ИПМ им.М.В.Келдыша РАН • Электронная библиотека

Препринты ИПМ • Препринт № 17 за 2015 г.

Власюк А.А., Орлов Ю.Н.

Точность идентификации выборочных распределений временных рядов в зависимости от типа распределения, нормы и длины выборки

Рекомендуемая форма библиографической ссылки: Власюк А.А., Орлов Ю.Н. Точность идентификации выборочных распределений временных рядов в зависимости от типа распределения, нормы и длины выборки // Препринты ИПМ им. М.В.Келдыша. 2015. № 17. 25 с. URL: http://library.keldysh.ru/preprint.asp?id=2015-17

О р д е н а Л е н и н а
ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
имени М.В.Келдыша
Р о с с и й с к о й а к а д е м и и н а у к

А.А. Власюк, Ю.Н. Орлов

Точность идентификации выборочных распределений временных рядов в зависимости от типа распределения, нормы и длины выборки

Москва — 2015

Власюк А.А., Орлов Ю.Н.

Точность идентификации выборочных распределений временных рядов в зависимости от типа распределения, нормы и длины выборки

Исследованы статистические методы определения типа выборочного распределения путем сравнения выборок с эталонным распределением, а также путем сравнения их между собой. Выяснено, что для различных типов эталонов – унимодальных, бимодальных и распределений с большой дисперсией – наилучшая идентификация достигается в нормах разных типов, причем тип наилучшей нормы меняется в зависимости от длины выборки. Определен оптимальный уровень разделения «свой-чужой» для указанных эталонов.

Ключевые слова: временной ряд, идентификация выборки, функция распределения, норма, длина выборки

Vlasyuk A.A., Orlov Yu.N.

Identification accuracy of sample distribution functions for time series depending on distribution type, norm and sample length

Statistical methods of sample distribution identification have been investigated for several types of general distributions – unimodal, bimodal and distributions with high dispersion. For various types of general distributions the optimal identification depends on norm and sample length. The optimal level of separation is determined as a result of statistical experiments.

Key words: time series, sample identification, distribution function, norm, sample length

Работа выполнена при поддержке гранта РФФИ, проект

№ 14-01-00145

Содержание

Введение и постановка задачи

При анализе временного ряда центральным является вопрос о стационарности его функции распределения как генеральной совокупности в соответствии с классической вероятностной схемой (см., напр., [1, 2]). Поскольку на практике наблюдается выборка значений ряда конечной длины, то исследователю доступны лишь выборочные функции распределения (далее ВФР). При сравнении двух ВФР, построенных по непересекающимся выборкам длины n, даже для стационарного ряда в общем случае наблюдается отличие одной ВФР от другой в норме, выбранной для такого сравнения. Целью исследования, предпринятого в настоящей работе, является нахождение оптимального уровня значимости, на котором возможно распознавание разных выборочных распределений в зависимости от нормы, длины выборки и расстояния между эталонами. Сравнение между собой пары независимых выборок в определенной норме будем называть статистическим экспериментом, а вычисленное расстояние – результатом эксперимента.

Существует много критериев (см., напр., справочник [3]), как параметрических, использующих оценки параметров распределений в некотором априори заданном классе функций, так и непараметрических, таких, как критерий Колмогорова-Смирнова [1, 3], которые при определенных формальных условиях относительно функциональной принадлежности генеральной совокупности дают вероятность того, что две наблюдаемых выборки взяты из одного и того же распределения. Если двигаться по временному ряду с шагом, равным длине окна n, то для каждой соседней пары выборок можно в соответствии с применяемым критерием получать значения вероятности того, что ВФР стационарна. Сами эти значения также образуют временной ряд, распределение которого, построенное по достаточно большому количеству статистических экспериментов, может дать представление о том, какова вероятность стационарности данных выборок в зависимости от длины выборки и от собственно критерия, который применяется для оценки стационарности. Последнее обстоятельство несколько затрудняет использование результатов такого статистического анализа в практической деятельности, поскольку априори неясно, выполнены ли требования, предъявляемые к временному ряду, для возможности применения того или иного критерия. Кроме того, даже если предположить, что применение, допустим, двух определенных критериев для анализа временного ряда корректно, то, получив в результате два – вообще говоря, различных – распределения вероятностей стационарности ВФР, трудно сделать вывод о том, какой критерий даст меньше ошибок на практике, причем вероятность стационарности не только оказывается субъективной величиной, зависящей от выбора критерия, но и сама имеет распределение вероятностей. Практически важным является именно доля ошибочных решений, т.е. доля ошибочных принятий выборки из «чужого» распределения за «свою», а также ложных

отклонений гипотезы о признании выборки «своей». Эта суммарная доля ошибочных решений зависит от того, с каким распределением сравнивается исходное, поэтому нельзя получить универсальное правило отбора «свой-чужой», пригодное для всех экспериментов сразу. В настоящей работе для изучения эффективности той или иной нормы при различении ВФР мы будем сравнивать между собой выборки, взятые только из двух различных генеральных совокупностей. Наряду со сравнением функций распределения мы также будем сравнивать между собой и выборочные плотности функций распределения (далее ВПФР).

Практический интерес представляет сравнение выборок из близких по типу распределений. Мы рассмотрим три варианта таких пар на отрезке $[0;1]$ (рис. 1): унимодальное распределение со сдвинутой модой; бимодальное распределение с модами, меняющимися местами; распределения с равными средними значениями и большими дисперсиями, такие как равномерное и арксинус. Интегральное распределение арксинуса существует в несобственном смысле на интервале $(0;1)$. Таким образом, в наших статистических экспериментах участвуют выборки, взятые из непрерывных распределений, плотности которых непрерывны на интервале $(0;1)$. На рис. 1 они представлены с шагом по классовым интервалам, равным 0,01.

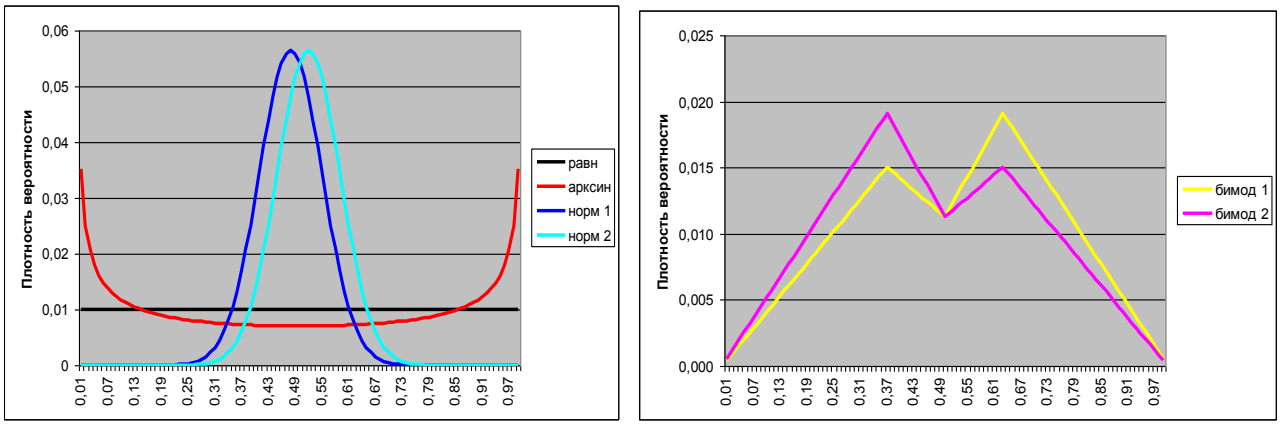

Рис. 1. Пары генеральных совокупностей: а) – унимодальные распределения и распределения с большой дисперсией; б) – бимодальные распределения;

Обозначим $f_{1,2}(x)$ непрерывные плотности первой и второй функций распределения из рассматриваемой пары, а через $F_{1,2}(x)$ сами функции распределения. Выборочные распределения строятся стандартным образом. Генерируется достаточно большое количество выборок длины n из равномерно распределенных случайных чисел. В наших экспериментах n меняется от 25 до 1000, а количество N независимых выборок длины n взято равным $N=5000$. После этого каждой равномерно распределенной выборке ставится в соответствие выборка с заданным генеральным распределением по следующему алгоритму. Пусть $y_k, 1\le k\le n,$ есть элемент равномерно

распределенной выборки. Тогда, например, число $x_{1,k} = F_1^{-1}(y_k)$ есть элемент случайной выборки из распределения F_1. По одной затравочной выборке из равномерного распределения строится только одна выборка – например, с распределением F_1. Для построения выборки с распределением F_2 берется новая равномерно распределенная выборка. Выборочные плотности функций распределения будем обозначать $f_{1,2}(x;n,l)$, $1 \le l \le N$, и аналогично $F_{1,2}(x;n,l)$ сами выборочные функции распределения.

Генеральные плотности в наших экспериментах являются непрерывными функциями. Однако их выборочные оценки, как и оценки функций распределения, могут быть представлены как в кусочно-непрерывном виде, так и в слаженном непрерывном. Поскольку ВПФР строятся неоднозначно и зависят от способа разбиения области изменения случайной величины на классовые интервалы, то имеется также и неоднозначность в восстановлении ВФР. Именно, ВФР можно оценивать естественным образом в соответствии с ее определением

$$F_1(x;n,l) = \frac{1}{n}\sum_{k=1}^{n}[x_{1,k} \le x], \qquad (1)$$

где

$$[z \le x] = \begin{cases} 1, z \le x \\ 0, z > x \end{cases},$$

и получать ВФР в виде ступенчатой кусочно-постоянной функции со скачками в точках $x = x_{1,k}$. С другой стороны, ВФР можно также определять через интеграл от ВПФР

$$F_1(x;n,l) = \int_0^x f_1(z;n,l)dz \qquad (2)$$

и получать непрерывную функцию независимо от того, каким способом была сделана оценка плотности. Чтобы на критерии близости между ВФР не влиял бы субъективный выбор классовых интервалов при построении ВПФР, далее ВФР определяется независимо от ВПФР по формуле (1).

Так как ВФР не зависит от мелкости разбиения, то при ее численном определении область изменения случайной величины разбивается на число интервалов, ширина которых много меньше ошибки в оценке генерального распределения по выборке длины n. Согласно критерию Колмогорова [1, 3], асимптотическая вероятность отклонения ВФР $F(x;n)$ от генеральной совокупности $F(x)$ дается формулой

$$\lim_{n\to\infty} P\left\{0 < \sqrt{n}\sup_x |F(x;n) - F(x)| < z\right\} = K(z) = \sum_{k=-\infty}^{\infty}(-1)^k \exp\left(-2k^2z^2\right), \qquad (3)$$

где $K(z)$ – табулированная функция Колмогорова. Отсюда следует, что неопределенность в ВФР не больше, чем $1/\sqrt{n}$, так что мелкость разбиения для

построения ВФР достаточно взять, например, в n раз меньшую, чем эта неопределенность. В условиях наших экспериментов для оценки ВФР достаточно взять 100 классовых интервалов равномерно на отрезке $[0;1]$.

Выборочные же плотности сильно зависят от мелкости разбиения отрезка $[0;1]$ для определенной длины выборки. Строго говоря, эту мелкость надо определять совместно с оптимальным критическим расстоянием между ВПФР, на котором «свою» плотность следует отделять от «чужой». В такой постановке это многопараметрическая оптимизационная задача. Мы будем решать более простую задачу оптимизации нормы и критического расстояния при фиксированной мелкости в зависимости от длины выборки. Именно, следуя результатам работы [4], число классовых интервалов в зависимости от длины выборки примем для всех типов плотностей распределений равным

$$N_{cl}(n) = \mathrm{int}\left(1{,}5 \cdot n^{1/3}\right). \qquad (4)$$

Внутри классового интервала ВПФР можно оценивать разными способами. Простейший вариант в виде гистограммы удобен для вычисления расстояния между ВПФР в норме суммируемых функций, т.е. в норме L1, но не удобен для определения расстояния между ВФР в норме C непрерывных функций, так как мелкости разбиения в этих случаях должны быть существенно разными. Для целей анализа статистических свойств расстояний от вида нормы будем использовать сглаженные оценки ВПФР по методу Парзена-Розенблатта [5] с квадратичным ядром.

Расстояние между двумя выборками длины n будем обозначать через $\rho_s^f(n)$, где нижний индекс отвечает виду нормы, а верхний – объекту сравнения: ВФР, ВПФР или выборочному моменту. Если требуется указать также и номер статистического эксперимента, то такое расстояние обозначается как $\rho_s^f(n,l)$, где l есть номер реализации пары выборок. Рассматриваются следующие виды расстояний между выборками.

1) Модуль разности между выборочными моментами. Простейший пример расстояния – это модуль разности между выборочными моментами первого порядка для двух независимых выборок.

$$\rho_1^M(n) = \left|\bar{x}_1(n) - \bar{x}_2(n)\right| = \frac{1}{n}\left|\sum_{k=1}^{n}\left(x_{1,k} - x_{2,k}\right)\right|. \qquad (5)$$

2) Норма C для ВФР или для ВПФР. Для выборочных распределений она определяется формулой

$$\rho_C^f(n) = \max_{x\in[0;1]}\left|f_1(x;n) - f_2(x,n)\right|; \quad \rho_C^F(n) = \max_{x\in[0;1]}\left|F_1(x;n) - F_2(x,n)\right|. \qquad (6)$$

3) Норма L1 для ВФР или для ВПФР, определяемая формулой

$$\rho_{L1}^f(n) = \int_0^1 \left|f_1(x;n) - f_2(x,n)\right|dx; \quad \rho_{L1}^F(n) = \int_0^1 \left|F_1(x;n) - F_2(x,n)\right|dx. \qquad (7)$$

4) Расстояние Хеллингера для ВПФР

$$\rho_{HE}^{f}(n)=2-2\int_{0}^{1}\sqrt{f_1(x;n)\,f_2(x,n)}dx. \qquad (8)$$

5) Квазирасстояние Кульбака-Лейблера или относительная энтропия выборок

$$\rho_{KL}^{f}(n)=\int_{0}^{1}f_1(x;n)\ln\left(\frac{f_1(x;n)}{f_2(x;n)}\right)dx. \qquad (9)$$

Расстояние (9) хотя и несимметрично, но неотрицательно. Поскольку же $\ln(1+x)\le x$, то расстояние Кульбака-Лейблера не меньше, чем расстояние Хеллингера, ибо $\ln(f/g)=2\ln\left(1+(\sqrt{f/g}-1)\right)\le 2\left(\sqrt{f/g}-1\right)$. Отсюда сразу следует, что $\rho_{KL}^{f}(n)\ge\rho_{HE}^{f}(n)$. Справедливы также оценки $\rho_{L1}^{f}(n)\ge\rho_{HE}^{f}(n)$, $\rho_{L1}^{f}(n)\ge\rho_{1}^{M}(n)$ и $\rho_{L1}^{f}(n)\ge\rho_{C}^{F}(n)$.

Возникает вопрос: в какой норме ошибка распознавания типа выборки наименьшая и как она зависит от длины выборки?

Применительно к трем парам генеральных совокупностей расстояния приведены в табл. 1. При этом мера множества пересечения пар генеральных плотностей выбрана для всех трех пар приблизительно одинаковой, равной 0,8.

Табл. 1. Расстояния между парами генеральных распределений

Норма	Унимодальные распределения	Бимодальные распределения	Равномерное и арксинус
ρ_1^M	0,04	0,04	0
ρ_{L1}^f	0,44	0,20	0,32
ρ_{L1}^F	0,04	0,04	0,05
ρ_C^f	0,019	0,004	0,025
ρ_C^F	0,22	0,10	0,08
ρ_{HE}^f	0,08	0,01	0,04
ρ_{KL}^f	0,16	0,02	0,06

Введем далее величину ε как уровень разделения двух выборок в той или иной норме. Именно, считаем, что если $\rho\le\varepsilon$, то сравниваемые выборки

принадлежат одной генеральной совокупности, а если $\rho > \varepsilon$, то разным. В следующих разделах приводятся результаты статистических экспериментов по отбору наилучшего уровня разделения в зависимости от нормы и длины выборки. На первом этапе решается задача определения того, из одной ли (не важно, какой) генеральной совокупности взяты две данные выборки. Рассматривается также вопрос о том, как меняется точность разделения выборок в зависимости от расстояния между эталонами. На втором этапе решается задача идентификации выборки путем сравнения ее с шестью эталонными распределениями, представленными на рис. 1.

Как было описано выше, сгенерируем $N = 5000$ выборок длиной 1000 из двух данных генеральных совокупностей. При каждой фиксированной длине выборки $10 \le n \le 1000$ сравним между собой «свои» пары и «чужие». «Своих» при каждой длине выборки имеется $N(N-1)$, а «чужих» N^2. Поскольку N достаточно велико, будем считать, что оба набора расстояний одинаково репрезентативны.

Обозначим $\rho^+(n)$ расстояние в определенной норме между «своими» выборками длины n, а $\rho^-(n)$ – между «чужими». По результатам экспериментов построим гистограмму распределения расстояний, разбив область изменения соответствующей нормы от 0 до $\rho_{\max}$ равномерно на 20 классовых интервалов. Для краткости будем обозначать соответствующие плотности распределения расстояний как $g_n^+(\rho)$ и $g_n^-(\rho)$. Интегральные функции распределения обозначим как

$$G_n^{\pm}(\rho) = \int_0^{\rho} g_n^{\pm}(r)dr. \qquad (10)$$

При равном количестве экспериментов по идентификации «своих» и «чужих» выборок суммарная ошибка равна

$$\alpha_n(\varepsilon) = \frac{1}{2}\int_0^{\varepsilon} g_n^-(r)dr + \frac{1}{2}\int_{\varepsilon}^{\rho_{\max}} g_n^+(r)dr. \qquad (11)$$

Оптимальный уровень разделения для выбранной нормы определяется как

$$\varepsilon^*(n) = \arg\min \alpha_n(\varepsilon). \qquad (12)$$

Представляет также интерес доля «своих» выборок, идентифицируемая на оптимальном для данной пары распределений и для данной нормы уровне разделения:

$$\beta(n) = G_n^+\left(\varepsilon^*(n)\right). \qquad (13)$$

Из (11) следует, что уровень значимости самоидентификации связан с ошибкой эксперимента $\alpha_n\left(\varepsilon^*(n)\right)$ следующим образом:

$$1 - \beta(n) = 2\alpha_n\left(\varepsilon^*(n)\right) - G_n^-\left(\varepsilon^*(n)\right). \qquad (14)$$

Последнее слагаемое в (14) – это ошибка непризнания «чужих» выборок за таковые.

Отметим, что величины ошибок не связаны прямо с величинами норм расстояний между эталонами (табл. 1), а определяются чувствительностью норм в задаче идентификации выборок. Целью работы является проведение численных экспериментов по определению этой чувствительности.

1. Идентификация выборок из унимодальных распределений

Примеры распределения расстояний между выборками 1-го типа (рис. 1-а) в различных нормах приведены на рис. 2-4.

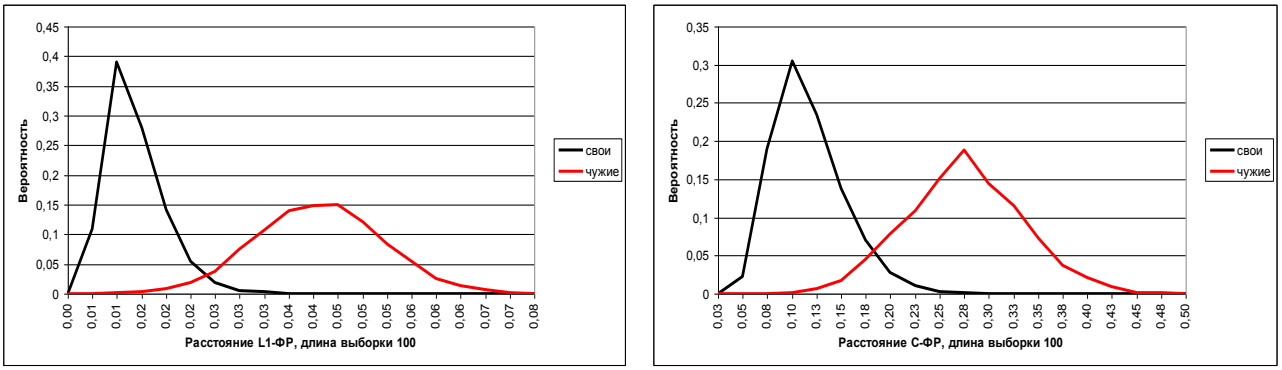

Рис. 2 (а, б). Распределение расстояний между «своими» и «чужими» выборками из унимодальных распределений; нормы L1 и C для ФР

Необычная интегральная норма L1-ФР между выборочными функциями распределения дает наилучшие среди прочих рассмотренных в работе норм результаты различения выборок из гауссоподобных распределений. Выборки длиной 25 различаются на уровне разделения 0,025 с ошибкой 0,22, выборки длиной 100 различаются на том же уровне с ошибкой 0,03, а на длине более 250 ошибка менее 0,001. Для второй по эффективности нормы С-ФР для выборок длиной 25 ошибка на оптимальном уровне разделения 0,18 составила 0,25, для выборок длиной 100 эта норма дает ошибку 0,06. Выборки длиной 300 на различаются с ошибкой 0,003.

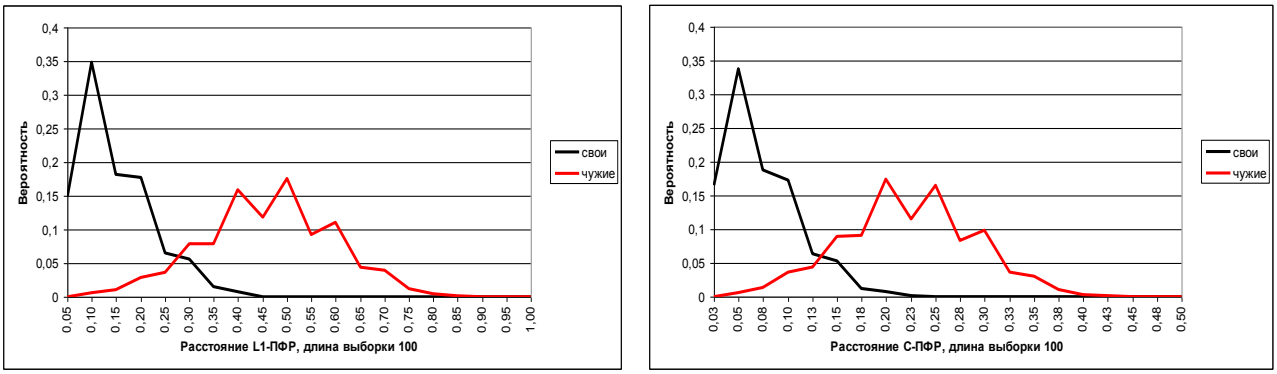

Рис. 3 (а, б). Распределение расстояний между «своими» и «чужими» выборками из унимодальных распределений; нормы L1 и C для ПФР

Из расстояний, вычисляемых по плотности распределения, на малых длинах (менее 100) наилучшей нормой является квазирасстояние Кульбака-Лейблера, на средних длинах (от 100 до 200) лучшая норма L1-ПФР, а на выборках более 250 лучшей является норма Хеллингера. На больших длинах нормы KL и С-ПФР дают наихудшие результаты.

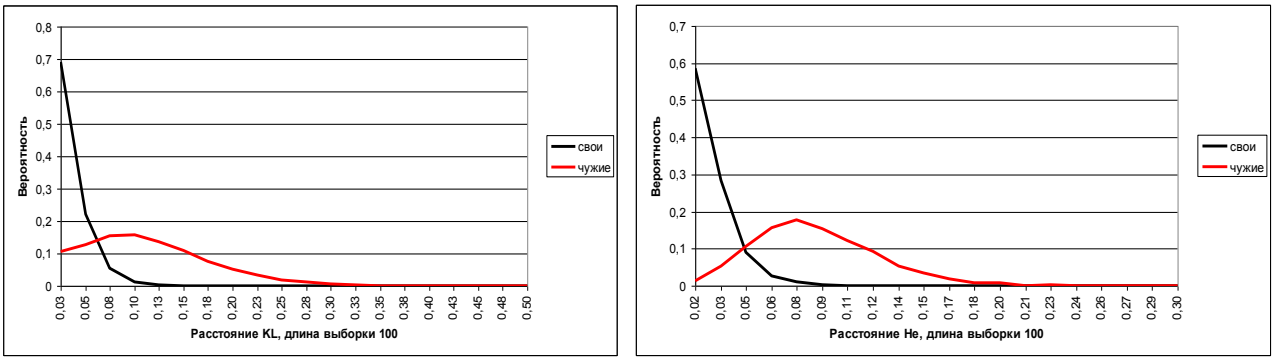

Рис. 4 (а, б). Распределение расстояний между «своими» и «чужими» выборками из унимодальных распределений, нормы KL и He

На рис. 5 показана доля ошибок $\alpha(\varepsilon)$ для выборок длины 100 в зависимости от уровня разделения, приведенного к единым величинам после нормировки его на максимальное значение расстояния между выборками в соответствующей норме. Обращает на себя внимание то, что сравнительная эффективность норм сильно зависит от уровня разделения. Представляет интерес, естественно, уровень разделения, при котором ошибка идентификации минимальна.

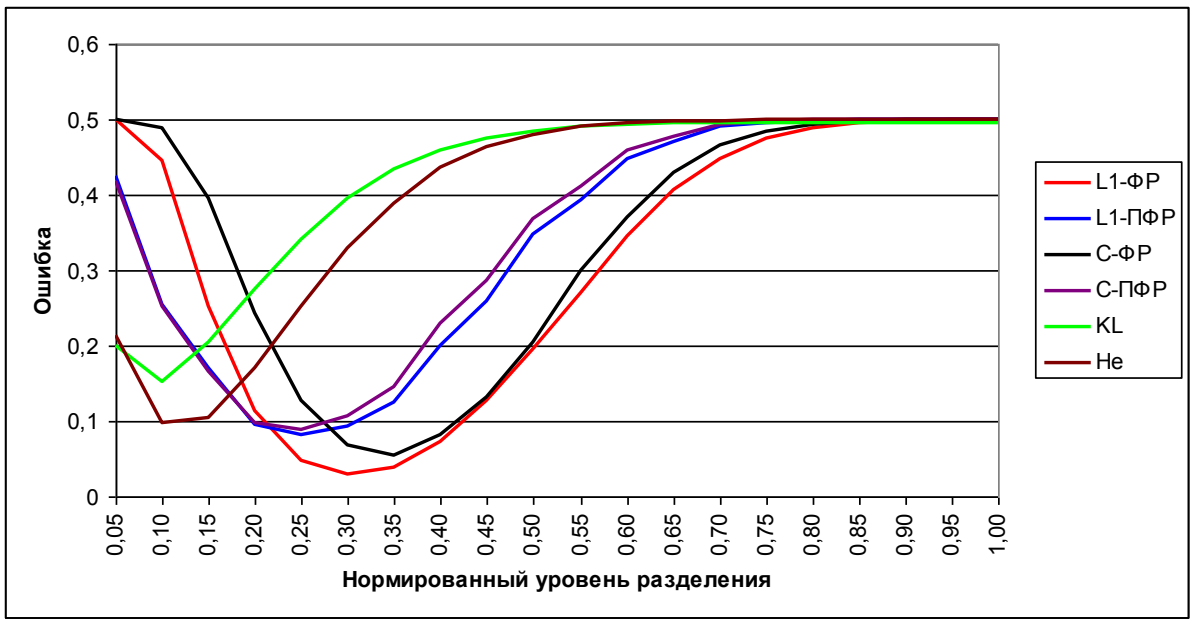

Рис. 5. Ошибка разделения выборок в зависимости от нормы и уровня разделения, унимодальные распределения, длина выборки 100

Зависимости минимальной ошибки в различных нормах от длины выборки приведены на рис. 6.

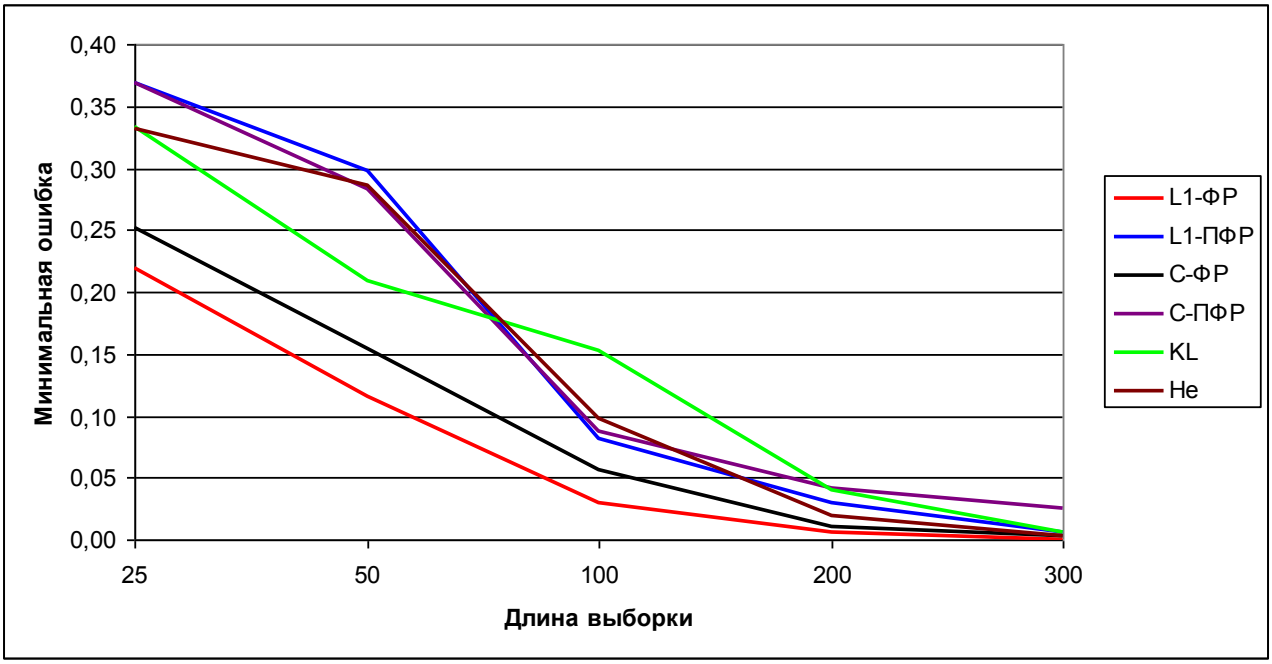

Рис. 6. Минимальная ошибка в зависимости от нормы и длины выборки

Как видно из рис. 6, наилучшей нормой для рассматриваемых гауссовых эталонов является норма L1 для ФР. Плотностные нормы заметно уступают в точности интегральным. Оптимальные уровни разделения примерно равны аргументам точек пересечения плотностей распределений расстояний «свой-чужой» (см. рис. 2-4) и не имеют выраженной зависимости от длины выборки, так как распределения расстояний с увеличением длины просто становятся более узкими, не меняя при этом существенно свою моду.

2. Идентификация выборок из бимодальных распределений

Распределения расстояний между выборками 2-го типа (рис. 1-б) похожи на представленные выше на рис. 2-4, поэтому для краткости они здесь не приводятся.

Лучшей нормой для разделения выборок бимодального типа является норма L1 для ФР, с которой конкурирует норма C для ФР. На малых длинах более точной является норма L1 для ФР, а на длинах, больших 250, – норма C для ФР. Следующей по точности идентификации здесь является норма L1 для ПФР. Худший результат получается в нормах He, KL и C для ПФР.

На рис. 7 показана доля ошибок $\alpha(\varepsilon)$ в зависимости от уровня разделения, приведенного к единым величинам после нормировки его на максимальное значение расстояния между выборками в соответствующей норме. Зависимость минимальной ошибки от длины выборки дана на рис. 8.

С увеличением длины выборки n минимальная ошибка для всех норм снижается примерно одинаково. Зависимость ошибки от длины выборки почти линейная.

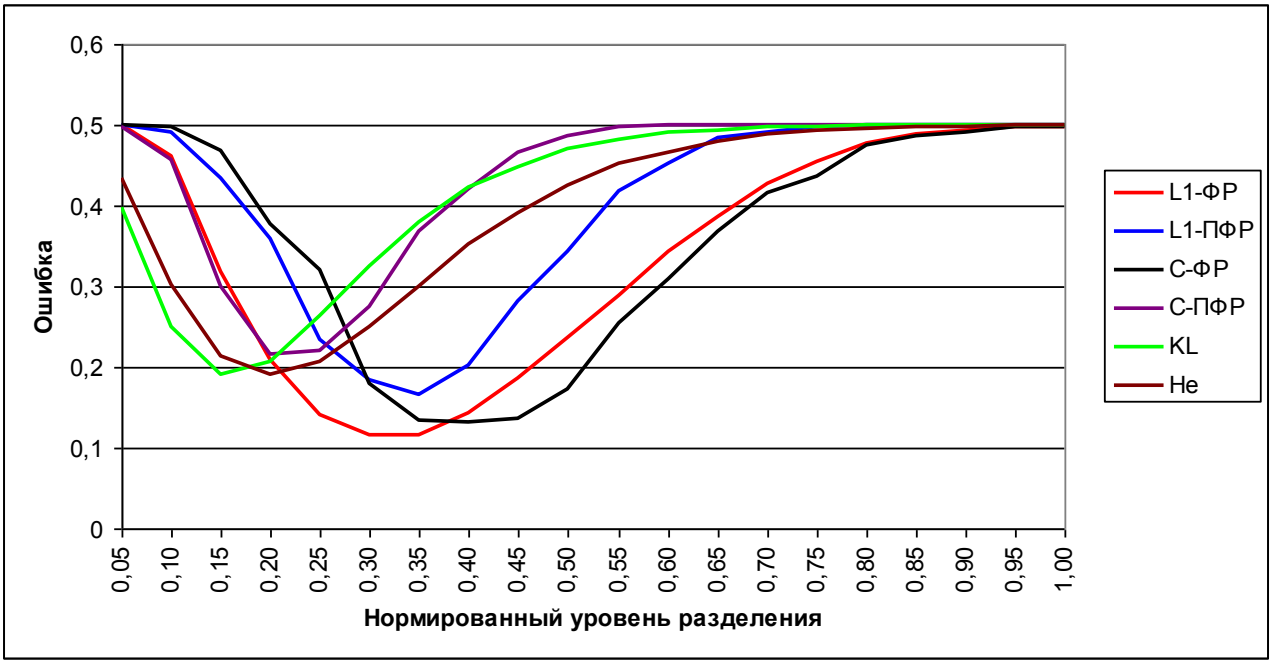

Рис. 7. Доля ошибок в зависимости от нормы и уровня разделения для бимодальных распределений, длина выборки 100

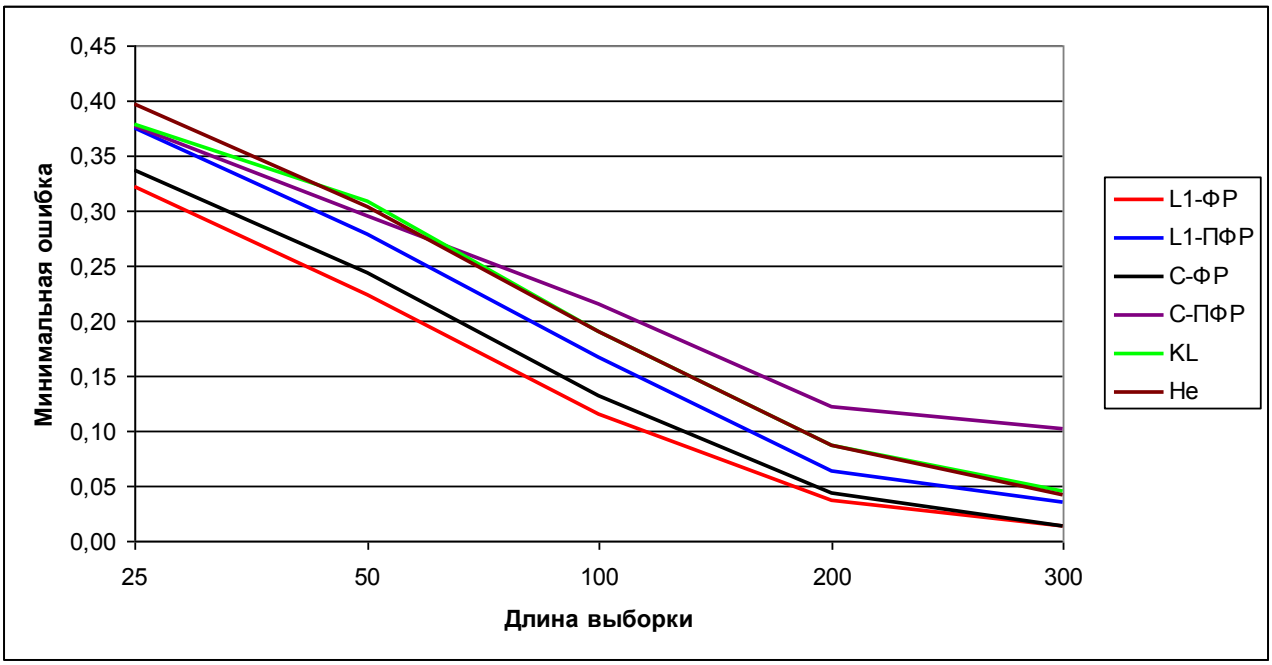

Рис. 8. Минимальная ошибка в зависимости от нормы и длины выборки

Различение выборок из бимодальных распределений приводит к несколько иной последовательности норм по их эффективности, чем для унимодальных распределений. В частности, на бимодальных распределениях норма С-ФР начинает выигрывать у нормы L1-ПФР на относительно больших выборках,

тогда как на унимодальных распределениях она всегда вторая. Расстояние KL для бимодальных распределений неэффективно, а для унимодальных может применяться на малых длинах выборок.

3. Идентификация выборок из распределений с большой дисперсией

Примеры распределения расстояний между выборками 3-го типа (рис. 1-а) в некоторых нормах приведены на рис. 9-11. Для этих распределений на малых и средних длинах выборок (до 250) оптимальной нормой является норма L1 для ФР, затем оптимальной становится норма L1 для ПФР, и норма KL.

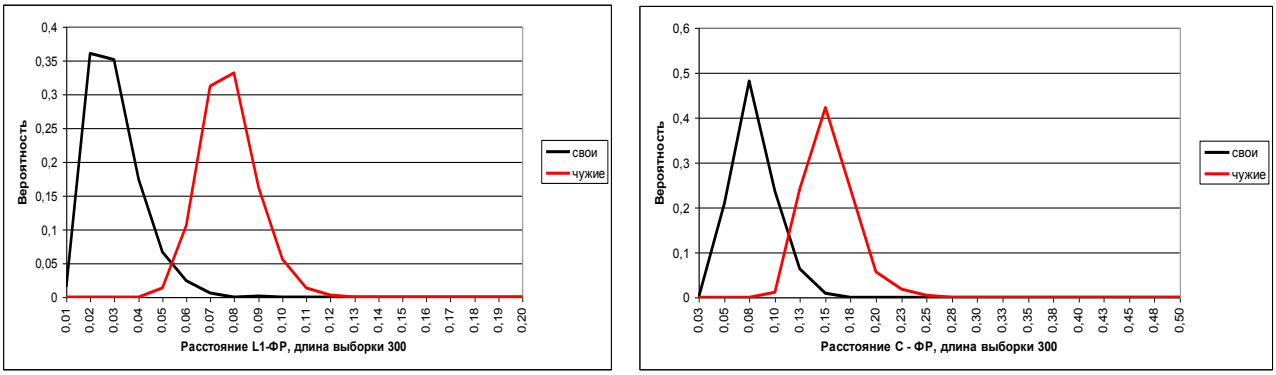

Рис. 9 (а, б). Распределение расстояний между «своими» и «чужими» выборками из распределений III типа; нормы L1 и C для ФР

Для нормы L1-ФР оптимальный уровень разделения выборок длин до 150 составил 0,35. Ошибка разделения составила 0,26 для выборок длиной 50, 0,15 для выборок длиной 100 и 0,07 для выборок длиной 150.

Норма C для ФР оказывается наихудшей для сравнения распределений с широкой дисперсией на всех длинах. Так, на длине 50 ошибка составила 0,32, а на длине 500 она все еще отлична от нуля и равна 0,02.

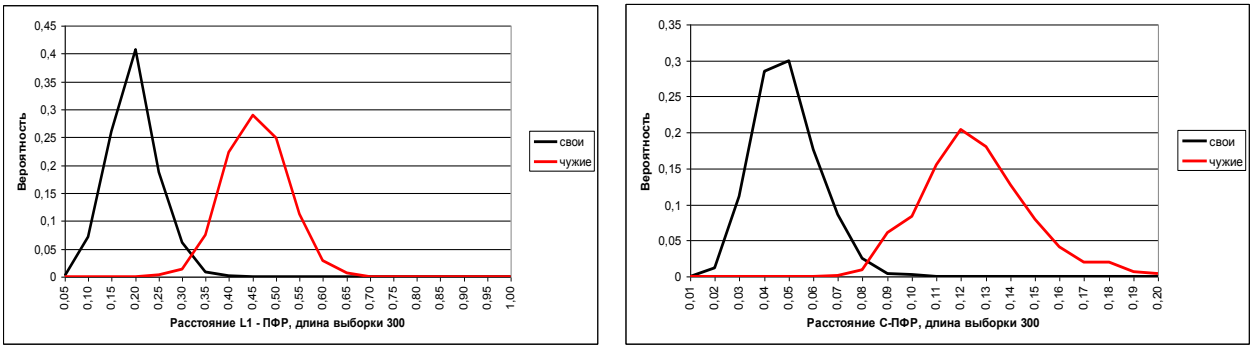

Рис. 10 (а, б). Распределение расстояний между «своими» и «чужими» выборками из распределений III типа; нормы L1 и C для ПФР

Норма L1 для ПФР на малых длинах выборок занимает второе место по эффективности разделения: ошибка составила 0,28 для выборок длины 50, 0,18 для выборок длины 100 и 0,07 для выборок 150. На длине более 200 эта норма становится самой эффективной, ошибка составляет менее 0,05. С дальнейшим

увеличением длин выборок эта норма уступает первенство расстоянию KL, а также и Хеллингеру, хотя ошибка во всех этих нормах достаточно мала и не превосходит 0,03.

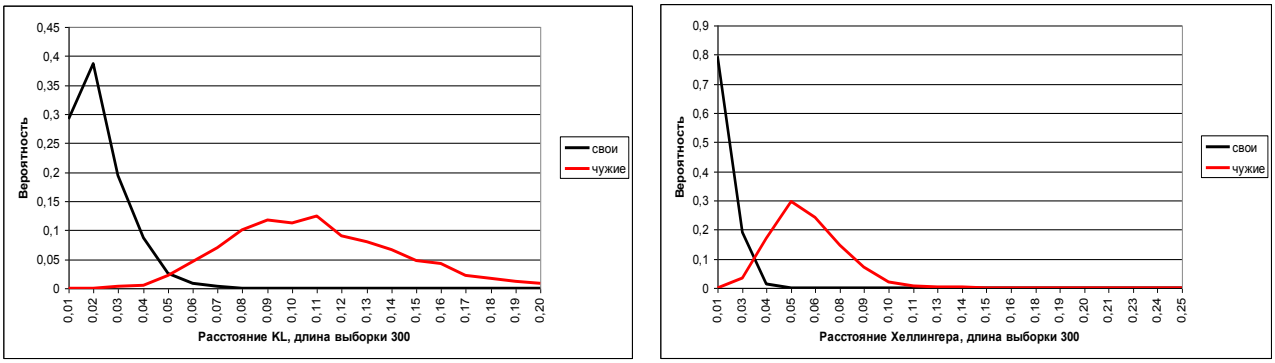

Рис. 11 (а, б). Распределение расстояний между «своими» и «чужими» выборками из распределений III типа; нормы KL и Не для ПФР

Расстояние KL эффективно на малых (что удивительно) и больших длинах. На средних длинах от 100 до 200, эта квазинорма для разделения неэффективна. В частности, на уровне разделения 0,05 при длинах более 300 ошибка становится менее 0,02.

Расстояние Хеллингера повышает свою эффективность с увеличением длины выборки и при длинах более 250 становится вторым-третьим.

На рис. 12 показана доля ошибок $\alpha(\varepsilon)$ в зависимости от уровня разделения, приведенного к единым величинам после нормировки его на максимальное значение расстояния между выборками в соответствующей норме.

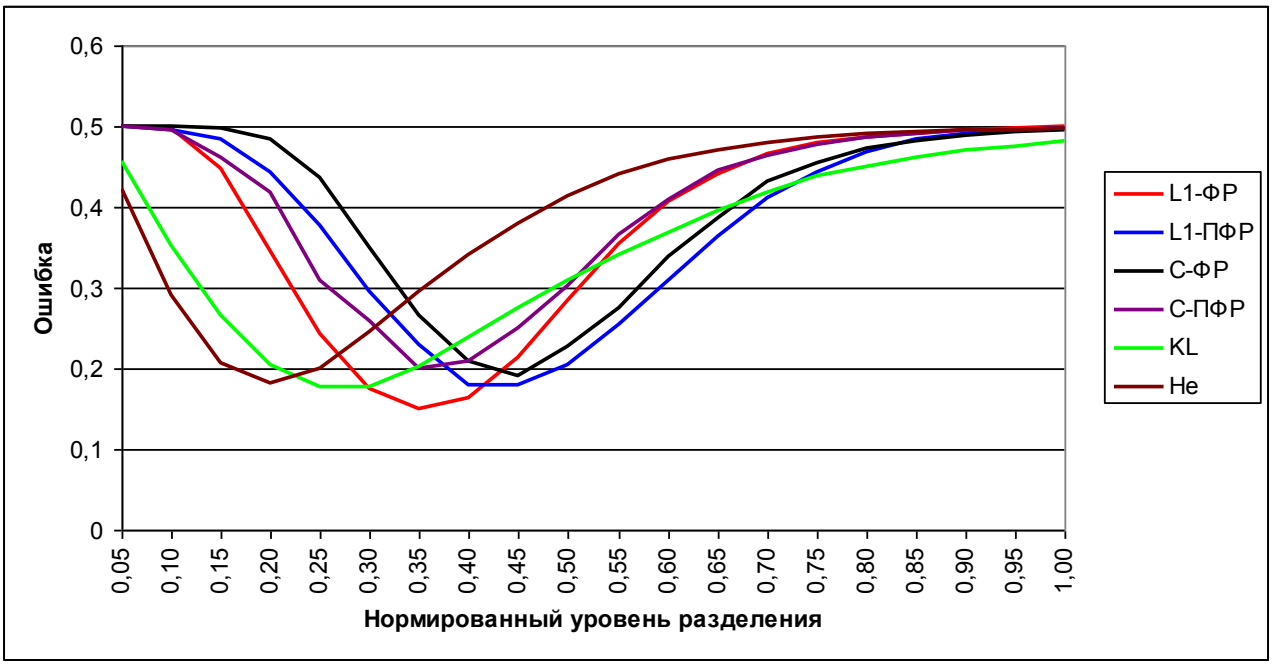

Рис. 12. Доля ошибок в зависимости от нормы и уровня разделения, длина выборки 100

Зависимости минимальной ошибки от длины выборки даны на рис. 13.

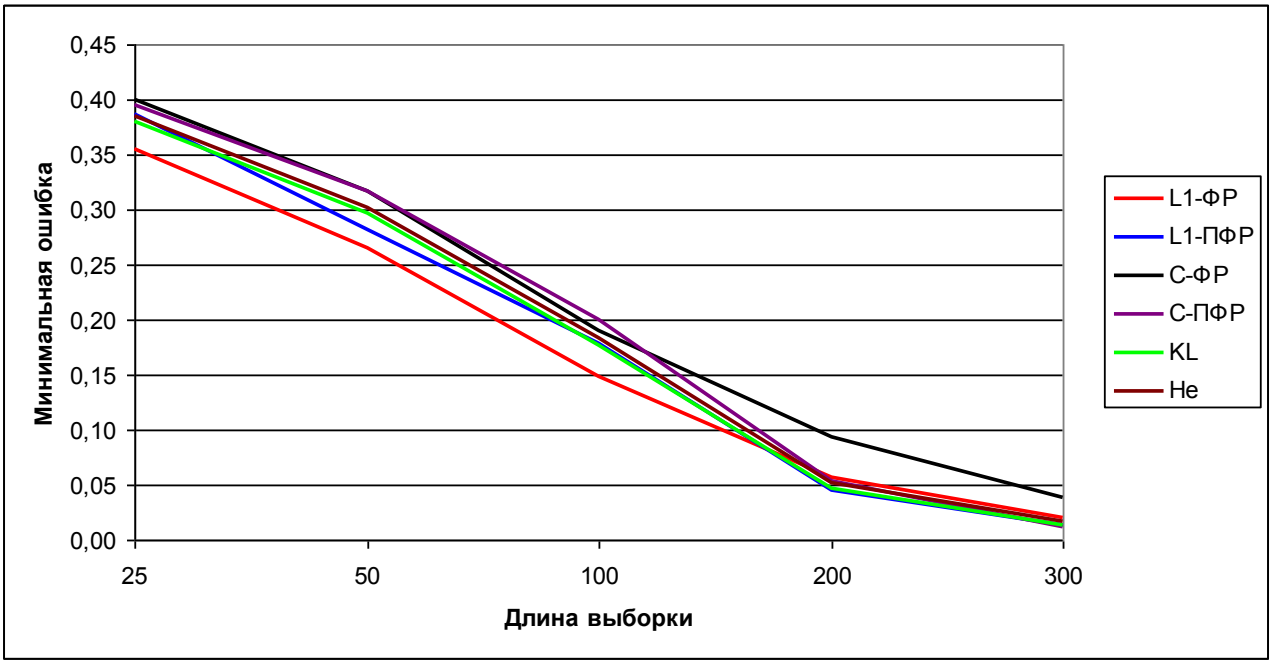

Рис. 13. Минимальная ошибка в зависимости от нормы и длины выборки

Как видно из рис. 13, до длин выборок 200 наилучший результат дает норма L1-ФР, но потом она уступает нормам L1-ПФР и KL, которые на малых длинах периодически занимают 2-е – 3-е места по эффективности, а на относительно больших выборках 1-е – 2-е места. Норма С-ФР, в отличие от предыдущих примеров, почти всегда наихудшая, либо наихудшей является норма С-ПФР.

В табл. 2 сведены результаты эффективности применения тех или иных норм для разделения выборок малых длин (до 100), средних (от 100 до 250) и относительно больших длин, более 250, взятых из трех рассматриваемых пар эталонных распределений.

Табл. 2. Сравнительная эффективность норм в задаче разделения выборок

	Малые длины	**Средние длины**	**Большие длины**
Унимодальные эталоны	**1. L1-ФР** **2. С-ФР** **3. KL, Не** **4. С-ПФР, L1-ПФР**	**1. L1-ФР** **2. С-ФР** **3. KL** **4. С-ПФР, L1-ПФР, Не**	**1. L1-ФР** **2. С-ФР, L1-ПФР, KL, Не** **3. С-ПФР**
Бимодальные эталоны	**1. L1-ФР** **2. С-ФР** **3. L1-ПФР, С-ПФР, KL** **4. Не**	**1. L1-ФР** **2. С-ФР** **3. L1-ПФР** **4. С-ПФР** **5. KL, Не**	**1. С-ФР** **2. L1-ФР** **3. L1-ПФР** **4. KL, Не** **5. С-ПФР**
Широкие эталоны	**1. L1-ФР** **2. KL** **3. L1-ПФР** **4. Не** **5. С-ФР, С-ПФР**	**1. L1-ФР** **2. L1-ПФР** **3. KL, Не** **4. С-ФР, С-ПФР**	**1. L1-ПФР, KL** **2. Не** **3. С-ПФР, L1-ПФР** **4. С-ФР**

4. Точность распознавания ВФР по ее близости к эталону

При решении задачи идентификации типа наблюдаемой выборки будем сравнивать ее с известными эталонными распределениями в соответствующих нормах. Возникает вопрос: какая норма оптимальна для этой цели? Та, в которой достигается наименьшая ошибка (11) статистического эксперимента по разделению распределений или какая-нибудь другая?

Занумеруем эталонные распределения (рис. 1) в порядке следования от 1 до 6. Соответствующие плотности и функции распределения обозначим $f_k(x)$ и $F_k(x)$, $k=1\div 6$. Имеющиеся в общей сложности 30 тыс. выборок длиной 1000 идентифицируются как выборки из той или иной генеральной совокупности. Метод идентификации – по наименьшему расстоянию в заданной норме. Именно, если расстояние между выборкой и k-ым эталоном минимально, то выборка признается взятой из этой совокупности. Какова ошибка идентификации выборки в зависимости от длины выборки и нормы? Для i-ой выборки длины n в определенной норме составляем разность

$$r_{ik}(n)=\| f_i(x;n)-f_k(x)\| \qquad (15)$$

и аналогично для матожидания или для интегральной функции распределения. Находим минимум полученных расстояний в зависимости от номера эталона. Если $r_{ik}(n)=\min$, то $i=k$, т.е. i-ая выборка взята из k-го эталона.

Для всех рассматриваемых эталонов распределение расстояний между ВФР и эталонной ФР в норме С дается формулой (3). Для других норм эти распределения можно построить по результатам экспериментов, но, в отличие от предыдущей задачи, эти распределения не позволяют определить ошибку идентификации выборки. Ошибкой здесь будем называть неправильное определение типа выборки. Доля ошибок в зависимости от нормы и длины выборки приведена в табл. 2 по типам пар выборок: I – гауссоподобные, II – бимодальные, III – равномерное и арксинус. Приведено также округленное среднее значение ошибки идентификации одной выборки среди шести эталонов по данным 5000 экспериментов.

Точность распознавания в той или иной норме зависит от типа распознаваемых выборок, а также от их длины. Оказалось, что на малых выборках унимодальные распределения лучше всего распознаются в норме L1 для ВФР, а остальные – в норме С по критерию Колмогорова-Смирнова.

Для выборок из бимодальных распределений на длинах 50 и 200 оптимальной является норма в С для ПФР, а на промежуточной длине 100 оптимальным оказалось квазирасстояние Кульбака-Лейблера. Для выборок с широким распределением наилучшей на всех длинах является опять-таки норма в С для ПФР. Хуже всех распознает выборки в сравнении с эталоном норма в L1. Расстояния Кульбака-Лейблера и Хеллингера имеют примерно равные точности, а норма в L1 для ПФР с увеличением длины выборки становится заметно хуже, чем норма в С. На длинах более 350 все методы дают безошибочное распознавание (разумеется, на тех примерах, которые рассматриваются в работе).

В среднем же, если тип распределения не известен, норма в С для ФР оказывается предпочтительнее остальных на любых длинах выборок. Это означает, что задача сравнения текущей выборки с эталоном принципиально отличается от задачи косвенного определения одинаковости эталона из сравнения между собой двух выборок. Для двух этих задач следует использовать разные нормы.

Табл. 3. Число ошибок идентификации
в зависимости от нормы и длины выборки, на 5000 экспериментов

Длина и тип пары выборки		ρ^f_{L1}	ρ^f_{HE}	ρ^f_{KL}	ρ^F_{C}	ρ^F_{L1}
50	I	20	6	10	4	1
	II	460	380	370	107	125
	III	510	360	350	228	352
	AVR	330	250	247	113	159
100	I	1	0	0	0	0
	II	364	70	2	10	12
	III	110	100	94	46	81
	AVR	158	57	32	19	31
200	I	0	0	0	0	0
	II	280	5	2	0	0
	III	17	12	12	2	12
	AVR	99	6	5	1	4
300	I	0	0	0	0	0
	II	240	0	0	0	0
	III	0	0	0	0	0
	AVR	80	0	0	0	0

Основной вклад в ошибку идентификации дают пары из близких распределений. Равномерное распределение на малых выборках иногда идентифицируется как унимодальное или бимодальное, однако следует

учитывать, что приведенные данные относятся к довольно близким между собой распределениям. С увеличением расстояния между эталонными плотностями в норме L1 увеличивается и точность методов распознавания.

5. Зависимость точности идентификации выборок от близости эталонов

Естественно, что для норм, заданных в виде функционалов от функций распределения, точность разделения выборок на своих и чужих увеличивается с уменьшением меры множества пересечения генеральных плотностей. Очевидно, при нулевой мере множества пересечения ошибка разделения выборок равна нулю для любых длин выборок. Если же мера множества пересечения близка к единице, то минимальная ошибка ожидается весьма высокой, также близкой к единице, но какой именно – зависит от нормы, используемой для сравнения выборок. Отметим, что для норм, представляющих собой модуль разности функционалов, как в случае сравнения средних значений, точность разделения не обязана зависеть от близости генеральных плотностей, и потому в общем случае такая «норма» не эффективна.

Рассмотрим, как ведет себя точность разделения выборок в зависимости от меры множества пересечения генеральных плотностей на примере бимодальных треугольных распределений, заданных на отрезке $[0;1]$ (рис. 1-б).

Для различных норм зависимость минимальной ошибки от длины выборки и меры пересечения показана на рис. 14. Сам же оптимальный уровень разделения выборок изменяется от 0,1 до 0,3 и имеет определенную тенденцию возрастания с уменьшением меры пересечения при фиксированной длине выборки.

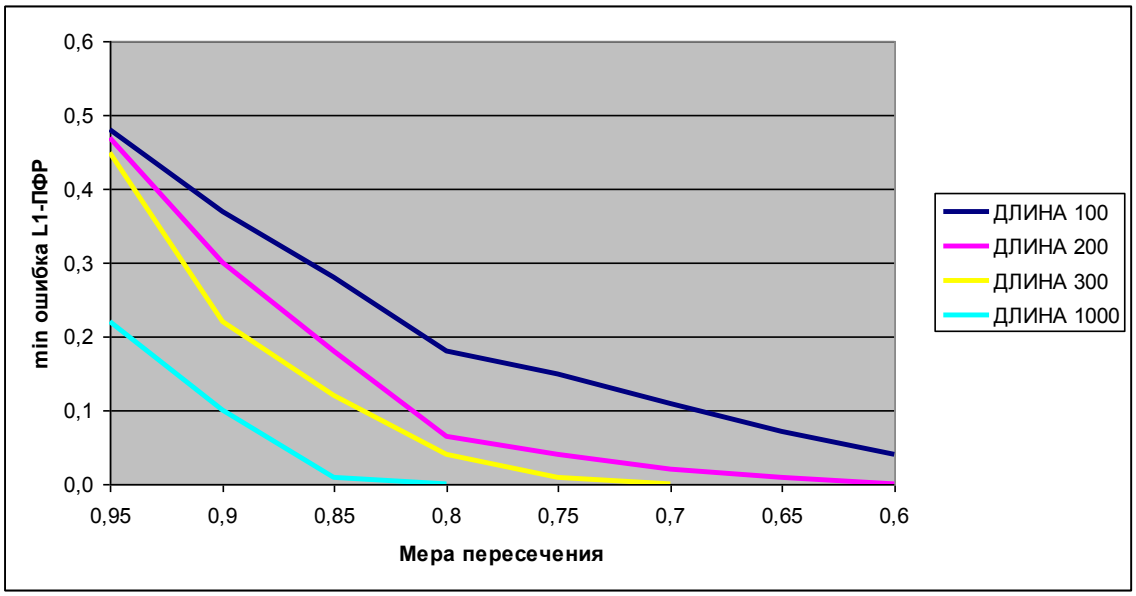

Рис. 14. Минимальная ошибка в зависимости от длины выборки и меры пересечения эталонных плотностей для нормы L1-ПФР

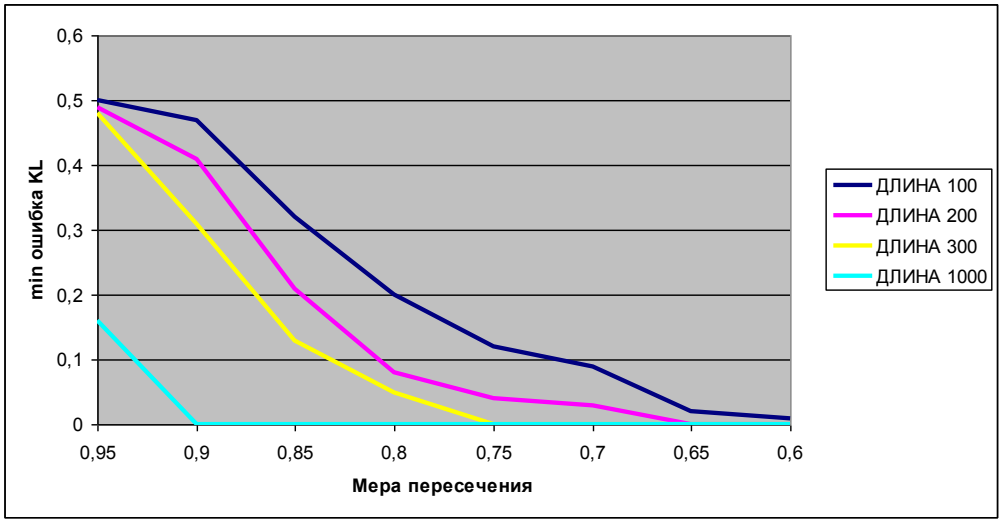

Рис. 15. Минимальная ошибка в зависимости от длины выборки и меры пересечения эталонных плотностей для нормы KL

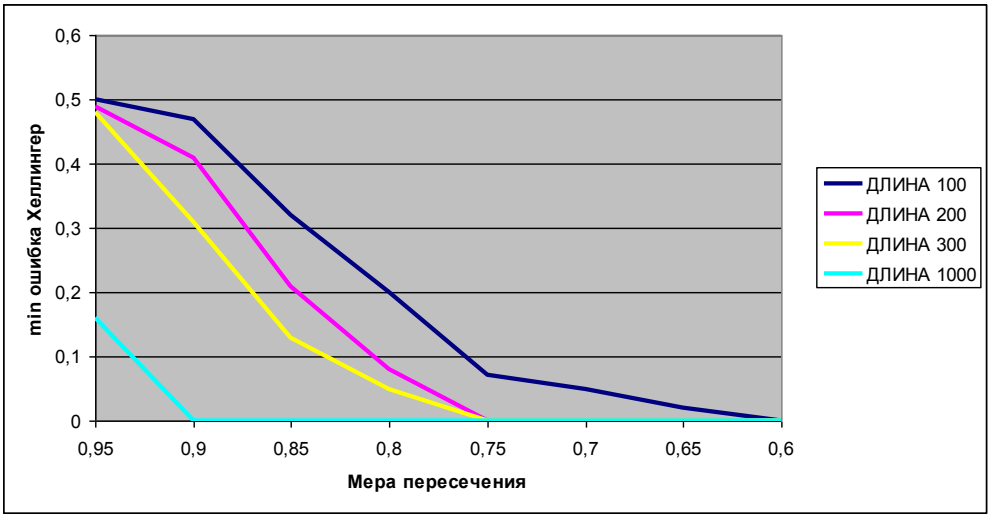

Рис. 16. Минимальная ошибка в зависимости от длины выборки и меры пересечения эталонных плотностей для нормы Хеллингера

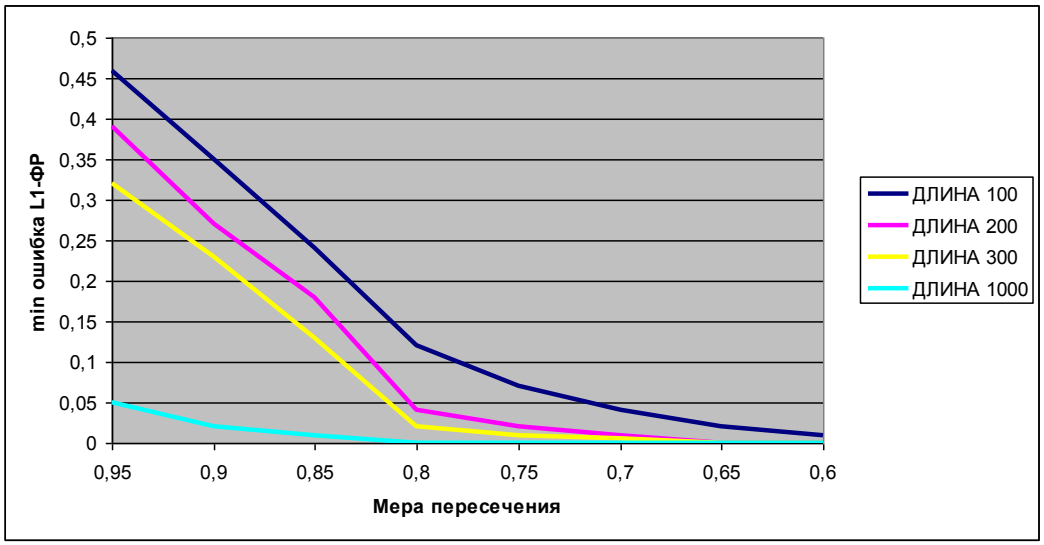

Рис. 17. Минимальная ошибка в зависимости от длины выборки и меры пересечения эталонных плотностей для нормы L1-ФР

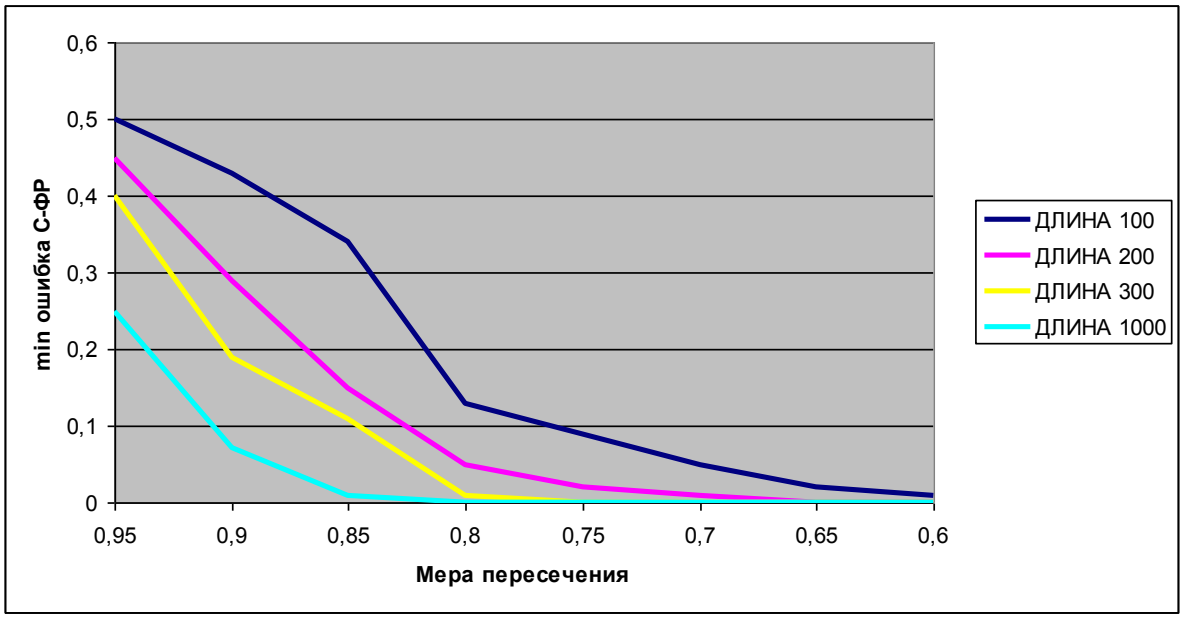

Рис. 18. Минимальная ошибка в зависимости от длины выборки и меры пересечения эталонных плотностей для нормы С-ФР

В целом тенденция изменения ошибки разделения для всех норм одинакова: с увеличением длины выборки ошибка снижается, как снижается она и при уменьшении общей площади под графиками эталонных плотностей. Однако четкого скэйлига указанных зависимостей не обнаруживается даже при достаточно большом количестве экспериментов по разделению выборок.

Результаты по оптимальному уровню разделения и оптимальной же норме в зависимости от близости сравниваемых распределений и длины выборки показывают, что эффективность норм существенно зависит от указанных параметров. Рассматривались нормы L1, L1-ФР, KL, *He* и С-ФР. Выяснилось, что при близких распределениях безусловным лидером по разделению выборок «свой-чужой» с наименьшей ошибкой как на малых длинах выборок, так и на больших является норма L1 для функций распределения, не имеющая вероятностного смысла. С увеличением расстояния между эталонами на малых и больших длинах оптимальной является норма Хеллингера, а на средних длинах конкурируют нормы для плотностей: L1-ПФР и KL. С дальнейшим расхождением эталонов на малых и средних длинах оптимальной является норма Хеллингера, а на больших длинах выборки разделяются с нулевой ошибкой сразу в нескольких нормах. Отметим, что норма С между ФР и норма L1 между ПФР лишь в одном варианте являются оптимальными. В табл. 4 сведены результаты оптимизации выбора нормы в зависимости от расстояния между эталонами для разделения выборок из бимодальных распределений. В итоге можно построить комбинированные уровни минимальной ошибки разделения «свой-чужой» в подходящих нормах в зависимости от близости эталонов и длин выборок. Эти уровни показаны на рис. 19.

Табл. 4. Оптимальные нормы в задаче разделения выборок в зависимости от длины выборки и близости между эталонами для бимодальных распределений

Длина выб. / Близость	100	200	300	1000
0,95	L1-ФР	L1-ФР	L1-ФР	L1-ФР
0,9	L1-ФР	L1-ФР	L1-ФР	He, KL
0,8	L1-ФР	L1-ФР	L1-ФР	Все
0,7	L1-ФР	He	He, KL	Все
0,6	He	Все, кроме L1-ПФР	Все	Все

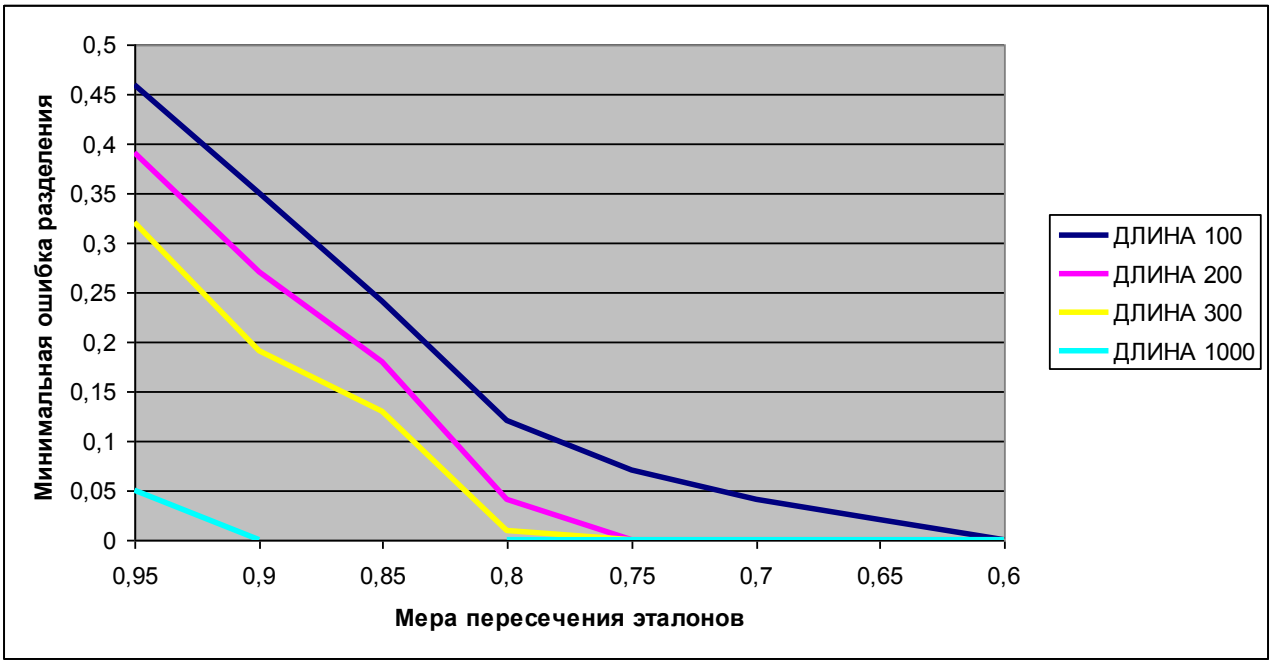

Рис. 19. Абсолютно минимальная ошибка разделения «свой-чужой» для бимодальных распределений

Близость распределений влияет на точность тех или иных норм в задаче разделения, но на это влияет также и тип распределений. Например, для унимодальных распределений на уровне близости 0,65 между эталонами того же вида, что и на рис. 1-а, получаем следующее. Интегральная норма L1-ФР между выборочными функциями распределения дает наилучшие среди прочих рассмотренных в работе норм результаты различения выборок из гауссоподобных распределений. Выборки длиной 25 различаются на уровне разделения 0,06 с ошибкой менее 0,08, а выборки длиной 100 различаются на уровне 0,05 уже безошибочно. Для второй по эффективности нормы C для выборок длиной 25 ошибка на оптимальном уровне разделения 0,35 составила 0,11. Впрочем, для выборок длиной 100 эта норма дает уже почти нулевую

ошибку различения 0,003 на уровне разделения 0,25. Выборки длиной 125 на уровне 0,3 различаются безошибочно.

Из расстояний, вычисляемых по плотности распределения, наилучший вариант различения выборок из бимодальных распределений имеет расстояние Хеллингера, близкий к нему результат получается для расстояния Кульбака-Лейблера, а наихудший вариант дает норма L1 для ПФР.

Важно заметить, что для выборок длиной более 200 безошибочное разделение возможно уже начиная с близости 0,8 (мера общей площади под графиками плотностей) между распределениями. Это означает, что для широкого класса нестационарных временных рядов, выборочные распределения которых имеют определенный вид (очень часто бимодальный или унимодальный), можно попытаться построить базис из эталонов, позволяющий с высокой точностью идентифицировать текущую ситуацию. В частности, таковыми являются многие ценовые ряды на финансовых рынках. Построение такого базиса для конкретных рядов является актуальной задачей, обсуждаемой в следующем разделе.

6. Пример идентификации выборок для ценовых рядов

Для минутных приростов цен закрытия курса рубль/доллар по данным [www.finam.ru] за 2013 г. мы выделяем пять основных эталонов-паттернов (рис. 20). Эти эталоны отвечают следующим состояниям валютного рынка: один – боковому движению, два – трендам вверх (медленному «1» и быстрому «2»); два – трендам вниз (также медленному «1» и быстрому «2»). Общая часть под графиками плотностей вероятностей эталонов меняется от 0,9 до 0,75, т.е. отвечает типовым расстояниям между эталонами, рассмотренными выше.

Эталон «flat» имеет максимум в точке ноль и примерно одинаково спадающие правые и левые части графика. Эталоны «up» характеризуются приподнятыми правыми «хвостами», а эталоны «down» – левыми, причем в медленном тренде мода, отвечающая нулевому приросту цены, превосходит моду распределения «flat». В быстром тренде мода смещена: для эталона «up» – вправо, а для эталона «down» – влево. В крайние классовые интервалы помещены все приросты, большие 10 по абсолютной величине.

Возникает вопрос, с какой точностью текущее распределение, построенное по выборке малой длины – например, за 30 мин или за час, - можно идентифицировать как распределение, относящееся к одному из пяти выделенных типов. Решение этой задачи позволит автоматизировать процесс соотнесения текущей выборки достаточно короткой длины с определенным поведением временного ряда.

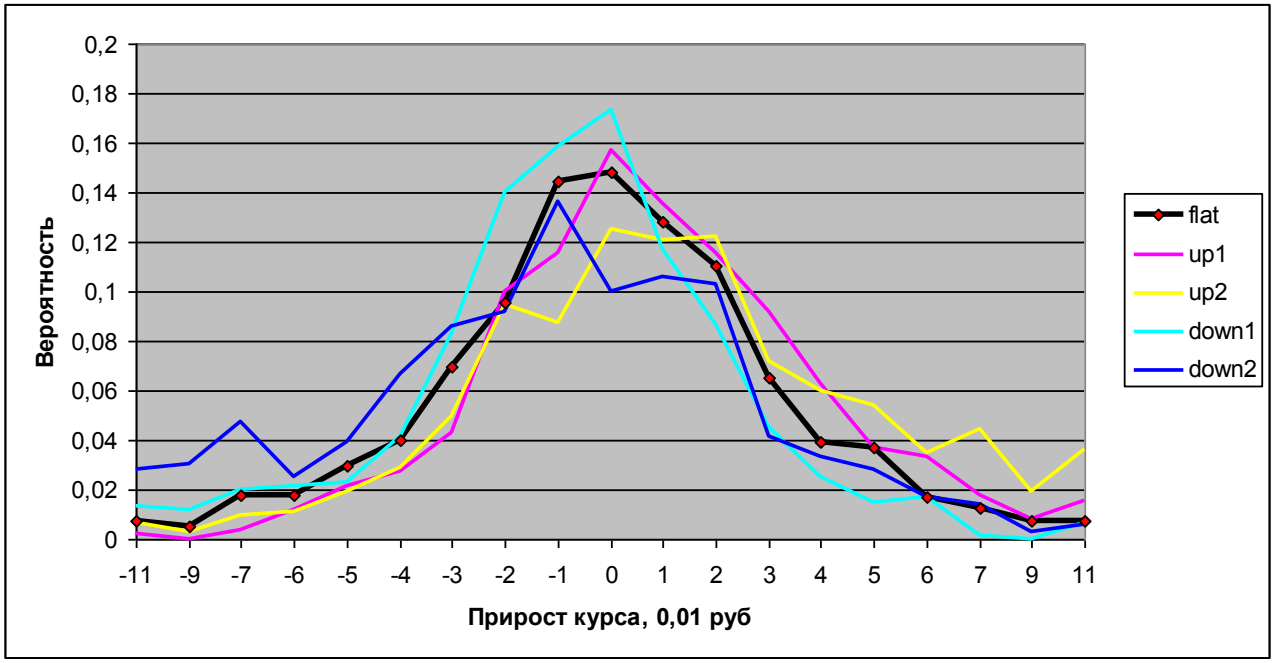

Рис. 20. Эталонные распределения приростов курса рубль/доллар

Значения площадей общих частей для пар эталонов приведены в табл. 5. Таблица симметрична, поэтому для удобства восприятия заполнена только ее верхняя треугольная часть.

Табл. 5. Площадь общей части пар эталонов

	flat	up1	up2	down1	down2
flat	**1**	**0,89**	**0,85**	**0,88**	**0,86**
up1		**1**	**0,89**	**0,80**	**0,77**
up2			**1**	**0,75**	**0,76**
down1				**1**	**0,84**
down2					**1**

Из рис. 20 видно, что данные эталоны ближе к унимодальному типу, поэтому из результатов предыдущего анализа следует, что на относительно малых выборках наиболее эффективной ожидается норма L1-ВФР, что подтверждается экспериментально. Поскольку, однако, выборки для сравнения с эталонами берутся из опыта, а не генерируются из этих самых эталонов, точность распознавания оказывается несколько хуже, чем в тестовых примерах раздела 4. Кроме того, эталоны в данном случае не являются стационарными генеральными совокупностями, а получены экспертным отбором определенных состояний временного ряда на длине 10 тыс., так что эталонное распределение – это по сути эмпирическое распределение примерно по длине 1,5 тыс. точек.

Подход к анализу нестационарного временного ряда с использованием эталонов состояния ряда опирается на гипотезу о том, что каждому типу поведения ряда (тренду вверх, тренду вниз, «боковику») отвечает определенный тип распределения приростов значений ряда. Тогда нестационарность может трактоваться как переход от одного состояния к другому, а сами состояния представляются в виде набора стационарных распределений. Распределение расстояний от выборок длины 30 до «своих» и «чужих» эталонов в норме L1-ФР для данного примера показано на рис. 21. Минимальная ошибка распознавания в этом примере равна 0,32, если критический уровень разделения принять равным 0,055.

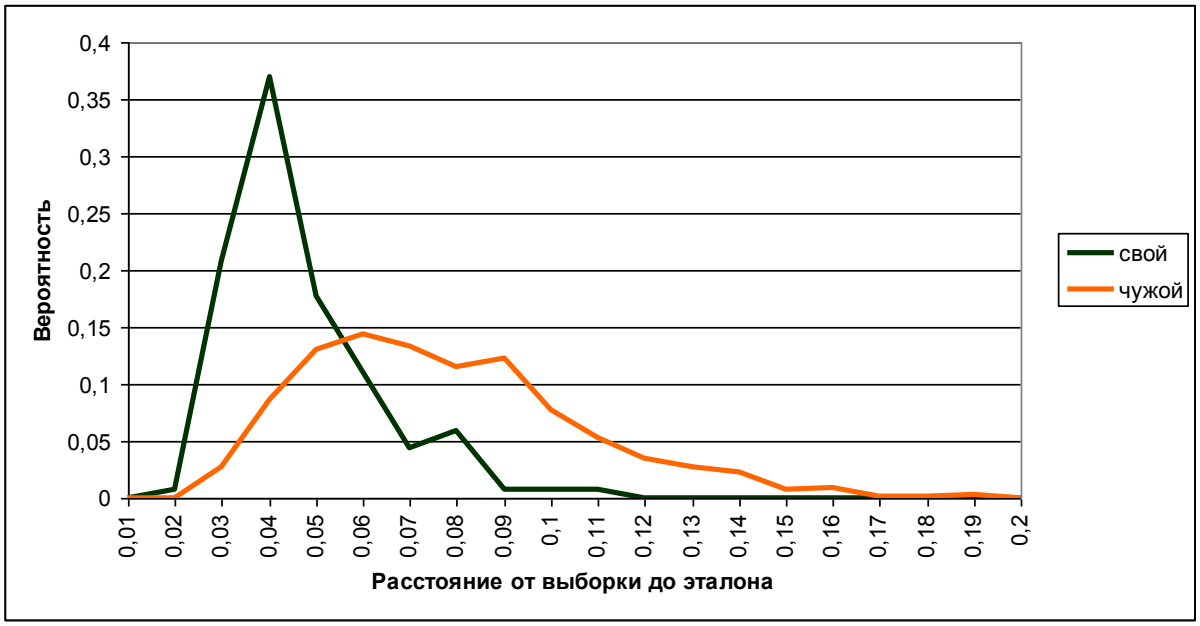

Рис. 21. Распределения расстояний от выборок до эталонов для нормы L1-ФР

Поскольку расстояния от выборок, трактуемых как «свои» по типу поведения ряда, до соответствующим образом отобранных эталонов сдвинуты влево и имеют распределение с заметно меньшей дисперсией, чем «чужие», гипотеза о сопоставлении распределения типу поведения ряда может считаться адекватной. Естественно, построение эталонов и их тестирование проводится на двух непересекающихся массивах данных.

Ошибка распознавания типа выборки по формуле (15) уменьшается с увеличением длины n выборки с достоверностью $R^2 = 0{,}99$ по формуле $err = 0{,}32 - 0{,}24\ln(n/30)$ и обращается в ноль (для данной серии экспериментов) при $n = 120$.

Чем больше общая площадь под графиками пары эталонов, тем выше ошибка распознавания выборки, отвечающей выделенному состоянию временного ряда («up», «down» или «flat»). Выборки «up1», «up2» и «flat» перепутываются примерно в 3 раза чаще, чем выборки «down1» и «down2», а выборки «up» и «down» вообще не перепутываются при распознавании.

В дальнейшем предполагается более детально изучить зависимость ошибки в наиболее эффективной норме для разных типов эталонных распределений, а также уточнить базис для набора актуальных временных рядов, не являющихся стационарными.

Литература

1. Гнеденко Б.В. Курс теории вероятностей. – М.: Физматлит, 1961. – 406 с.
2. Прохоров Ю.В., Розанов Ю.А. Теория вероятностей: основные понятия, предельные теоремы, случайные процессы. – М.: Наука, 1967. – 496 с.
3. Кобзарь А.И. Прикладная математическая статистика. – М.: Физматлит, 2006. – 816 с.
4. Орлов Ю.Н. Оптимальное разбиение гистограммы для оценивания выборочной плотности функции распределения нестационарного временного ряда // Препринты ИПМ им. М.В. Келдыша. 2013. № 14. 20 с.
URL: http://library.keldysh.ru/preprint.asp?id=2013-14
5. Parzen E. On Estimation of a Probability Density Function and Mode // The Annals of Mathematical Statistics, 1962. Vol. 33. P. 1065-1076.