Некоторые задачи проверки статистических гипотез и их применения

Выпускная квалификационная работа

Лунев Иван Сергеевич, группа 422

Санкт-Петербургский Государственный Университет
Прикладная математика и информатика
Вычислительная стохастика и статистические модели

Научный руководитель: к.ф.-м.н., доцент В. В. Некруткин
Рецензент: к.ф.-м.н., доцент А. И. Коробейников

Санкт-Петербург
2018 г.

Общая проблема

Две задачи — проверка равенства нулю коэффициента корреляции Пирсона и частного коэффициента корреляции.

Классические критерии — в предположении гауссовости выборок.

Пакеты Statistica, SPSS, R и т. д.

Проблемы:

1. Когда классические критерии применимы, если распределения не гауссовские?
2. Построение критериев, применимых в общей ситуации.

Проверка гипотезы $\rho = 0$. Гауссовские выборки

Если вектор (x, y) имеет невырожденное двумерное гауссовское распределение, критерий хорошо известен [Robb J. Muirhead, 2009]:

Нулевая гипотеза: $H_0 : \{\rho = 0\}$.
Статистика:

$$\mu(\mathbf{X}_n, \mathbf{Y}_n) := \sqrt{n-2}\ \frac{\widehat{\rho}_n}{\sqrt{1-\widehat{\rho}_n^2}},$$

где $\widehat{\rho}_n$ — выборочный коэффициент корреляции.

Распределение статистики:

$$\mathcal{L}_{H_0}\big(\mu(\mathbf{X}_n, \mathbf{Y}_n)\big) = \mathrm{t}(n-2).$$

Классический критерий

Получаем критерий с уровнем значимости α, который отвергает нулевую гипотезу, если $|\mu(\mathbf{X}_n, \mathbf{Y}_n)| \geq T_{n-2,1-\alpha/2}$, где $T_{m,\gamma}$ — γ-квантиль распределения Стьюдента с m степенями свободы.

$(\mathbf{X}_n, \mathbf{Y}_n)$ – двумерная независимая повторная выборка, соответствующая вектору (x, y) с непрерывными координатами, причем четвертые моменты x и y конечны. Обозначим

$$x^* = (x - \mathbb{E}x)/\sigma_x, \quad y^* = (y - \mathbb{E}y)/\sigma_y,$$

$$\sigma^2_{f,\rho} = \mathbb{D}\left(\rho({x^*}^2 + {y^*}^2) - 2x^*y^*\right)/4,$$

$$\mu(\mathbf{X}_n, \mathbf{Y}_n) = \frac{\sqrt{n}}{\sigma_{f,\rho}}\left(\widehat{\rho}_n - \rho\right),$$

$\widehat{\rho}_n$ — выборочный коэффициент корреляции.

Утверждение (ход доказательства, например, в [E. Lehman, 1999]).

Если $|\rho| \neq 1$ и $n \to \infty$

$$\mathcal{L}(\mu(\mathbf{X}_n, \mathbf{Y}_n)) \Rightarrow \mathrm{N}(0, 1).$$

Для гауссовской выборки $\sigma^2_{f,\rho} = (1 - \rho^2)^2$.

Проверка гипотезы $\rho = 0$. Общий вид критерия

Нулевая гипотеза: $H_0 : \{\rho = 0\}$.

Статистика:

$$\mu(\mathbf{X}_n, \mathbf{Y}_n) = \frac{\sqrt{n}}{\widehat{\sigma_f}}\widehat{\rho}_n, \quad \text{где } \widehat{\sigma_f^2} \text{ — состоятельная оценка } \sigma^2_{f,0}.$$

Критерий

Пусть $\alpha \in (0, 1)$. Критерий отвергает гипотезу H_0, если

$$|\mu(\mathbf{X}_n, \mathbf{Y}_n)| \geq C_{1-\alpha/2}\,,$$

где C_β – квантиль уровня β распределения $\mathrm{N}(0, 1)$.

Проверка гипотезы $\rho = 0$. Две оценки $\sigma^2_{f,\rho}$ — два критерия

«Общий» критерий:

$$\widehat{\sigma}^2_{f,\rho} = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\rho_n}\left(\left(\frac{x_i-\overline{x}}{\overline{s}_n(x)}\right)^2+\left(\frac{y_i-\overline{y}}{\overline{s}_n(y)}\right)^2\right)-2\,\frac{x_i-\overline{x}}{\overline{s}_n(x)}\,\frac{y_i-\overline{y}}{\overline{s}_n(y)}\right)^2\Big/4$$

— оценка состоятельна для любого ρ (генеральные моменты заменены на выборочные).

«Модифицированный» критерий:

$$\widehat{\sigma}^2_{f,0} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})^2(y_i-\overline{y})^2}{n\overline{s}^2_n(x)\overline{s}^2_n(y)}$$

— оценка состоятельна при $\rho = 0$.

Утверждение

Оба критерия асимптотически точны и состоятельны при $n \to \infty$.

Предположим, что $\rho = 0$. Рассмотрим предельную дисперсию

$$\sigma^2_{f,0} = \mathbb{D}\left(x^* y^*\right),$$

где $x^* = (x - \mathbb{E}x)/\sigma_x$, $y^* = (y - \mathbb{E}y)/\sigma_y$.

Утверждение

Классический критерий, примененный к негауссовской выборке, будет асимптотически точным и состоятельным тогда и только тогда, когда $\sigma^2_{f,0} = 1$, что эквивалентно равенству $\rho(x^2, y^2) = 0$.

Следствие

Если случайные величины x и y независимы, то $\sigma^2_{f,0} = 1$.

Ошибки первого рода, $\alpha = 0.05$. Независимые x, y.

$\rho = 0$, $\sigma^2_{f,0} = 1$: классический критерий асимптотически точен.

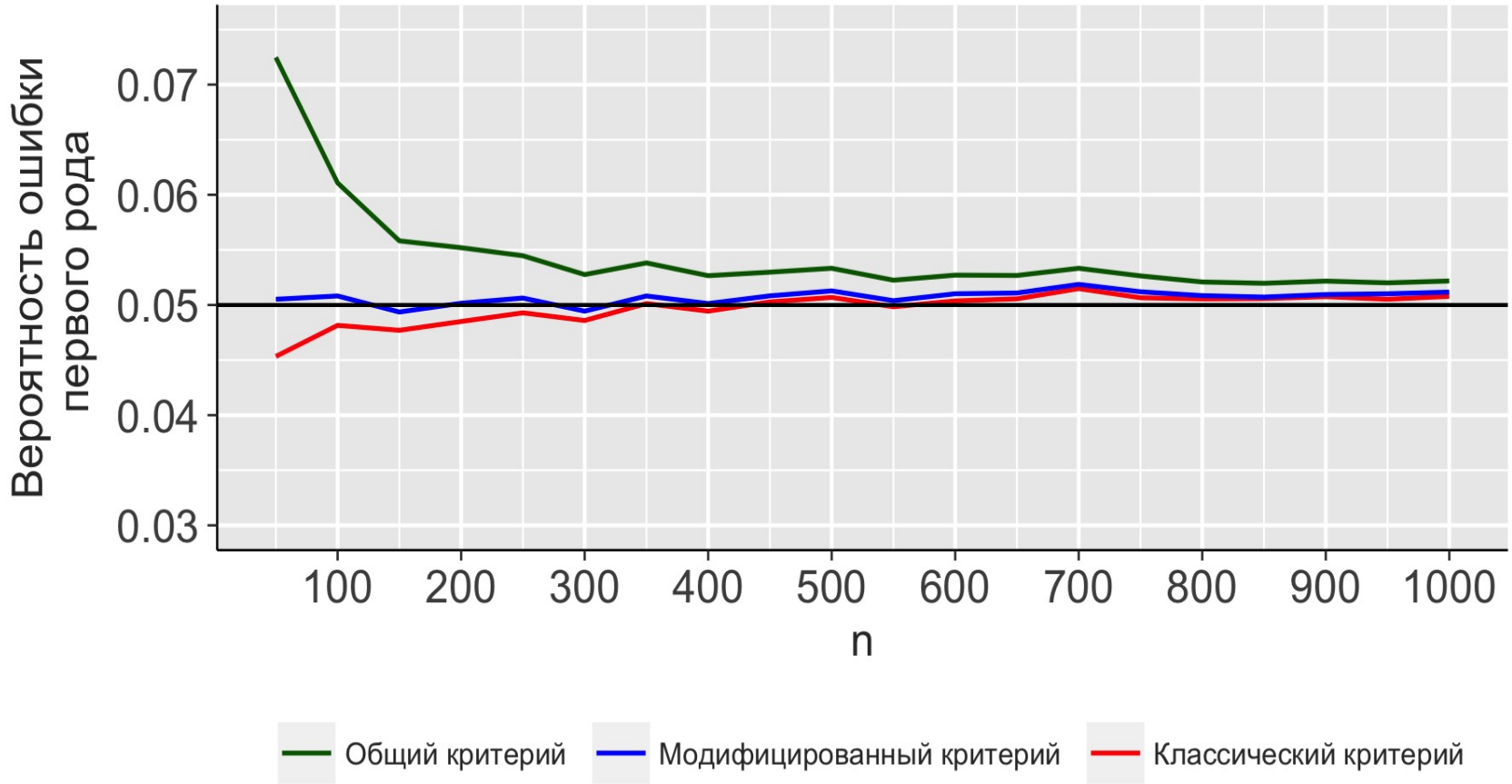

Рис.: Равномерно распределенные независимые случайные величины. Оценки вероятностей ошибок первого рода для трех критериев при $\alpha = 0.05$, $N = 10^5$

Ошибки первого рода, $\alpha = 0.05$. Равномерное распределение в круге

$\rho = 0$, $\sigma^2_{f,0} = 2/3$: классический критерий асимптотически консервативен.

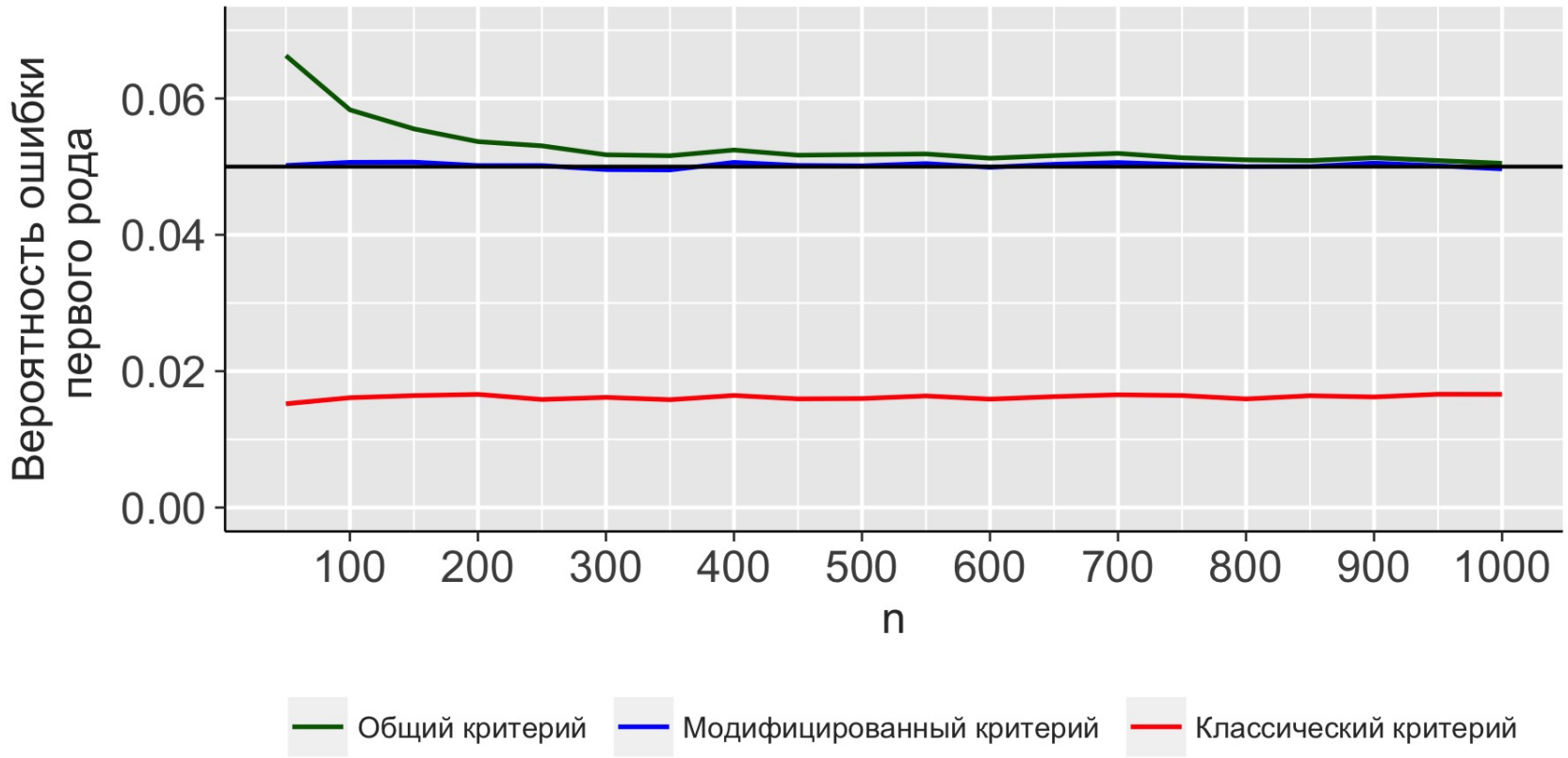

Рис.: Равномерное распределение в круге. Оценки вероятностей ошибок первого рода для трех критериев при $\alpha = 0.05$, $N = 10^5$

Ошибки первого рода, $\alpha = 0.05$. Круговая симметрия с $r^2 \in \mathrm{G}(1/2, 1)$.

$\rho = 0$, $\sigma^2_{f,0} = 1.5$: классический критерий асимптотически либерален.

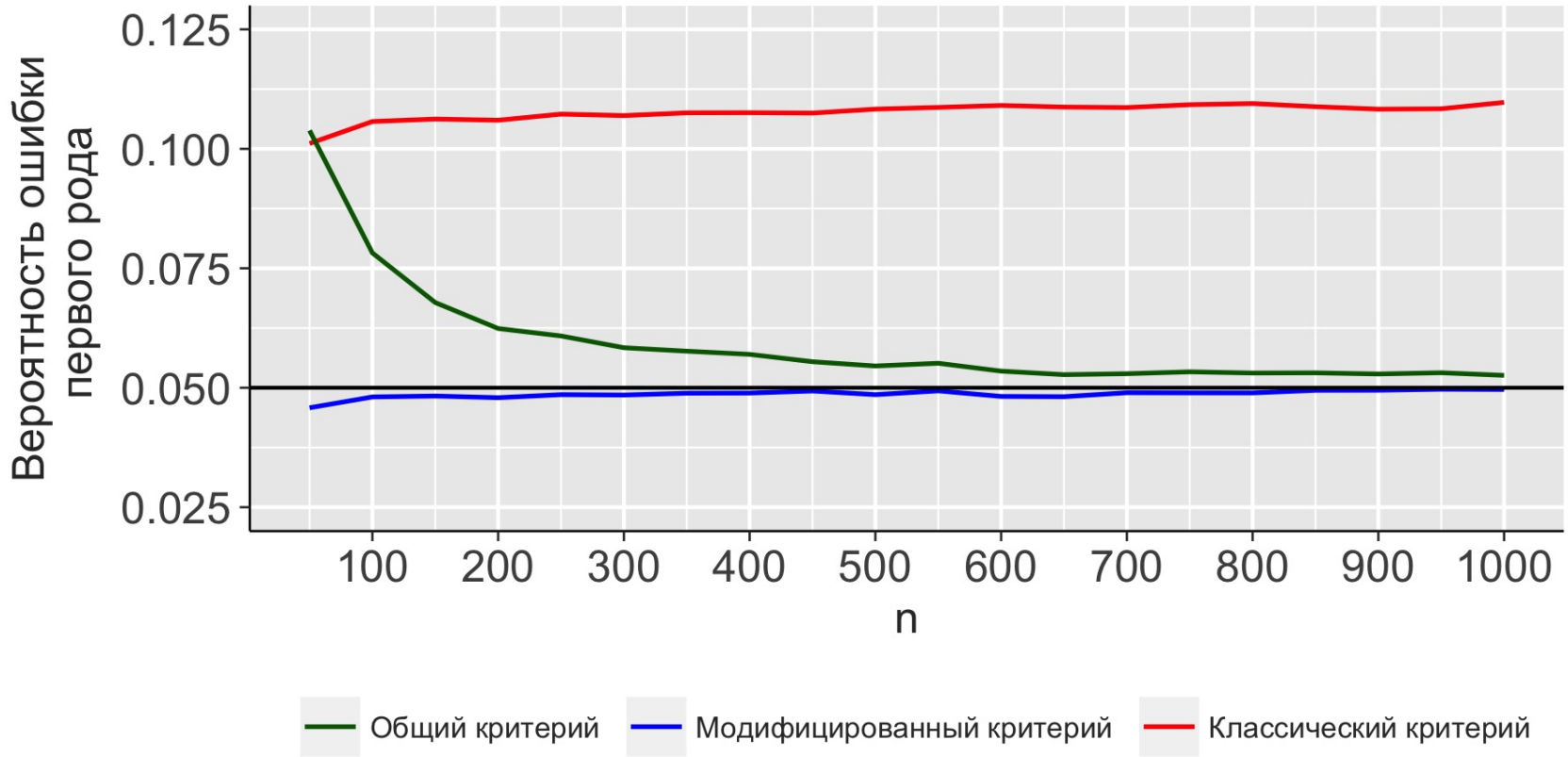

Рис.: Круговая симметрия с $r^2 \in \mathrm{G}(1/2, 1)$. Оценки вероятностей ошибок первого рода для трех критериев при $\alpha = 0.05$, $N = 10^5$

Ошибки первого рода, $\alpha = 0.05$. Пример зависимости с $\sigma^2_{f,0} = 1$

Равномерная смесь равномерных распределений на трех окружностях с радиусами 1, $\sqrt{2}$ *и* $\sqrt{6+\sqrt{39}}$.

Зависимость, $\rho = 0$, $\sigma^2_{f,0} = 1$: классический критерий асимптот. точен.

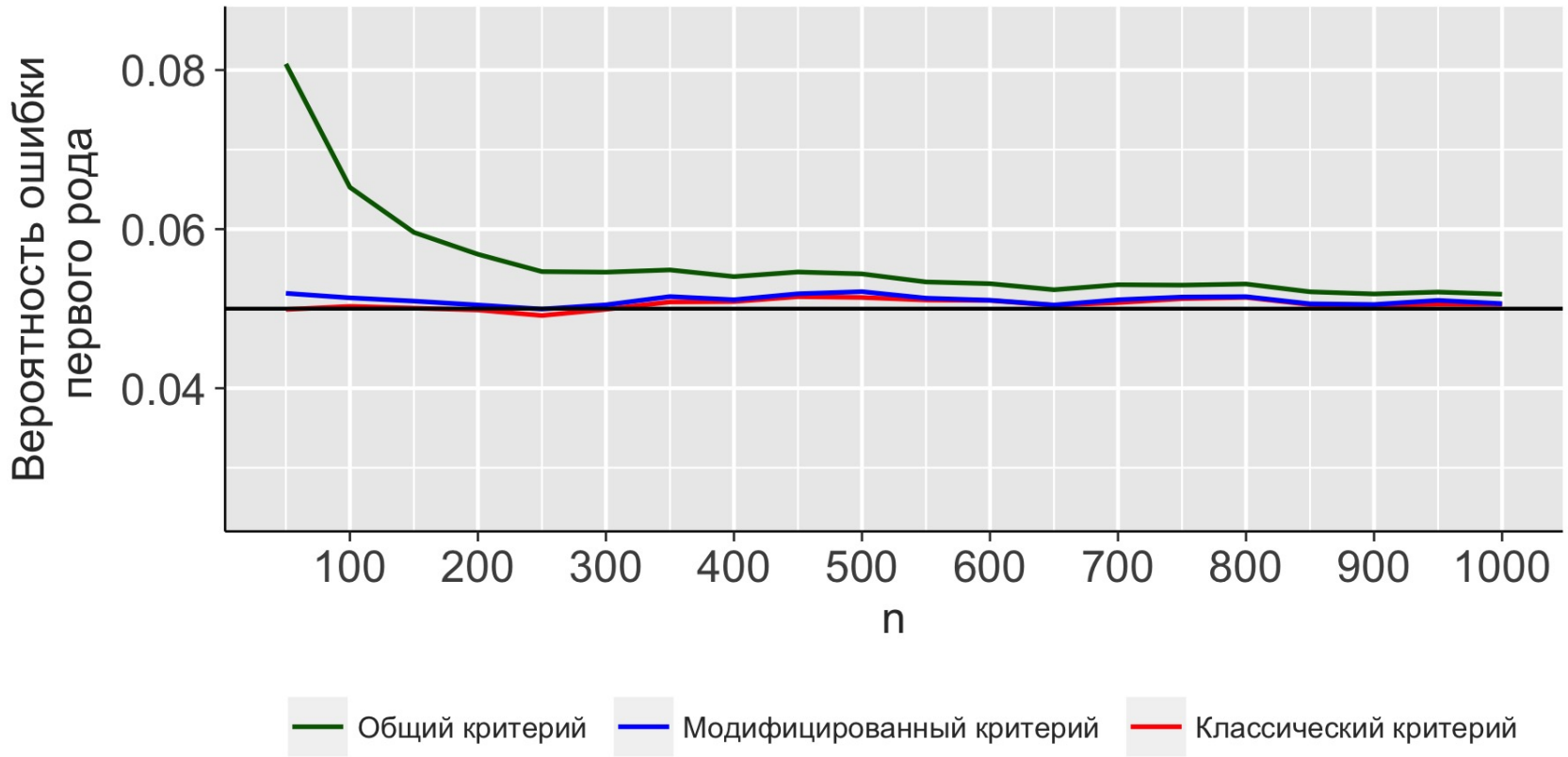

Рис.: Смесь равномерных распределений на трех окружностях. Оценки вероятностей ошибок первого рода для трех критериев при $\alpha = 0.05$; $N = 10^5$.

Частный коэффициент корреляции: общая постановка задачи

Случайные величины x, y, z с конечными вторыми моментами, $|\rho_{x,z}| < 1$, $|\rho_{y,z}| < 1$.

Линейные регрессии x на z и y на z с остатками ϵ_1 и ϵ_2.

$$\rho_{x,y|z} = \rho(\epsilon_1, \epsilon_2) = \frac{\rho_{x,y} - \rho_{x,z}\rho_{y,z}}{\sqrt{1-\rho_{x,z}^2}\sqrt{1-\rho_{y,z}^2}} :$$

частный коэффициент корреляции между x, y при исключении влияния z.

(X_n, Y_n, Z_n) – трехмерная повторная выборка объема n.

Нулевая гипотеза

$$H_0 : \{\rho_{x,y|z} = 0\}.$$

Частный коэффициент корреляции: критерии для проверки нулевой гипотезы

(x, y, z) — невырожденное гауссовское распределение: классический критерий (например, [ван дер Варден, 1960]).

Те же вопросы:

1. Когда классический критерий применим, если распределение не гауссовское?
2. Построение критериев, применимых в общей ситуации.

Ответ

Так как частный коэффициент сводится к коэффициенту корреляции Пирсона между остатками ϵ_1 и ϵ_2, то все результаты про коэффициент корреляции Пирсона переносятся на частный коэффициент корреляции.

$(\mathbf{X}_n, \mathbf{Y}_n, \mathbf{Z}_n)$ – независимая повторная выборка из распределения вектора (x, y, z) с непрерывными распределениями x, y, z и конечными четвертыми моментами. Пусть x^*, y^* и z^* — стандартизированные (x, y, z) и $\rho^2_{x,z} \neq 1$, $\rho^2_{y,z} \neq 1$.

$$\sigma^2_{xy|z} = \frac{1}{(1-\rho^2_{x,z})(1-\rho^2_{y,z})}\mathbb{D}\Big(x^*y^* + z^{*2}\rho_{x,z}\rho_{y,z} - x^*z^*\rho_{y,z} - y^*z^*\rho_{x,z}\Big),$$

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

Утверждение

Если $\rho_{x,y|z} = 0$, тогда

$$\mathcal{L}\big(\sqrt{n}\widehat{\rho}_{x,y|z}\big) \Rightarrow \mathrm{N}(0, \sigma^2_{x,y|z}).$$

Замечание

Если вектор (x, y, z) имеет гауссовское распределение, то $\sigma^2_{xy|z} = 1$.

Оценка $\sigma^2_{x,y|z}$ при $\rho_{xy|z} = 0$.

Пусть $\widehat{\rho}_{x,z}$ и $\widehat{\rho}_{y,z}$ — выборочные коэффициенты корреляции между x, z и y, z. Кроме того, положим

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

Тогда при $\rho_{xy|z} = 0$

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

— состоятельная оценка $\sigma^2_{x,y|z}$.

Утверждение

Критерий, порожденный оценкой $\widehat{\sigma}^2_{x,y\,|\,z}$, является асимптотически точным и состоятельным при $n \to \infty$.

Заключение

Таким образом, в работе были получены следующие результаты.

1. Для коэффициента корреляции Пирсона
   - Предложены два асимптотического критерия для проверки гипотезы равенства нулю коэффициента корреляции Пирсона, доказаны их асимптотические точность и состоятельность;
   - С помощью вычислительных экспериментов проведено сравнение этих критериев;
   - Проанализированы свойства классического (гауссовского) критерия при его применении к негауссовской выборке.
2. Для частного коэффициента корреляции
   - Все теоретические результаты работы, относящиеся к коэффициенту корреляции Пирсона, перенесены на частный коэффициент корреляции.