СЧЕТНАЯ ПАЛАТА РОССИЙСКОЙ ФЕДЕРАЦИИ

МЕТОДИКА
ПО ОЦЕНКЕ НАДЕЖНОСТИ ДАННЫХ

УТВЕРЖДЕНА
Коллегией Счетной палаты
Российской Федерации
(протокол от 15 марта 2022 г.
N 12К (1537)

(в ред. Постановления Коллегии от 23.05.2023 N 33К /9)

МЕТОДИКА ПО ОЦЕНКЕ НАДЕЖНОСТИ ДАННЫХ

1. РАЗРАБОТАНА Департаментом исследований и методологии аппарата Счетной палаты Российской Федерации.

2. ДАТА ВСТУПЛЕНИЯ В СИЛУ:

3. РАЗРАБОТАНА ВПЕРВЫЕ.

1. Общие положения

1.1. Методика по оценке надежности данных (далее - Методика) определяет порядок оценки надежности данных в ходе проведения аналитических расчетов [1] в целях недопущения использования ненадежных данных в процессе получения доказательств при проведении Счетной палатой Российской Федерации контрольных и экспертно-аналитических мероприятий (далее - мероприятия, Счетная палата).

---

1 Аналитический расчет согласно Методике по документированию аналитических расчетов, это проверка гипотез мероприятия при помощи методов математического моделирования, математической статистики, эконометрики и машинного обучения.

1.2. Методика разработана с учетом положений стандартов внешнего государственного аудита (контроля) Счетной палаты [2], положений руководства ИНТОСАИ для высших органов аудита [3], а также с учетом опыта высших органов аудита, стандартов по качеству и управлению данными [4].

---

2 Стандарт внешнего государственного аудита (контроля) СГА 101 "Общие правила проведения контрольного мероприятия" (пункты 6.2.4, 7.2.5), Стандарт внешнего государственного аудита (контроля) СГА 102 "Общие правила проведения экспертно-аналитических мероприятий" (пункты 5.2.2.2, 6.3).
(в ред. Постановления Коллегии Счетной палаты РФ от 23.05.2023 N 33К /9)

3 Руководство ИНТОСАИ GUID 3910 "Основные концепции аудита достижения результатов" (пункт 88).

4 DAMA-DMBOK2: Свод знаний по управлению данными (Table 29 Common Dimensions of Data Quality, подраздел 1.3.3 Data Quality Dimensions), Национальный

стандарт Российской Федерации ГОСТ Р 57773-2017 "Пространственные данные. Качество данных", Government Accountability Office (2019) "Assessing data reliability", INTOSAI Development initiative, Performance Audit ISSAI Implementation Handbook (August 2021).

2. Порядок определения и оформления оценки надежности данных

2.1. Данные являются надежными, если они обладают следующими характеристиками:

1) **применимость для достижения целей мероприятия** - использование данных позволит достичь указанных целей;

Пример

В рамках одного из мероприятий была поставлена цель по оценке эффективности субсидирования отрасли. Для проведения оценки эффективности субсидирования необходимо наличие двух групп: группа получателей субсидии и контрольная группа, т. е. группа фирм, не получавших субсидии, но обладающих схожими характеристиками. Были собраны данные по финансовым показателям фирм и субсидиям, предоставляемым из федерального бюджета. При этом собранные данные оказались неприменимы для проведения оценки эффективности, так как отсутствовала возможность сформировать контрольную группу для сравнения. В этом случае данные являются ненадежными для целей мероприятия, поскольку их использование не позволяет достичь цели мероприятия по оценке эффективности субсидий.

2) **репрезентативность** - характеристики выборки соответствуют характеристикам генеральной совокупности данных (совокупность всех объектов (единиц), относительно которых предполагается делать выводы при изучении конкретной задачи) за необходимый временной период;

Пример

Для цели анализа расходов малоимущих семей в России репрезентативными данными можно считать данные специально разработанного опроса малоимущих семей, основанного на выборке, структура которой в части основных характеристик (пол, возраст, число членов семьи, наличие детей и т. д.) соответствует структуре всей совокупности малоимущих семей в стране. Такая организация выборки может позволить с учетом весов наблюдений распространить закономерности, выявленные для выборки, на генеральную совокупность. При этом данные административных реестров, содержащие данные получающих ежемесячное пособие на ребенка из малоимущей семьи, могут не характеризовать все малоимущие семьи с детьми, так как не содержат информации о тех, кто не обращался за пособием.

Другой, экстремальный, пример: нельзя считать единственное наблюдение репрезентативно представляющим генеральную совокупность.

3) **точность** - данные соответствуют истинным значениям рассматриваемых показателей;

Пример

Для цели анализа годовых доходов домохозяйств отклонение в десятки рублей можно считать точным значением.

4) **полнота** - данные не содержат пропущенных значений;

Пример

При анализе данных по получателям субсидий, предоставляемым из федерального бюджета, выяснилось, что за определенный год реестр субсидий содержит не все платежи по субсидиям (20 % бюджетных ассигнований, доведенных казначейскими уведомлениями). Поэтому такие данные были

признаны неполными.

5) **уникальность** - данные не содержат повторяющихся записей;

Пример

В выгрузке финансовых показателей фирм из информационно-аналитической системы СПАРК обнаружены дублирующиеся строки, то есть данные могут рассматриваться как надежные данные только после удаления повторяющихся значений.

6) **валидность (доменная согласованность)** - значения соответствуют области допустимых значений;

Пример

При анализе данных по экспорту от Федеральной таможенной службы выяснилось, что один из ИНН содержал букву "ё". Такой ИНН был удален, потому что ИНН не может содержать буквы.

Репрезентативность, точность, полнота, уникальность и валидность (доменная согласованность) также являются характеристиками качества данных.

2.2. Оценка надежности проводится для объединенного набора данных, используемых для достижения целей мероприятия.

Пример

При анализе данных по экспорту использовались три набора данных: данные по стоимостному объему экспорта, финансовые характеристики фирм (выручка, количество сотрудников и прочие), данные о получателях субсидий на компенсацию затрат на транспортировку экспортной продукции. Для цели анализа эти наборы были объединены по ИНН и году в один набор. Оценку надежности следует проводить для последнего объединенного набора, а не исходных трех наборов.

2.3. Результаты оценки надежности данных включаются в состав рабочей документации мероприятия.

Шаблон оценки надежности данных и пример его заполнения приведены в приложениях N 1 и N 2 к настоящей Методике.

2.4. Допустимо использовать результаты оценки надежности данных, полученные в ходе проведения иных мероприятий, если оценивалась надежность таких же данных и на момент проведения соответствующего мероприятия такая оценка является актуальной.

2.5. При признании данных ненадежными не допускается их использование, поскольку такие данные не позволят сформировать достаточные и надлежащие доказательства для достижения целей мероприятия.

2.6. В случае признания данных, необходимых для проверки гипотез мероприятия, ненадежными объекту аудита (контроля) или иному источнику таких данных могут быть направлены предложения (рекомендации) об улучшении процесса сбора, обработки и выгрузки данных с целью повышения конкретных характеристик их надежности, которые также подлежат включению в отчет о результатах мероприятия в порядке, предусмотренном стандартами внешнего государственного аудита (контроля) СГА 101 "Общие правила проведения контрольного мероприятия" и СГА 102 "Общие правила проведения экспертно-аналитических мероприятий".

3. Способы оценки надежности данных

3.1. Оценка надежности данных проводится следующими способами:

тестирование (профилирование) данных;

анализ документации, описывающей сбор, обработку и выгрузку данных;

интервью с сотрудниками, отвечающими за источник данных;

выборочная сверка с первичным источником данных;

иной способ оценки надежности данных.

3.2. **Тестирование (профилирование) данных.**

Тестирование (профилирование) данных осуществляется путем:

1) установления соответствия количества и наименования переменных тому, что содержится в описании данных;

Пример

В рамках одного из мероприятий в Росстате были запрошены расширенные версии выборочного наблюдения доходов населения и участия в социальных программах (ВНДН) Росстата за последние несколько лет. При сравнении переменных в полученной базе за 2017 год с описанием публичной версии этой же базы на сайте Росстата выяснилось, что наряду с наличием в полученной базе дополнительных переменных, в ней отсутствуют три расчетные переменные:
R_I_DOP ("Общая сумма заработка (дохода) за выполнение дополнительной работы (наряду с основной) до выплаты подоходного налога");
R_I_DOP_NALOG ("Величина подоходного налога на сумму заработка (дохода), получаемого "на руки", за выполнение дополнительной работы (наряду с основной)");
R_H_SOBST ("Доходы от собственности").
В ходе интервью с сотрудниками Росстата было выяснено, что в полученном наборе данных последняя переменная есть, но она называется иначе - R_H_DOX_SOBST. В результате переменные R_I_DOP и R_I_DOP_NALOG не были использованы в расчетах, переменная R_H_DOX_SOBST была использована в расчетах.

2) обнаружение пропущенных записей: отсутствие целых записей (для отдельных объектов наблюдения или временных периодов) или значений переменных;

Пример

В рамках одного из мероприятий решалась задача анализа влияния различных факторов на один из параметров государственной политики в сфере образования, в том числе социально-экономических условий в муниципалитетах. Для решения этой задачи предполагалось использовать информацию из базы данных муниципальных образований Росстата. Эта база содержала большое количество пропущенных значений. В частности, данные о численности населения за 2016-2018 годы имелись лишь для 11,5-12 тыс. муниципалитетов из примерно 20 тыс. общего количества муниципальных образований, о численности городского населения - примерно для 3,3-4 тыс. муниципальных образований, о среднемесячной начисленной заработной плате работников крупных и средних предприятий и некоммерческих организаций - около 2,2-2,4 тыс., и т. д. В результате эти данные были признаны неполными, а следовательно, и ненадежными, поэтому было принято решение отказаться от их использования.

3) обнаружение дублирования записей: повторение целых записей или переменных;

4) обнаружение значений, которые являются недопустимыми или не ожидаемыми;

Пример

Недопустимые данные:
Заработная плата не может быть отрицательной;
ИНН может быть либо 10-, либо 12-значным;
В базе характеристик школ по формам ОО-1 и ОО-2, полученной по запросу у Министерства просвещения Российской Федерации, имеются записи о школах, в которых адрес записан, например, как "г. Махачкала", "г. Махачкала, ул.", "641493", "sch115@school.ru", "634021, ул. Лебедева, 92" и т. п.
Допустимые, но не ожидаемые данные:
Значение "1990" в данных за 2010-2020 годы;
Адрес "г. Москва, ул. Зубовская, д. 2" в данных по Республике Коми.

Если недопустимых или неожидаемых значений мало, можно попытаться скорректировать их вручную, например, сверив их с каким-либо надежным источником (например, при возможности, с первоисточником). Если корректировка таких значений представляется трудоемкой, то их следует исключить из рассмотрения.

В то же время если подобных записей достаточно много и после их исключения выборка сильно сокращается, то данные следует признать ненадежными и отказаться от их использования, поскольку свойства оставшейся части выборки могут значительно отличаться от свойств генеральной совокупности.

5) обнаружение неправдоподобно больших или маленьких значений показателей;

Пример

В качестве примера можно рассмотреть исследовательскую задачу, связанную с анализом фактического рабочего времени населения на данных Российского мониторинга экономического положения и здоровья населения, проводимого НИУ ВШЭ (далее - РМЭЗ). На основе ответов респондентов строится переменная, отражающая количество рабочих часов в месяц. Если, например, человек работает по 8 часов 5 дней в неделю, то за месяц количество отработанных часов в среднем составит около 170 часов. Если, например, человек работает по совместительству, или его рабочий день не нормирован, то даже если человек работает по 14 часов в день 31 день в месяц, то количество рабочих часов не превысит 434. В то же время база РМЭЗ за 2000-2019 годы включает наблюдения, в которых это количество превышает 500, 600 и даже 700 часов.

Как правило, подобные наблюдения удаляют из выборки. Порог для такого удаления либо устанавливают в абсолютном виде (например, удалить из базы наблюдения, в которых количество рабочих часов в месяц превышает 400), либо в относительном (например, удаляют наблюдения, в которых количество рабочих часов попадает в 1 % самых больших значений).

Если количество таких отбрасываемых наблюдений невелико, то данные признаются надежными и могут быть использованы в анализе.

6) проверка корректности результатов объединения данных (если для достижения целей мероприятия возникла необходимость объединить наборы данных);

Пример

1. База РМЭЗ состоит из двух частей: база, содержащая информацию (ответы на вопросы) об индивидуумах-членах домохозяйства, и база, содержащая информацию (ответы на вопросы) о домохозяйствах. Часто возникает необходимость провести анализ объединенной базы. Объединение этих баз производится по уникальному ключу - коду домохозяйства. При этом в базе РМЭЗ встречаются наблюдения (домохозяйства), для которых заполнена информация только об индивидуальных характеристиках членов домохозяйства, но отсутствуют характеристики домохозяйства как целого. В такой ситуации при слиянии баз отсутствует взаимно-однозначное соответствие между ключами, имеющимися в двух базах. В результате в объединенной базе могут возникнуть наблюдения, в которых часть информации отсутствует, либо (в зависимости от параметров слияния баз) часть информации, полезной для анализа только индивидуальных

характеристик, может быть утеряна.

2. При проведении мероприятия возникла необходимость произвести слияние двух расширенных баз ВНДН за 2017 год, полученных по запросу от Росстата - с индивидуальными характеристиками и с характеристиками домохозяйств. При этом если в публичных версиях ВНДН идентификационный номер домохозяйства был уникален в пределах базы и мог бы служить ключом для слияния двух таких баз, то в расширенной версии ВНДН идентификационный номер домохозяйства не был уникальным - он был уникальным только в пределах региона. По этой причине возникла необходимость создать новый идентификатор домохозяйств, уникальный в пределах всей базы, и производить слияние баз по этому новому идентификатору.

Таким образом, рекомендуется обращать внимание: по каким ключам производится слияние баз, насколько полно производится это слияние и какая часть информации может быть утеряна при слиянии. Если эта часть невелика, то данные признаются надежными и могут быть использованы в анализе.

7) иные варианты тестирования на усмотрение оценивающего.

Тестирование данных является **основным способом оценки надежности данных**, с которого необходимо начинать оценку надежности данных, и может быть дополнено применением иных способов, указанных в пункте 3.1 настоящей Методики.

При этом если один из способов оценки позволяет сделать вывод о ненадежности данных, продолжать оценку прочими способами уже не требуется.

3.3. **Анализ документации, описывающей сбор, обработку и выгрузку данных,** состоит в изучении документации, содержащей информацию о следующих процессах:

1) сбор данных;

2) обработка данных, в том числе с описанием процедуры расчета показателей;

3) ввод новых данных в базу данных и контроль за их внесением (например, в поле, содержащее значение "Год" не допустимо включение буквенных значений);

4) выгрузка данных.

Пример

В ходе одного из мероприятий решалась задача прогноза уровня бедности с использованием данных ВНДН. Для этого требовалось проверить веса наблюдений на предмет обеспечения репрезентативности выборки. По итогам изучения методики, используемой Росстатом для вычисления выборочных весов, в контексте существующих подходов других национальных и международных статистических служб, а также после изучения научной литературы был сделан вывод о соответствии методики Росстата лучшим практикам и возможности адаптации этой методики для изменения весов наблюдений домохозяйств с целью включения демографической динамики в качестве одного из факторов в долгосрочный прогноз уровня бедности населения.

3.4. **Интервью с сотрудниками, отвечающими за источник данных.**

В ходе интервью следует получить пояснения по вопросам, возникшим в процессе тестирования данных или анализа документации, описывающей сбор, обработку и выгрузку данных, а также при необходимости воспользоваться примерным перечнем дополнительных вопросов, приведенных в приложении N 3 к настоящей Методике.

Пример

В ходе одного из мероприятий при тестировании данных ВНДН было выявлено несоответствие между суммой доходов членов домохозяйства и полным доходом домохозяйства. Интервью с сотрудниками Росстата, отвечающими за составление ВНДН, выявило, что в состав полного дохода домохозяйства включаются доходы, которые не могут быть отнесены на счет кого-либо из членов этого домохозяйства. Например, доход от реализации продукции приусадебного участка является доходом всех членов домохозяйства в равной степени. Как результат, содержимое переменных дохода домохозяйства было признано корректным, а полученная в интервью информация позволила скорректировать алгоритм расчета модельных показателей.

3.5. **Выборочная сверка с первичным источником данных.**

В случае наличия доступа к первичной информации, на основании которой были получены оцениваемые данные, необходимо посредством выборочной сверки данных с данными из первичного источника, оценить степень точности, с которой эти данные отражают первичную информацию, объем которой определяется в зависимости от результатов иных способов оценки надежности данных.

Приложение N 1
к Методике по оценке надежности
данных, утвержденной Коллегией
Счетной палаты Российской Федерации
(протокол от 15 марта 2022 г. N 12К (1537)

Шаблон оценки надежности данных

1	**Мероприятие (пункт плана работы Счетной палаты)**	
2	**Цель, вопрос, гипотеза мероприятия**	
3	**Оценивающий (Ф.И.О., должность)**	
4	**Описание оцениваемых данных и их источника**	
5	**Способы, использованные для оценки надежности данных (в правой колонке при наличии приводится ссылка на документацию)**	
51	Тестирование данных	
52	Анализ документации, описывающей сбор, обработку и выгрузку данных	
53	Интервью с сотрудниками, отвечающими за источник данных	

54	Выборочная сверка с первичным источником данных	
55	Иной способ оценки надежности данных	
6	**Описание характеристик данных и выявленных ограничений для достижения целей мероприятия (в правой колонке при наличии описываются ограничения)**	
61	Применимость для достижения целей мероприятия	
62	Репрезентативность	
63	Точность	
64	Полнота	
65	Уникальность	
66	Валидность (доменная согласованность)	
7	**Вывод о надежности данных (выберите нужный вариант)**	
	Данные являются надежными для достижения целей мероприятия. Данные являются ненадежными для целей достижения мероприятия и исключены из анализа.	

Приложение N 2
к Методике по оценке надежности
данных, утвержденной Коллегией
Счетной палаты Российской Федерации
(протокол от 15 марта 2022 г. N 12К (1537)

Пример заполненного шаблона оценки надежности данных

1	**Мероприятие (пункт плана работы Счетной палаты Российской Федерации)**
	ЭАМ "Анализ эффективности механизмов государственной поддержки организаций в сфере массовых коммуникаций в 2018-2019 годах и истекшем периоде 2020 года" (пункт 2.4.10.2 Плана работы Счетной палаты Российской Федерации на 2021 год)
2	**Цель, вопрос, гипотеза мероприятия**
	Цель мероприятия:

	оценить взаимосвязь между получением государственной поддержки из средств федерального бюджета и динамикой развития компаний Вопрос мероприятия: проанализировать взаимосвязи между объемами государственной поддержки и темпами роста выручки компаний, основных фондов. Провести сравнительный анализ получателей поддержки со схожими компаниями, не получавшими такую поддержку и выявить статистически значимые различия в динамике их развития за проверяемый период Гипотеза мероприятия: государственной поддержкой в виде субсидий Роспечати пользуется незначительное количество компаний, владеющих СМИ. Основной объем субсидий приходится на 3 компании	
3	**Оценивающий (Ф.И.О., должность)**	
	Иванов Иван Иванович, ведущий инспектор Департамента исследований и методологии аппарата Счетной палаты Российской Федерации	
4	**Описание оцениваемых данных и их источника**	
	Из ГИИС "Электронный бюджет" был выгружен реестр соглашений по следующим параметрам: код ГРБС (по бюджетной классификации): 135; за 2018-2020 года; по состоянию на 2 апреля 2021 года. Данные содержатся в файле "Реестр соглашений_135_ 02.04.2021.xls" и охватывают 2018-2020 годы. Для проверки гипотезы мероприятия релевантны следующие колонки на вкладках: вкладка "Общие сведения": "Уникальный номер"; "Наименование получателя субсидии"; "ИНН получателя"; "Дата принятия соглашения/НПА"; "Размер субсидии, бюджетных инвестиций, межбюджетного трансферта (средств) в рублевом эквиваленте"; вкладка "Платежные документы": "Уникальный номер"; "Дата платежного документа"; "Сумма перечисления / возврата в рублевом эквиваленте"; вкладка "Платежные документы_1": "Уникальный номер"; "Дата платежного документа"; "Сумма перечисления / возврата в рублевом эквиваленте".	
5	**Способы, использованные для оценки надежности данных (в правой колонке при наличии приводится ссылка на документацию)**	
51	Тестирование данных	Файл "Data analysis.ipynb"
52	Анализ документации, описывающей сбор, обработку и выгрузку данных	Не использовался
53	Интервью с сотрудниками,	Не использовался

	отвечающими за источник данных	
54	Выборочная сверка с первичным источником данных	Не использовался
55	Иной способ оценки надежности данных: выборочная сверка данных реестра соглашений с реестром отчетов федеральных органов исполнительной власти, проведенная ФКУ "ЦЭАИТ СП"	Электронное письмо от 12 августа 2021 года, 19:10. Тема "Вопросы по реестру соглашений Роспечати (ГРБС 135) за 2018, 2019 и 2020 годы"
6	**Описание характеристик данных и выявленных ограничений для целей мероприятия (в правой колонке при наличии описываются ограничения)**	
61	Применимость для достижения целей мероприятия	Ограничений нет
62	Репрезентативность	Ограничений нет
63	Точность	Выборочная сверка с первичным источником данных не проводилась
64	Полнота	Данные могут содержать не все платежи по субсидиям: в 2018 году реестр содержит 94 % бюджетных ассигнований, доведенных казначейскими уведомлениями; в 2019 году - 89 %; в 2020 году - 88 %.
65	Уникальность	Ограничений нет
66	Валидность (доменная согласованность)	Ограничений нет
7	**Вывод о надежности данных**	
	Данные являются надежными для достижения целей мероприятия.	

Приложение N 3
к Методике по оценке надежности
данных, утвержденной Коллегией
Счетной палаты Российской Федерации
(протокол от 15 марта 2022 г. N 12К (1537)

Примерный перечень дополнительных вопросов для интервью с сотрудниками, отвечающими за источник данных

1. Опишите порядок внесения новых данных в базу данных, в том числе:

а) систему проверки соответствия внесенной в базу данных информации исходной (например, форматно-логический контроль, дублирование ввода, валидация введенных значений на соответствие

требуемому типу данных, валидация значений посредством выпадающих списков вместо ввода текста для категориальных данных);

б) систему контроля полноты ввода данных (например, сигнализирование о пропуске полей, наличие разделения переменных на обязательные и необязательные);

в) систему перекрестной валидации внесенных данных (например, проверка правильности следования дат наблюдений, сигнализирование о выбросах значений).

2. Опишите процедуру корректировки (изменения) данных в базе данных.

3. Опишите систему контроля корректности обработки данных.

4. Опишите систему контроля корректности создаваемых файлов.

5. Проходят ли пользователи, которые обладают правами ввода и изменения данных, обучение по работе с базой данных до начала работы в ней?

6. Учитывается ли качество данных в базе данных при измерении показателей эффективности сотрудников, отвечающих за ввод или изменение данных в базе данных? Если да, то, каким образом?