ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ОБРАЗОВАНИЯ
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет информатики, математики и компьютерных наук

**Программа подготовки бакалавров по направлению
01.03.02 Прикладная математика и информатика**

Антонов Илья Витальевич

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

Проверка условной независимости в трехмерном распределении Бернулли

Рецензент
д.ф.-м.н., проф.
ФИО

Научный руководитель
д.ф.-м.н., проф.
П. А. Колданов

Нижний Новгород, 2024

Содержание

Введение

Обзор по теме исследования и актуальность В современных задачах биоинформатики, информационного поиска, обработки речи и изображений возникает необходимость изучения взаимосвязей между большим количеством случайных величин. Методологию для решения такой проблемы предоставляют графические модели. Графической моделью [6] называется семейство вероятностных распределений, определенное в терминах ориентированного или неориентированного графа. В таком графе вершины соответствуют случайным величинам, а ребра отображают некоторые условные зависимости между случайными величинами. Основным назначением графических моделей является создание аппарата, упрощающего вычисление совместного распределения, маргинальных и условных вероятностей.

Особым образом в литературе выделяют графические модели с попарным марковским свойством (pairwise Markov property) [7]. В таких графических моделях отсутствие ребра между парой вершин обозначает условную независимость пары случайных величин при условии остальных случайных величин. Наиболее изученной графической моделью с таким свойством является гауссовская графическая модель [4], в которой рассматриваемые случайные величины имеют многомерное нормальное распределение. Процедуры идентификации гауссовской графической модели по наблюдениям, основанные на проверке равенства нулю частного коэффициента корреляции Пирсона, приводятся в работах [4; 5]. Однако, эти процедуры оказываются неустойчивыми к отклонению от многомерного нормального распределения. В частности, при таком отклонении тесты проверки индивидуальных гипотез перестают контролировать уровень значимости. Кроме того, многомерное нормальное распределение не всегда является адекватной моделью для описания реальных данных. Поэтому вопрос построения устойчивой графической модели встает остро.

Естественным обобщением многомерного нормального распределения является эллиптическое распределение [1]. Стоит отметить, что класс многомерных нормальных распределений содержится в классе эллиптических распределений. Одним из параметров эллиптического распределения выступа-

ет S – матрица формы (shape matrix), которая при существовании вторых моментов компонент эллиптического случайного вектора пропорциональна ковариационной матрице. В работе [10] для построения эллиптической графической модели, как устойчивого аналога гауссовской графической модели, вводится K – матрица обобщенных частных корреляций – как функция от матрицы S. Ноль в матрице обобщенных частных корреляций обозначает условную некоррелированность пары случайных величин при условии остальных случайных величин. То есть в отличии от гауссовской графической модели, в эллиптической графической модели отсутствие ребра обозначает не условную независимость, а условную некоррелированность. В работе [10] рассматривают $\hat{K}$ – оценку матрицы K – как функцию от $\hat{S}$ – оценки матрицы S. Для $\hat{S}$ вводятся некоторые предположения, при выполнении которых вектор $Vec(\hat{K})$, составленный из элементов $\hat{K}$, сходится по распределению к многомерному нормальному распределению с некоторыми параметрами. Затем приводятся примеры таких оценок $\hat{S}$, для которых эти предположения выполнены при условии существовании некоторых моментов нормы разности эллиптического случайного вектора и его вектора средних. Таким образом, наличие оценок для K и их асимптотических распределений приводит к статистической теории идентификации эллиптической графической модели.

Другое направление построения устойчивой графической модели можно связать с переходом от рассмотрения случайных величин X_i к рассмотрению индикаторных случайных величин $I_i = I(a_i < X_i < b_i)$. Условные зависимости между такими случайными величинами естественно приводят к графической модели. Совместное распределение индикаторных случайных величин описывается многомерным распределением Бернулли [3; 9]. Настоящая выпускная квалификационная работа посвящена проверке условной независимости в трехмерном распределении Бернулли. В частности, эта теория может быть использована для задачи идентификации графической модели с попарным марковским свойством в трехмерном распределении Бернулли. Предполагается, что рассмотрение трехмерного случая позволит понять перспективность исследований общего многомерного случая.

Постановка задачи Задачей выпускной квалификационной работы является построение тестов φ уровня α проверки гипотезы $H : X \perp\!\!\!\perp Y \mid Z$

по повторной выборке

$$\begin{pmatrix} X_1 \\ Y_1 \\ Z_1 \end{pmatrix}, \begin{pmatrix} X_2 \\ Y_2 \\ Z_2 \end{pmatrix}, \ldots, \begin{pmatrix} X_n \\ Y_n \\ Z_n \end{pmatrix}$$

из распределения $(X, Y, Z)^T$ с трехмерным распределением Бернулли, а также сравнение этих тестов при отклонении от условной независимости.

1 Теоретическая часть

1.1 Условная независимость в трехмерном распределение Бернулли

Определим трехмерное распределение Бернулли [3; 9].

Определение 1.1.1. *Случайный вектор* $(X, Y, Z)^T$ *имеет трехмерное распределение Бернулли, если множество его возможных значений:*

$$\begin{pmatrix}0\\0\\0\end{pmatrix}, \begin{pmatrix}0\\0\\1\end{pmatrix}, \begin{pmatrix}0\\1\\0\end{pmatrix}, \ldots, \begin{pmatrix}1\\1\\1\end{pmatrix}$$

и заданы $P(X = x, Y = y, Z = z) = p_{xyz} \geq 0, \sum_{x=0}^{1} \sum_{y=0}^{1} \sum_{z=0}^{1} p_{xyz} = 1.$

Приведем определение понятия условной независимости [7].

Определение 1.1.2. *Пусть* $(X, Y, Z)^T$ *– дискретный случайный вектор. Говорят, что случайные величины* X *и* Y *условно независимы при условии* Z*, и пишут* $X \perp\!\!\!\perp Y \mid Z$*, если:*

$$P(X = x, Y = y \mid Z = z) = P(X = x \mid Z = z)P(Y = y \mid Z = z)$$

при любых x, y *и* z *для которого* $P(Z = z) > 0$*.*

Сформулируем критерий условной независимости в трехмерном распределении Бернулли.

Теорема 1.1.1. *Пусть* $(X, Y, Z)^T$ *– случайный вектор, имеющий трехмерное распределение Бернулли, в котором* $P(Z = 0) > 0$*. Случайные величины* X *и* Y *условно независимы при условии* Z *тогда и только тогда, когда*

$$p_{00z}p_{11z} = p_{01z}p_{10z}$$

для всех $z \in \{0, 1\}$*.*

Доказательство. Пусть $X \perp\!\!\!\perp Y \mid Z$. Значит, для любых $x \in \{0, 1\}$, $y \in \{0, 1\}$ и $z \in \{0, 1\}$ выполнено условие:

$$P(X = x, Y = y \mid Z = z) = P(X = x \mid Z = z)P(Y = y \mid Z = z) \qquad (1.1.1)$$

После домножения (1.1.1) на $P(Z=z)^2 > 0$ получаем эквивалентное условие:

$$P(X=x, Y=y, Z=z)P(Z=z) = P(X=x, Z=z)P(Y=y, Z=z) \tag{1.1.2}$$

Найдем следующие вероятности:

$$P(X=x, Z=z) = p_{x0z} + p_{x1z},\ P(Y=y, Z=z) = p_{0yz} + p_{1yz}$$

$$P(Z=z) = p_{00z} + p_{01z} + p_{10z} + p_{11z}$$

Тогда условие (1.1.2) перепишем в следующем виде:

$$p_{xyz}(p_{00z} + p_{01z} + p_{10z} + p_{11z}) = (p_{x0z} + p_{x1z})(p_{0yz} + p_{1yz})$$

Это условие выполняется для всех $x \in \{0,1\}$, $y \in \{0,1\}$, $z \in \{0,1\}$. Пусть z фиксировано. Если $x=0$ и $y=0$, то:

$$p_{00z}(p_{00z} + p_{01z} + p_{10z} + p_{11z}) = (p_{00z} + p_{01z})(p_{00z} + p_{10z}) \Leftrightarrow p_{00z}p_{11z} = p_{01z}p_{10z}$$

Если $x=0$ и $y=1$, то:

$$p_{01z}(p_{00z} + p_{01z} + p_{10z} + p_{11z}) = (p_{00z} + p_{01z})(p_{01z} + p_{11z}) \Leftrightarrow p_{00z}p_{11z} = p_{01z}p_{10z}$$

Если $x=1$ и $y=0$, то:

$$p_{10z}(p_{00z} + p_{01z} + p_{10z} + p_{11z}) = (p_{10z} + p_{11z})(p_{00z} + p_{10z}) \Leftrightarrow p_{00z}p_{11z} = p_{01z}p_{10z}$$

Если $x=1$ и $y=1$, то:

$$p_{11z}(p_{00z} + p_{01z} + p_{10z} + p_{11z}) = (p_{10z} + p_{11z})(p_{01z} + p_{11z}) \Leftrightarrow p_{00z}p_{11z} = p_{01z}p_{10z}$$

Таким образом, из $X \perp\!\!\!\perp Y \mid Z$ следует $p_{00z}p_{11z} = p_{01z}p_{10z}$ для всех $z \in \{0,1\}$.

Поскольку в вышеприведенных рассуждениях все переходы равносильные, мы также доказали, что из из условия $p_{00z}p_{11z} = p_{01z}p_{10z}$ для всех $z \in \{0,1\}$ следует $X \perp\!\!\!\perp Y \mid Z$. □

Покажем, что существует случайный вектор $(X,Y,Z)^T$ с трехмерным распределением Бернулли, в котором $X \perp\!\!\!\perp Y \mid Z$.

Пример 1.1.1. *Пусть* $(X, Y, Z)^T$ *имеет трехмерное распределение Бернулли с вероятностями* $p_{000} = 0.15$, $p_{001} = 0.1$, $p_{010} = 0.3$, $p_{011} = 0.1$, $p_{100} = 0.05$, $p_{101} = 0.1$, $p_{110} = 0.1$, $p_{111} = 0.1$. *Заметим, что:*

$$p_{000}p_{110} = p_{010}p_{100} = 0.015$$

$$p_{001}p_{111} = p_{011}p_{101} = 0.01$$

Следовательно из теор. 1.1.1 следует, что $X \perp\!\!\!\perp Y \mid Z$.

1.2 Частный коэффициент корреляции Пирсона в трехмерном распределении Бернулли

В данном разделе исследуем свойства частного коэффициента корреляции Пирсона в трехмерном распределении Бернулли.

Для случайного вектора $(X, Y, Z)^T$ определим ковариационную матрицу:

$$\Sigma = \begin{pmatrix} \sigma_{XX} & \sigma_{XY} & \sigma_{XZ} \\ \sigma_{YX} & \sigma_{YY} & \sigma_{YZ} \\ \sigma_{ZX} & \sigma_{ZY} & \sigma_{ZZ} \end{pmatrix}$$

где $\sigma_{XY} = E[(X - EX)(Y - EY)]$. Остатками от X и Y при регрессии на Z называются случайные величины:

$$X' = (X - EX) - \frac{\sigma_{XZ}}{\sigma_{ZZ}}(Z - EZ)$$

$$Y' = (Y - EY) - \frac{\sigma_{YZ}}{\sigma_{ZZ}}(Z - EZ)$$

Согласно работе [2], частный коэффициент корреляции Пирсона определяется как коэффициент корреляции Пирсона между остатками, другими словами:

$$\rho^{XY \cdot Z} = \frac{E(X'Y')}{\sqrt{E(X'^2)E(Y'^2)}}$$

Приведем соотношения, которые справедливы для $\rho^{XY \cdot Z}$ в произвольном распределении.

Лемма 1.2.1. *Пусть* $(X, Y, Z)^T$ *– произвольный случайный вектор. Тогда:*

$$\rho^{XY \cdot Z} = \frac{\sigma_{XY}\sigma_{ZZ} - \sigma_{XZ}\sigma_{YZ}}{\sqrt{\sigma_{XX}\sigma_{ZZ} - \sigma_{XZ}^2}\sqrt{\sigma_{YY}\sigma_{ZZ} - \sigma_{YZ}^2}} = \frac{\rho_{XY} - \rho_{XZ}\rho_{YZ}}{\sqrt{1 - \rho_{XZ}^2}\sqrt{1 - \rho_{YZ}^2}}$$

где $\rho_{XY},\ \rho_{XZ},\ \rho_{YZ}$ *– коэффициенты корреляции Пирсона между случайными величинами* X *и* Y*,* X *и* Z*,* Y *и* Z *соответственно.*

Доказательство.

$$E(X'Y') = E[((X - EX) - \frac{\sigma_{XZ}}{\sigma_{ZZ}}(Z - EZ))((Y - EY) - \frac{\sigma_{YZ}}{\sigma_{ZZ}}(Z - EZ))] =$$

$$= \sigma_{XY} - \frac{\sigma_{YZ}}{\sigma_{ZZ}}\sigma_{XZ} - \frac{\sigma_{XZ}}{\sigma_{ZZ}}\sigma_{YZ} + \frac{\sigma_{XZ}\sigma_{YZ}}{\sigma_{ZZ}\sigma_{ZZ}}\sigma_{ZZ} = \frac{\sigma_{XY}\sigma_{ZZ} - \sigma_{XZ}\sigma_{YZ}}{\sigma_{ZZ}}$$

Тогда:

$$\rho^{XY \cdot Z} = \frac{E(X'Y')}{\sqrt{E(X'^2)E(Y'^2)}} = \frac{\frac{\sigma_{XY}\sigma_{ZZ} - \sigma_{XZ}\sigma_{YZ}}{\sigma_{ZZ}}}{\sqrt{\frac{\sigma_{XX}\sigma_{ZZ} - \sigma_{XZ}\sigma_{XZ}}{\sigma_{ZZ}}}\sqrt{\frac{\sigma_{YY}\sigma_{ZZ} - \sigma_{YZ}\sigma_{YZ}}{\sigma_{ZZ}}}} =$$

$$= \frac{\sigma_{XY}\sigma_{ZZ} - \sigma_{XZ}\sigma_{YZ}}{\sqrt{\sigma_{XX}\sigma_{ZZ} - \sigma_{XZ}^2}\sqrt{\sigma_{YY}\sigma_{ZZ} - \sigma_{YZ}^2}} =$$

$$= \frac{\sigma_{ZZ}\sqrt{\sigma_{XX}\sigma_{YY}}\left(\frac{\sigma_{XY}}{\sqrt{\sigma_{XX}\sigma_{YY}}} - \frac{\sigma_{XZ}}{\sqrt{\sigma_{XX}\sigma_{ZZ}}}\frac{\sigma_{YZ}}{\sqrt{\sigma_{YY}\sigma_{ZZ}}}\right)}{\sqrt{\sigma_{XX}\sigma_{ZZ} - \sigma_{XZ}^2}\sqrt{\sigma_{YY}\sigma_{ZZ} - \sigma_{YZ}^2}} =$$

$$= \frac{\rho_{XY} - \rho_{XZ}\rho_{YZ}}{\sqrt{1 - \frac{\sigma_{XZ}^2}{\sigma_{XX}\sigma_{ZZ}}}\sqrt{1 - \frac{\sigma_{YZ}^2}{\sigma_{YY}\sigma_{ZZ}}}} = \frac{\rho_{XY} - \rho_{XZ}\rho_{YZ}}{\sqrt{1 - \rho_{XZ}^2}\sqrt{1 - \rho_{YZ}^2}}$$

□

Для дальнейших рассуждений используем следующие обозначения:

$$p_{x**} = P(X = x),\ p_{*y*} = P(Y = y),\ p_{**z} = P(Z = z)$$

$$p_{xy*} = P(X = x, Y = y),\ p_{x*z} = P(X = x, Z = z),\ p_{*yz} = P(Y = y, Z = z)$$

Найдем значение выражения $\sigma_{XY}\sigma_{ZZ} - \sigma_{XZ}\sigma_{YZ}$ в трехмерном распределении Бернулли.

Лемма 1.2.2. *Пусть* $(X, Y, Z)^T$ *– случайный вектор, имеющий трехмерное распределение Бернулли. Тогда:*

$$\sigma_{XY}\sigma_{ZZ} - \sigma_{XZ}\sigma_{YZ} = p_{**0}(p_{001}p_{111} - p_{011}p_{101}) + p_{**1}(p_{000}p_{110} - p_{010}p_{100})$$

Доказательство. Легко проверить, что $\sigma_{ZZ} = p_{**1}(1 - p_{**1})$. Найдем соотношение для σ_{XY}. Воспользуемся формулой $\sigma_{XY} = E(XY) - E(X)E(Y)$.

$$E(XY) = 1 \cdot p_{11*} + 0 \cdot (p_{00*} + p_{01*} + p_{10*}) = p_{11*}$$

Таким образом, $\sigma_{XY} = p_{11*} - p_{1**}p_{*1*}$. Аналогично, $\sigma_{XZ} = p_{1*1} - p_{1**}p_{**1}$ и $\sigma_{YZ} = p_{*11} - p_{*1*}p_{**1}$. Преобразуем выражение $\sigma_{XY}\sigma_{ZZ} - \sigma_{XZ}\sigma_{YZ} =$

$$= (p_{11*} - p_{1**}p_{*1*})p_{**1}(1 - p_{**1}) - (p_{1*1} - p_{1**}p_{**1})(p_{*11} - p_{*1*}p_{**1}) =$$

$$= p_{11*}p_{**1} - p_{11*}p_{**1}p_{**1} - p_{1**}p_{*1*}p_{**1} - p_{1*1}p_{*11} + p_{1*1}p_{*1*}p_{**1} + p_{1**}p_{**1}p_{*11} =$$

$$= (p_{111}p_{**1} + p_{110}p_{**1}) - p_{11*}p_{**1}p_{**1} - p_{1**}p_{*1*}p_{**1} -$$

$$-p_{1*1}p_{*11} + p_{1*1}p_{*1*}p_{**1} + p_{1**}p_{**1}p_{*11} =$$

$$= (p_{111}p_{**1} - p_{1*1}p_{*11}) + p_{**1}(p_{110} - p_{11*}p_{**1} - p_{1**}p_{*1*} + p_{1*1}p_{*1*} + p_{1**}p_{*11}) \quad (1.2.1)$$

Заметим, что:

1. $p_{110} - p_{11*}p_{**1} = p_{110} - p_{110}p_{**1} - p_{111}p_{**1} = p_{110}(1 - p_{**1}) - p_{111}p_{**1} =$
   $= p_{110}p_{**0} - p_{111}p_{**1}$

2. $-p_{1**}p_{*1*} + p_{1*1}p_{*1*} + p_{1**}p_{*11} = -(p_{1*0} + p_{1*1})(p_{*10} + p_{*11}) + p_{1*1}(p_{*10} + p_{*11}) +$
   $+(p_{1*0} + p_{1*1})p_{*11} = -p_{1*0}p_{*10} + p_{1*1}p_{*11}$

Учитывая вышеприведенные соотношения, запишем (1.2.1):

$$(p_{111}p_{**1} - p_{1*1}p_{*11}) + p_{**1}((p_{110}p_{**0} - p_{1*0}p_{*10}) - (p_{111}p_{**1} - p_{1*1}p_{*11})) =$$

$$= (1 - p_{**1})(p_{111}p_{**1} - p_{1*1}p_{*11}) + p_{**1}(p_{110}p_{**0} - p_{1*0}p_{*10}) =$$

$$= p_{**0}(p_{111}p_{**1} - p_{1*1}p_{*11}) + p_{**1}(p_{110}p_{**0} - p_{1*0}p_{*10}) \quad (1.2.2)$$

Также заметим, что:
$p_{11z}p_{**z} - p_{1*z}p_{*1z} = p_{11z}(p_{00z} + p_{01z} + p_{10z} + p_{11z}) - (p_{10z} + p_{11z})(p_{01z} + p_{11z}) =$
$= p_{00z}p_{11z} - p_{01z}p_{10z}$. Тогда в (1.2.2) имеем:

$$\sigma_{XY}\sigma_{ZZ} - \sigma_{XZ}\sigma_{YZ} = p_{**0}(p_{001}p_{111} - p_{011}p_{101}) + p_{**1}(p_{000}p_{110} - p_{010}p_{100})$$

□

Вышеприведенное соотношение для $\sigma_{XY}\sigma_{ZZ} - \sigma_{XZ}\sigma_{YZ}$ позволяет доказать следующую теорему.

Теорема 1.2.1. *Пусть $(X, Y, Z)^T$ – случайный вектор, имеющий трехмерное распределение Бернулли. Если $X \perp\!\!\!\perp Y \mid Z$, то $\rho^{XY \cdot Z} = 0$.*

Доказательство. Пусть $X \perp\!\!\!\perp Y \mid Z$. Тогда по теор. 1.1.1: $p_{000}p_{110} = p_{010}p_{100}$ и $p_{001}p_{111} = p_{011}p_{101}$. Используя лемм. 1.2.2, имеем:

$$\sigma_{XY}\sigma_{ZZ} - \sigma_{XZ}\sigma_{YZ} = p_{**0}(p_{001}p_{111} - p_{011}p_{101}) + p_{**1}(p_{000}p_{110} - p_{010}p_{100}) = 0$$

Следовательно, $\rho^{XY \cdot Z} = 0$. □

Таким образом, равенство нулю частного коэффициента корреляции Пирсона является необходимым условием условной независимости. Однако, это условие не является достаточным, так как в обратную сторону теор. 1.2.1 неверна. Приведем контрпример.

Пример 1.2.1. *Пусть $p_{000} = 0.15$, $p_{001} = 0.1$, $p_{010} = 0.1$, $p_{011} = 0.15$, $p_{100} = 0.1$, $p_{101} = 0.15$, $p_{110} = 0.15$, $p_{111} = 0.1$. Тогда $p_{**0} = 0.5$, $p_{**1} = 0.5$ и*

$$\sigma_{XY}\sigma_{ZZ} - \sigma_{XZ}\sigma_{YZ} = p_{**1}(p_{000}p_{110} - p_{010}p_{100}) + p_{**0}(p_{001}p_{111} - p_{011}p_{101}) =$$

$$= 0.5 \cdot (0.15 \cdot 0.15 - 0.1 \cdot 0.1) + 0.5 \cdot (0.1 \cdot 0.1 - 0.15 \cdot 0.15) = 0$$

Следовательно $\rho^{XY \cdot Z} = 0$. Однако, случайные величины X и Y условно зависимы при условии Z поскольку:

$$p_{000}p_{110} - p_{010}p_{100} = 0.15 \cdot 0.15 - 0.1 \cdot 0.1 = 0.0125 \neq 0$$

$$p_{001}p_{111} - p_{011}p_{101} = 0.1 \cdot 0.1 - 0.15 \cdot 0.15 = -0.0125 \neq 0$$

Для оценки частного коэффициента корреляции Пирсона можно использовать выборочный частный коэффициент корреляции Пирсона:

$$r^{XY \cdot Z} = \frac{r_{XY} - r_{XZ}r_{YZ}}{\sqrt{1 - r_{XY}^2}\sqrt{1 - r_{YZ}^2}}$$

где r_{XY}, r_{XZ}, r_{YZ} – выборочные коэффициенты корреляции Пирсона между случайными величинами X и Y, X и Z, Y и Z соответственно. Известно

[1], что в трехмерном нормальном распределении при истинности гипотезы $H^{\text{Partial}} : \rho^{XY \cdot Z} = 0$ статистика:

$$T^{\text{Partial}} = \sqrt{n-3}\frac{r^{XY \cdot Z}}{\sqrt{1-(r^{XY \cdot Z})^2}}$$

имеет распределение Стьюдента с $n-3$ степенями свободы, где n – количество наблюдений. Тогда тест уровня α проверки гипотезы $H^{\text{Partial}} : \rho^{XY \cdot Z} = 0$ против альтернативы $K^{\text{Partial}} : \rho^{XY \cdot Z} \neq 0$ имеет вид:

$$\varphi^{\text{Partial}}(t^{\text{Partial}}) = \begin{cases} 1, & |t^{\text{Partial}}| > C \\ 0, & |t^{\text{Partial}}| \leq C \end{cases}$$

где константа C удовлетворяет уравнению $P_{H^{\text{Partial}}}(T^{\text{Partial}} > C) = 1 - \alpha/2$. В разделе 2.2 с помощью численных экспериментов будет проверено контролирует ли тест φ^{Partial} уровень значимости для гипотезы $H^{\text{Partial}} : \rho^{XY \cdot Z} = 0$ в трехмерном распределении Бернулли.

1.3 РНМН тест проверки необходимого условия условной независимости

Запишем вид трехмерного распределения Бернулли в экпоненциальной форме:

$$P(X = x, Y = y, Z = z) = p_{000}^{(1-x)(1-y)(1-z)} \dots p_{111}^{xyz} =$$

$$= \exp\Bigg\{ \ln(p_{000}) + \ln\left(\frac{p_{001}p_{111}p_{010}p_{100}}{p_{011}p_{101}p_{000}p_{110}}\right) xyz + \ln\left(\frac{p_{100}}{p_{000}}\right) x + \ln\left(\frac{p_{010}}{p_{000}}\right) y +$$

$$+ \ln\left(\frac{p_{001}}{p_{000}}\right) z + \ln\left(\frac{p_{000}p_{110}}{p_{010}p_{100}}\right) xy + \ln\left(\frac{p_{000}p_{101}}{p_{001}p_{100}}\right) xz + \ln\left(\frac{p_{000}p_{011}}{p_{001}p_{010}}\right) yz \Bigg\}$$

Среди параметров, стоящих при статистиках xyz, x, y, z, xy, xz, yz, выделим параметр, связанный с условной независимостью.

Теорема 1.3.1. *Пусть $(X, Y, Z)^T$ – случайный вектор, имеющий трехмерное распределение Бернулли, и $\theta = \ln\left(\frac{p_{001}p_{111}p_{010}p_{100}}{p_{011}p_{101}p_{000}p_{110}}\right)$. Если выполнено одно из условий:*

- $X \perp\!\!\!\perp Y \mid Z$

- $X \perp\!\!\!\perp Z \mid Y$
- $Y \perp\!\!\!\perp Z \mid X$

то параметр θ *принимает значение* 0.

Доказательство. Результаты теор. 1.1.1 можно обобщить следующим образом:

$$X \perp\!\!\!\perp Z \mid Y \Leftrightarrow p_{000}p_{101} = p_{001}p_{100} \text{ и } p_{010}p_{111} = p_{011}p_{110}$$

$$Y \perp\!\!\!\perp Z \mid X \Leftrightarrow p_{000}p_{011} = p_{001}p_{010} \text{ и } p_{100}p_{111} = p_{101}p_{110}$$

1. Пусть $X \perp\!\!\!\perp Y \mid Z$, тогда по теор. 1.1.1 выполнено: $p_{000}p_{110} = p_{010}p_{100}$ и $p_{001}p_{111} = p_{011}p_{101}$. Отсюда следует, что $\theta = \ln(1) = 0$.
2. Пусть $X \perp\!\!\!\perp Z \mid Y$, тогда из вышеприведенного $p_{000}p_{101} = p_{001}p_{100}$ и $p_{010}p_{111} = p_{011}p_{110}$. Отсюда следует, что $\theta = \ln(1) = 0$.
3. Пусть $Y \perp\!\!\!\perp Z \mid X$, тогда из вышеприведенного $p_{000}p_{011} = p_{001}p_{010}$ и $p_{100}p_{111} = p_{101}p_{110}$. Отсюда следует, что $\theta = \ln(1) = 0$.

□

Таким образом, нулевое значение параметра θ является необходимым условием наличия хотя бы одной условно независимой пары случайных величин в трехмерном распределении Бернулли. В частности, нулевое значение параметра θ является необходимым условием $X \perp\!\!\!\perp Y \mid Z$. Для проверки гипотезы $H^{\text{Theta}} : \theta = 0$ используем теорию РНМН тестов [8] в многопараметрическом экспоненциальном семействе. Пусть

$$\begin{pmatrix} X_1 \\ Y_1 \\ Z_1 \end{pmatrix}, \begin{pmatrix} X_2 \\ Y_2 \\ Z_2 \end{pmatrix}, \ldots, \begin{pmatrix} X_n \\ Y_n \\ Z_n \end{pmatrix}$$

повторная выборка из распределения случайного вектора $(X, Y, Z)^T$. Совместное распределение повторной выборки имеет вид:

$$P(X_1 = x_1, Y_1 = y_1, Z_1 = z_1, \ldots, X_n = x_n, Y_n = y_n, Z_n = z_n) =$$

$$= \prod_{i=1}^{n} P(X_i = x_i, Y_i = y_i, Z_i = z_i) =$$

$$= \exp\Bigg\{ \ln(p_{000})n + \ln\left(\frac{p_{001}p_{111}p_{010}p_{100}}{p_{011}p_{101}p_{000}p_{110}}\right)\sum_{i=1}^{n} x_i y_i z_i +$$

$$+ \ln\left(\frac{p_{100}}{p_{000}}\right)\sum_{i=1}^{n} x_i + \ln\left(\frac{p_{010}}{p_{000}}\right)\sum_{i=1}^{n} y_i + \ln\left(\frac{p_{001}}{p_{000}}\right)\sum_{i=1}^{n} z_i +$$

$$+ \ln\left(\frac{p_{000}p_{110}}{p_{010}p_{100}}\right)\sum_{i=1}^{n} x_i y_i + \ln\left(\frac{p_{000}p_{101}}{p_{001}p_{100}}\right)\sum_{i=1}^{n} x_i z_i + \ln\left(\frac{p_{000}p_{011}}{p_{001}p_{010}}\right)\sum_{i=1}^{n} y_i z_i \Bigg\}$$

Пусть

$$U = \sum_{i=1}^{n} X_i Y_i Z_i,\ T_1 = \sum_{i=1}^{n} X_i Y_i,\ T_2 = \sum_{i=1}^{n} X_i Z_i,$$

$$T_3 = \sum_{i=1}^{n} Y_i Z_i,\ T_4 = \sum_{i=1}^{n} X_i,\ T_5 = \sum_{i=1}^{n} Y_i,\ T_6 = \sum_{i=1}^{n} Z_i$$

Обозначим $T = (T_1, \ldots, T_6)$, $t = (t_1, \ldots, t_6)$. Согласно [8] РНМН тест уровня α проверки гипотезы $H^{\text{Theta}} : \theta = 0$ против альтернативы $K^{\text{Theta}} : \theta \neq 0$ имеет вид:

$$\varphi^{\text{Theta}}(u,t) = \begin{cases} 1,\ u < C_1(t) \text{ или } u > C_2(t) \\ \gamma_i,\ u = C_i(t),\ i = 1,2 \\ 0,\ C_1(t) < u < C_2(t) \end{cases}$$

где константы $C_i(t)$ и $\gamma_i(t)$ определяются из системы уравнений:

$$\begin{cases} E_{\theta=0}[\varphi^{\text{Theta}}(U,T) \mid T = t] = \alpha \\ E_{\theta=0}[U\varphi^{\text{Theta}}(U,T) \mid T = t] = \alpha E_{\theta=0}[U \mid T = t] \end{cases}$$

Приведем распределение статистики U при условии $T = t$.

Лемма 1.3.1. *Пусть* $k_1(u) = u$, $k_2(u) = t_1 - u$, $k_3(u) = t_2 - u$, $k_4(u) = t_3 - u$, $k_5(u) = t_4 - t_1 - t_2 + u$, $k_6(u) = t_5 - t_1 - t_3 + u$, $k_7(u) = t_6 - t_2 - t_3 + u$, $k_8(u) = n - u + t_1 + t_2 + t_3 - t_4 - t_5 - t_6$. *Тогда*

$$P_{\theta=0}(U = u \mid T = t) = \frac{(\prod_{i=1}^{8} k_i(u)!)^{-1}}{\sum_{s\in\mathcal{D}}(\prod_{i=1}^{8} k_i(s)!)^{-1}}$$

где $\mathcal{D} = \{s \in \mathbb{Z} : 0 \leq k_i(s) \leq n$ *для всех* $i = 1 \ldots, 8\}$.

Доказательство. Найдем совместное распределение статистик $(U, T_1, \ldots, T_6)$:

$$P(U=u, T=t) = P(U=u, T_1=t_1, T_2=t_2, T_3=t_3, T_4=t_4, T_5=t_5, T_6=t_6) =$$

$$= P\Bigg(\sum_{i=1}^{n} X_iY_iZ_i = u, \sum_{i=1}^{n} X_iY_i = t_1, \sum_{i=1}^{n} X_iZ_i = t_2, \sum_{i=1}^{n} Y_iZ_i = t_3,$$

$$\sum_{i=1}^{n} X_i = t_4, \sum_{i=1}^{n} Y_i = t_5, \sum_{i=1}^{n} Z_i = t_6\Bigg) =$$

$$= P\Bigg(\sum_{i=1}^{n} X_iY_iZ_i = u, \sum_{i=1}^{n} X_iY_i(1-Z_i) = t_1 - u, \sum_{i=1}^{n} X_i(1-Y_i)Z_i = t_2 - u,$$

$$\sum_{i=1}^{n}(1-X_i)Y_iZ_i = t_3 - u, \sum_{i=1}^{n} X_i(1-Y_i)(1-Z_i) = t_4 - t_1 - t_2 + u,$$

$$\sum_{i=1}^{n}(1-X_i)Y_i(1-Z_i) = t_5 - t_1 - t_3 + u, \sum_{i=1}^{n}(1-X_i)(1-Y_i)Z_i = t_6 - t_2 - t_3 + u,$$

$$\sum_{i=1}^{n}(1-X_i)(1-Y_i)(1-Z_i) = n - u + t_1 + t_2 + t_3 - t_4 - t_5 - t_6\Bigg) = \frac{n!}{\prod_{i=1}^{8} k_i(u)!} \times$$

$$\times p_{111}^{u} p_{110}^{t_1-u} p_{101}^{t_2-u} p_{011}^{t_3-u} p_{100}^{t_4-t_1-t_2+u} p_{010}^{t_5-t_1-t_3+u} p_{001}^{t_6-t_2-t_3+u} p_{000}^{n-u+t_1+t_2+t_3-t_4-t_5-t_6}$$

Тогда условное распределение статистики U при условии $T = t$ можно записать как:

$$P(U=u \mid T=t) = \frac{P(U=u, T=t)}{P(T=t)} = \frac{P(U=u, T=t)}{\sum_{s \in \mathcal{D}} P(U=s, T=t)} =$$

$$= \frac{(\prod_{i=1}^{8} k_i(u)!)^{-1} \left(\dfrac{p_{001}p_{010}p_{100}p_{111}}{p_{000}p_{011}p_{101}p_{110}}\right)^{u}}{\sum_{s \in \mathcal{D}} (\prod_{i=1}^{8} k_i(s)!)^{-1} \left(\dfrac{p_{001}p_{010}p_{100}p_{111}}{p_{000}p_{011}p_{101}p_{110}}\right)^{s}}$$

При истинности гипотезы $\theta = 0$ параметр $\dfrac{p_{001}p_{010}p_{100}p_{111}}{p_{000}p_{011}p_{101}p_{110}} = 1$. Следовательно:

$$P_{\theta=0}(U=u \mid T=t) = \frac{(\prod_{i=1}^{8} k_i(u)!)^{-1}}{\sum_{s \in \mathcal{D}} (\prod_{i=1}^{8} k_i(s)!)^{-1}}$$

□

Так как φ^{Theta} является РНМН тестом, и в частности несмещенным, уровня α проверки гипотезы $H^{\text{Theta}} : \theta = 0$, и из $X \perp\!\!\!\perp Y \mid Z$ следует $\theta = 0$, то φ^{Theta} также является несмещенным тестом уровня α проверки гипотезы $H : X \perp\!\!\!\perp Y \mid Z$.

1.4 РНМН тест проверки независимости в условном распределении

Пусть $(X, Y, Z)^T$ – случайный вектор, имеющий трехмерное распределение Бернулли. При фиксированном $Z = z$:

$$P(X = x, Y = y \mid Z = z) = \frac{p_{xyz}}{p_{00z} + p_{01z} + p_{10z} + p_{11z}} = q_{xy}$$

Приведем определение случайного вектора с двумерным распределением Бернулли [3].

Определение 1.4.1. *Случайный вектор $(X, Y)^T$ имеет двумерное распределение Бернулли, если множество его возможных значений:*

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

и заданы $P(X = x, Y = y) = q_{xy} \geq 0, \sum_{x=0}^{1} \sum_{y=0}^{1} q_{xy} = 1$.

Таким образом, $(X, Y)^T$ при условии $Z = z$ имеет двумерное распределении Бернулли. Изложим теорию проверки независимости в двумерном распределении Бернулли, которая в разд. 1.5 будет использована для проверки условной независимости.

В работе [3] показано, что:

$$P(X = x, Y = y) = \exp\left\{\ln(q_{00}) + \ln\left(\frac{q_{10}}{q_{00}}\right) x + \ln\left(\frac{q_{01}}{q_{00}}\right) y + \ln\left(\frac{q_{00}q_{11}}{q_{01}q_{10}}\right) xy\right\}$$

Также, в [3] доказано, что X и Y независимы тогда и только тогда, когда

$$\theta = \ln\left(\frac{q_{00}q_{11}}{q_{01}q_{10}}\right) = 0$$

Пусть

$$\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \begin{pmatrix} X_2 \\ Y_2 \end{pmatrix}, \ldots, \begin{pmatrix} X_n \\ Y_n \end{pmatrix}$$

повторная выборка из распределения случайного вектора $(X, Y)^T$. Тогда:

$$P(X_1 = x_1, Y_1 = y_1, \ldots, X_n = x_n, Y_n = y_n) = \prod_{i=1}^{n} P(X_i = x_i, Y_i = y_i) =$$

$$= \exp\left\{\ln(q_{00})n + \ln\left(\frac{q_{10}}{q_{00}}\right)\sum_{i=1}^{n} x_i + \ln\left(\frac{q_{01}}{q_{00}}\right)\sum_{i=1}^{n} y_i + \ln\left(\frac{q_{00}q_{11}}{q_{01}q_{10}}\right)\sum_{i=1}^{n} x_i y_i\right\}$$

Пусть

$$U = \sum_{i=1}^{n} X_i Y_i,\ T_1 = \sum_{i=1}^{n} X_i,\ T_2 = \sum_{i=1}^{n} Y_i$$

Обозначим $T = (T_1, T_2)$, $t = (t_1, t_2)$. Согласно [8] РНМН тест уровня α проверки гипотезы $H^{\text{Independence}} : \theta = 0$ против альтернативы $K^{\text{Independence}} : \theta \neq 0$ имеет вид:

$$\varphi^{\text{Independence}}(u,t) = \begin{cases} 1,\ u < C_1(t) \text{ или } u > C_2(t) \\ \gamma_i,\ u = C_i(t),\ i = 1,2 \\ 0,\ C_1(t) < u < C_2(t) \end{cases}$$

где константы $C_i(t)$ и $\gamma_i(t)$ определяются из системы уравнений:

$$\begin{cases} E_{\theta=0}[\varphi^{\text{Independence}}(U,T) \mid T = t] = \alpha \\ E_{\theta=0}[U\varphi^{\text{Independence}}(U,T) \mid T = t] = \alpha E_{\theta=0}[U \mid T = t] \end{cases}$$

Приведем распределение статистики U при условии $T = t$.

Лемма 1.4.1. *Пусть* $k_1(u) = u$, $k_2(u) = t_1 - u$, $k_3(u) = t_2 - u$, $k_4(u) = n - t_1 - t_2 + u$. *Тогда*

$$P_{\theta=0}(U = u \mid T = t) = \frac{(\prod_{i=1}^{4} k_i(u)!)^{-1}}{\sum_{s\in\mathcal{D}}(\prod_{i=1}^{4} k_i(s)!)^{-1}}$$

где $\mathcal{D} = \{s \in \mathbb{Z} : 0 \leq k_i(s) \leq n$ *для всех* $i = 1\ldots,4\}$.

Доказательство лемм. 1.4.1 не приводится, поскольку оно полностью аналогично доказательству лемм. 1.3.1.

Стоит отметить, что $\ln\left(\frac{q_{00}q_{11}}{q_{01}q_{10}}\right) = \ln\left(\frac{p_{00z}p_{11z}}{p_{01z}p_{10z}}\right)$. Поэтому, проверяя гипотезу о параметре $\ln\left(\frac{q_{00}q_{11}}{q_{01}q_{10}}\right)$ в условном распределении, мы проверяем гипотезу о параметре $\ln\left(\frac{p_{00z}p_{11z}}{p_{01z}p_{10z}}\right)$ в трехмерном распределении Бернулли.

1.5 Проверка условной независимости по подвыборкам из условных распределений

Приведем трактовку опр. 1.1.2 для трехмерного распределения Бернулли.

Определение 1.5.1. *В трехмерном распределении Бернулли случайные величины X и Y условно независимы при условии Z, если выполнены следующие условия:*

- *X и Y независимы при условии $Z = 0$*
- *X и Y независимы при условии $Z = 1$*

Используя опр. 1.5.1, сформулируем индивидуальные гипотезы:

- H_0 : X и Y независимы при условии $Z = 0$
- H_1 : X и Y независимы при условии $Z = 1$

Тогда гипотеза об условной независимости имеет вид $H = H_0 \cap H_1$. Естественным образом, гипотезу H_0 необходимо проверять по наблюдениям $(x_i, y_i, z_i)^T$, в которых $z_i = 0$. Поскольку $(X, Y)^T$ при условии $Z = z$ имеет двумерное распределение Бернулли, то в качестве теста для H_0 можно использовать $\varphi_0 = \varphi_0^{\text{Independence}}$, приведенный в разд. 1.4. Аналогичные рассуждения справедливы и для гипотезы H_1. Учитывая озвученные соображения, построим тест проверки гипотезы H, контролирующий вероятность ошибки первого рода на уровне α. Пусть

$$\begin{pmatrix} X_1 \\ Y_1 \\ Z_1 \end{pmatrix}, \begin{pmatrix} X_2 \\ Y_2 \\ Z_2 \end{pmatrix}, \ldots, \begin{pmatrix} X_n \\ Y_n \\ Z_n \end{pmatrix}$$

повторная выборка из распределения случайного вектора $(X, Y, Z)^T$. Обозначим $\mathbf{Z} = (Z_1, \ldots, Z_n)$ и $\mathbf{z} = (z_1, \ldots, z_n)$. Покажем, что в условном распределении при $\mathbf{Z} = \mathbf{z}$ наблюдения

$$\Xi = \begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \begin{pmatrix} X_2 \\ Y_2 \end{pmatrix}, \ldots, \begin{pmatrix} X_n \\ Y_n \end{pmatrix}$$

являются независимыми.

Лемма 1.5.1.

$$P(X_1 = x_1, Y_1 = y_1, \ldots, X_n = x_n, Y_n = y_n \mid \mathbf{Z} = \mathbf{z}) =$$

$$= \prod_{i=1}^{n} P(X_i = x_i, Y_i = y_i \mid \mathbf{Z} = \mathbf{z})$$

Доказательство. С одной стороны:

$$P(X_1 = x_1, Y_1 = y_1, Z_1 = z_1, \ldots, X_n = x_n, Y_n = y_n, Z_n = z_n) =$$

$$= P(\mathbf{Z} = \mathbf{z}) P(X_1 = x_1, Y_1 = y_1, \ldots, X_n = x_n, Y_n = y_n \mid \mathbf{Z} = \mathbf{z})$$

С другой стороны:

$$P(X_1 = x_1, Y_1 = y_1, Z_1 = z_1, \ldots, X_n = x_n, Y_n = y_n, Z_n = z_n) =$$

$$= \prod_{i=1}^{n} P(X_i = x_i, Y_i = y_i, Z_i = z_i) =$$

$$= \prod_{i=1}^{n} P(X_i = x_i, Y_i = y_i \mid Z_i = z_i) P(Z_i = z_i) =$$

$$= P(\mathbf{Z} = \mathbf{z}) \prod_{i=1}^{n} P(X_i = x_i, Y_i = y_i \mid \mathbf{Z} = \mathbf{z})$$

□

Пусть также $\mathbf{Z} = \mathbf{z}$ фиксированы. Разобьем выборку Ξ на две подвыборки Ξ_0 и Ξ_1, такие что:

$$\Xi_0 = \begin{pmatrix} X_{i_1} \\ Y_{i_1} \end{pmatrix}, \begin{pmatrix} X_{i_2} \\ Y_{i_2} \end{pmatrix}, \ldots, \begin{pmatrix} X_{i_{n_0}} \\ Y_{i_{n_0}} \end{pmatrix}$$

где $Z_{i_k} = z_{i_k} = 0$ для всех i_k при $k = \overline{1, n_0}$ и

$$\Xi_1 = \begin{pmatrix} X_{j_1} \\ Y_{j_1} \end{pmatrix}, \begin{pmatrix} X_{j_2} \\ Y_{j_2} \end{pmatrix}, \ldots, \begin{pmatrix} X_{j_{n_1}} \\ Y_{j_{n_1}} \end{pmatrix}$$

где $Z_{j_k} = z_{j_k} = 1$ для всех j_k при $k = \overline{1, n_1}$. Причем $n = n_0 + n_1$. Отметим, что разбиение $\Xi = \Xi_0 \sqcup \Xi_1$ определяется лишь набором $\mathbf{Z} = \mathbf{z}$.

Сформулируем следующую теорему.

Теорема 1.5.1. *Пусть* $\mathbf{Z} = \mathbf{z}$ *– фиксированы. Тогда выборка* Ξ_0 *является повторной выборкой из распределения* $(X, Y)^T$ *при условии* $Z = 0$.

Доказательство. По лемм. 1.5.1:

$$P(X_1 = x_1, Y_1 = y_1, \ldots, X_n = x_n, Y_n = y_n \mid \mathbf{Z} = \mathbf{z}) =$$

$$= \prod_{i=1}^{n} P(X_i = x_i, Y_i = y_i \mid \mathbf{Z} = \mathbf{z})$$

Просуммировав обе части вышепредставленного равенства по всем возможным значениям $x_{j_1}, y_{j_1}, \ldots, x_{j_{n_1}}, y_{j_{n_1}}$ получаем:

$$P(X_{i_1} = x_{i_1}, Y_{i_1} = y_{i_1}, \ldots, X_{i_{n_0}} = x_{i_{n_0}}, Y_{i_{n_0}} = y_{i_{n_0}} \mid \mathbf{Z} = \mathbf{z}) =$$

$$= \prod_{k=1}^{n_0} P(X_{i_k} = x_{i_k}, Y_{i_k} = y_{i_k} \mid \mathbf{Z} = \mathbf{z})$$

Значит

$$\Xi_0 = \begin{pmatrix} X_{i_1} \\ Y_{i_1} \end{pmatrix}, \begin{pmatrix} X_{i_2} \\ Y_{i_2} \end{pmatrix}, \ldots, \begin{pmatrix} X_{i_{n_0}} \\ Y_{i_{n_0}} \end{pmatrix}$$

независимые наблюдения при условии $\mathbf{Z} = \mathbf{z}$.

Покажем, что $(X_{i_k}, Y_{i_k})^T$ при условии $\mathbf{Z} = \mathbf{z}$ распределен также как и $(X, Y)^T$ при условии $Z = 0$.

$$P(X_{i_k} = x_{i_k}, Y_{i_k} = y_{i_k} \mid \mathbf{Z} = \mathbf{z}) = P(X_{i_k} = x_{i_k}, Y_{i_k} = y_{i_k} \mid Z_{i_k} = z_{i_k}) =$$

$$= P(X_{i_k} = x_{i_k}, Y_{i_k} = y_{i_k} \mid Z_{i_k} = 0) = P(X = x_{i_k}, Y = y_{i_k} \mid Z = 0)$$

□

Аналогично показывается, что Ξ_1 является повторной выборкой из распределения $(X, Y)^T$ при условии $Z = 1$.

Сформулируем теорему.

Теорема 1.5.2. *Пусть* Ξ_0 *и* Ξ_1 *– подвыборки, полученные разбиением случайной выборки* Ξ. *Пусть* φ_0 *и* φ_1 *– рандомизированные тесты проверки гипотез* H_0 *и* H_1 *по повторным выборкам* Ξ_0 *и* Ξ_1 *соответственно. Введем события:*

$$A_0 = \{\textit{отвергнуть гипотезу } H_0 \textit{ рандомизированным тестом } \varphi_0\}$$

$$A_1 = \{\textit{отвергнуть гипотезу } H_1 \textit{ рандомизированным тестом } \varphi_1\}$$

Пусть φ_0 и φ_1 тесты уровня α_0 и α_1 при любом объеме наблюдений в подвыборках Ξ_0 и Ξ_1 соответственно, то есть:

$$P_{H_0 \cap H_1, \mathbf{Z}=\mathbf{z}}(A_0) = \alpha_0$$

$$P_{H_0 \cap H_1, \mathbf{Z}=\mathbf{z}}(A_1) = \alpha_1$$

Тогда $P_{H_0 \cap H_1}(A_0 \cap A_1) = P_{H_0 \cap H_1}(A_0) P_{H_0 \cap H_1}(A_1)$.

Доказательство. Пусть $\mathbf{Z} = \mathbf{z}$ фиксировано. Обозначим через

$$T_0 = (X_{i_1}, Y_{i_1}, \ldots, X_{i_{n_0}}, Y_{i_{n_0}})^T,\ T_1 = (X_{j_1}, Y_{j_1}, \ldots, X_{j_{n_1}}, Y_{j_{n_1}})^T$$

случайные векторы наблюдений, используемые в тестах φ_0 и φ_1 соответственно. Распишем следующую вероятность

$$P_{H_0 \cap H_1, \mathbf{Z}=\mathbf{z}}(A_0 \cap A_1) =$$

$$= \sum_{t_0} \sum_{t_1} P_{H_0 \cap H_1, \mathbf{Z}=\mathbf{z}}(A_0 \cap A_1 \mid T_0 = t_0, T_1 = t_1) P_{H_0 \cap H_1, \mathbf{Z}=\mathbf{z}}(T_0 = t_0, T_1 = t_1)$$

Отметим, что $P_{H_0 \cap H_1, \mathbf{Z}=\mathbf{z}}(A_0 \cap A_1 \mid T_0 = t_0, T_1 = t_1) = \varphi_0(t_0)\varphi_1(t_1)$, поскольку для того, чтобы при известных значениях статистик t_0 и t_1 отвергнуть гипотезы H_0 и H_1, в рандомизированном тесте нужно провести два испытания с вероятностью успеха $\varphi_0(t_0)$ и $\varphi_1(t_1)$. Постулируется, что такие испытания независимые. Тогда:

$$P_{H_0 \cap H_1, \mathbf{Z}=\mathbf{z}}(A_0 \cap A_1) = \sum_{t_0} \sum_{t_1} \varphi_0(t_0)\varphi_1(t_1) P_{H_0 \cap H_1, \mathbf{Z}=\mathbf{z}}(T_0 = t_0, T_1 = t_1) =$$

$$= \sum_{t_0} \sum_{t_1} \varphi_0(t_0)\varphi_1(t_1) P_{H_0 \cap H_1, \mathbf{Z}=\mathbf{z}}(T_0 = t_0) P_{H_0 \cap H_1, \mathbf{Z}=\mathbf{z}}(T_1 = t_1) =$$

$$= \sum_{t_0} \left[\varphi_0(t_0) P_{H_0 \cap H_1, \mathbf{Z}=\mathbf{z}}(T_0 = t_0) \left(\sum_{t_1} \varphi_1(t_1) P_{H_0 \cap H_1, \mathbf{Z}=\mathbf{z}}(T_1 = t_1) \right) \right] =$$

$$= \sum_{t_0} \varphi_0(t_0) P_{H_0 \cap H_1, \mathbf{Z}=\mathbf{z}}(T_0 = t_0) \alpha_1 = \alpha_0 \alpha_1$$

По формуле полной вероятности:

$$P_{H_0 \cap H_1}(A_0) = \sum_{\mathbf{z}} P_{H_0 \cap H_1, \mathbf{Z}=\mathbf{z}}(A_0) P_{H_0 \cap H_1}(\mathbf{Z} = \mathbf{z}) = \sum_{\mathbf{z}} \alpha_0 P_{H_0 \cap H_1}(\mathbf{Z} = \mathbf{z}) = \alpha_0$$

Аналогично, $P_{H_0 \cap H_1}(A_1) = \alpha_1$. Также по формуле полной вероятности:

$$P_{H_0 \cap H_1}(A_0 \cap A_1) = \sum_{\mathbf{z}} P_{H_0 \cap H_1, \mathbf{Z}=\mathbf{z}}(A_0 \cap A_1) P_{H_0 \cap H_1}(\mathbf{Z} = \mathbf{z}) =$$

$$= \sum_{\mathbf{z}} \alpha_0 \alpha_1 P_{H_0 \cap H_1}(\mathbf{Z} = \mathbf{z}) = \alpha_0 \alpha_1$$

Таким образом, $P_{H_0 \cap H_1}(A_0 \cap A_1) = P_{H_0 \cap H_1}(A_0) P_{H_0 \cap H_1}(A_1) = \alpha_0 \alpha_1$. □

Применим теор. 1.5.2 для проверки условной независимости в трехмерном распределении Бернулли. Положим индивидуальные гипотезы:

- $H_0 : X$ и Y независимы при условии $Z = 0$
- $H_1 : X$ и Y независимы при условии $Z = 1$

Для проверки гипотез H_0 и H_1 будем использовать тесты $\varphi_0 = \varphi_0^{\text{Independence}}$ и $\varphi_1 = \varphi_1^{\text{Independence}}$ уровня α_0 и α_1 по повторным выборкам Ξ_0 и Ξ_1 соответственно. Тогда гипотеза об условной независимости имеет вид $H = H_0 \cap H_1$ и тест проверки условной независимости можно определить как:

$$\varphi^{\text{Subsamples}} = \begin{cases} 1, & \text{наступило событие } A_0 \cup A_1 \\ 0, & \text{иначе} \end{cases}$$

Пусть далее $\alpha_0 = \alpha_1 = \gamma$. Тогда

$$P_{H_0 \cap H_1}(\varphi^{\text{Subsamples}} = 1) = P_{H_0 \cap H_1}(A_0 \cup A_1) =$$

$$= P_{H_0 \cap H_1}(A_0) + P_{H_0 \cap H_1}(A_1) - P_{H_0 \cap H_1}(A_0 \cap A_1) = 2\gamma - \gamma^2$$

Нетрудно проверить, что для контроля $P_{H_0 \cap H_1}(\varphi^{\text{Subsamples}} = 1) = \alpha$ достаточно положить уровень значимости $\gamma = 1 - \sqrt{1 - \alpha}$ на тестах проверки индивидуальных гипотез H_0 и H_1.

Покажем, что тест $\varphi^{\text{Subsamples}}$ является несмещенным.

Теорема 1.5.3. *Тест $\varphi^{Subsamples}$ уровня α проверки гипотезы $H : X \perp\!\!\!\perp Y \mid Z$ является несмещенным.*

Доказательство. Положим $\Theta_H = \{\theta : p_{000}p_{110} = p_{010}p_{100}$ и $p_{001}p_{111} = p_{011}p_{101}\}$, $\Theta_K = \{\theta : p_{000}p_{110} \neq p_{010}p_{100}$ или $p_{001}p_{111} \neq p_{011}p_{101}\}$.

Отметим, что $\varphi_0 = \varphi_0^{\text{Independence}}$ и $\varphi_1 = \varphi_1^{\text{Independence}}$ – несмещенные тесты проверки гипотез H_0 и H_1 уровня γ соответственно. Причем, при истинности гипотез H_0 и H_1 на тестах φ_0 и φ_1 вероятность ошибки первого рода в точности равна γ. Поэтому, $E_\theta(\varphi_0) = P_\theta(A_0) \geq \gamma$ и $E_\theta(\varphi_1) = P_\theta(A_1) \geq \gamma$ для любого $\theta \in \Theta_H \cup \Theta_K$.

Пусть $\theta \in \Theta_H$, тогда $E_\theta(\varphi^{\text{Subsamples}}) = P_\theta(\varphi^{\text{Subsamples}} = 1) = \alpha$ поскольку $\varphi^{\text{Subsamples}}$ тест уровня α.

Пусть $\theta \in \Theta_K$. Как и в доказательстве теор. 1.5.2 можно показать, что $P_{\theta,\mathbf{Z}=\mathbf{z}}(A_0 \cap A_1) = P_{\theta,\mathbf{Z}=\mathbf{z}}(A_0)P_{\theta,\mathbf{Z}=\mathbf{z}}(A_1)$. Следовательно:

$$P_{\theta,\mathbf{Z}=\mathbf{z}}(A_0 \cup A_1) = P_{\theta,\mathbf{Z}=\mathbf{z}}(A_0) + P_{\theta,\mathbf{Z}=\mathbf{z}}(A_1) - P_{\theta,\mathbf{Z}=\mathbf{z}}(A_0)P_{\theta,\mathbf{Z}=\mathbf{z}}(A_1) \geq \gamma + \gamma - \gamma^2 = \alpha$$

Тогда:

$$E_\theta(\varphi^{\text{Subsamples}}) = P_\theta(\varphi^{\text{Subsamples}} = 1) = P_\theta(A_0 \cup A_1) =$$
$$= \sum_{\mathbf{z}} P_{\theta,\mathbf{Z}=\mathbf{z}}(A_0 \cup A_1) P_\theta(\mathbf{Z} = \mathbf{z}) \geq \sum_{\mathbf{z}} \alpha P_\theta(\mathbf{Z} = \mathbf{z}) = \alpha$$

□

2 Экспериментальная часть

2.1 Способ вычисления вероятностей для РНМН теста на ЭВМ

Для нахождения порогов в РНМН тестах возникает необходимость подсчета вероятностей вида:

$$P_{\theta=0}(U = u \mid T = t) = \frac{(\prod_{i=1}^{p} k_i(u)!)^{-1}}{\sum_{s\in\mathcal{D}}(\prod_{i=1}^{p} k_i(s)!)^{-1}}$$

где $\mathcal{D}$ – некая область допустимых значений, $k_i(u) : \mathcal{D} \to \{0, \ldots, n\},\ i = \overline{1, p}$. Основную проблему в этой формуле представляют факториалы, вычисление которых затруднительно на ЭВМ. Предложим методологию, которая поможет обойти эту проблему.

Пусть $f(i) = \sum_{j=1}^{i} \ln(j)$. Тогда $\ln(n!) = f(n)$. Учитывая это, запишем:

$$\frac{(\prod_{i=1}^{p} k_i(u)!)^{-1}}{\sum_{s\in\mathcal{D}}(\prod_{i=1}^{p} k_i(s)!)^{-1}} = \frac{\exp\{-\ln(\prod_{i=1}^{p} k_i(u)!)\}}{\sum_{s\in\mathcal{D}} \exp\{-\ln(\prod_{i=1}^{p} k_i(s)!)\}} =$$

$$= \frac{\exp\{-\sum_{i=1}^{p} f(k_i(u))\}}{\sum_{s\in D} \exp\{-\sum_{i=1}^{p} f(k_i(s))\}}$$

Полученное выражение удобно с позиции того, что оно не требует подсчета факториалов и ЭВМ умеют эффективно вычислять функцию:

$$\mathrm{softmax}(x, i) = \frac{\exp\{x_i\}}{\sum_{j=1}^{N} \exp\{x_j\}},\ x = (x_1, \ldots, x_N)$$

Это происходит благодаря свойству:

$$\mathrm{softmax}(x, i) = \frac{\exp\{x_i\}}{\sum_{j=1}^{N} \exp\{x_j\}} = \frac{\exp\{x_i - C\}}{\sum_{j=1}^{N} \exp\{x_j - C\}},\ \text{где } C = \max_{1\le j\le N} x_j$$

за счет которого удается избежать переполнения вещественного типа данных, связанного с вычислением экспоненты.

2.2 Сравнение тестов

Кратко напомним тесты, которые будут сравниваться.

1. φ^{Theta} – РНМН тест уровня $\alpha = 0.05$ проверки гипотезы $H^{\text{Theta}} : \theta = 0$, где $\theta = \ln\left(\dfrac{p_{001}p_{111}p_{010}p_{100}}{p_{011}p_{101}p_{000}p_{110}}\right)$. Поскольку из $X \perp\!\!\!\perp Y \mid Z$ следует $\theta = 0$, то φ^{Theta} также является несмещенным тестом уровня $\alpha = 0.05$ проверки гипотезы $H : X \perp\!\!\!\perp Y \mid Z$.

2. $\varphi^{\text{Subsamples}}$ – несмещенный тест уровня $\alpha = 0.05$ для проверки гипотезы $H : X \perp\!\!\!\perp Y \mid Z$. Гипотеза H отвергается, если РНМН-тестами уровня $\gamma = 1 - \sqrt{1-\alpha}$ отвергается хотя бы одна гипотеза $H_z : \theta_z = 0$, где $\theta_z = \ln\left(\dfrac{p_{00z}p_{11z}}{p_{01z}p_{10z}}\right)$. Напомним, что гипотеза $H_z : \theta_z = 0$ проверяется по по подвыборке из условного распределения $(X, Y)^T$ при условии $Z = z$.

3. φ^{Partial} – точный тест уровня $\alpha = 0.05$ проверки гипотезы $H^{\text{Partial}} : \rho^{XY \cdot Z} = 0$ в трехмерном нормальном распределении.

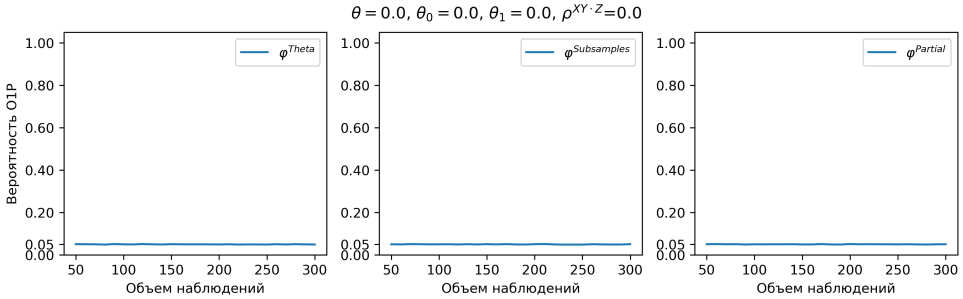

Рис. 1: Графики зависимости вероятности ошибки 1 рода (ОШР) от количества наблюдений, $p_{000} = 0.125, p_{001} = 0.125, p_{010} = 0.125, p_{011} = 0.125, p_{100} = 0.125, p_{101} = 0.125, p_{110} = 0.125, p_{111} = 0.125$. Гипотеза $H : X \perp\!\!\!\perp Y \mid Z$ верна. Вероятность оценивается по 10^5 экспериментам.

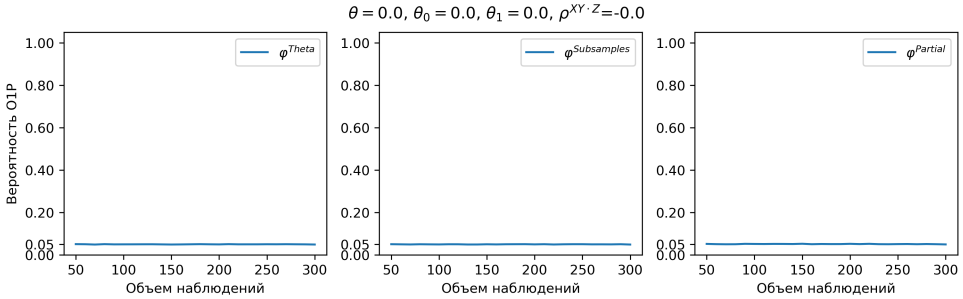

Рис. 2: Графики зависимости вероятности ошибки 1 рода (ОШР) от количества наблюдений, $p_{000} = 0.15, p_{001} = 0.1, p_{010} = 0.3, p_{011} = 0.1, p_{100} = 0.05, p_{101} = 0.1, p_{110} = 0.1, p_{111} = 0.1$. Гипотеза $H : X \perp\!\!\!\perp Y \mid Z$ верна. Вероятность оценивается по 10^5 экспериментам.

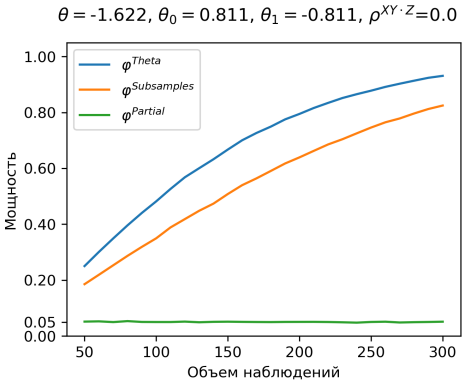

Рис. 3: График зависимости мощности от количества наблюдений, $p_{000} = 0.15, p_{001} = 0.1, p_{010} = 0.1, p_{011} = 0.15, p_{100} = 0.1, p_{101} = 0.15, p_{110} = 0.15, p_{111} = 0.1$. Гипотеза $H : X \perp\!\!\!\perp Y \mid Z$ не верна, однако верна гипотеза $H^{\text{Partial}} : \rho^{XY \cdot Z} = 0$. Мощность оценивается по 10^5 экспериментам.

Из (рис. 1) и (рис. 2) видно, что для гипотезы $H : X \perp\!\!\!\perp Y \mid Z$ тесты φ^{Theta}, $\varphi^{\text{Subsamples}}$, контролируют вероятность ошибки первого рода на уровне $\alpha = 0.05$. Этот результат полностью согласуется с теорией из разд. 1.3, разд. 1.5.

(рис. 1), (рис. 2), (рис. 3) показывают, что тест φ^{Partial} контролирует вероятность ошибки первого рода на уровне $\alpha = 0.05$ для гипотезы $H^{\text{Partial}} : \rho^{XY \cdot Z} = 0$ в трехмерном распределении Бернулли. Этот результат является неожиданным, поскольку тест φ^{Partial} теоретически обоснован лишь для трехмерного нормального распределения. Поскольку из $X \perp\!\!\!\perp Y \mid Z$ следует $\rho^{XY \cdot Z} = 0$, то тест φ^{Partial} также контролирует вероятность ошибки первого рода на уровне $\alpha = 0.05$ и для гипотезы $H : X \perp\!\!\!\perp Y \mid Z$, что показано на (рис. 1), (рис. 2). Однако, стоит отметить, что φ^{Partial} проверяет необходимое условие условной независимости. Поэтому может возникнуть ситуация как на (рис. 3), когда гипотеза $H : X \perp\!\!\!\perp Y \mid Z$ не верна, но тест φ^{Partial} не распознает отклонение от условной независимости, поскольку контролирует вероятность ошибки первого рода на уровне $\alpha = 0.05$ для гипотезы $H^{\text{Partial}} : \rho^{XY \cdot Z} = 0$.

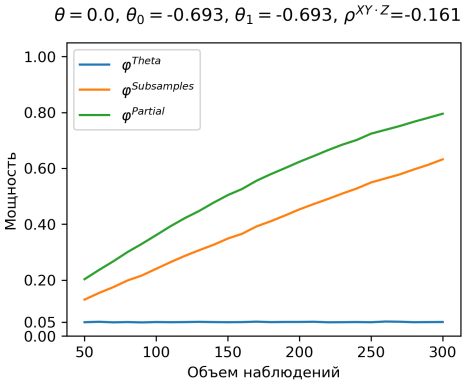

Рис. 4: График зависимости мощности от количества наблюдений, $p_{000} = 0.15, p_{001} = 0.05, p_{010} = 0.3, p_{011} = 0.1, p_{100} = 0.1, p_{101} = 0.1, p_{110} = 0.1, p_{111} = 0.1$. Гипотеза $H : X \perp\!\!\!\perp Y \mid Z$ не верна, однако верны гипотезы $H^{\text{Theta}} : \theta = 0$ и $H' : Y \perp\!\!\!\perp Z \mid X$. Мощность оценивается по 10^5 экспериментам.

Напомним, что тест φ^{Theta} проверяет необходимое условие условной независимости. Так на (рис. 4) гипотеза $H : X \perp\!\!\!\perp Y \mid Z$ не верна, но тест φ^{Theta} не распознает отклонение от условной независимости и контролирует вероятность ошибки первого рода на уровне $\alpha = 0.05$ для гипотезы $H^{\text{Theta}} : \theta = 0$.

Отметим, что φ^{Theta} – несмещенный тест уровня α проверки гипотезы $H : X \perp\!\!\!\perp Y \mid Z$. Так как по теор. 1.5.3 тест $\varphi^{\text{Subsamples}}$ является несмещенным тестом уровня α проверки гипотезы $H : X \perp\!\!\!\perp Y \mid Z$, и на (рис. 4) тест $\varphi^{\text{Subsamples}}$ мощнее теста φ^{Theta}, то тест φ^{Theta} не является РНМН тестом проверки гипотезы $H : X \perp\!\!\!\perp Y \mid Z$, хотя является РНМН тестом проверки гипотезы $H^{\text{Theta}} : \theta = 0$.

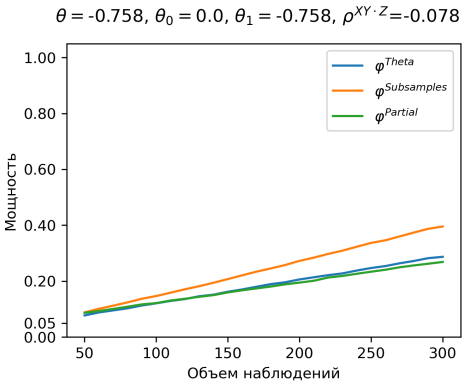

Рис. 5: График зависимости мощности от количества наблюдений, $p_{000} = 0.15, p_{001} = 0.06, p_{010} = 0.3, p_{011} = 0.16, p_{100} = 0.05, p_{101} = 0.08, p_{110} = 0.1, p_{111} = 0.1$. Гипотеза $H : X \perp\!\!\!\perp Y \mid Z$ не верна, однако X и Y независимы при условии $Z = 0$. Мощность оценивается по 10^5 экспериментам.

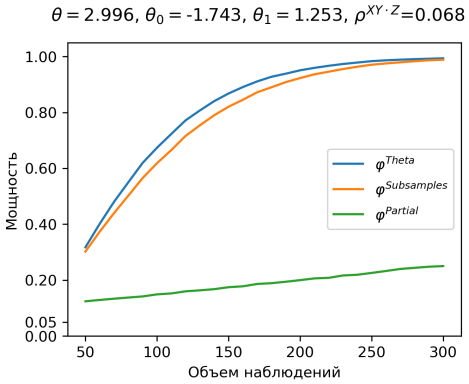

Рис. 6: График зависимости мощности от количества наблюдений, $p_{000} = 0.03, p_{001} = 0.1, p_{010} = 0.04, p_{011} = 0.08, p_{100} = 0.3, p_{101} = 0.1, p_{110} = 0.07, p_{111} = 0.28$. Гипотеза $H : X \perp\!\!\!\perp Y \mid Z$ не верна. Мощность оценивается по 10^5 экспериментам.

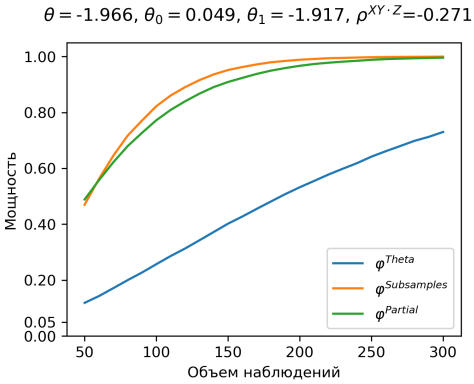

Рис. 7: График зависимости мощности от количества наблюдений, $p_{000} = 0.21, p_{001} = 0.12, p_{010} = 0.04, p_{011} = 0.34, p_{100} = 0.1, p_{101} = 0.12, p_{110} = 0.02, p_{111} = 0.05$. Гипотеза $H : X \perp\!\!\!\perp Y \mid Z$ не верна. Мощность оценивается по 10^5 экспериментам.

(рис. 3), (рис. 4), (рис. 5), (рис. 6), (рис. 7) показывают, что, вообще говоря, рассматриваемые тесты нельзя упорядочить по мощности. Кроме того, несмещенные тесты $\varphi^{\text{Subsamples}}$ и φ^{Theta} уровня α проверки гипотезы $H : X \perp\!\!\!\perp Y \mid Z$ также нельзя упорядочить по мощности. Поэтому вопрос построения РНМН теста проверки гипотезы $H : X \perp\!\!\!\perp Y \mid Z$ остается открытым.

Заключение

В настоящей выпускной квалификационной работы были получены следующие результаты:

1. Сформулирован и доказан критерий условной независимости в трехмерном распределении Бернулли.

2. Доказано, что равенство нулю частного коэффициента корреляции Пирсона $\rho^{XY \cdot Z}$ является необходимым, но не достаточным, условием условной независимости X и Y при условии Z в трехмерном распределении Бернулли.

3. Эмпирически, при объеме наблюдений $50 \leq n \leq 300$, показано, что тест φ^{Partial} является тестом уровня α как для гипотезы $H^{\text{Partial}} : \rho^{XY \cdot Z} = 0$, так и для гипотезы $H : X \perp\!\!\!\perp Y \mid Z$ в трехмерном распределении Бернулли.

4. В экспоненциальной форме записи трехмерного распределения Бернулли найден параметр $\theta = \ln\left(\dfrac{p_{001}p_{111}p_{010}p_{100}}{p_{011}p_{101}p_{000}p_{110}}\right)$, равенство нулю которого является необходимым условием выполнения одного из условий:

   - $X \perp\!\!\!\perp Y \mid Z$
   - $X \perp\!\!\!\perp Z \mid Y$
   - $Y \perp\!\!\!\perp Z \mid X$

5. Построен РНМН-тест φ^{Theta} уровня α проверки гипотезы $H^{\text{Theta}} : \theta = 0$. Показано, что φ^{Theta} является несмещенным тестом уровня α проверки гипотезы $H : X \perp\!\!\!\perp Y \mid Z$, однако не является РНМН-тестом проверки этой же гипотезы.

6. Построен тест $\varphi^{\text{Subsamples}}$ уровня α проверки гипотезы $H : X \perp\!\!\!\perp Y \mid Z$. Доказано, что этот тест является несмещенным.

Результаты настоящей работы могут быть использованы для задачи идентификации графической модели с попарным марковским свойством в трехмерном распределении Бернулли.

Список литературы

1. *Anderson T.* An Introduction to Multivariate Statistical Analysis. — Wiley-Interscience, 2003.
2. *Cramér H.* Mathematical methods of statistics. — Princeton University Press, 1946.
3. *Dai B.*, *Ding S.*, *Wahba G.* Multivariate Bernoulli distribution // Bernoulli. — 2013. — Т. 19, № 4. — С. 1465—1483.
4. *Drton M.*, *Perlman M. D.* Model selection for Gaussian concentration graphs // Biometrika. — 2004. — Т. 91, № 3. — С. 591—602.
5. *Drton M.*, *Perlman M. D.* Multiple testing and error control in Gaussian graphical model selection // Statistical Science. — 2007. — Т. 22, № 3. — С. 430—449.
6. *Jordan M. I.* Graphical Models // Statistical Science. — 2004. — Т. 19, № 1. — С. 140—155.
7. *Lauritzen S. L.* Graphical models. — Clarendon Press, 1996.
8. *Lehmann E. L.* Testing statistical hypotheses. — Wiley, 1986.
9. *Teugels J. L.* Some representations of the multivariate Bernoulli and binomial distributions // Journal of Multivariate Analysis. — 1990. — Т. 32, № 2. — С. 256—268.
10. *Vogel D.*, *Fried R.* Elliptical graphical modelling // Biometrika. — 2011. — Т. 98. — С. 935—951.