Статистические методы анализа информации постепенно вытесняются Big Data и ERP-системами. Но эти подходы дорогостоящие и требуют специальных навыков. Аудитору могут понадобиться более простые и дешевые инструменты для обнаружения нетипичных операций в объемных учетных данных. С помощью предлагаемого инструмента[1] можно быстро сформировать выборку.

Евгений ЗВЕРЕВ, *CIA, внутренний аудитор, член ассоциации «ИВА»*
Сергей КАБАРДИН, *внутренний аудитор, член ассоциации «ИВА»*

Распределение Бенфорда: простой и доступный инструмент анализа финансово-экономических данных

В четвертом номере журнала за 2018 г. была опубликована статья Е. Зверева и А. Никифорова «Распределение Бенфорда: выявление нестандартных элементов в больших совокупностях финансовой информации». Для журнала это первая публикация, описывающая применение распределения Бенфорда[2] как эффективного статистического аудиторского инструмента, позволяющего выявлять в совокупностях финансово-экономической информации нестандартные элементы[3], которые обычно являются индикаторами («красными флагами») рисков, присущих аудируемой совокупности, особенно мошеннических[4].

Это несложный аудиторский инструмент статистического анализа информации, выгруженной из соответствующей базы учетных данных.

[1] Метод может использоваться не только банками, но и инвестиционными фондами, страховыми компаниями, управляющими компаниями, в периметр которых входят промышленные предприятия, то есть любыми хозяйствующими субъектами, работающими с финансовой информацией.

[2] Распределение Бенфорда — специфическое частотное распределение значимых цифр в больших числовых совокупностях. См.: Benford F. The law of anomalous numbers // Proceedings of the American Philosophical Society. 1938. Vol. 78. P. 551–572.

[3] Здесь под элементами совокупности понимаются величины транзакций, остатки или обороты по счетам, стоимость активов и т.п., иначе — объекты учета.

[4] Зверев Е., Никифоров А. Учетные данные с нестандартным поведением: практические способы выявления признаков мошенничества // Внутренний контроль в кредитной организации. 2020. № 1. С. 37–42.

Евгений ЗВЕРЕВ
Сергей КАБАРДИН

Авторами настоящей статьи создан программный инструмент на базе Excel, позволяющий извлечь учетные бухгалтерские данные из стандартных отчетов «1С Бухгалтерия», конвертировать их в удобный формат и проанализировать на предмет выявления нестандартных значений, применяя распределение Бенфорда.

Описание индикатора риска

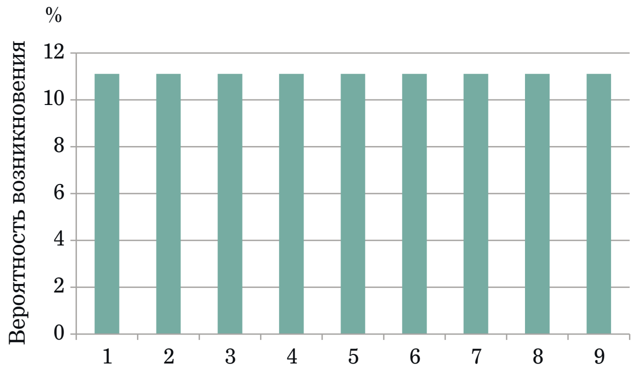
Поскольку с первой публикации прошло практически пять лет, необходимо напомнить суть самого распределения Бенфорда (далее также — Распределение).

Цифры складываются в числа, а совокупности состоят из чисел. Распределение Бенфорда — это распределение частоты появления определенных одиночных цифр/пар цифр/троек цифр на определенном месте внутри числа, обычно на первом месте. На первый взгляд все цифры в числах (1, 2, 3 и т.д.) равнозначны, интуитивно хочется предположить, что цифры в числах возникают с одинаковой частотой, а вероятность появления той или иной цифры в числе должна быть одинакова для любой цифры (рис. 1).

Но на практике так получается далеко не всегда. Американский астроном Саймон Ньюком в 1881 г. обнаружил, что страницы книги, содержащие логарифмические таблицы, истрепаны больше там, где содержатся логарифмы чисел, начинающихся с единицы, и практически девственно чисты для чисел, начинающихся на 9. Эта закономерность была подтверждена инженером Френком Бенфордом в 1938 г. Он проанализировал около 200 логарифмических таблиц с различными данными, и оказалось, что единица является первой

Рисунок 1

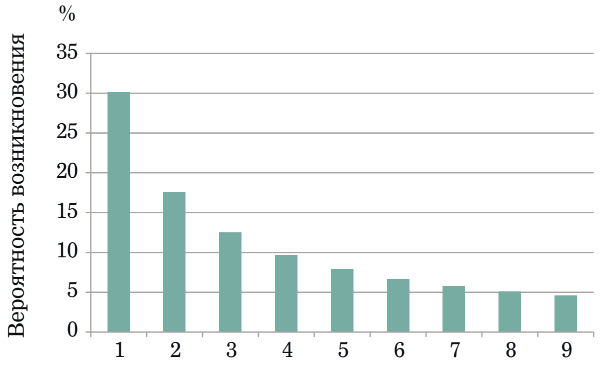
Равномерное распределение



значащей цифрой с вероятностью не 1/9, как думается, а около 1/3 (рис. 2).

Рисунок 2

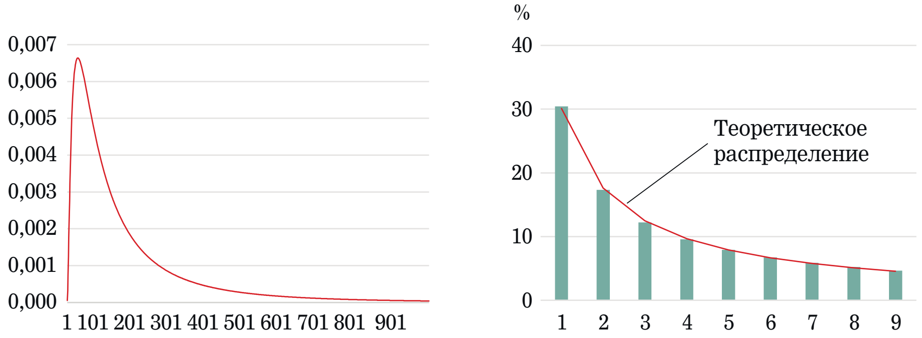
Распределение Бенфорда для первой значащей цифры числа



Распределение Бенфорда применимо к множеству совокупностей, элементы которых могут расти экспоненциально, то есть темп роста величины пропорционален ее текущему значению (примеры: распространение вируса, рост стоимости активов и пр.). Иными словами, распределение Бенфорда описывает числа, подчиняющиеся логнормальному[1] распределению (рис. 3).

Рисунок 3

Графики логнормального распределения и распределения Бенфорда



[1] Логарифмическое нормальное распределение — это статистическое распределение логарифмических значений из соответствующего нормального распределения. Логарифмическое нормальное распределение можно преобразовать в нормальное распределение и наоборот, используя соответствующие логарифмические вычисления.

Евгений ЗВЕРЕВ
Сергей КАБАРДИН

Для аудиторов особенно важно, что Распределению обычно подчиняются совокупности, отражающие финансово-экономическую деятельность, поэтому отклонение от него является инструментом выявления в них нестандартных элементов. Нетипичные операции, «выпадающие» из текущей логики бизнеса, иначе говоря, нестандартные (неестественные) учетные данные[1], с большой вероятностью будут отклоняться от Распределения.

Определение частот «появления» значащих цифр в анализируемых числовых совокупностях и сравнение их частотами Распределения позволяет выявлять подозрительные (нестандартные) транзакции для дальнейшего детального анализа (рис. 4, 5).

Рисунок 4

Упрощенная модель применения Распределения для анализа данных

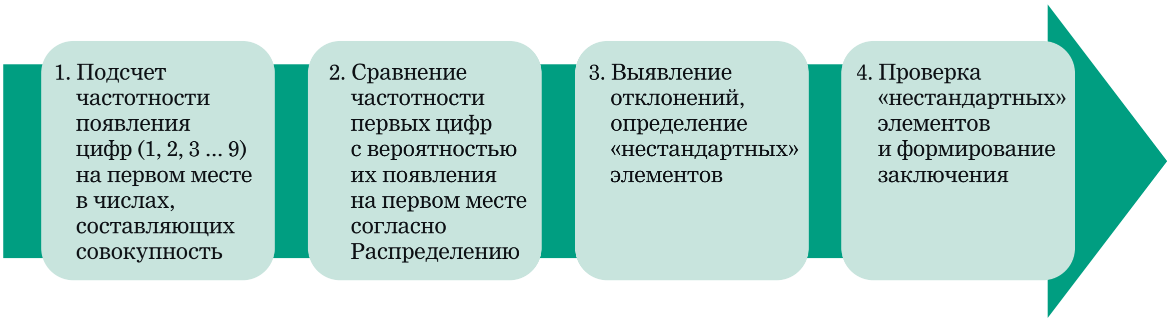
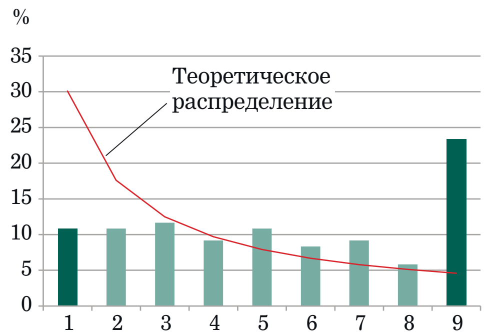


Рисунок 5

Графический результат анализа данных (шаги 1-3)



[1] Естественные финансовые учетные данные (остатки на счетах кредитных карт и (или) транзакции по ним, суммы бухгалтерских проводок, страховых выплат, гарантийного ремонта, выставленных счетов к оплате, объемы поставок, суммы в налоговых декларациях, списки стоимости покупок или денежных поступлений и пр.) также подчиняются Распределению, причем оно фактически выполняется, даже если значения конвертируются из одной валюты в другую.

Распределение Бенфорда: простой и доступный инструмент анализа финансово-экономических данных

Практика применения описываемого инструмента выявила особенность: индикатором риска выступает не одна цифра с максимальным абсолютным отклонением от Распределения (на рис. 5 это 9), а пара цифр с самыми большими абсолютными отклонениями относительно него и противоположными относительно него. Например, индикатором риска для совокупности рис. 5 является пара *9 (акцептор) – 1 (донор)*. Для этого примера отклонение фактических финансово-экономических данных от Распределения по первой значащей цифре указывает на искусственность количества указанных первых цифр, что одних цифр (9) почему-то много, а других цифр (1) почему-то мало. Отклонение предупреждает о наличии нестандартных данных в анализируемой совокупности финансово-экономической информации.

Тест на основе отклонения финансово-экономических данных от Распределения — «красный флаг» возможного мошенничества. Ведь что такое финансово-экономические данные? Это числа, которые формируются в ходе финансово-хозяйственной деятельности и в идеале должны ей полностью соответствовать, но они собираются, структурируются и анализируются работником и именно он может воздействовать на них.

Рассматриваемый индикатор риска прежде всего интересен аудиторам, проверяющим хозяйствующие субъекты, не обладающие учетными ERP-системами, в которых каждый элемент совокупности финансово-экономической информации имеет большое количество дополнительных информационных полей. Анализируя дополнительную информацию с помощью инструментария Excel по работе с таблицами, аудитор может выявить их нестандартное поведение.

Предлагаемый метод является *математическим, вероятностным*, не требующим дополнительной информации. Нужны только стоимостные данные объектов учета, которые можно выгрузить в виде стандартного отчета из 1С.

Описание модели, реализующей индикатор риска

Базовые требования к анализируемой совокупности

Экспоненциальная функция и геометрическая прогрессия с экспоненциальным основанием одинаковы, поэтому Распределению подчиняются цифры в числах, формирующих геометрическую прогрессию: доходы физических лиц, размеры банковских вкладов, цены

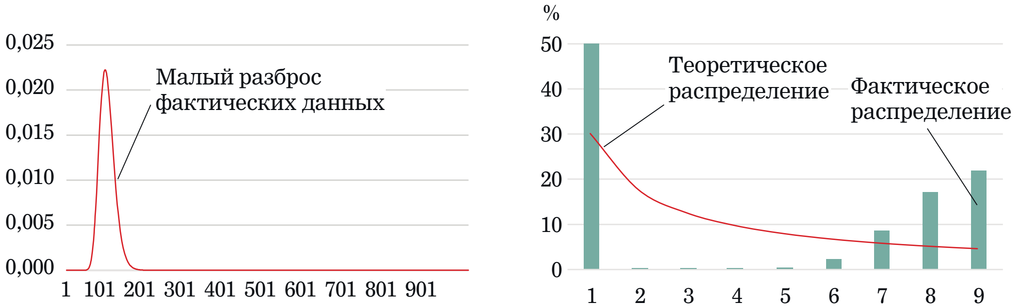
Евгений ЗВЕРЕВ
Сергей КАБАРДИН

активов, размеры заработной платы, размеры посевных площадей, численность населения городов, длина запросов в поисковике и многое другое. Однако далеко не всякая совокупность, даже из перечисленных, соответствует Распределению.

Для того чтобы совокупность описывалась Распределением, размерности ее элементов должны охватывать несколько порядков: единицы, десятки, сотни, тысячи и т.д. Кроме того, они не должны концентрироваться вокруг среднего значения, иными словами, совокупность должна быть «рыхлой». На рис. 6 представлена иллюстрация распределения первой цифры для величин элементов совокупности, лежащих в интервале от 1 до 200. Фактическое распределение не является логнормальным, распределение Бенфорда *не выполняется*.

Рисунок 6

Иллюстрация ограничения метода по диапазону размерности значений элементов



Степень «рыхлости»/«плотности» группирования элементов совокупности вокруг среднего значения оценивается с помощью коэффициента вариации случайной величины (характеристика рассеивания).

Коэффициент вариации — это применяемая в статистике величина, равная отношению стандартного (среднеквадратичного) отклонения случайной величины к ее математическому ожиданию (среднее значение):

$$C_v = \sigma / k,$$

где σ — стандартное отклонение элементов совокупности;
k — среднее значение.

Распределение Бенфорда: простой и доступный инструмент анализа финансово-экономических данных

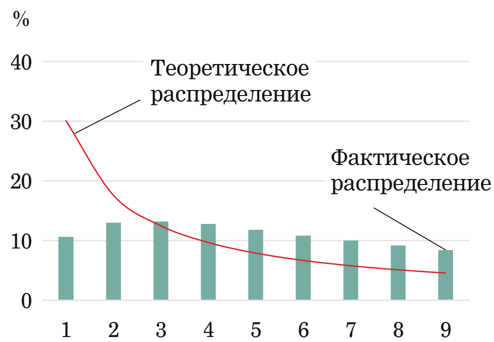
Если коэффициент вариации меньше 10%, то степень рассеивания считается незначительной, от 10 до 20% — средней, если он больше 20% и меньше или равен 33% — значительной. Нет смысла ориентироваться на результаты анализа совокупности с использованием Распределения при значении коэффициента вариации менее 20%.

Минимальная размерность элементов совокупности, анализируемых с помощью Распределения, составляет:
— по первой цифре элементов — десятки;
— по первым двум цифрам — тысячи;
— по первым трем цифрам — десятки тысяч;
— по второй цифре — сотни.

Совокупность не должны быть ограничена по количеству элементов. На рис. 7 представлен пример Распределения для первой цифры при ограничении количества элементов на уровне 100 единиц. Столбики диаграммы — это фактическая частота появления цифр при указанном ограничении. Распределение на 100 элементах можно рассматривать как логнормальное, но распределение Бенфорда *не выполняется*.

Рисунок 7

Иллюстрация ограничения метода по количеству элементов в совокупности



Минимальное количество элементов совокупности для тестирования составляет:
— по первой цифре — 110;
— по первым двум цифрам — 1100;
— по первым трем цифрам — 11 000;
— по второй цифре — 54.

Евгений ЗВЕРЕВ
Сергей КАБАРДИН

Оценка статистической значимости модели

Культура статистического анализа предполагает оценку статистической значимости[1] как в части соответствия фактической совокупности самому Распределению (теория), так и в части оценки достоверности полученных результатов. Для этих целей используются:

1. Критерий согласия Пирсона, или критерий согласия χ^2 (*хи-квадрат*)[2], для оценки того, насколько фактическое распределение элементов анализируемой совокупности в целом соответствует теоретическому Распределению. Для этого подсчитывается критическое значение с целью проверки нулевой[3] гипотезы — статистическое распределение элементов фактической совокупности близко к логнормальному (соответствует ему).

Аргументами являются: уровень значимости результата — 0,05 (соответствует 90% доверительной вероятности) и степень свободы — 8. Если выполняется неравенство $\chi^2_{кр.} > \chi^2$, то фактическая совокупность не является логнормальной — распределение Бенфорда *не выполняется*.

Мы использовали встроенные функции Excel:

— ХИ2.ТЕСТ(фактический интервал; теоретический интервал) — фактическое значение критерия согласия χ^2;

— ХИ2.ОБР.ПХ(0,05;8) — $\chi^2_{кр.}$, критическое значение критерия согласия.

Для дополнительной проверки фактической совокупности «на логнормальность» проведен повторный анализ фактического распределения с условием изъятия из расчета критерия согласия цифр с самой большой и самой малой частотностями (см. рис. 5) — 9 и 1. Если результат повторился и для 7 цифр ($\chi^2_{кр.} > \chi^2$, уровень значимости результата прежний, но степень свободы равняется 6), то фактическая совокупность гарантированно не является логнормальной, выявление аномальных значений с использованием Распределения бесперспективно.

---

1 Статистическая значимость (statistical significance) — статистические критерии для оценки получаемых результатов. Эти критерии позволяют оценить вероятность того, что такие результаты могли появиться чисто случайно. Характеристика статистически значимых дается результатам, вероятность случайного появления которых равна или ниже некоторого общепринятого уровня. Традиционно статистически значимый уровень 5% (или ниже) — вероятность случайного получения результата; это обычно выглядит как $p < 0,05$ (или $p < 0,01$) и означает, что если бы данное исследование повторили 100 раз, случайного появления таких результатов можно было бы ожидать менее чем в 5 случаях (или менее чем в 1 случае соответственно).

2 Метод, позволяющий оценить статистическую значимость различий двух или нескольких относительных показателей (частот, долей). Является наиболее часто употребляемым критерием для проверки гипотезы о принадлежности наблюдаемой выборки некоторому теоретическому закону распределения.

3 Нулевая (основная) гипотеза — статистическая гипотеза, подлежащая проверке. Здесь это гипотеза о том, что фактическая совокупность является логнормальной.

Для совокупностей, подтвердивших свою логнормальность со второго раза, подсчитывается значение p-value[1]. Аргументами являются критерий χ^2 и степень свободы 6. Чем выше значение p-value, тем ближе фактическая совокупность к логнормальному распределению. Уровень значимости 60% и более можно считать очень хорошим.

2. Z-тест[2] с заданным уровнем доверительной вероятности для подтверждения достоверности выявленных нестандартных элементов совокупности. Предполагается, что отклонение фактического распределения от теоретического возможно в результате воздействия большого количества случайных факторов, то есть распределяется по нормальному закону.

Если выполняется неравенство Z-тест > $Z_{\text{кр.}}$, считается, что отклонение частоты появления цифры превышает допустимую случайную величину, то есть полученный результат не случаен (статистически значим). Такие элементы фактической совокупности следует отбирать для детальной проверки. Рекомендуется выбирать значение доверительной вероятности в диапазоне от 80 до 99%, поскольку *увеличение* доверительной вероятности (уровня значимости результата) ограничивает количество элементов совокупности, подлежащих более детальному анализу.

Формулы теоретического распределения Бенфорда

Наиболее эффективны для выявления нестандартных элементов в фактических совокупностях, подчиняющихся логнормальному распределению, следующие формулы:

1. Распределение для первой значащей цифры числа:

$$P(n_1) = \log_{10}(1 + 1 / n_1),$$

где n_1 — первая значащая цифра числа, от 1 до 9;
P — вероятность.

Графически оно представлено на рис. 2.

2. Распределение для второй цифры числа:

$$P(n_2) = \sum_{1}^{9} \log_{10}(1 + 1 / (10 \times n_1 + n_2)),$$

где n_1 — первая значащая цифра числа, от 1 до 9;
n_2 — вторая значащая цифра числа, от 0 до 9;
P — вероятность.

[1] P-значение, p-уровень значимости, p-критерий – вероятность получить для имеющегося логнормального распределения фактических значений такое же или лучшее значение критерия согласия.

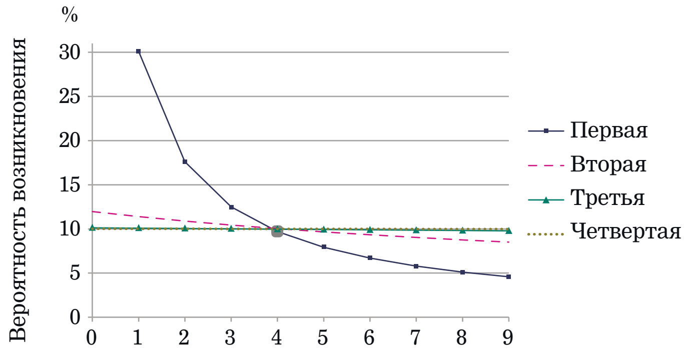
[2] Зверев Е., Никифоров А. Распределение Бенфорда: выявление нестандартных элементов в больших совокупностях финансовой информации // Внутренний контроль в кредитной организации. 2018. № 4. С. 42-56.

Евгений ЗВЕРЕВ
Сергей КАБАРДИН

Сравнение Распределений для первой, второй, третьей и четвертой значащих цифр представлено на рис. 8. Из сравнения видно, что делать тестирование следует по первой и второй цифрам, поскольку третья и четвертая цифры (и все последующие) распределяются практически равномерно, что не позволит эффективно выявлять нестандартные элементы фактической совокупности финансово-экономических данных.

Рисунок 8

Распределение Бенфорда для первой, второй, третьей и четвертой цифр



3. Распределение для первой пары цифр числа:

$$P(n_1 n_2) = \log_{10}(1 + 1/(n_1 n_2)),$$

где $n_1 n_2$ — первая-вторая значащие цифры числа, от 10 до 99;
P — вероятность.

Графически оно представлено на рис. 9.

4. Распределение для первой тройки цифр числа:

$$P(n_1 n_2 n_3) = \log_{10}(1 + 1/(n_1 n_2 n_3)),$$

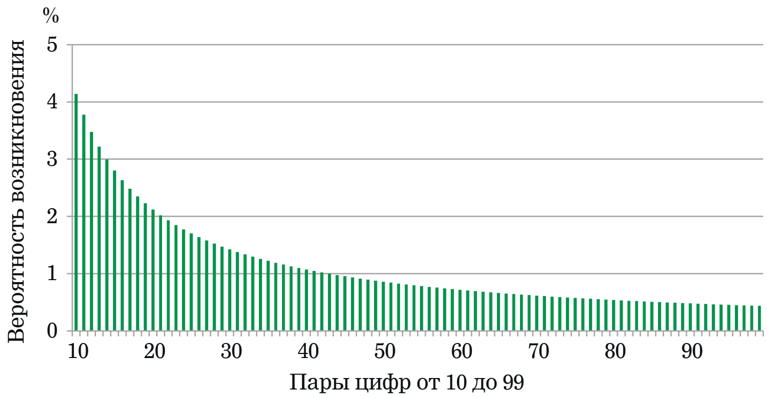
где $n_1 n_2 n_3$ — первая-вторая-третья значащие цифры числа, от 100 до 999;
P — вероятность.

Графически оно близко к тому, что показано на рис. 9.

При анализе распределения первых двух и первых трех цифр получается существенно больше значений, что позволяет более корректно выявлять отклонения фактической частоты распределения цифр от Распределения. Особенно это важно при анализе очень больших совокупностей, содержащих элементы с большими значениями.

Рисунок 9

Распределение Бенфорда (теоретическое) для первых двух цифр



Методология применения модели

Модель применяется следующим образом:

1. Выбираются данные для анализа. Распределению подчиняются логнормальные совокупности, но для корректного тестирования фактическая совокупность должна обладать определенными свойствами, обусловленными ее логарифмической природой. Последовательность проверки на «пригодность» данных показана на рис. 10.
2. По представленным формулам рассчитывается частотность цифр. Формулы применяются в зависимости от характера чисел, составляющих анализируемую совокупность.
3. Полученное фактическое распределение сравнивается с распределением Бенфорда (теоретическим).
4. Отбираются элементы с максимальным отклонением от теоретического Распределения. При этом каждый отобранный элемент должен быть статистически значим.
5. По отобранным элементам проводится детальная проверка.

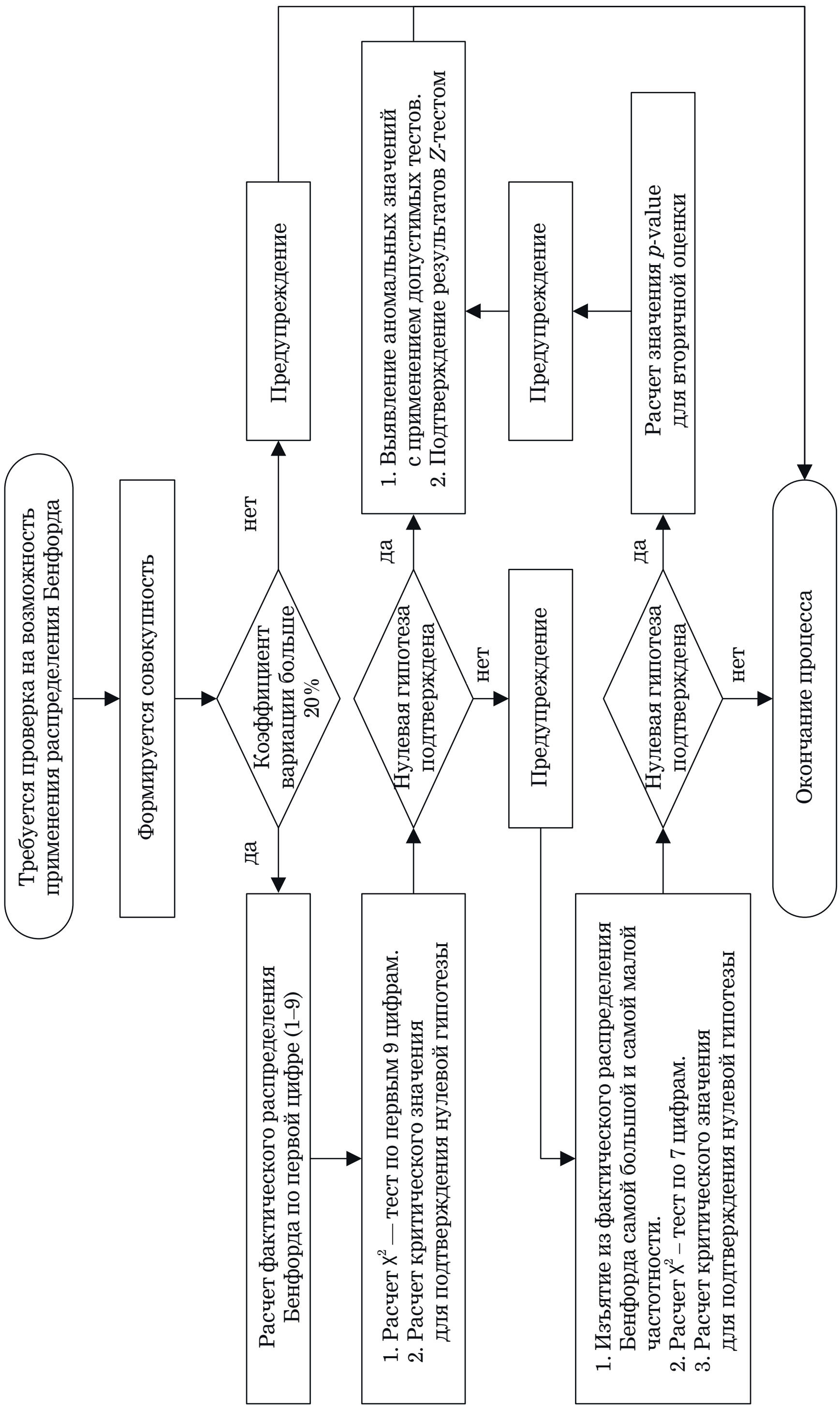
Пример выявления индикатора риска с помощью Распределения

Индикатор риска был выявлен при анализе распределения премиальных выплат в производственной компании (3000 сотрудников). Иными словами, рассматривалась совокупность *естественных финансовых учетных данных*, которая предварительно была проверена на пригодность к тестированию (см. рис. 10). Выполнено три

Евгений ЗВЕРЕВ
Сергей КАБАРДИН

Рисунок 10

Проверка на пригодность выборки к тестированию



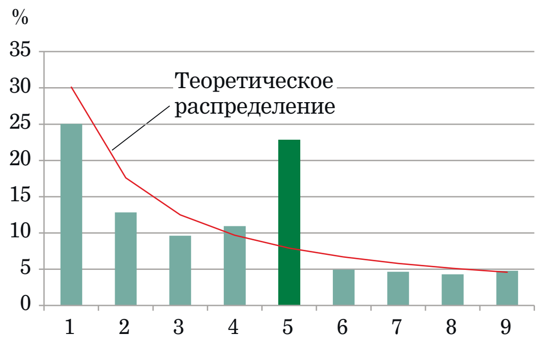
теста: анализ распределения частоты «выпадения» первой цифры, второй цифры и пары первых двух цифр.

Анализ по первой значащей цифре

Самый статистически значимый результат (Z-тест) по первой значащей цифре — 5, значительно меньший у 4. Остальные цифры (кроме 9) стали «донорами» для цифр 4 и 5 (рис. 11).

Рисунок 11

Анализ совокупности премиальных выплат по первой значащей цифре

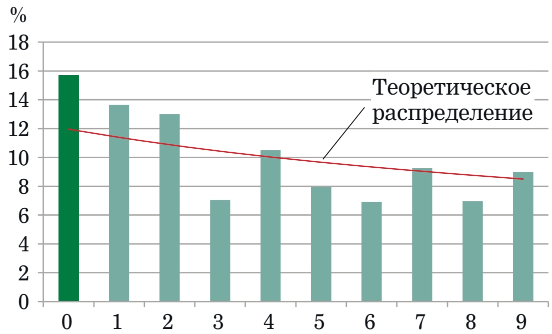


Анализ по второй значащей цифре

Самый статистически значимый результат (Z-тест) по второй цифре — 0, «доноры» — цифры 3, 5, 6, 8 (рис. 12).

Рисунок 12

Анализ совокупности премиальных выплат по второй значащей цифре



Евгений ЗВЕРЕВ
Сергей КАБАРДИН

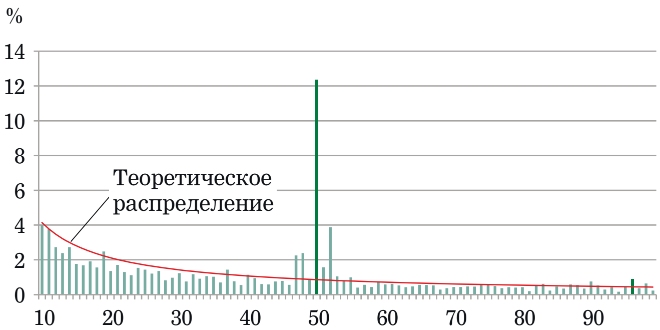
Для рассматриваемой совокупности сработали индикаторы риска на 5 и на 0, таким образом, аудитору документально нужно просмотреть премиальные выплаты вокруг 50. Можно сразу сказать о наличии большого количества небольших чисел (цифры на 50), иначе говоря, небольших премий.

Анализ по первой паре значащих цифр

Рассчитывалась частота «выпадения» первой пары цифр (от 10 до 99) каждого элемента (рис. 13, таблица).

Рисунок 13

Анализ совокупности премиальных выплат по первой паре цифр



Таблица

Цифровой анализ совокупности премиальных выплат по первой паре цифр

№	Первая пара цифр	Абсолютное отклонение от теоретического Распределения, %	Значение	
			Z-тест	$Z_{кр.}$ [1]
1	15	–1,0	4,61	4,417
2	47	1,3	10,36	
3	48	1,5	11,62	
4	50	11,5	91,98	
5	51	0,7	8,87	
6	52	3,0	24,76	
7	96	1,0	4,86	

[1] При доверительной вероятности 99%.

Анализ совокупности по первой паре цифр также указал на 50, но дополнительно было выявлено отклонение на 96. *Анализ по первой и второй цифрам проблемы с 9 и 6 не выявил.* Отклонение от Распределения по двум первым цифрам 50 и 96 указывает на проблемы в формировании совокупности премиальных выплат, искусственность количества элементов на 50 и на 96, то есть мы имеем быструю аналитическую оценку риска внешнего воздействия на данную совокупность.

Дальнейший подробный анализ результатов тестирования выявил, что было необоснованно выдано небольшое количество крупных премий (числа на 96, большие цифры) и, чтобы выровнять «баланс» распределения премий по величине, крупные выплаты уравновесили большим количеством мелких, просто «рассыпав» цифры рядом с 50.

Рекомендации

По отклонениям от теоретического распределения Бенфорда можно сделать вывод о возможном проведении нетипичных (подозрительных) операций, которые нарушают стандартную логику функционирования бизнеса, а тем самым нарушают логнормальное распределение. Примеры: разбиение крупных платежей на мелкие, ограничение платежей, проведение транзакций, не характерных для бизнеса. Оценка соответствия элементов совокупности финансово-экономических данных Распределению является необходимым первым шагом при ее исследовании на наличие риска мошенничества. Естественно, результаты тестирования не всегда верны и могут указывать на ложные элементы, но нельзя отрицать, что они — важный дополнительный инструмент для аудитора.

Для тестирования финансово-экономических совокупностей на основе Распределения рекомендуем:

1. Выбирать данные в денежном выражении, которые отражают специфику бизнеса (*естественные финансовые учетные данные*). Нельзя смешивать данные, относящиеся к разным учетным системам (например, данные производственного и бухгалтерского учета), разным организационным структурам (например, данные по филиалу и головной компании) или бизнес-процессам (например, данные закупок и продаж).

2. Поверять элементы отобранных совокупностей:

— на количество;

— «разброс» этих данных относительно среднего значения;

— размерность;

— отсутствие отрицательных и нулевых значений (поскольку логнормальное распределение по определению *положительное*).

3. Проверять данные на «логнормальность» с помощью критерия согласия Пирсона с доверительной вероятностью 90%.

4. Сочетать и комбинировать результаты разных тестов: тестов по первой цифре, по второй цифре, совместного анализа частот первой и второй цифр, тестов по первым двум цифрам, по первым трем цифрам.

5. Выбирать доверительную вероятность для *Z*-теста в пределах от 80 до 99%. Варьируя доверительную вероятность, можно устанавливать количество транзакций для более детального анализа, исходя из наличия ресурсов и трудоемкости его проведения.

6. Не забывать, несмотря на простоту применения Распределения, о том, что существуют данные, не подчиняющиеся ему, а также объемы данных, размер которых недостаточен для применения этого статистического метода.

Все перечисленные в статье ограничения, рекомендации и расчетные формулы реализованы авторами в виде программного комплекса[1] на базе Excel, который позволяет:

— конвертировать данные стандартных отчетов 1С в вид, пригодный для автоматизированной аналитической обработки;

— контролировать и отбирать информацию для тестирования;

— анализировать данные из отчетов 1С на соответствие Распределению.

[1] Программный комплекс расположен для ознакомительных целей в свободном доступе на сайте ассоциации «Институт внутренних аудиторов» на условиях «как есть».