Некоторые свойства статистического теста «Book Stack»

Бзикадзе Андрей Важевич, гр. 422

Санкт-Петербургский государственный университет
Математико-механический факультет
Кафедра статистического моделирования

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

СтатМод

Санкт-Петербург
2015г.

«Book Stack»-преобразование

- $\mathcal{A} = \{a_1, \dots, a_S\}$ — множество «книг»;
- Стопка «книг»;
- Начальный порядок;
- Из стопки случайная книга перекладывается наверх;

- $\{\eta_i\}_{i=1}^{\infty}$ — последовательность названий случайных книг;
- $\{\xi_i\}_{i=1}^{\infty}$ — последовательность положений случайных книг.

Пример

$S = 5$

$i = 1$	$i = 2$	$i = 3$	$i = 4$
$\eta_i = 4$	$\eta_i = 4$	$\eta_i = 2$	
$\xi_i = 2$	$\xi_i = 1$	$\xi_i = 5$	

a_3	a_4	a_4	a_2
a_4	a_3	a_3	a_4
a_1	a_1	a_1	a_3
a_5	a_5	a_5	a_1
a_2	a_2	a_2	a_5

Обзор литературы

Применение «Book Stack»-преобразования:

- Рябко Б.Я. (1980): алгоритм сжатия данных под названием «метод стопки книг»;
- Bentley J.L., Sleator D.D., Tarjan R.E., Wei V.K. (1986): тот же алгоритм под названием «Move To Front»;
- Рябко Б.Я., Монарев В.А., Пестунов А.И., Шокин Ю.И., Стогниенко В.С. (2003–2004): тест для проверки свойств генераторов случайных чисел под названием «Book Stack».

Равносильны:

- $\mathrm{H}_0 : \eta_i$ независимы и равномерно распределены на $\{1, 2, \ldots, S\}$;
- $\mathrm{H}_0^* : \xi_i$ независимы и равномерно распределены на $\{1, 2, \ldots, S\}$.

Два теста:

- χ^2-тест к исходной выборке;
- χ^2-тест к преобразованной выборке.

Против каких альтернатив критерий «после» Book Stack будет более мощным, чем этот же критерий «до»?

Равносильны:

- $\mathrm{H}_0 : \eta_i$ независимы и равномерно распределены на $\{1, 2, \ldots, S\}$;
- $\mathrm{H}_0^* : \xi_i$ независимы и равномерно распределены на $\{1, 2, \ldots, S\}$.

Два теста:

- χ^2-тест к исходной выборке;
- χ^2-тест к преобразованной выборке.

Против каких альтернатив критерий «после» Book Stack будет более мощным, чем этот же критерий «до»?

Выбор альтернативной гипотезы и пример

Альтернатива: $\mathrm{H}_1 : \{\eta_i\}$ — независимы и одинаково, но **не равномерно**, распределены.

Моделирование:

- Заданное дискретное распределение на $\{1, 2, \ldots, S\}$;
- Вихрь Мерсенна;
- n — размер выборки, m — количество выборок;
- Критерий χ^2 с $S-1$ степенью свободы: m штук P-значений.

Цель:

- Сравнение мощности критерия χ^2 «до» и «после» преобразования «Book Stack».

Моделирование: мощность χ^2-тестов

Обозначим: $p_k = \mathbb{P}(\eta_i = k) > 0$ для $k = 1, 2, \ldots, S$.

Параметры: $S = 10$, $n = 10^4$, $m = 100$, $p_1 = 0.11$, $p_i = 0.099$.

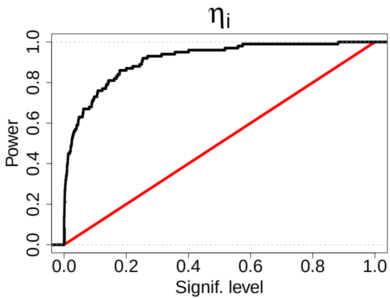


Рис. : Мощность критерия χ^2 до «Book Stack».
P-значение $= 4.4 \cdot 10^{-16}$.

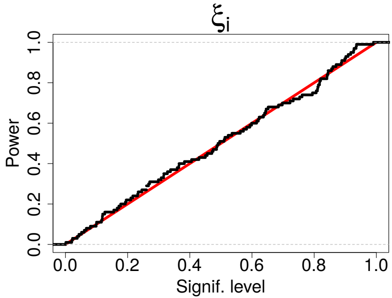
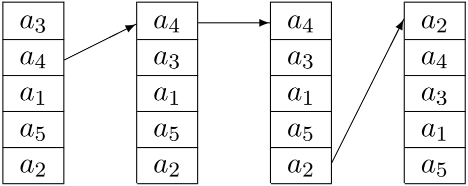


Рис. : Мощность критерия χ^2 после «Book Stack».
P-значение $= 0.903$.

Предложение

Пусть $\{\eta_i\}_{i\geqslant 1}$ *—* ***независимы и одинаково распределены*** *на* ***множестве*** $\{1, 2, \dots, S\}$. *Тогда последовательность состояний всей стопки —* ***эргодическая однородная марковская цепь.***



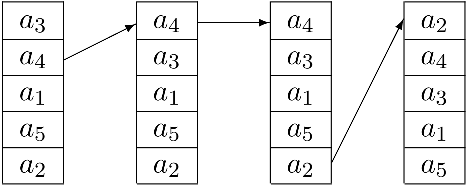
- Начальный порядок: $\Xi_0 = (3, 4, 1, 5, 2)^{\mathrm{T}}$;
- $\Xi_1 = \Xi_2 = (4, 3, 1, 5, 2)^{\mathrm{T}}$;
- $\Xi_3 = (2, 4, 3, 1, 5)^{\mathrm{T}}$.

Начальный порядок книг — начальное распределение

Марковская цепь. Пример.

Предложение

Пусть $\{\eta_i\}_{i\geqslant 1}$ *—* ***независимы и одинаково распределены*** *на множестве* $\{1, 2, \ldots, S\}$. *Тогда последовательность состояний всей стопки —* ***эргодическая однородная марковская цепь.***



- Начальный порядок: $\Xi_0 = (3, 4, 1, 5, 2)^{\mathrm{T}}$;
- $\Xi_1 = \Xi_2 = (4, 3, 1, 5, 2)^{\mathrm{T}}$;
- $\Xi_3 = (2, 4, 3, 1, 5)^{\mathrm{T}}$.

Начальный порядок книг — начальное распределение.

Стационарное распределение ОМЦ

Предложение

Фиксируем $S \in \mathbb{N}$. ***Стационарное распределение*** *ОМЦ преобразования «Book Stack» задается следующим вектором-строкой:*

$$\pi_S = (\pi_{1,2,\ldots,S}, \pi_{1,2,\ldots,S,S-1}, \ldots, \pi_{S,S-1,\ldots 1}) \in \mathbb{R}^{S!},$$

$$\text{где } \pi_{i_1,i_2,\ldots,i_S} = \frac{\prod\limits_{k=1}^{S-1} p_{i_k}}{\prod\limits_{k=1}^{S-2}\left(1 - \sum\limits_{j=1}^{k} p_{i_j}\right)}.$$

Предложение

Пусть начальное распределение полки $\{\Xi_i\}_{i\geqslant 0}$ *— стационарное распределение* $(\pi_{1,2,\ldots,S}, \ldots, \pi_{S,S-1,\ldots,1})$.
Тогда

1. *последовательность* $\{\xi_i\}_{i\geqslant 1}$ *является* ***стационарной «в узком смысле»***;
2. *для любого* $j \in 1 : S$

$$s_j \stackrel{\text{def}}{=} \mathbb{P}(\xi_i = j) = \sum_{k=1}^{S} p_k \sum_{\substack{\alpha\in\mathfrak{S}_S \\ \alpha_j=k}} \pi_\alpha,$$

где $\mathfrak{S}_S$ *— множество перестановок порядка* S.

Сходимость к стационарному распределению

Теорема

Пусть $\xi_1, \ldots \xi_n, \ldots$ — последовательность положений случайных книг. Обозначим

$$\tau_k = \tau_k(n) = \mathbb{I}_k(\xi_1) + \ldots + \mathbb{I}_k(\xi_n),$$

где $\mathbb{I}_A$ — индикатор множества A и $1 \leqslant k \leqslant S$. Тогда

$$\mathbb{E}\left(\frac{\tau_k}{n} - s_k\right)^2 = \mathrm{O}(1/n).$$

Сходимость по вероятности сохраняется и при группировке, то есть при разбиении $\{1, 2, \ldots, S\}$ на дизъюнктные подмножества.

Теорема

Пусть $\{\eta_i\}$ *независимы и одинаково распределены с распределением, задаваемым набором вероятностей* $\{p_i\}_{i=1}^S$, *где* $\max_i p_i > \min_i p_i > 0$. *Тогда*

$$\sum_{i=1}^{S}(p_i - 1/S)^2 \geqslant \sum_{i=1}^{S}(s_i - 1/S)^2.$$

При $n \to \infty$ статистика χ^2 после «Book Stack» не больше, чем до преобразования.

Аналогичное «выравнивание» имеет место для энтропии и расстояния по вариации.

Параметр $S = 8$, $n = m = 10^3$, $p_1 = 0.165$, $p_i = 0.119$.

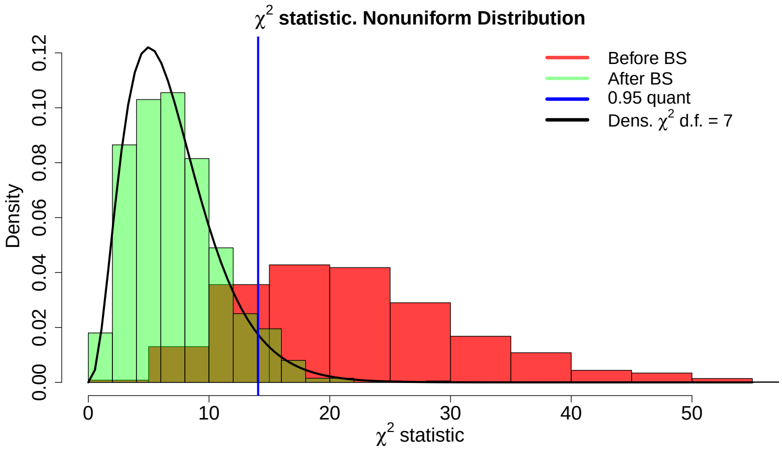


Рис. : Гистограмма распределения статистики критерия χ^2 до и после «Book Stack». Теоретическое значение 0.95 квантили — 14.07.

Параметры: $S = 8$, $n = m = 10^3$, $p_1 = 0.25$, $p_i = 0.107$.

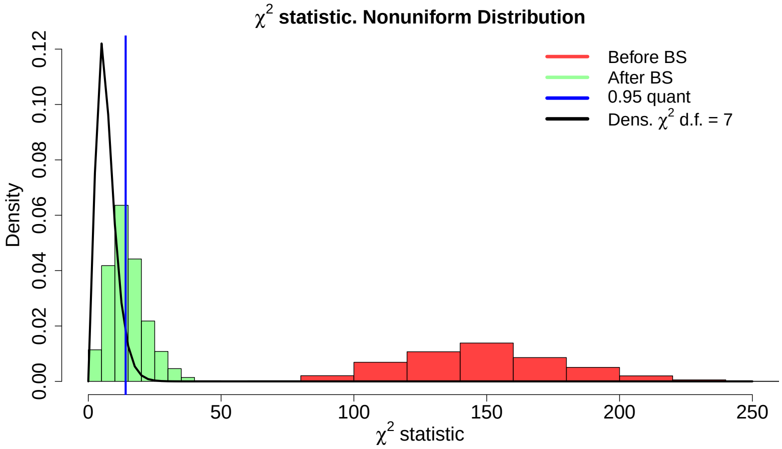


Рис. : Гистограммы распределения статистики критерия χ^2 до и после «Book Stack». Теоретическое значение 0.95-квантили — 14.07.

Итог и другая альтернатива

Итог:
Если рассматривать в качестве альтернативы независимые и одинаково, но не равномерно, распределенные случайные величины η_i, то при больших объемах выборки

- Критерий χ^2 к ξ_i не мощнее, чем к η_i.

Другая альтернатива:

- $\mathrm{H}_1 : \{\eta_i\}$ — конечная однородная марковская цепь со стационарным равномерным распределением на множестве $\{1, 2, \ldots, S\}$.

Только вычислительные эксперименты.

Модель марковской цепи и статистические тесты

Модель марковской цепи:

- Задано: $p < 1$;
- Матрица переходных вероятностей: $\mathbf{P} = (p_{ij})$, где $p_{ii} = p$ и $p_{ij} = (1-p)/(S-1)$ при $i \neq j$.

Статистические тесты:

- Критерий χ^2 для исходной марковской цепи $\{\eta_i\}$;
- Критерий χ^2 для $\{\xi_i\}$;
- Двумерный критерий χ^2 для $\left\{(\eta_i, \eta_{i+1})^{\mathrm{T}}\right\}$ (i — нечетное);
- Двумерный критерий χ^2 для $\left\{(\xi_i, \xi_{i+1})^{\mathrm{T}}\right\}$ (i — нечетное).

Параметры: $S = 8$, $n = 10^4$, $p = 1/S + 0.01$, $m = 10^3$.

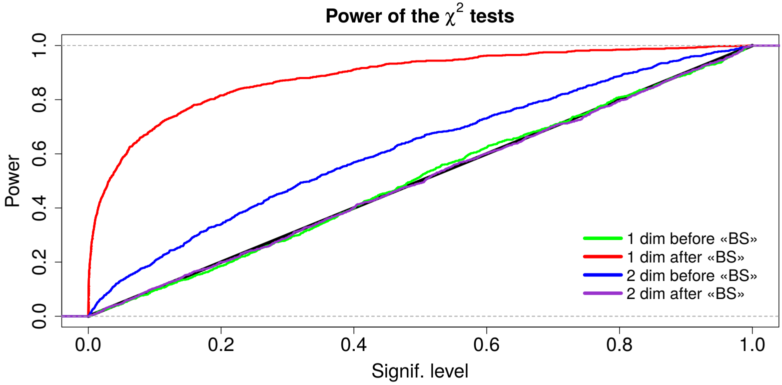


Рис. : Мощности одномерного/двумерного критериев χ^2 до/после «Book Stack».

Выводы

- Если рассматривать в качестве альтернативы независимые и одинаково, но **не равномерно**, распределенные случайные величины η_i, то применение «Book Stack»-теста **не является оправданным**.
- **Перспективным** представляется изучение «Book Stack»-теста для альтернатив, связанных с **зависимостью** η_i.