позволяет оценить риск потери конфиденциальной информации, а также помогает определить меры для снижения вероятности ее потери.

Литература

1. Ульянов, В. Анализ рисков в области информационной безопасности, PC Week/RE №40 (598) 30 октября – 5 ноября 2007, URL: http://www.pcweek.ru/security/article/detail.php?ID=103317 (дата обращения: 28.09.11)

2. Осадчая, И.А., Берестнева, О.Г. Кластерный анализ социально-психологических данных на базе пакета Novospark // Материалы VIII Всероссийской научно-практической конференции «Технологии Microsoft в теории и практике программирования». – Томск : Национальный исследовательский Томский политехнический университет, 2011. – С. 15–18.

3. Гмурман, В.Е. Теория вероятностей и математическая статистика. – 2003. – С. 43–56.

4. Мутанов, Г.М., Куликова, В.П. Математическое моделирование экономических процессов. – Алматы, 2006. – С. 78–92.