

Замечание о некоторых классических критериях математической статистики

И. С. Лунев, В. В. Нежруткин

Санкт-Петербургский государственный университет,
Российская Федерация, 199034, Санкт-Петербург, Университетская наб., 7–9

Для цитирования: Лунев И. С., Нежруткин В. В. Замечание о некоторых классических критериях математической статистики // Вестник Санкт-Петербургского университета. Математика. Механика. Астрономия. 2019. Т. 6 (64). Вып. 2. С. 221–231.
<https://doi.org/10.21638/11701/spbu01.2019.204>

В работе изучаются асимптотические свойства стандартного статистического критерия (иногда называемого t -критерием для коэффициента корреляции) для проверки гипотезы о значимости коэффициента корреляции Пирсона между случайными величинами x и y . Несмотря на то, что этот критерий обоснован только при условии гауссовости совместного распределения x и y , он является очень распространенным и включен в большинство статистических пакетов. При этом предположение о гауссовости распределений на практике, как правило, не выполняется, и поэтому возникает задача описания области применимости t -критерия при больших объемах выборки. В работе доказано, что при выполнении некоторых дополнительных условий этот критерий является асимптотически точным, если x и y независимы, в то время как обычной некоррелированности для этого бывает недостаточно. Также построен асимптотически точный и состоятельный критерий в случае, когда независимость отсутствует. Вычислительные эксперименты свидетельствуют о его применимости на практике. Кроме того, эти результаты перенесены с соответствующими изменениями на частный коэффициент корреляции.

Ключевые слова: коэффициент корреляции Пирсона, частный коэффициент корреляции, критерии значимости, асимптотический анализ.

1. Введение. Хорошо известно, что некоторые классические критерии математической статистики основаны на предположении о гауссовости исходных распределений. Часто эти критерии восходят к результатам сэра Рональда Фишера, полученным еще в первой четверти прошлого столетия (см., например, [1]).

Такого рода процедуры обычно опираются на точные распределения соответствующих статистик. Конечно, на практике предположение о гауссовости выборок, как правило, не выполняется, и распределения этих статистик могут быть совершенно другими.

Тем не менее, во многих статистических пакетах (например, Statistica, R или SAS, из отечественных пакетов можно упомянуть STADIA, см. [2]) представлены именно эти классические критерии, и часто никаких альтернативных процедур не предусмотрено.

В связи с этим возникает вопрос об условиях применимости классических статистических критериев в том случае, когда выборки не являются гауссовскими. Кроме того, если эти условия не выполняются, желательно иметь альтернативный критерий для проверки соответствующей гипотезы.

В настоящей работе эти задачи решаются для двух распространенных на практике критериев. А именно, рассматриваются гипотезы о равенстве нулю коэффициента корреляции Пирсона и частного коэффициента корреляции. Для обеих гипотез есть общеупотребительные классические критерии (они описаны в разделах 2 и 3), обоснованные в случае выборок из невырожденных гауссовских распределений, для краткости они в дальнейшем будут называться соответственно N_{corr} - и N_{pcorr} -критериями.

Во втором разделе работы (см. предложение 1 и замечание 1) показано, что при выполнении некоторых условий (их обсуждению посвящено замечание 3) N_{corr} -критерий является асимптотически точным, если рассматриваемые случайные величины являются не просто некоррелированными, но и независимыми.

Кроме того, в предложении 2 и замечании 2 приводятся достаточно универсальные асимптотически точные и состоятельные критерии проверки равенства нулю коэффициента корреляции, не требующие предположения о гауссовости выборки.

В разделе 4 помещены некоторые результаты численных экспериментов, подтверждающие практическую значимость полученных результатов.

Заметим, что нас интересует только случай $\rho = 0$ (проверяется именно эта гипотеза, а вопросы, связанные с ошибками второго рода рассматриваются только в самом общем виде — является ли критерий состоятельным или нет, см. предложение 2 и замечание 2). Поэтому мы опускаем обсуждение особенностей, связанных с оценением ненулевых коэффициентов корреляции (и проверкой соответствующих гипотез). В частности, здесь не рассматриваются вопросы применения так называемого z -преобразования Фишера (см., например, [3, §69]), являющегося методом «стабилизации дисперсии» при $\rho \neq 0$ для гауссовских выборок.

Проверке гипотезы о равенстве нулю частного коэффициента корреляции $\rho_{x,y|z}$ между двумя случайными величинами x и y при исключении влияния третьей величины z посвящен раздел 3 работы. Поскольку $\rho_{x,y|z}$ есть ничто иное, как обычный коэффициент корреляции между остатками линейных регрессий x на z и y на z , то все результаты о коэффициенте корреляции Пирсона переносятся с соответствующими изменениями и на этот случай.

2. Коэффициент корреляции Пирсона. 2.1. Асимптотические свойства N_{corr} -критерия. Пусть случайный вектор $(x, y)^T$ обладает непрерывными распределениями координат, причем дисперсии $\sigma_x^2 = Dx$ и $\sigma_y^2 = Dy$ предполагаются конечными.

Рассмотрим n независимых копий $(x_1, y_1)^T, \dots, (x_n, y_n)^T$ этого вектора и будем проверять нулевую гипотезу

$$H_0 : \rho_{x,y} = 0, \quad (1)$$

где $\rho_{x,y}$ — коэффициент корреляции между x и y .

Если вектор $(x, y)^T$ имеет невырожденное двумерное гауссовское распределение, критерий для проверки гипотезы (1) хорошо известен. А именно, имеет место следующее утверждение (см., например, [3, §67] или [4, следств. 5.1.2]).

Теорема 1. Пусть $(x_1, y_1)^T, \dots, (x_n, y_n)^T$ — независимые случайные одинаково распределенные гауссовские векторы, причем $\rho_{x,y} = 0$. Положим

$$\nu_n = \sqrt{n-2} \frac{\hat{\rho}_n}{\sqrt{1-\hat{\rho}_n^2}}, \quad (2)$$

где $\widehat{\rho}_n$ – выборочный коэффициент корреляции

$$\widehat{\rho}_n = \widehat{\rho}_{x,y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})/n}{\overline{s}_n(x)\overline{s}_n(y)}, \quad (3)$$

а $\overline{s}_n(x)$ и $\overline{s}_n(y)$ – выборочные стандарты. Тогда случайная величина ν_n имеет распределение Стьюдента с $n - 2$ степенями свободы.

Определение 1. Благодаря утверждению теоремы 1 получаем критерий с уровнем значимости α , который отвергает нулевую гипотезу, если $|\nu_n| \geq T_{n-2, 1-\alpha/2}$, где $T_{n,\gamma}$ – γ -квантиль распределения Стьюдента с n степенями свободы. Для краткости назовем этот критерий N_{corr} -критерием.

Перейдем теперь к негауссовскому случаю. Основным утверждением, на которое опирается все дальнейшее повествование, является следующий факт.

Теорема 2. Рассмотрим n независимых копий $(x_1, y_1)^T, \dots, (x_n, y_n)^T$ случайного вектора $(x, y)^T$, причем предположим, что распределения x и y непрерывны и что их четвертые моменты конечны. Обозначим $\rho = \rho_{x,y}$ и

$$\sigma_\rho^2 = D(\rho(x_*^2 + y_*^2) - 2x_*y_*)/4, \quad (4)$$

где $x_* = (x - Ex)/\sigma_x$, $y_* = (y - Ey)/\sigma_y$. Тогда при $n \rightarrow \infty$

$$\mathcal{L}(\sqrt{n}(\widehat{\rho}_n - \rho)) \Rightarrow N(0, \sigma_\rho^2), \quad (5)$$

где $\widehat{\rho}_n$ – выборочный коэффициент корреляции (3).

Доказательство теоремы 2 можно найти в [5, теор. 8], оно основано на применении так называемого «дельта-метода» (см., например, [6, теор. 5.4.6] или [7, теор. 1.12]). Конечно, если $\rho = 0$, то предельная дисперсия (4) имеет вид

$$\sigma_0^2 := D(x_*y_*). \quad (6)$$

Следующее утверждение посвящено условиям, при которых N_{corr} -критерий являлся асимптотически точным для проверки нулевой гипотезы (1), то есть когда вероятность ошибки первого рода критерия стремится к заданному значению α .

Предложение 1. Пусть выполнены условия теоремы 2. Тогда для того, чтобы N_{corr} -критерий являлся асимптотически точным для проверки гипотезы (1), необходимо и достаточно, чтобы $\sigma_0^2 = 1$.

ДОКАЗАТЕЛЬСТВО. Согласно теореме 2, $\mathcal{L}_{H_0}(\sqrt{n}\widehat{\rho}_n) \Rightarrow N(0, \sigma_0^2)$, поэтому распределение статистики ν_n , определенной в (2), имеет такой же слабый предел. Осталось заметить, что квантили распределения Стьюдента с n степенями свободы сходятся к квантилям распределения $N(0, 1)$ при $n \rightarrow \infty$. \square

Замечание 1. В частности, если в условиях теоремы 2 случайные величины x и y независимы, то N_{corr} -критерий будем асимптотически точным.

Этим фактом можно объяснить популярность N_{corr} -критерия на практике. Так как, проверяя гипотезу о равенстве нулю коэффициента корреляции, пользователи обычно имеют в виду предварительную проверку независимости рассматриваемых

случайных величин, то при достаточно больших объемах выборки классический N_{corr} -критерий может служить для проверки именно этой гипотезы.

Вычислительные эксперименты (результаты некоторых из них помещены в раздел 4, см. рис. 1, *a* и *б*) в целом подтверждают эти рассуждения: объемы выборок, для которых вероятность ошибок первого рода достаточно близка к уровню значимости, оказываются для N_{corr} -критерия вполне реалистичными.

Следует отметить, что N_{corr} -критерий не будет состоятельным, если в качестве альтернативы рассматривать зависимость случайных величин x и y . Действительно, пусть распределение вектора $(x, y)^T$ обладает круговой симметрией, а для полярного радиуса $r = \sqrt{x^2 + y^2}$ выполнены условия $Er^2 > 0$ и $Er^4 < \infty$. Тогда $\rho_{x,y} = 0$ и, как нетрудно показать, $\sigma_0^2 = 0.5Er^4 / (Er^2)^2$. Таким образом, для любой положительной случайной величины r^2 , удовлетворяющей условию $Er^4 = 2(Er^2)^2$, имеет место равенство $\sigma_0^2 = 1$. Примером может служить случайная величина r^2 , равномерно распределенная на множестве $\{1, 2, 6 + \sqrt{39}\}$, для которой x и y некоррелированы, но зависимы.

2.2. Асимптотический критерий для негауссовских выборок. Как и прежде, мы будем предполагать выполнение условий теоремы 2.

Будем по-прежнему рассматривать нулевую гипотезу $H_0 : \rho = 0$, а в качестве альтернативы H_ρ возьмем любое распределение, удовлетворяющее условиям теоремы 2, для которого $0 < |\rho| < 1$. Задача состоит в построении асимптотически точного и состоятельного критерия для проверки гипотезы H_0 против альтернативы H_ρ .

Определение 2. Пусть для оценки $\hat{\sigma}^2 = \hat{\sigma}^2(n)$ асимптотической дисперсии σ_0^2 , определенной в (6), равенство $P_{H_0}(\hat{\sigma}^2 > 0) = 1$ выполняется для любого достаточно большого n . Положим $\mu_n = \sqrt{n}\hat{\rho}_n/\hat{\sigma}$. Наконец при $\alpha \in (0, 1)$ введем $\hat{\sigma}^2$ -критерий, который отвергает нулевую гипотезу, если $|\mu_n| \geq C_{1-\alpha/2}$, где C_γ — γ -квантиль стандартного нормального распределения.

Предложение 2. 1. Если

$$\hat{\sigma}^2 \xrightarrow{P_{H_0}} \sigma_0^2 \quad \text{при } n \rightarrow \infty, \quad (7)$$

то $\hat{\sigma}^2$ -критерий является асимптотически точным для проверки нулевой гипотезы.

2. Если $P_{H_\rho}(\hat{\sigma}^2 > 0) = 1$ и существует такое M , что $P_{H_\rho}(\hat{\sigma}^2 < M) \rightarrow 1$ при $n \rightarrow \infty$, то $\hat{\sigma}^2$ -критерий является состоятельным против альтернативы H_ρ .

Доказательство. 1. Ввиду результата теоремы 2 асимптотическая точность критерия не требует специального доказательства.

2. Возьмем $\delta > 0$ и рассмотрим случай $\rho > 0$ (для отрицательных значений ρ все аналогично). Обозначив B_n событие $\{\hat{\sigma}^2 < M\}$, получим, что

$$\begin{aligned} P_{H_\rho}(|\mu_n| > \delta) &\geq P_{H_\rho}\left(\frac{\sqrt{n}}{\sigma_\rho}(\hat{\rho}_n - \rho) < -\delta\hat{\sigma} + \frac{\sqrt{n}}{\sigma_\rho}\rho, B_n\right) + \\ &+ P_{H_\rho}\left(\frac{\sqrt{n}}{\sigma_\rho}(\hat{\rho}_n - \rho) < -\delta\hat{\sigma} + \frac{\sqrt{n}}{\sigma_\rho}\rho, B_n^c\right) = J_1(n) + J_2(n). \end{aligned}$$

Далее будем иметь $J_2(n) \leq P_{H_\rho}(B_n^c) \rightarrow 0$ и

$$J_1(n) \geq P_{H_\rho}\left(\frac{\sqrt{n}}{\sigma_\rho}(\hat{\rho}_n - \rho) < -\delta\sqrt{M} + \frac{\sqrt{n}}{\sigma_\rho}\rho\right),$$

что стремится к 1 ввиду теоремы 2. Утверждение доказано. \square

Обсудим варианты выбора статистики $\hat{\sigma}^2$. Если взять в качестве $\hat{\sigma}^2$ оценку

$$\hat{\sigma}_\rho^2 := \frac{1}{n} \sum_{i=1}^n \left(\hat{\rho}_n \left(\left(\frac{x_i - \bar{x}}{\bar{s}_n(x)} \right)^2 + \left(\frac{y_i - \bar{y}}{\bar{s}_n(y)} \right)^2 \right) - 2 \frac{x_i - \bar{x}}{\bar{s}_n(x)} \frac{y_i - \bar{y}}{\bar{s}_n(y)} \right)^2 / 4, \quad (8)$$

которая является выборочным аналогом правой части (4), то все условия предложения 2 будут выполнены.

Поскольку статистика (8) является состоятельной оценкой предельной дисперсии σ_ρ^2 как при выполнении нулевой гипотезы, так и для альтернативы, то ее можно использовать не только для проверки гипотезы H_0 , но и при построении доверительных интервалов для неизвестного значения ρ . Нас, однако, интересует только случай $\rho = 0$, поэтому оценку (8) можно упростить. Возьмем в качестве $\hat{\sigma}^2$ статистику

$$\hat{\sigma}_0^2 := \frac{\sum_{i=1}^n (x_i - \bar{x})^2 (y_i - \bar{y})^2}{n \bar{s}_n^2(x) \bar{s}_n^2(y)}. \quad (9)$$

Поскольку при выполнении нулевой гипотезы $\hat{\sigma}_0^2$ является выборочным аналогом σ_0^2 , то сходимость (7) выполнена автоматически. Более того, для любой альтернативы H_ρ имеет место сходимость $\hat{\sigma}_0^2 \xrightarrow{P_{H_\rho}} E_{H_\rho} x_*^2 y_*^2$, что обеспечивает выполнение условий второго пункта предложения 2.

Определение 3. Назовем G_{corr} -критерием $\hat{\sigma}$ -критерий с $\hat{\sigma}^2 = \hat{\sigma}_\rho^2$, а M_{corr} -критерием $\hat{\sigma}$ -критерий с $\hat{\sigma}^2 = \hat{\sigma}_0^2$.

Замечание 2. Таким образом, и G_{corr} -критерий с оценкой предельной дисперсии (8), и M_{corr} -критерий с оценкой (9) являются асимптотически точными и состоятельными для проверки гипотезы о значимости коэффициента корреляции.

Использование статистики (9), однако, представляется более предпочтительным, так как она при выполнении нулевой гипотезы выглядит «менее случайной».

Численное моделирование (см. рис. 1, 6 и 2 в разделе 4) это подтверждает: во всех проведенных экспериментах оказалось, что для достижения приемлемой близости вероятности ошибки первого рода к заданной величине уровня критерия α M_{corr} -критерий требует в несколько раз меньшие объемы выборок, чем G_{corr} -критерий.

3. Частный коэффициент корреляции. Пусть у случайных величин x , y и z есть конечные вторые моменты и ненулевые дисперсии σ_x^2 , σ_y^2 и σ_z^2 , причем $|\rho_{x,z}| \neq 1$ и $|\rho_{y,z}| \neq 1$. Рассмотрим остатки

$$\varepsilon_1 = x - Ex - \rho_{x,z} \frac{\sigma_x}{\sigma_z} (z - Ez), \quad \varepsilon_2 = y - Ey - \rho_{y,z} \frac{\sigma_y}{\sigma_z} (z - Ez) \quad (10)$$

наилучших (в смысле среднеквадратического отклонения) линейных аппроксимаций случайных величин x и y с помощью случайной величины z .

Коэффициент корреляции $\rho_{\varepsilon_1, \varepsilon_2}$ между этими остатками называется *частным коэффициентом корреляции между x и y при исключении влияния z* и обозначается $\rho_{x,y|z}$. Прямые вычисления показывают, что

$$\rho_{x,y|z} = \frac{\rho_{x,y} - \rho_{x,z}\rho_{y,z}}{\sqrt{1 - \rho_{x,z}^2}\sqrt{1 - \rho_{y,z}^2}}. \quad (11)$$

Ситуация, когда $\rho_{x,y} \neq 0$, а $\rho_{x,y|z} = 0$, обычно интерпретируется таким образом, что наблюдаемая зависимость между x и y объясняется исключительно их зависимостью от z . Отсюда возникает задача проверки гипотезы о том, что $\rho_{x,y|z} = 0$.

Так как $\rho_{x,y|z}$ является коэффициентом корреляции Пирсона для регрессионных остатков (10), то все основные утверждения раздела 2 переносятся и на этот случай. Кратко остановимся на соответствующих результатах.

Если $(x_1, y_1, z_1)^T, \dots, (x_n, y_n, z_n)^T$ — независимые копии случайного вектора $(x, y, z)^T$, имеющего невырожденное гауссовское распределение, то критерий для проверки такой гипотезы снова хорошо известен. Для этой цели рассматривается статистика

$$\beta_n = \sqrt{n-3} \hat{\rho}_{x,y|z} / \sqrt{1 - \hat{\rho}_{x,y|z}^2}, \quad (12)$$

где $\hat{\rho}_{x,y|z}$ — выборочный коэффициент частной корреляции:

$$\hat{\rho}_{x,y|z} = \frac{\hat{\rho}_{x,y} - \hat{\rho}_{x,z}\hat{\rho}_{y,z}}{\sqrt{1 - \hat{\rho}_{x,z}^2}\sqrt{1 - \hat{\rho}_{y,z}^2}}. \quad (13)$$

Поскольку статистика (12) при выполнении нулевой гипотезы $H_0 : \rho_{x,y|z} = 0$ имеет распределение Стьюдента с $(n-3)$ степенями свободы (см., например, [3, § 68]), то критерий для проверки гипотезы H_0 строится совершенно аналогично критерию N_{corr} (см. определение 1). Для удобства назовем этот критерий N_{pcorr} -критерием.

В общем случае имеет место следующий аналог теоремы 2.

Теорема 3. Пусть $(x_1, y_1, z_1)^T, \dots, (x_n, y_n, z_n)^T$ — независимые копии случайного вектора $(x, y, z)^T$, причем предположим, что распределения x , y и z непрерывны, четвертые моменты x , y и z конечны, а $\rho_{x,z}^2 \neq 1$, $\rho_{y,z}^2 \neq 1$. Обозначим x_* , y_* и z_* результат центрирования и нормировки x , y и z . Кроме того, положим

$$u_* = \frac{x_* - \rho_{x,z}z_*}{\sqrt{1 - \rho_{x,z}^2}}, \quad v_* = \frac{y_* - \rho_{y,z}z_*}{\sqrt{1 - \rho_{y,z}^2}}.$$

Тогда $\mathcal{L}(\sqrt{n}(\hat{\rho}_{x,y|z} - \rho_{x,y|z})) \Rightarrow N(0, \sigma_{xy|z}^2)$ при $n \rightarrow \infty$, где коэффициент $\hat{\rho}_{x,y|z}$ определен в (13), а

$$\sigma_{xy|z}^2 = D(\rho_{u,v}(u_*^2 + v_*^2) - 2u_*v_*)/4. \quad (14)$$

ДОКАЗАТЕЛЬСТВО. Так как x_* , y_* и z_* центрированы и нормированы, то

$$\rho_{x,y|z} = \rho(x_* - \rho_{x,z}z_*, y_* - \rho_{y,z}z_*) = E u_* v_*.$$

Ссылка на формулу (4) теоремы 2 завершает доказательство. \square

Приведем теперь аналог предложения 1.

Предложение 3. Пусть выполнены условия теоремы 3. Для того чтобы $N_{рсорт}$ -критерий был асимптотически точным для проверки гипотезы $\rho_{x,y|z} = 0$, необходимо и достаточно, чтобы $D(u_*v_*) = 1$. Достаточным условием для этого является независимость регрессионных остатков (10).

Повторяя с естественными изменениями рассуждения раздела 2, построим асимптотический критерий для проверки гипотезы $\rho_{x,y|z} = 0$. Пусть $\hat{\rho}_{x,z}$ и $\hat{\rho}_{y,z}$ — выборочные коэффициенты корреляции между x, z и y, z . Кроме того, положим

$$\hat{u}_* = \frac{\frac{x_i - \bar{x}}{\bar{s}_n(x)} - \hat{\rho}_{x,z} \frac{z_i - \bar{z}}{\bar{s}_n(z)}}{\sqrt{1 - \hat{\rho}_{x,z}^2}}, \quad \hat{v}_* = \frac{\frac{y_i - \bar{y}}{\bar{s}_n(y)} - \hat{\rho}_{y,z} \frac{z_i - \bar{z}}{\bar{s}_n(z)}}{\sqrt{1 - \hat{\rho}_{y,z}^2}}$$

и определим статистики

$$\hat{\sigma}_{x,y|z,\rho}^2 = \frac{1}{n} \sum_{i=1}^n \left(\hat{\rho}_{x,y|z} (\hat{u}_*^2 + \hat{v}_*^2) - 2\hat{u}_*\hat{v}_* \right)^2 / 4 \quad (15)$$

и

$$\hat{\sigma}_{x,y|z,0}^2 = \frac{1}{n} \sum_{i=1}^n (\hat{u}_*\hat{v}_*)^2, \quad (16)$$

где выборочный частный коэффициент корреляции $\hat{\rho}_{x,y|z}$ определен в (13).

Далее, рассмотрим статистику $\hat{\tau}^2$ такую, что $P_{H_0}(\hat{\tau}^2 > 0) = 1$. Положив

$$\gamma_n = \sqrt{n} \hat{\rho}_{x,y|z} / \hat{\tau}$$

и взяв $\alpha \in (0, 1)$, введем $\hat{\tau}$ -критерий для проверки гипотезы $\rho_{x,y|z} = 0$ точно так же, как это делалось для коэффициента корреляции Пирсона (см. определение 2).

Предложение 4. Если в условиях теоремы 3 $\hat{\tau}^2$ имеет вид (15) или (16), то $\hat{\tau}$ -критерий будет асимптотически точным и состоятельным.

Замечание 3. Остановимся на ограничениях, при которых получены результаты настоящей работы. Требование конечности четвертых моментов в теоремах 2 и 3 является естественным, так как оно обеспечивает конечность предельных дисперсий (4) и (14). Что касается требования непрерывности распределений случайных величин x и y в теореме 2, а также x, y и z в теореме 3, то оно не является обязательным и введено только для удобства изложения.

Действительно, если в распределении случайных величин x или y есть дискретная компонента, то выборочный коэффициент корреляции (3) оказывается для любого n определенным с вероятностью, меньшей 1. Поэтому, например, сходимость (5) приобретает вид

$$\mathcal{L}\left(\sqrt{n}(\hat{\rho}_n - \rho) \mid \bar{s}_n(x)\bar{s}_n(y) > 0\right) \Rightarrow N(0, \sigma_\rho^2),$$

что влечет за собой соответствующие изменения во всех дальнейших рассуждениях.

Более того, и традиционная формулировка теоремы о «дельта-методе», упомянутой в разделе 2, требует коррекции. Все эти изменения не влияют на основные выводы работы, но сильно загромождают ее изложение. Поэтому в статье использованы более сильные ограничения на распределения рассматриваемых случайных величин, делающие все рассуждения значительно прозрачнее.

4. Вычислительные эксперименты. В вычислительных экспериментах при различных объемах выборки и различных совместных распределениях случайных величин x и y с нулевым коэффициентом корреляции ρ производится сравнение (в смысле близости оценок вероятностей ошибок первого рода $\hat{\alpha}$ с заданным уровнем значимости $\alpha = 0.05$) трех критериев: N_{corr} -критерия (см. определение 1), а также G_{corr} -критерия и M_{corr} -критерия (см. определение 3). Число повторностей при оценке вероятности ошибок первого рода выбрано равным $N = 10^5$, а моделирование проводилось методом зависимых выборок при $n = 50$ (50) 1000.

Так как погрешности оценок искомых вероятностей при таком числе повторностей не влияют на общие выводы из проведенных экспериментов, то доверительные интервалы для вероятностей ошибок первого рода графически не изображаются.

Общий вывод из всех проведенных (а не только представленных ниже) вычислительных экспериментов может быть сформулирован следующим образом.

M_{corr} -критерий, по всей видимости, может быть использован на практике при проверке значимости коэффициента корреляции для объемов выборок, начиная с n порядка $50 \div 100$, если выборки производятся из распределений с компактными носителями, и начиная с $n \approx 100 \div 200$, если эти распределения обладают не слишком «тяжелыми хвостами». Такие объемы выборок являются вполне реалистичными для практики.

Аналогичный вывод можно сделать об использовании N_{corr} -критерия в случае независимости случайных величин x и y .

4.1. Независимые координаты. Здесь представлено сравнение оценок вероятностей ошибок первого рода $\hat{\alpha}$ с уровнем значимости $\alpha = 0.05$ для описанных выше критериев в случае, когда случайные величины x, y независимы. В этом случае $\sigma_0^2 = 1$, и для всех критериев вероятности этих ошибок должны быть близки к уровням значимости при достаточно больших n .

На рис. 1, *a* и *б* показаны оценки вероятностей ошибок первого рода для равномерного (как примера распределений с компактным носителем) и показательного (имеющего более «тяжелый» хвост, чем нормальное) распределений для трех описанных выше критериев.

Как и ожидалось, результаты моделирования показывают, что «классический» N_{corr} -критерий, как и M_{corr} -критерий с оценкой предельной дисперсии (9) имеют примерно одинаковые и адекватные результаты при достаточно реалистичных объемах выборок.

В то же время G_{corr} -критерий с оценкой предельной дисперсии (8) демонстрирует существенное допредельное завышение вероятности ошибок первого рода над заданным уровнем значимости.

4.2. Зависимые, но некоррелированные переменные. Рис. 1, *в* и *г* являются иллюстрациями к ситуации, когда случайные величины x, y являются зависимыми, но некоррелированными.

На рис. 1, *в* показаны результаты сравнения оценок вероятностей ошибок первого рода $\hat{\alpha}$ с заданным уровнем значимости $\alpha = 0.05$ для трех описанных выше

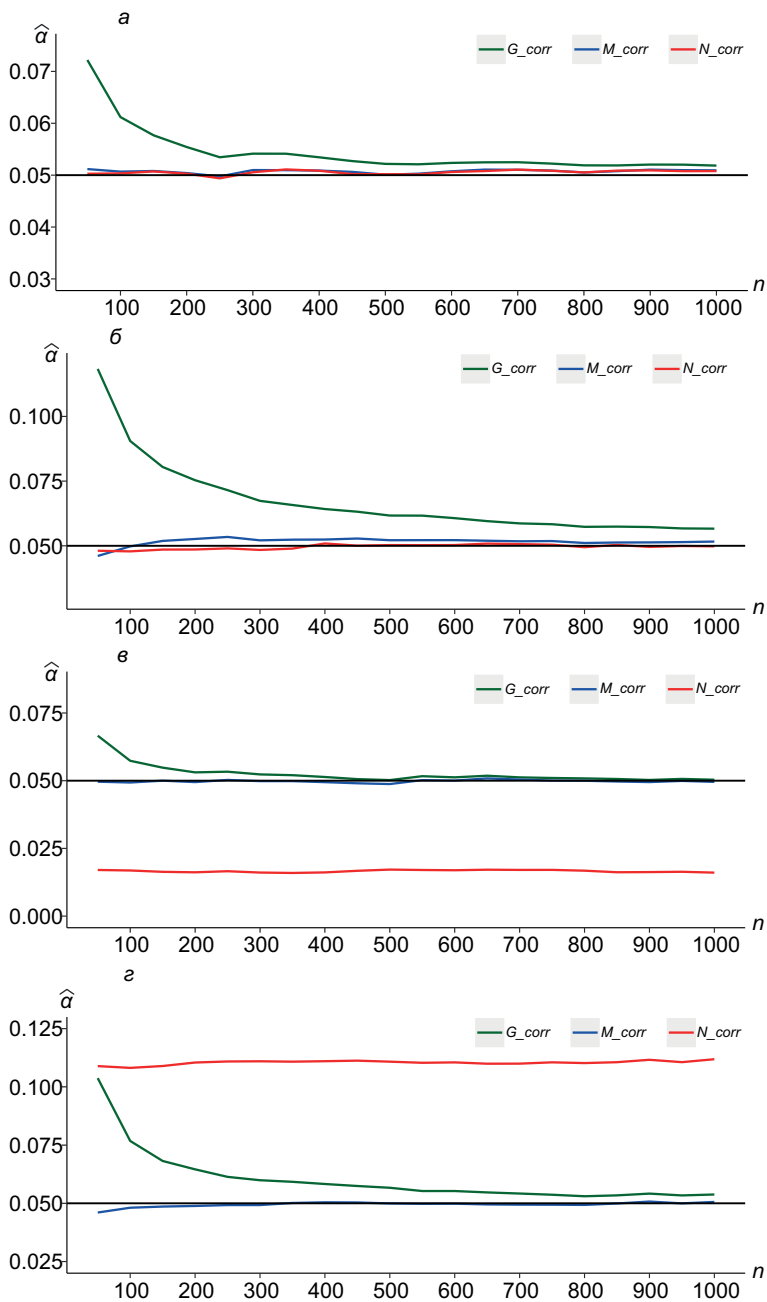


Рис. 1. Оценки вероятностей ошибок первого рода $\hat{\alpha}$ для критериев G_{corr} , M_{corr} и N_{corr} при $\alpha = 0.05$ и различных объемах выборки n : а — равномерное распределение, независимость; б — экспоненциальное распределение, независимость; в — равномерное распределение в круге, некоррелированность; г — круговая симметрия, $r^2 \in \text{Gamma}(1/2, 1)$, некоррелированность.

критериев в случае, когда выборка осуществляется из равномерного распределения в круге.

Здесь $\sigma_0^2 = 2/3 \neq 1$ и можно ожидать, что N_{corr} -критерий продемонстрирует значительное расхождение между заданным уровнем значимости и вероятностью ошибки первого рода, а M_{corr} -критерий с оценкой предельной дисперсии (9) окажется приемлемо точным при умеренно большом объеме выборки.

Моделирование подтверждает эти ожидания, причем для M_{corr} -критерия близость вероятности ошибки первого рода к уровню значимости наблюдается во всем диапазоне рассматриваемых объемов выборок.

В примере, представленном на рис. 1, *г*, считается, что распределение вектора $(x, y)^T$ обладает круговой симметрией, а квадрат полярного радиуса $r^2 = x^2 + y^2$ имеет гамма-распределение с параметром формы $k = 1/2$. Тогда, как нетрудно видеть, $\sigma_0^2 = 1.5$.

Результаты в целом аналогичны показанным на рис. 1, *в*: вероятности ошибок первого рода для N_{corr} -критерия сильно отличаются от заданного уровня значимости, в то время как M_{corr} -критерий демонстрирует хорошую точность при относительно небольших объемах выборки.

Отметим, что в обоих экспериментах видна медленная скорость сходимости распределения статистики G_{corr} -критерия к предельному нормальному распределению.

5. Заключение. Таким образом, в настоящей работе показано, что общеупотребительный классический критерий проверки значимости коэффициента корреляции, обоснованный в случае выборок из невырожденных гауссовских распределений, является асимптотически точным в случае, когда рассматриваемые случайные величины являются не просто некоррелированными, но и независимыми.

Также представлен достаточно универсальный асимптотически точный и состоятельный M_{corr} -критерий проверки равенства нулю коэффициента корреляции, не требующий этих предположений.

Проведенные вычислительные эксперименты позволяют надеяться, что этот критерий может быть использован на практике при объемах выборок, начиная с $n \approx 50 \div 100$, если выборки производятся из распределений с компактными носителями, и начиная с $n \approx 100 \div 200$, если эти распределения обладают не слишком «тяжелыми хвостами».

Аналогичные результаты получены для частного коэффициента корреляции.

Авторы благодарят обоих анонимных рецензентов, чьи замечания, несомненно, способствовали улучшению нашей работы.

Литература

1. Фишер Р. Статистические методы для исследователей. М.: Госстатиздат, 1958.
2. Тюрин Ю. Н., Макаров А. А. Анализ данных на компьютере. М.: ИНФРА-М, 2002.
3. ван дер Варден Б. Л. Математическая статистика. М.: ИИЛИ, 1960.
4. Muirhead R. J. Aspects of multivariate statistical theory. New York: Wiley-Interscience, 2005.
5. Ferguson T. S. A course in large-sample theory. London: Chapman & Hall, 1996.
6. Lehmann E. L. Elements of large-sample theory. New York: Springer-Verlag, 1999.
7. Shao Ju. Mathematical statistics. New York: Springer-Verlag, 1999.

Статья поступила в редакцию 2 июля 2018 г.;
после доработки 2 декабря 2018 г.;
рекомендована в печать 20 декабря 2018 г.

Контактная информация:

Лунев Иван Сергеевич — студент; ivan.lunev96@gmail.com

Некруткин Владимир Викторович — канд. физ.-мат. наук, доц.; v.nekrutkin@spbu.ru, vnekr@statmod.ru

A remark on certain classical criteria of mathematical statistics

I. S. Lunev, V. V. Nekrutkin

St. Petersburg State University, Universitetskaya nab., 7–9, St. Petersburg, 199034, Russian Federation

For citation: Lunev I. S., Nekrutkin V. V. A remark on certain classical criteria of mathematical statistics. *Vestnik of Saint Petersburg University. Mathematics. Mechanics. Astronomy*, 2019, vol. 6 (64), issue 2, pp. 221–231. <https://doi.org/10.21638/11701/spbu01.2019.204> (In Russian)

The paper is devoted to asymptotical features of the standard statistical test for the significance of Pearson correlation between random variables x and y . Although this test (sometimes called t -test for the correlation coefficient) is only justified if the distribution of (x, y) is Gaussian, it is very widespread and is included in many statistical packages. However, in practice the Gaussian assumption usually fails. Thus, it is worth to describe the area of the correct application of t -test for big sample sizes. It is proven that under certain additional conditions, this test is asymptotically precise if x and y are independent, while the simple lack of correlation is not sufficient for such a feature. Also, the general asymptotically precise and consistent test for the significance of correlation coefficient is proposed. Computational experiments show that this test can be successfully applied in practice. Under modifications these results are transferred to the partial correlation coefficient.

Keywords: Pearson correlation coefficient, partial correlation, significance criteria, asymptotical analysis.

References

1. Fisher R., *Statistical methods for research workers* (5th ed., Edinburgh, Oliver & Boyd, 1934).
2. Tyurin Yu. N., Makarov A. A., *Data analysis on a computer* (INFRA-M Publ., Moscow, 2002). (In Russian)
3. van der Waerden B. L., *Mathematical statistics* (Springer-Verlag, New York, 1969).
4. Muirhead R. J., *Aspects of multivariate statistical theory* (Wiley-Interscience, New York, 2005).
5. Ferguson T. S., *A course in large-sample theory* (Chapman & Hall, London, 1996).
6. Lehmann E. L., *Elements of large-sample theory* (Springer-Verlag, New York, 1999).
7. Shao Ju., *Mathematical statistics* (Springer-Verlag, New York, 1999).

Received: July 2, 2018

Revised: December 2, 2018

Accepted: December 20, 2018

Author's information:

Ivan S. Lunev — ivan.lunev96@gmail.com

Vladimir V. Nekrutkin — v.nekrutkin@spbu.ru, vnekr@statmod.ru