УТВЕРЖДЁН
Методологическим комитетом
общества с ограниченной ответственностью
«Национальные Кредитные Рейтинги»
Протокол № 129 от «04» июня 2026 г.

Алгоритмы расчёта и оценки ключевых метрик моделей кредитных рейтингов

Москва

2026

Оглавление

1. Общие положения

Данный методологический документ дополняет Основные понятия, используемые обществом с ограниченной ответственностью «Национальные Кредитные Рейтинги» в методологической и рейтинговой деятельности (далее — Основные понятия), прежде всего, в части принципов разработки и валидации методологий.

Документ содержит описание статистических методов и подходов к расчёту и оценке ключевых метрик, а также соответствующие пороговые значения, используемые НКР при проверке качества методологий как на этапе разработки, так и на этапе валидации.

В случае малого числа кредитных рейтингов или иных признаков нерелевантности соответствующей статистики НКР (например, из-за существенного пересмотра методологии) оценивает ключевые характеристики методологии исходя из соответствующих метрик, рассчитанных на основе тестовых рейтингов.

Пределы погрешности расчётов показателей и метрик зависят от используемой версии MS Excel и (или) Python. НКР исходит из того, что для всех используемых версий MS Excel и (или) Python пределы погрешностей расчёта существенно ниже уровней, при которых они могли бы оказать влияние на решения, принимаемые при разработке, проверке качества и применении методологий.

2. Список терминов и определений

Ниже приведены термины, которые отсутствуют в Основных понятиях либо определения которых в настоящем документе отличаются от Основных понятий.

Бенчмарк — используемый в методологии критерий, позволяющий преобразовать исходное значение показателя в его оценку в определенном диапазоне.

Биномиальный тест — статистический тест, используемый для проверки гипотезы о равенстве наблюдаемых частот дефолтов и ожидаемых вероятностей дефолта.

Бэк-тестирование — оценка качества методологии или проекта методологии, основанная на имитации её применения к датам в прошлом или к уже присвоенным кредитным рейтингам.

Доверительный интервал — интервал, внутри которого с заданной вероятностью находится значение оцениваемого по определенной выборке параметра.

Индекс стабильности системы (PSI) — коэффициент, характеризующий степень различий в структуре двух наборов данных (например, двух временных срезов).

Индекс обусловленности — коэффициент, характеризующий степень мультиколлинеарности, наблюдаемой в наборе данных, рассчитывается на основе матрицы парных корреляций.

Индекс Херфиндаля — Хиршмана (HHI) — показатель, характеризующий степень концентрации (равномерности), наблюдаемых для ряда данных.

Коэффициент корреляции — коэффициент, характеризующий взаимосвязь двух рядов данных.

Коэффициент парной корреляции Пирсона — коэффициент, отражающий величину линейной связи между двумя рядами данных.

Коэффициент ранговой парной корреляции Кендалла — коэффициент, отражающий величину ранговой связи между двумя рядами данных.

Критерий Колмогорова — Смирнова — статистический тест, который позволяет проверить гипотезу о принадлежности двух распределений (например, распределений дефолтных и устойчивых объектов) одному закону распределения путем выявления максимального расхождения между двумя распределениями и оценки достоверности этого расхождения.

Мультиколлинеарность — наличие линейной зависимости внутри набора данных.

Ошибка второго рода — ошибка бинарной классификации, заключающаяся в пропуске события (объект классифицирован как устойчивый, а на самом деле допустил Дефолт). Как правило, для национальной рейтинговой шкалы для Российской Федерации под ошибкой второго рода понимается дефолт объекта рейтинга с рейтингом BBB.ru и выше на горизонте три года.

Ошибка первого рода — ошибка бинарной классификации, заключающаяся в ложном срабатывании (объект классифицирован как дефолтный, но является устойчивым). Как правило, для национальной рейтинговой шкалы для Российской Федерации под ошибкой первого рода понимается отсутствие дефолта объекта рейтинга с рейтингом CCC.ru и ниже на горизонте три года.

Тестовый рейтинг — оценка объекта рейтинга, полученная в рамках бэк-тестирования методологии. Тестовый рейтинг не является кредитным рейтингом и базируется преимущественно на публичных данных. При присвоении тестового рейтинга оцениваются все ключевые факторы соответствующей методологии (включая качественные факторы), хотя в силу недостатка публичной информации некоторые алгоритмы могут быть упрощены.

Accuracy ratio (AR, коэффициент Джини) — отношение площади между ROC-кривой и диагональю к площади между идеальной (соответствующей алгоритму, точно различающему дефолтные и устойчивые объекты) ROC-кривой и диагональю. AR также оценивает кумулятивный профиль точности дифференциации объектов рейтинга, предусматривающий оценку показателя кривой соответствия между долями дефолтных объектов наблюдений в общем числе объектов наблюдений и накопленными долями объектов наблюдений в общем числе объектов наблюдений (CAP-кривая).

AUROC — показатель площади под кривой соответствия между долями дефолтных объектов наблюдений в общем числе объектов наблюдений и долями недефолтных объектов наблюдений в общем числе объектов наблюдений, отражает тесноту линейной связи между рядом данных и дефолтностью.

Bootstrap (бутстреп) — исследование распределения статистик вероятностных распределений, основанное на многократной генерации подвыборок с возвращением на базе имеющейся выборки.

CAP-кривая — график, позволяющий оценить качество бинарной классификации, отображает соотношение между долей верно классифицированных объектов и накопленной долей объектов от общего количества объектов.

P-значение (p-value) — вероятность получить наблюдаемый или более экстремальный результат при условии, что нулевая гипотеза верна.

ROC-кривая — график, позволяющий оценить качество бинарной классификации, отображает соотношение между долей верно классифицированных объектов и долей объектов от общего количества объектов, классифицированных ошибочно при варьировании порога решающего правила.

VIF (фактор инфляции дисперсии) — показатель, характеризующий мультиколлинеарность одной переменной относительно других.

U-тест (U-критерий Манна — Уитни) — критерий, используемый в рамках Положения для проверки гипотезы об отличии AUROC от 0,5. Этот же тест применим для AR (проверка гипотезы об отличии AR от 0) в связи с линейной связью между показателями AUROC и AR.

2.1. Некоторые общепринятые обозначения

$\overline{x}$	выборочное среднее
$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$	биномиальный коэффициент, число способов выбрать k объектов из n (без повторений)
$\Gamma(x) = \int_0^{+\infty} t^{x-1}e^{-t}dt$	гамма-функция Эйлера
$\hat{y}$	предсказанное значение функции y
$\mathbb{R}$	множество вещественных (действительных) чисел
$\mathbb{R}^n$	вещественные векторы длины n (упорядоченные наборы из n чисел)
A^{-1}	обратная к A матрица, т. е. такая что $AA^{-1} = A^{-1}A = I$ (единичная матрица)
$sign(x)$	функция вещественного аргумента, 1 для положительного аргумента, 0 для случая, если аргумент равен 0, -1 для отрицательного аргумента

3. Оценка дифференцирующей и прогностической способностей методологии

3.1. Оценка дифференцирующей способности методологии

НКР обеспечивает проверяемость достоверности кредитных рейтингов, в том числе на основе исторических данных, посредством выявления отклонений между предпосылками и допущениями, используемыми в методологии, и фактической информацией о неплатежах рейтингуемых лиц. Для этого агентство применяет следующие подходы, направленные на установление высокой способности Методологии дифференцировать объекты рейтинга в зависимости от наличия (отсутствия) дефолта рейтингуемого лица:

– Оценка показателя площади под кривой соответствия между долями дефолтных объектов наблюдений в общем числе объектов наблюдений и долями недефолтных объектов наблюдений в общем числе объектов наблюдений (далее — AUROC).
– Оценка кумулятивного профиля точности дифференциации объектов рейтинга, предусматривающая оценку показателя кривой соответствия между долями дефолтных объектов наблюдений в общем числе объектов наблюдений и накопленными долями объектов наблюдений в общем числе объектов наблюдений (далее — AR).
– Проверка критерия Колмогорова — Смирнова, свидетельствующая о различии в распределениях дефолтных и недефолтных объектов наблюдений.

При наличии достаточной статистики дефолтов в рамках оценки дифференцирующей способности методологии проводится анализ примеров ошибок первого и второго рода, который должен включать выявление ключевых факторов, которые могли затруднить правильную классификацию.

При оценке дифференцирующей способности методологии НКР руководствуется пороговыми значениями, приведёнными в таблице 1[1]. В отчётах о проверке качества методологии термины «отличная» и «хорошая» в отношении дифференцирующей способности методологии могут также заменяться термином «высокая».

Таблица 1. **Пороговые значения для оценки дифференцирующей способности методологии**

Дифференцирующая способность методологии	**Модуль AR**	**AUROC**	**Критерий Колмогорова — Смирнова**
слабая	< 0,4	< 0,7	< 0,15
приемлемая	[0,4; 0,6)	[0,7; 0,8)	[0,15; 0,3)
хорошая	[0,6; 0,7]	[0,8; 0,85]	[0,3; 0,4]
отличная	> 0,7	> 0,85	> 0,4

Дополнительно для AUROC и (или) AR оцениваются величины статистической ошибки и доверительных интервалов, кроме случаев, когда размер или структура выборки не позволяют этого сделать. Как правило, оценки рассчитываются с использованием алгоритма bootstrap либо на основе прямого расчёта доверительных интервалов на основе аналитических подходов при их релевантности. Число итераций при bootstrap обычно равно 10 000, но в любом случае не менее 1 000. При построении доверительных интервалов в рамках оценки дифференцирующей способности методологии в целом используется уровень доверия 0,95.

Как правило, оценки дифференцирующей способности методологии осуществляются для наборов дефолтных объектов на горизонте 1 год и 3 года. В качестве целевой переменной при методологии в целом, как правило, используются флаги (метки) фактически допущенных

[1] Пороговые значения установлены на основании российского и зарубежного опыта. См., например, Model Risk for Acceptable, but Imperfect, Discrimination and Calibration in Basel PD and LGD Models. WORKING PAPER SERIES No. 92 / April 2022 Henry Penikas.

дефолтов; для оценки компонентов модели более низкого уровня, как правило, используются флаги, отражающие наличие как дефолтов, так и синтетических дефолтов.

В случае малого числа дефолтных объектов в выборке, применяемой для разработки и (или) валидации методологии, указанные выше методы могут оказаться неприменимыми или неподходящими в силу высокой статистической ошибки. В этом случае основные выводы в отношении дифференцирующей способности методологии могут быть сделаны на основании:

- расчёта AR, критерия Колмогорова — Смирнова с использованием оценок собственной кредитоспособности объектов рейтинга и данных, дополненных информацией о синтетических дефолтах;
- сопоставления с иными аналогами оценки кредитного качества (например, кредитными спредами для методологий, отражающих ожидаемый уровень потерь).

При использовании в расчётах тестовых рейтингов и их компонентов, как правило, учитывается временной разрыв (лаг) между отчётной датой и принятием кредитным рейтинговым агентством (далее — КРА) решения о кредитном рейтинге. Этот лаг включает, прежде всего, время, необходимое для подготовки финансовой отчётности, её передачи КРА и её последующей обработки.

3.2. Оценка прогностической способности методологии

Чтобы обеспечить соответствие мнения НКР о способности рейтингуемых лиц исполнять принятые на себя финансовые обязательства и (или) о кредитном риске их отдельных финансовых обязательств или финансовых инструментов фактической информации о такой способности (таком кредитном риске), агентство проводит оценку прогностической способности методологии. Для этого НКР проводит биномиальный тест, определяющий для каждого уровня кредитного рейтинга степень соответствия единой для всех методологий ожидаемой вероятности дефолта фактической частоте дефолтов, рассчитанной для данной методологии. При проведении биномиального теста используется уровень значимости 5% (уровень доверия 95%).

4. Оценка стабильности кредитных рейтингов

4.1. Общие подходы

НКР оценивает стабильность кредитных рейтингов, присвоенных рейтингуемому лицу, в течение календарного года (при условии неизменности в течение таких периодов способности рейтингуемого лица исполнять принятые на себя финансовые обязательства с учётом факторов внешнего влияния) следующими способами, направленными на установление факта высокой стабильности кредитных рейтингов:

- сравнение индекса стабильности системы (далее — PSI), демонстрирующего стабильность модели, используемой в Методологии, на горизонте 1 год, с используемыми НКР пороговыми значениями такого индекса за период не менее

3 последних календарных лет, предшествующих дате сравнения;

- анализ матрицы миграции кредитных рейтингов по уровням кредитного рейтинга за период не менее 3 последних календарных лет, предшествующих дате анализа;
- ретроспективный анализ влияния изменений методологии НКР на кредитные рейтинги, присвоенные НКР, за период не менее 3 последних календарных лет, предшествующих дате анализа;
- иные способы, применение которых обосновано (например, тест Фишера для малых и небольших сравниваемых наборов наблюдений).

Указанные выше методы, как правило, применяются также для оценки стабильности тестовых рейтингов, поэтому далее НКР использует также термин «стабильность рейтингов» как синоним терминов «стабильность тестовых рейтингов» и «стабильность кредитных рейтингов».

4.2. Оценка стабильности рейтингов в разрезе отдельных периодов

НКР использует PSI в качестве базовой метрики для оценки стабильности рейтингов в разрезе отдельных периодов, однако это может быть сопряжено с рядом трудностей (чувствительность к методу заполнения пустых категорий, изменению состава объектов рейтинга и т.д.), которые чаще всего проявляются в случае малых и небольших сравниваемых наборов наблюдений (срезов). В этой связи НКР рассчитывает PSI в разрезе рейтинговых категорий, а также проводит оценку обоснованности применения для отдельных периодов (срезов наблюдений) альтернативных метрик, характеризующих стабильность рейтингов.

Оценка обоснованности применения для отдельных периодов альтернативных метрик, характеризующих стабильность рейтингов, проводится на основе таблицы 2.

Таблица 2. **Оценка обоснованности применения для отдельных периодов альтернативных метрик, характеризующих стабильность рейтингов**

Минимальное число наблюдений в сравниваемых срезах	**Доля общих объектов рейтинга в сравниваемых срезах**	**Чувствительность расчёта PSI к методу заполнения пустых категорий**	**Обоснованный подход к оценке отдельного периода**
>= 100	50% и менее	низкая	Стандартная формула PSI
>= 100	> 50%	низкая	Расчёт PSI по объектам, общим для обоих срезов (PSI like-to-like)
>= 100	–	высокая	Тест Фишера
(100;3)	–	–	Тест Фишера
<=3	–	–	Оценка стабильности рейтингов невозможна

Пороговые значения для PSI приведены в п. 4.4. Интерпретация результатов теста Фишера основана на сравнении полученного p-value с широко используемыми пороговыми значениями для уровня значимости. Если p-value <=0,05, то стабильность рейтингов для рассматриваемого периода признаётся низкой, если p-value в интервале (0,05;0,15] — умеренной, в остальных случаях стабильность рейтингов оценивается как высокая.

При оценке чувствительности PSI к методу заполнения пустых категорий применяется следующий алгоритм:

- проводится расчёт двух вариантов PSI — с заполнением пустых категорий константами 0,01 и 0,001.
- если разница между полученными выше PSI составляет менее 10% или менее 0,02, то чувствительность расчёта PSI к методу заполнения пустых категорий признаётся низкой, и при расчёте PSI для пустых категорий используется значение 0,01. В противном случае НКР признаёт, что чувствительность расчёта PSI к методу заполнения пустых категорий является высокой.

4.3. Оценка стабильности рейтингов для методологии в целом

После оценки обоснованности применения для отдельных периодов альтернативных метрик, характеризующих стабильность рейтингов, НКР проводит оценку стабильности рейтингов для выборки целом.

Если оценка стабильности рейтингов исходя из таблицы 2 возможна более чем для 80% периодов, оценка стабильности рейтингов по выборке в целом проводится на основе таблицы 3.

Таблица 3. **Оценка стабильности рейтингов по выборке в целом, если оценка стабильности рейтингов возможна для не менее чем 80% периодов**

Оценка стабильности для периода	**Доля периодов (от количества периодов, для которых оценка стабильности возможна), для которых определена указанная оценка**	**Оценка стабильности рейтингов по всей выборке**
высокая	>75%	высокая
низкая	>50%	низкая
все другие случаи		умеренная

Если доля периодов, для которых возможна оценка стабильности рейтингов исходя из таблицы 3, менее 80%, то НКР делает вывод в отношении стабильности рейтингов на основе метрик, рассчитанных на основе агрегации (за все рассматриваемые периоды) данных из однолетних матриц миграции рейтингов по уровням рейтинга, а именно:

- доля рейтингов без изменения (в течение 1 года) по выборке в целом,
- доля рейтингов, изменившихся на 3 и более уровней (в течение 1 года) по выборке в целом.

Таблица 4. **Оценка стабильности рейтингов по выборке в целом, если оценка стабильности рейтингов возможна для не более 80% периодов**[2]

Доля рейтингов без изменения	**Доля рейтингов, изменившихся на 3 и более уровней**		
	0,1 и ниже	**(0,1; 0,15)**	**0,15 и выше**
0,65 и выше	высокая	высокая	приемлемая
[0,45;0,65)	высокая	приемлемая	низкая
Ниже 0,45	приемлемая	низкая	низкая

Интерпретация результатов теста Фишера основана на сравнении полученного p-value с широко используемыми пороговыми значениями для уровня значимости. Если p-value <=0,05, то стабильность рейтингов для рассматриваемого периода признаётся низкой, если p-value в интервале (0,05; 0,15] — умеренной, в остальных случаях стабильность рейтингов оценивается как высокая.

4.4. Пороговые значения для PSI

Поскольку значения PSI чувствительны к числу категорий и наблюдений в сравниваемых срезах, для оценки PSI и PSI like-to-like для отдельного периода НКР использует пороговые значения, установленные на основе в т.ч. предельно допустимых уровней значимости для PSI[3] и зависящие от размеров сравниваемых выборок и числа корзин (категорий). Пороговые значения в таблице 5, превышающие 0,1, рассчитаны для уровня значимости 0,15, семи категорий и распределения хи-квадрат (пороговые значения указаны в долях).

При поиске порогового значения PSI в таблице 5 выбираются строка и столбец, ближайшие к размерам сравниваемых выборок. Если же 2 строки (столбца) равноудалены от фактического размера сравниваемой выборки, то выбираются строка (столбец), соответствующие большему размеру выборки. Далее выбранное пороговое значение (пересечение выбранных выше строки и столбца в таблице 5) сравнивается с фактическим значением PSI по следующим правилам:

- Если фактическое значение PSI ⩾ 0,25, то уровень стабильности рейтингов (для рассматриваемого периода) оценивается как «низкий».
- Если фактическое значение PSI лежит между пороговым значением из таблицы 5 и 0,25, то уровень стабильности рейтингов (для рассматриваемого периода) оценивается как «приемлемый».
- Если фактическое значение PSI не превышает порогового значения из таблицы 5, то уровень стабильности рейтингов (для рассматриваемого периода) оценивается как

[2] Указанные пороговые значения установлены с учётом стабильности кредитных рейтингов, присвоенных российскими КРА, а также меньшей стабильности Тестовых рейтингов по сравнению с кредитными рейтингами. По оценкам НКР, доля рейтингов без изменения на горизонте 1 год для разных КРА в период с 01.01.18 по 01.01.24 составляет ~60-85%.

[3] Уровни значимости, при которых начинает приниматься гипотеза об отличии PSI от нуля (т. е. различии структуры сравниваемых временных срезов), хотя до этого принималась гипотеза о равенстве PSI нулю (т. е. схожести структуры сравниваемых временных срезов). См. Yurdakul B. Statistical Properties of Population Stability Index, 2018.

«высокий».

Таблица 5. **Пороговые значения для PSI**

		Размер среза 1										
		100	110	120	130	140	150	160	170	180	190	200
Размер среза 2	100	0,19	0,18	0,17	0,17	0,16	0,16	0,15	0,15	0,15	0,14	0,14
	110	0,18	0,17	0,16	0,16	0,15	0,15	0,14	0,14	0,14	0,14	0,13
	120	0,17	0,16	0,16	0,15	0,15	0,14	0,14	0,13	0,13	0,13	0,13
	130	0,17	0,16	0,15	0,15	0,14	0,14	0,13	0,13	0,13	0,12	0,12
	140	0,16	0,15	0,15	0,14	0,13	0,13	0,13	0,12	0,12	0,12	0,11
	150	0,16	0,15	0,14	0,14	0,13	0,13	0,12	0,12	0,12	0,11	0,11
	160	0,15	0,14	0,14	0,13	0,13	0,12	0,12	0,11	0,11	0,11	0,11
	170	0,15	0,14	0,13	0,13	0,12	0,12	0,11	0,11	0,11	0,11	0,1
	180	0,15	0,14	0,13	0,13	0,12	0,12	0,11	0,11	0,1	0,1	0,1
	190	0,14	0,14	0,13	0,12	0,12	0,11	0,11	0,11	0,1	0,1	0,1
	200	0,14	0,13	0,13	0,12	0,11	0,11	0,11	0,1	0,1	0,1	0,1

5. Оценка иных характеристик методологии, отдельных показателей методологии и их групп

Оценка концентрации (равномерности) распределения по рейтинговой шкале осуществляется посредством расчёта Индекса Херфиндаля — Хиршмана (далее — HHI). Концентрация признаётся низкой, если HHI < 1 500, умеренной, если HHI находится в диапазоне [1 500; 2 500], высокой — при значении HHI выше 2 500[4].

Дополнительно НКР может проверять корректность распределения тестовых рейтингов по рейтинговой шкале, что важно для выборок с преобладанием наблюдений без дефолта. Для этой цели НКР может использовать корреляцию Кендалла с эталонными рангами и аналогичные метрики, где эталонные ранги отражают имеющуюся информацию о кредитном качестве объектов рейтинга, с учётом в т.ч. оценок НКР и иных кредитных рейтинговых агентств. Корреляция с эталонными рангами может использоваться также при оценке показателей методологии и иных компонентов модели.

Как правило, НКР оценивает дифференцирующую способность отдельных показателей (или их групп), оценивая значимость отличий AR показателя (групп показателей) от нуля (на основе U-теста). Такая проверка, в отличие от использования фиксированных пороговых значений, позволяет учесть размер выборки, используемой для оценки AR.

[4] Указанные пороговые значения установлены с учётом международного опыта, см., например, https://corporatefinanceinstitute.com/resources/valuation/herfindahl-hirschman-index-hhi/

Если не обосновано иное, при рассмотрении компонентов методологии и соответствующей ей модели (показатели, субфакторы, факторы, оценка собственной кредитоспособности) для оценки статистической значимости используется уровень значимости 0,1 (уровень доверия 0,9), в то время как при оценке методологии (модели) в целом, если не обосновано иное, используется уровень значимости 0,05 (уровень доверия 0,95). Использование при тестировании компонентов модели менее строгого уровня значимости позволяет снизить вероятность исключения из модели показателей, которые способны улучшить работу модели в целом, но демонстрируют не очень высокие индивидуальные метрики (например, в силу особенностей выборки).

В рамках анализа отдельных показателей может проводиться оценка оптимальности существующих бенчмарков / функций преобразования значений в оценку с точки зрения максимизации дифференцирующей способности (либо иной метрики) отдельного показателя или группы показателей.

Наличие корреляции и мультиколлинеарности по группам показателей (субфакторов, факторов) и модели в целом, как правило, оценивается с помощью расчётов:

- матриц корреляций (для непрерывных рядов данных — на основе коэффициента парной корреляции Пирсона; если хотя бы один из рядов дискретный — дополнительно на основе коэффициента ранговой парной корреляции Кендалла, если не обоснована нерелевантность данной метрики);
- оценок мультиколлинеарности факторов на основе пошаговых VIF-тестов и общего коэффициента мультиколлинеарности, рассчитываемого как индекс обусловленности.

Использование на одном уровне агрегации рядов данных с коэффициентами парной корреляции свыше 0,7 и (или) значениями VIF свыше 10 и (или) значениями индекса обусловленности свыше 30 требует дополнительного обоснования[5].

6. Метрики, характеризующие дифференцирующую и прогностическую способности методологии

6.1. Алгоритм расчёта AUROC и AR

Интервал значений анализируемой переменной разбивается на части (для дискретных переменных, как правило, берутся все возможные значения).

Для каждой граничной точки из шага (1) рассчитываются FPR (false positive rate) и TPR (true positive rate):

$$FPR = \frac{FP}{TN + FP}, \qquad TPR = \frac{TP}{TP + FN},$$

где:

[5] Указанные пороговые значения установлены с учётом международного опыта, см., например, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6900425/

FP — False Positive, количество наблюдений, которые были неправильно классифицированы как дефолтные (ошибки первого рода),
TP — True Positive, количество наблюдений, которые были правильно классифицированы как дефолтные,
FN — False Negative, количество наблюдений, которые были неправильно классифицированы как недефолтные (ошибки второго рода),
TN — True Negative, количество наблюдений, которые были правильно классифицированы как недефолтные.

Тогда:

$$AUROC = \int_0^1 R(x)dx,$$

где:
$R(x)$ — кривая, построенная по точкам из шага (2), т. н. ROC-кривая, где по оси OX отложены FPR, а по оси OY — TPR.

Accuracy ratio (AR) рассчитывается по формуле:

$$AR = 2 * AUROC - 1$$

6.2. Оценка стандартной ошибки и доверительных интервалов AUROC и AR

Выделяются два способа: аналитический и численный. НКР имеет право выбора используемого алгоритма исходя из специфики выборки.

6.2.1. Аналитический способ

1. Высчитывается стандартная ошибка для AUROC по формуле:

$$SE = \sqrt{\frac{AUROC(1 - AUROC) + (N_1 - 1)(Q_1 - AUROC^2) + (N_2 - 1)(Q_2 - AUROC^2)}{N_1 N_2}}$$

где:
$Q_1 = \frac{AUROC}{2 - AUROC}$,
$Q_2 = \frac{2AUROC^2}{1 + AUROC}$,
N_1 — количество дефолтных наблюдений,
N_2 — количество недефолтных наблюдений.

2. Для построения доверительного интервала необходимо вычислить коэффициент Z — квантиль стандартного нормального распределения при заданном уровне значимости, деленном на 2 ($\frac{\alpha}{2}$).

3. Рассчитываются нижний и верхний пределы доверительного интервала:

$$AUROC_{low} = AUROC - Z * SE,$$

$$AUROC_{upper} = AUROC + Z * SE$$

6.2.2. **Численный способ**

Способ основан на многократной генерации выборок методом Монте-Карло на базе имеющейся выборки. Как правило, количество итераций берется равным 10 000.

1. Исходная выборка: имеется исходная выборка данных, состоящая из N наблюдений.
2. Генерация подвыборок:
    a) выбор M случайных объектов (компаний) из исходной выборки с возвращением;
    b) у каждого объекта (компании) имеется N_M наблюдений, на основании этих наблюдений формируется подвыборка для дальнейшего вычисления искомой статистики.
3. Для каждой подвыборки вычисляется искомая статистика, в данном случае — AUROC.
4. Вычисление стандартной ошибки AUROC как среднеквадратического отклонения распределения значений, полученных на шаге (3).
5. Вычисление доверительного интервала для AUROC, через квантили распределения значений, полученных в шаге (3).

Аналитический и численный способы расчёта стандартной ошибки AR и доверительных интервалов аналогичны расчётам для AUROC, имея в виду равенство:

$$AR = 2 * AUROC - 1$$

6.3. U-статистика Манна — Уитни (U-Test Mann — Whitney)

НКР использует U-тест для проверки гипотезы об отличии AUROC от 0,5 (эквивалентно проверке гипотезы об отличии AR от 0). Данный тест основан на сопоставлении двух выборок и выявлении статистически значимых различий между ними. Для указанных выше целей в качестве сравниваемых выборок выступают набор наблюдений с флагом дефолта и набор наблюдений без флага дефолта.

Итак, пусть имеются две выборки объёмом n_1 и n_2, каждому элементу которых приписан некоторый числовой параметр. Мы желаем понять, насколько велико различие между выборками по уровню этого параметра.

Алгоритм расчёта:

1. Элементы обеих выборок упорядочиваются в один список по возрастанию параметра, в случае совпадения параметров у нескольких элементов, их номера в списке усредняются.
2. Подсчитывается отдельно сумма порядковых номеров, пришедшихся на долю элементов первой выборки — R_1 и отдельно для второй — R_2.

3. Рассчитывается значение U-статистики Манна — Уитни (используется минимум из двух значений U_1 и U_2)

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

4. Далее значение статистики U сравнивается с критическим значением.

5. При больших выборках для расчёта $pvalue$ можно пользоваться тем, что распределение величины

$$\left[\frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12} - \frac{T}{(n_1+n_2)(n_1+n_2-1)}}}\right]$$

близко к стандартному нормальному, где $T = \sum(t_i^3 - t_i)$ — сумма по размерам групп повторов.

6. Нулевая гипотеза H_0 (различие отсутствует) отвергается, если $\alpha > pvalue$, где α — заданный уровень значимости.

6.4. Тест Колмогорова-Смирнова

НКР использует тест Колмогорова-Смирнова, чтобы сравнить эмпирические функции распределения (ECDFs) двух выборок и определить, происходят ли они из одного и того же базового распределения. Вычисляется статистика Колмогорова-Смирнова как максимальное расстояние по вертикали между двумя ECDF.

Нулевая гипотеза: оба вектора данных получены из одного распределения.

Альтернативная гипотеза: векторы данных получены из различных распределений.

Формула расчёта:

$$D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)|,$$

где:

$\sup_x$ — супремум функции (т. е. максимальное значение),

$F_{1,n}$ и $F_{2,m}$ — эмпирические функции распределения.

Критическое значение для критерия Колмогорова-Смирнова:

$$\sqrt{-\ln\left(\frac{\alpha}{2}\right) \cdot \frac{1+\frac{m}{n}}{2m}},$$

где:

α — уровень значимости,

n и m — размеры первого и второго вектора.

Если статистика теста превышает это значение, нулевая гипотеза отклоняется.

6.5. Биномиальный тест

Пусть дана выборка, полученная из некоторой схемы Бернулли (на каждом шаге проводится испытание с некой фиксированной вероятностью успеха, испытания независимы). Двусторонний биномиальный тест позволяет проверить гипотезу H_0: «наблюдаемая частота дефолтов $= p$».

1. По имеющейся выборке из n испытаний с k успехами вычисляется P-значение ($pvalue$)

$$P(X < k) = \sum_{i=0}^{k} \binom{n}{k} p^k (1-p)^{n-k}$$

2. Полученное значение $pvalue$ сравнивается с заданным уровнем значимости, что позволяет сделать вывод, может ли быть принята нулевая гипотеза:

$$pvalue \leq a \rightarrow H_0 \text{ отвергается}$$
$$pvalue > a \rightarrow \text{нет оснований отвергнуть } H_0$$

6.6. Особенности определения флагов дефолта

Для выставления флагов дефолта для каждого отдельного наблюдения (тестового рейтинга объекта на дату), на соответствующем горизонте (12, 24, 36 месяцев, иные горизонты) определяется зафиксировано ли событие дефолта. Если событие дефолта присутствует, то выставляется флаг 1, если отсутствует — флаг 0.

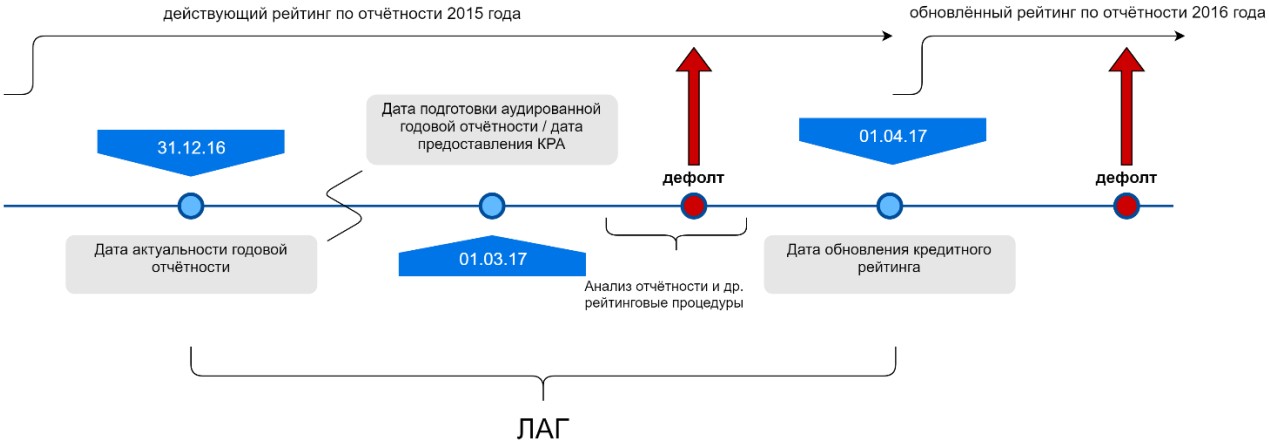
При работе с тестовыми рейтингами, как правило, учитывается временно́й лаг, который отражает, прежде всего, время, необходимое для подготовки финансовой отчётности и её обработки рейтинговым агентством. Это позволяет обеспечить сопоставимость кредитных рейтингов и тестовых рейтингов.

Особенности учёта лагов могут существенно влиять на флаги дефолта и, как следствие, на оценки AUROC и AR. Так, если получение и обработка отчётности рейтинговым агентством занимают 3 месяца после завершения года, то рейтинг, учитывающий итоги 2016 года, появится не ранее 1 апреля 2017 года. А рейтинг, фактически действующий на 1 января 2017 года, в лучшем случае будет отражать отчётность объекта рейтинга на 1 октября 2016 года. Эта особенность учитывается, через смещения окна наблюдения за дефолтом на соответствующий лаг: к примеру, в случае трехмесячного лага тестовый рейтинг, рассчитанный по данным за 2016 год, сопоставлялся с дефолтами, которые произошли в период с 1 апреля 2017 по 1 апреля 2020 г. (дефолты, зафиксированные до 1 апреля 2017 года в этом случае должны быть «предсказаны» тестовыми рейтингами, рассчитанными на более ранних данных — по итогам 2013, 2014 или 2015 года).

Периоды времени, необходимые для получения и обработки отчётности, существенно различаются и зависят от типа объекта, используемые при разработке лаги раскрываются в соответствующих отчётах о проверке качества методологий.

При анализе фактически присвоенных кредитных рейтингов лаги не используются.

Рисунок 1. **Иллюстрация сдвига окна наблюдения за дефолтом**



7. Показатели, характеризующие корреляцию и мультиколлинеарность

7.1. Парная корреляция Пирсона

Для выборок $x_1, \dots, x_n$ и $y_1, \dots, y_n$ коэффициент корреляции Пирсона вычисляется по формуле:

$$r_{x,y} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \cdot \sum(y_i - \overline{y})^2}}$$

Для проверки нулевой гипотезы (корреляция равна нулю) вычисляется статистика критерия по формуле:

$$T = \frac{r_{x,y} \cdot \sqrt{n-2}}{\sqrt{1 - r_{x,y}^2}}$$

p-value вычисляется как $P(|\xi| \geq |T|)$, где ξ — случайная величина с плотностью распределения Стьюдента, а именно:

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}\left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} \qquad \text{где} \quad \nu = n - 2 \quad \text{(степени свободы)}$$

Таким образом:

$$pvalue = 2(1 - \int_{-\infty}^{|T|} f(s)ds)$$

Нулевая гипотеза отвергается, если $pvalue < \alpha$, где α — выбранный уровень значимости.

7.2. **Парная ранговая корреляция Кендалла** ($r_{kendall}$)

Пусть $x_1, \dots, x_n$ и $y_1, \dots, y_n$ — две выборки. Будем говорить, что пары (x_i, y_i) и (x_j, y_j) $(i \neq j)$ согласованы, если $sign(x_i - x_j) = sign(y_i - y_j)$. Тогда коэффициент корреляции Кендалла вычисляется как

$$r_{kendall} = \frac{N_T - N_F}{N} = 1 - \frac{2 \cdot N_F}{\binom{n}{2}},$$

где:

N_T — количество согласованных пар,

N_F — количество не согласованных пар,

N — всевозможное количество пар.

Для проверки гипотезы об отличии корреляции от 0 вычисляется статистика критерия по формуле:

$$\tau = \frac{r_{kendall}}{\sqrt{(2(2n+5)/9n(n-1))}}$$

Нулевая гипотеза (об отсутствии корреляции) отвергается, если $|\tau|$ превышает

$$(1 - \alpha/2) \text{ — квантиль стандартного нормального распределения}$$

Аппроксимация работает удовлетворительно при $n \geq 10$. Так как $pvalue$ по определению вычисляется как $pvalue = P(|\xi| \geq |\tau|) = 2\big(1 - \Phi(|\tau|)\big)$, где ξ — стандартная нормальная случайная величина, то описанный критерий в точности означает, что нулевая гипотеза отвергается при $pvalue < \alpha$.

При необходимости доверительный интервал для корреляции Кендалла рассчитывается методом bootstrap: из исходной выборки формируется большое количество подвыборок с повторами, на каждой из них считается метрика, далее для получившегося распределения оценивается доверительный интервал с использованием соответствующих процентилей.

Если в данных много повторений (ties) нужно использовать ранговую корреляцию Кенделла с поправкой на повторы (Kendell τ-b). В нём учитываются пары, связанные (с повторениями) либо по x, либо по y:

$$r_{kendall\ \tau-b} = \frac{N_T - N_F}{\sqrt{(N_T + N_F + T_x)(N_T + N_F + T_y)}},$$

где:

T_x — количество пар, связанных по x,

T_y — количество пар, связанных по y.

Для проверки гипотезы об отличии корреляции от 0 вычисляется статистика критерия по формуле:

$$\tau_b = \frac{r_{kendall\ \tau-b}}{\sqrt{Var}}$$

где

$$Var = \frac{1}{18}(n(n-1)(2n+5) - x_1 - y_1) + \frac{2x_{tie}y_{tie}}{n(n-1)} + \frac{x_0 y_0}{9n(n-1)(n-2)}$$

x_1, y_1 — основные поправки на повторы,

x_{tie}, y_{tie} — количество пар связанных по x и по y соответственно,

x_0, y_0 — количество упорядоченных троек внутри одинаковых рангов.

7.3. Тест VIF (Variance Inflation Factor)

Пусть поставлена задача построения по выборке (X_i, y_i) где $X_i \in \mathbb{R}^n, y_i \in \mathbb{R}$ линейной регрессии $\hat{y} = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$, тогда в целях проверки на мультиколлинеарность для каждой компоненты x_j можно вычислить

$$VIF_j = \frac{1}{(1 - R_j^2)},$$

где:

R_j^2 — коэффициент детерминации, полученный из линейной регрессии, в которой x_j трактуется как целевая переменная, а $x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n$ — как предикторы.

Дополнительно вектор $VIF \in \mathbb{R}^n$ можно найти как $VIF = diag(A^{-1})$, где A — матрица попарных корреляций Пирсона X_i.

7.4. Condition index (индекс обусловленности, $cond_{index}$)

Пусть поставлена задача построения по выборке (X_i, y_i) где $X_i \in \mathbb{R}^n, y_i \in \mathbb{R}$ линейной регрессии $\hat{y} = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$, тогда в целях проверки на мультиколлинеарность можно вычислить индекс обусловленности по следующему алгоритму:

1. Вычисляется матрица парных корреляций X_{corr}.
2. Проверяется определитель матрицы попарных корреляций, при $\det(X_{corr}) = 0$ расчёт прекращается, мультиколлинеарность считается полной, в противном случае — переход на шаг (3).
3. Поиск собственных значений матрицы парных корреляций $\lambda_i, i = 1, 2, 3, \ldots, N$.
4. Поиск индекса обусловленности $cond\_index$.

$$cond_{index} = \sqrt{\frac{\max \lambda}{\min \lambda}},$$

где λ пробегает собственные значения матрицы.

8. Метрики качества линейной регрессии

8.1. Общие подходы к подбору параметров линейной регрессии

Линейная регрессия — метод статистического анализа, который используется для изучения и моделирования зависимости между одной зависимой переменной (также

называемой откликом или целевой переменной) и одной или несколькими независимыми переменными (также называемыми предикторами или объясняющими переменными).

Линейная регрессия может строиться НКР как с использованием расчётных инструментов Excel (надстройка «Анализ данных»), так и с использованием иных расчётных инструментов.

Основная цель линейной регрессии — найти линейную функцию, которая как можно лучше описывает связь между этими переменными. Эта функция имеет вид:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m,$$

где:

$\hat{y}$ — предсказанное значение зависимой переменной,

β_0 — свободный член (иногда называется интерцептом),

$\beta_1, \ldots, \beta_m$ — коэффициенты регрессии, которые измеряют влияние соответствующих независимых переменных,

$x_1, \ldots, x_m$ — значения независимых переменных.

Алгоритм расчёта:

1. Определение уравнения регрессии. Задаётся уравнение регрессии в линейной форме.
2. Оценка коэффициентов. Коэффициенты $\beta_1, \ldots, \beta_m$ оцениваются с помощью метода наименьших квадратов (МНК). Этот метод минимизирует сумму квадратов остатков (разности между наблюдаемыми значениями y и предсказанными значениями $\hat{y}$:

$$\text{Сумма квадратов остатков (RSS)} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Чтобы минимизировать RSS, решается система уравнений, известная как нормальные уравнения:

$$X^T X\beta = X^T y,$$

где:

X — матрица независимых переменных,

y — вектор зависимых переменных,

β — вектор коэффициентов.

3. Оценка точности модели. После получения коэффициентов проверяется точность модели с использованием метрик, таких как коэффициент детерминации R^2 или среднеквадратичная ошибка (MSE), стандартные ошибки коэффициентов регрессии (SE).

$$MSE = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2},$$

где:

$\hat{y}$ — предсказанное значение зависимой переменной.

SE коэффициентов регрессии:

$$SE(\beta)_i = \frac{SE(MSE \text{ в частных случаях})}{\sqrt{n} \cdot S(x_j)}$$

где:

SE — стандартная ошибка регрессии,

$S(x_j)$ — стандартное отклонение каждой независимой переменной.

8.2. Проверка статистической значимости коэффициента детерминации регрессии

В целях обоснования статистической значимости коэффициента детерминации регрессии НКР может использовать один из эквивалентных вариантов расчёта:

1. Рассчитывается F-статистика, полученное значение сравнивается с критическим значением распределения Фишера со степенями свободы *(k, n-k-1) или (k, n-k),* когда свободный член не учитывается, и выбранным уровнем значимости.

   F-статистика может быть рассчитана по формуле:

   $$F_{stat} = \frac{\frac{R^2}{k}}{\frac{(1-R^2)}{n-k-1}},$$

   где:

   k — количество независимых переменных (т.е. исключая константу),

   n — объём выборки.

   Если $F_{stat} > F_{\text{крит}}$, то НКР признаёт коэффициент детерминации регрессии *статистически значимым* (эквивалентно отклонению нулевой гипотезы о равенстве нулю всех коэффициентов при независимых переменных регрессии).

   Если $F_{stat} \leq F_{\text{крит}}$, то НКР признаёт коэффициент детерминации регрессии *статистически незначимым* (эквивалентно принятию нулевой гипотезы о равенстве нулю всех коэффициентов при независимых переменных регрессии).

2. Расчёт *p-value* для F-статистики и её сравнение с уровнем значимости:

   $$pvalue = P(F_{stat} \geq F_{\text{крит}} | H_0),$$
   $$F_{stat} > F_{\text{крит}} \Leftrightarrow pvalue < \alpha$$

   Если $pvalue < \alpha$, то НКР признаёт коэффициент детерминации регрессии статистически значимым (эквивалентно отклонению нулевой гипотезы о равенстве нулю всех коэффициентов при независимых переменных регрессии).

   Если $pvalue \geq \alpha$, то НКР признаёт коэффициент детерминации регрессии статистически незначимым (эквивалентно принятию нулевой гипотезы о равенстве нулю всех коэффициентов при независимых переменных регрессии).

8.3. Проверка статистической значимости коэффициентов регрессии

В целях обоснования статистической значимости коэффициентов регрессии НКР использует расчёт p-value.

Для каждой независимой переменной рассчитывается её собственное значение $p\text{-}value$ для t-статистики, которое сравнивается с данным уровнем α. В случае если $pvalue < \alpha$, принимается альтернативная гипотеза о значимости коэффициента линейной регрессии. В противном случае принимается нулевая гипотеза о равенстве нулю (или незначимости) коэффициента регрессии.

t-статистика для $\hat{\beta}_j$:

$$t = \hat{\beta}_j / SE(\hat{\beta}_j)$$

p-value (двусторонний тест):

$$p = 2 \times P(T > |t|),$$

где:

$T \sim t(n - k - 1)$

9. Индекс Херфиндаля — Хиршмана (HHI)

$$HHI = 10000 \cdot \sum_{i=1}^{n} S_i^2,$$

где:

S_i — доля объектов в i-м уровне рейтинга.

10. Метрики, характеризующие стабильность кредитных рейтингов

10.1. Матрица миграции[6] рейтингов

Алгоритм расчёта:

1. Определяется шаг для вычисления переходов (как правило, 1 год).
2. Рассчитываются переходы из состояния i в состояние j (при условии наличия информации об обоих состояниях на соответствующие даты) для каждого шага отдельно.
3. Для каждой пары $\{i; j\}$ оценивается вероятность перехода за один шаг из состояния i в состояние j, с помощью отношения

$$\frac{N_{ij}}{N_{total\ i}},$$

где:

$N_{total\ i}$ — сумма переходов из i-го состояния за рассматриваемый период времени.

6 Матрица переходных вероятностей Марковского случайного процесса с дискретным временем.

10.2. Расчёта PSI

1. Разделение данных на базовый и тестовый периоды: необходимо разделить данные на две выборки — базовую и тестовую. Например, в случае использования PSI для оценки стабильности кредитных рейтингов в течение 1 года базовая выборка — это рейтинги на дату T, тестовая — рейтинги на дату T+1 год.
2. Расчёт процентного распределения: для каждой выборки необходимо вычислить процентное распределение в разрезе возможных значений анализируемого показателя.
3. Формула расчёта PSI:

$$PSI = \sum\left((P_{test} - P_{base})\right) * \ln {}^{P_{test}}/_{P_{base}}$$

где:
P_{test} — процентное распределение значений переменной в тестовой выборке,
P_{base} — процентное распределение значений переменной в базовой выборке.

10.3. Тест Фишера

Для оценки стабильности рейтингов НКР использует описанный ниже приближенный тест Фишера для мультиноминальной матрицы (расчёты через хи-квадрат)[7].

Пусть даны два категориальных распределения по C категориям:

$$O = \begin{bmatrix} O_{1,1} & O_{1,2} & \dots & O_{1,C} \\ O_{2,1} & O_{2,2} & \dots & O_{2,C} \end{bmatrix}$$

где:

- первая строка $(O_{1,j})$ — базовое распределение (ожидаемое),
- вторая строка $(O_{2,j})$ — текущее распределение (наблюдаемое).

Удаляются категории, в которых суммарное количество наблюдений равно нулю.

$$J = \{j, O_{1,j} + O_{2,j} > 0\}$$
$$O' = O_{\cdot,J}$$

Обозначим:

$$R_i = \sum_j O'_{i,j}$$
$$C_j = \sum_i O'_{i,j}$$
$$N = \sum_{i,j} O'_{i,j}$$

[7] Описание точного теста Фишера см. https://ru.wikipedia.org/wiki/%D0%A2%D0%BE%D1%87%D0%BD%D1%8B%D0%B9_%D1%82%D0%B5%D1%81%D1%82_%D0%A4%D0%B8%D1%88%D0%B5%D1%80%D0%B0.
В отношении приближенного теста см. https://arxiv.org/abs/2004.00973, https://doi.org/10.1002/9781118445112.stat06165

$$E_{i,j} = \frac{R_i C_j}{N}$$

Случай r×c, (r > 2 или c > 2):

Если $\min_{i,j} E_{i,j} \geq 5$ используется стандартный асимптотический χ^2-тест Пирсона:

$$\chi^2 = \sum_{i,j} \frac{\left(O'_{i,j} - E_{i,j}\right)^2}{E_{i,j}}$$

со степенями свободы

$$df = (r-1)(c-1)$$

В противном случае:

Пусть χ^2_{obs} — наблюдаемая статистика. Методом Монте-Карло генерируются B выборок (с сохранением маргинальных сумм, т. е. сумм по строкам и столбцам исходной матрицы), вычисляется статистика χ^2_b, и тогда

$$p = \frac{1}{B} \sum_{b=1}^{B} I(\chi^2_b \geq \chi^2_{obs})$$

– $I(\cdot)$ — индикаторная функция, которая равна 1, если условие выполняется, и 0 иначе.

Далее величина p сравнивается с пороговой аналогично точному тесту.