Отчет по тестовому заданию для проекта Change point detection in CI performance data

Сайфулин Дмитрий

28 сентября 2022

Содержание

1 Введение

Тестовое задание посвящено анализу функций сдвига, в зарубежной литературе их называют `shift function`. Функция свдига идейно очень просто конструируется. Берем две выборки из каких-нибудь распределений, для каждой выборки вычисляем оценку квантилей. Далее считаем разность между каждым квантилем одной выборки и другой.

Обычно, когда идет речь о сравнении двух выборок, люди сразу думают в сторону различных известных тестов, как, например, критерий согласия Колмогорова-Смирнова. Однако стоит отметить, что если критерий оказался значимым для каких-то двух выборок, то мы знаем, что выборки разные, но не знаем, как именно и в чем это выражается. Таким образом, функции сдвига оказываются полезными в данном вопросе.

2 Литература

В ходе работы мне удалось использовать различные источники. Первыми из них был статьи Doksum, K. (1974), Doksum, K.A. (1977), Doksum, K.A. Sievers, G.L. (1976). В них были изложены различные эвристики отображения на графике разности квантилей для двух распределений. Однако основная работа, на которую я опирался была Wilcox, R.R. (1995) Comparing Two Independent Groups Via Multiple Quantiles. В ней Wilcox использовал оценку Harrell-Davis для квантилей ($Q_{HD}(p)$). Оригинальная формула выглядит так:

$$Q_{HD}(p) = \sum_{i=1}^{n} W_i \cdot x_{(i)}$$

$$W_i = I_{i/n}(a,b) - I_{(i-1)/n}(a,b), \text{ где } a = p(n+1), b = (1-p)(n+1)$$

В этой формуле $I_t(a,b)$ – неполная бета-функция, а $x_{(i)}$ это i-тая порядковая статистика. Я посчитал, что линейную интерполяцию использовать довольно скучно, поэтому выбрал этот метод.

3 Работа с данными

В задании было сказано поэкспериментировать с различными выборками из различных распределений. Для всех распределений я генерировал 10^4 наблюдений. Нельзя сказать, что это много, но для симуляции вполне сойдет.

Итак, я рассмотрел 4 случая.

1. Два равномерных распределения с разными отрезками

   Самый базовый и неинтересный случай. Первая выборка из $\mathbb{U}[11, 17]$, вторая из $\mathbb{U}[0, 900]$.

2. Распределение Коши и бета-распределение

   Здесь решил взять более интересные распределения. Первая выборка $\sim \mathrm{C}(0, 0.05)$. Вторая выборка $\sim \mathrm{Beta}(3, 3)$.

3. Два бимодальных распределения Далее я решил поработать с мультимодальными рапсределениями. В университете у меня не было особого опыта в генерации, поэтому гугл помог. Для первой выборки я взял смешанное рапсределение двух нормальных:

   $$0.5\,\mathcal{N}(4, 2) + 0.5\,\mathcal{N}(20, 2)$$

   Для второй выборки я решил подвинуть «купола» друг к друга, поэтому:

   $$0.5\,\mathcal{N}(7, 2) + 0.5\,\mathcal{N}(15, 2)$$

   Генерацию таких распределений в `R` я подсмотрел у австралийского разработчика Brendan Gregg здесь.

   Однако далее я задумался. Генерить долями легко и приятно. Но мне же потом нужно будет доставать «правильные» значения квантилей для таких распределений. В целом можно было реализовать свою функцию, но я решил гуглить. Таким образом, наткнулся на пакет `gendist`. Там есть функция для генерации выборки `rmixt` и для определения квантилей `qmixt`.

4. Два унимодальных скошенных вправо распределения

   Для данного случая я решил взять ассиметричные рапсределения. Для первой выборки это распределение Вейбулла $\mathbb{W}(1, 1.5)$, а для второй экспоненциальное распределение $\mathrm{Exp}(2)$

4 Результаты

4.1 Случай 1

Рассмотрим первый случай, где обе выборки взяты из равномерного распределения.

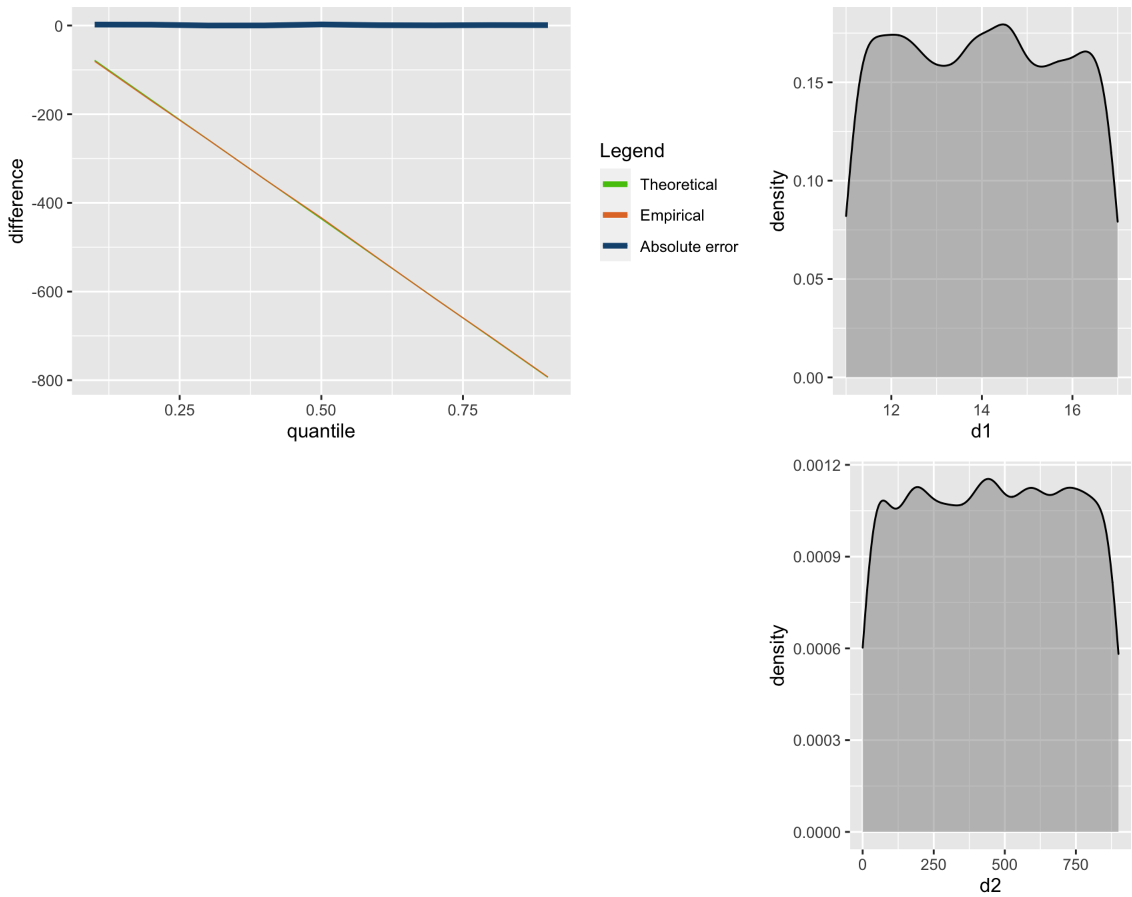

Рис. 1: Плотности равномерных распределений и функция сдвига

Справа на рисунке отображены плотности распределения, чтобы визуально понимать, какие распределения рассматриваются.

На первом графике есть три линии: разность квантилей из распределений, разность оценок квантилей по выборкам (Q_{HD}) и абсолютная разность между двумя разницами. Из-за разномерностей по оси Y довольно плохо видно скачки синей линии абсолютной разности. Однако также стоит отметить, что порядок величины значения квантилей разнится от распределения к распределению. Поэтому далее рассмотрим вариант, который может помочь это нивелировать.

Попробуем проанализировать функции свдига используя *относительные* величины. Для каждого случая построим следующую статистику:

$$\texttt{relative}(p) = \frac{|Q(p) - Q_{HD}(p)|}{|Q(p)|}$$

Далее для каждого случая построим линию, проходящую через каждую точку квантиль-статистика и отметим границу в 1%.

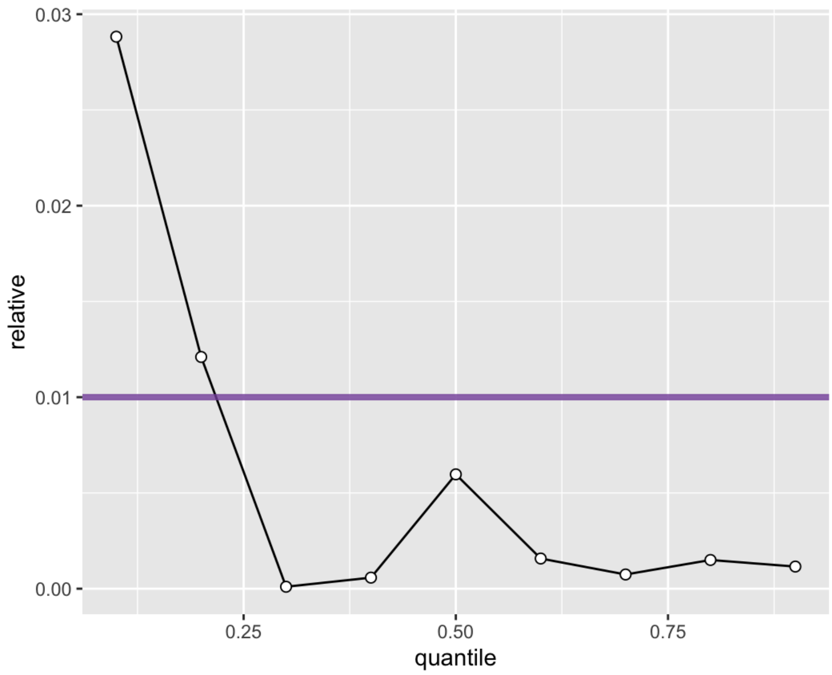

Рис. 2: Сравнение эмпирической и теоретической функции сдвига для двух равномерных распределений

На рисунке 2 видно, что значения статистики для отрезка квантилей $[0.1, 0.2]$ оказались выше 1%, тогда как все остальные отрезки ниже. Однако если бы мы выбрали отметку 0.5%, то медиана тоже бы вылезла из критической области.

Перейдем к следующему случаю.

4.2 Случай 2

Здесь выбраны более хитрые распредления.

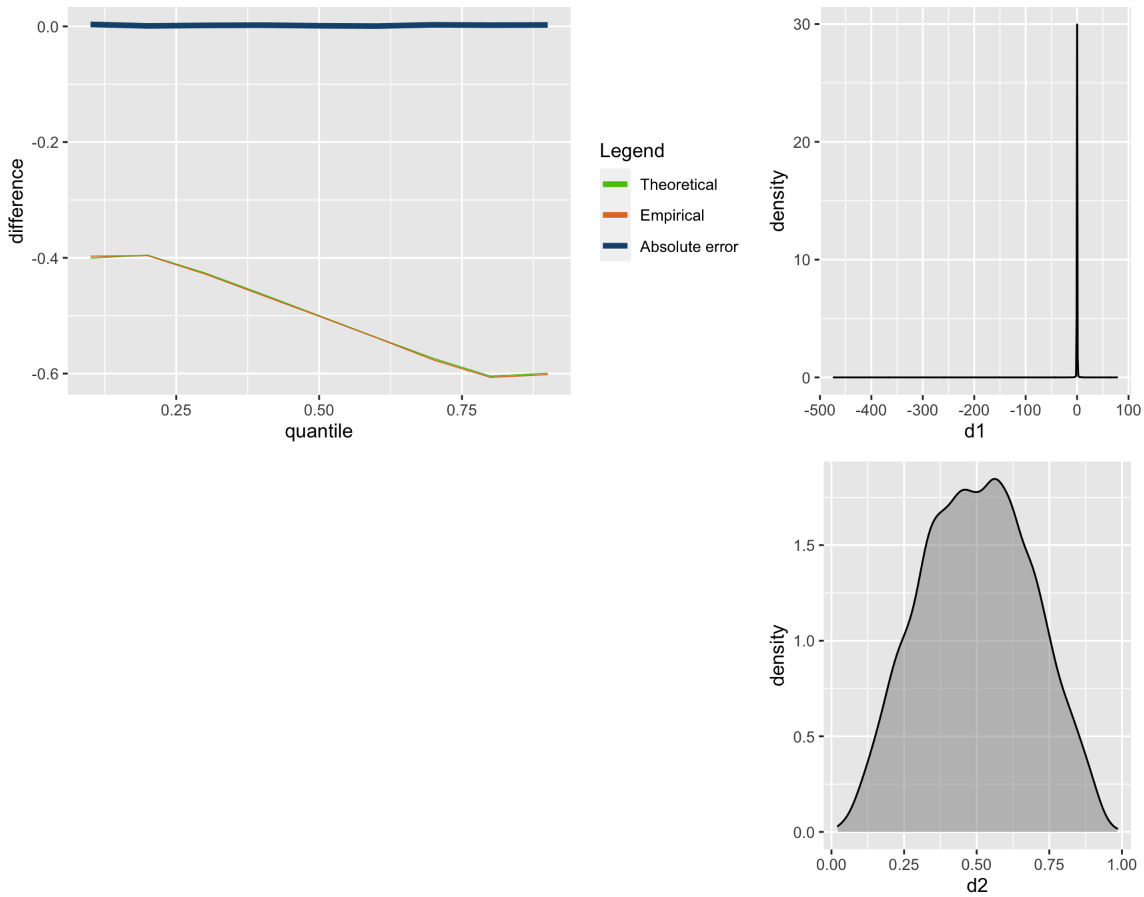

Рис. 3: Плотности распределения Коши и бета-распределения; функция сдвига

Если посмотреть на оранжевую и зеленую линии, то можно видеть, что они переплетаются между собой, словно веревки. В первом случае у нас были почти идентичные линии, здесь же видны скачки. Опять же, из-за масштаба синяя линия не выглядит скачкообразной, однако стоит проанализировать относительные значения.

Здесь график выглядит более интересным, однако все значения оказались ниже 1%, вопреки тому, что мы видели на рисунке 3.

Если же мы посомтрим в сторону критического значения достоверности $= 0.5\%$, то обросится только отрезок $[0.1, 0.12]$.

Однако если выставить отметку в 0.25%, то в «достоверную» зону попадают лишь отрезки квантилей $[0.2, 0.25]$ и $[0.5, 0.6]$.

Идем далее.

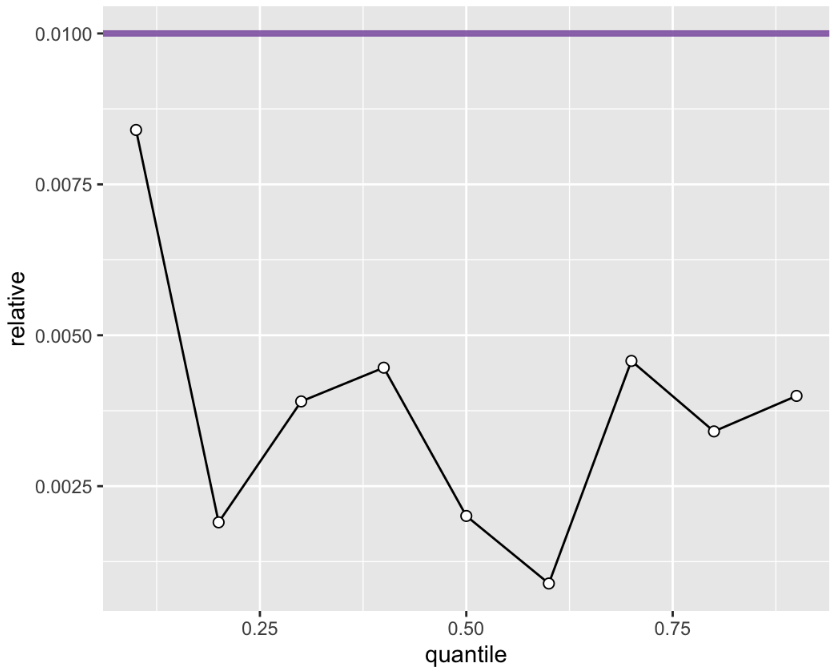

Рис. 4: Сравнение эмпирической и теоретической функции сдвига для коши и бета-распределений

4.3 Случай 3

Здесь мы будем анализировать выборки из бимодальных распределений.

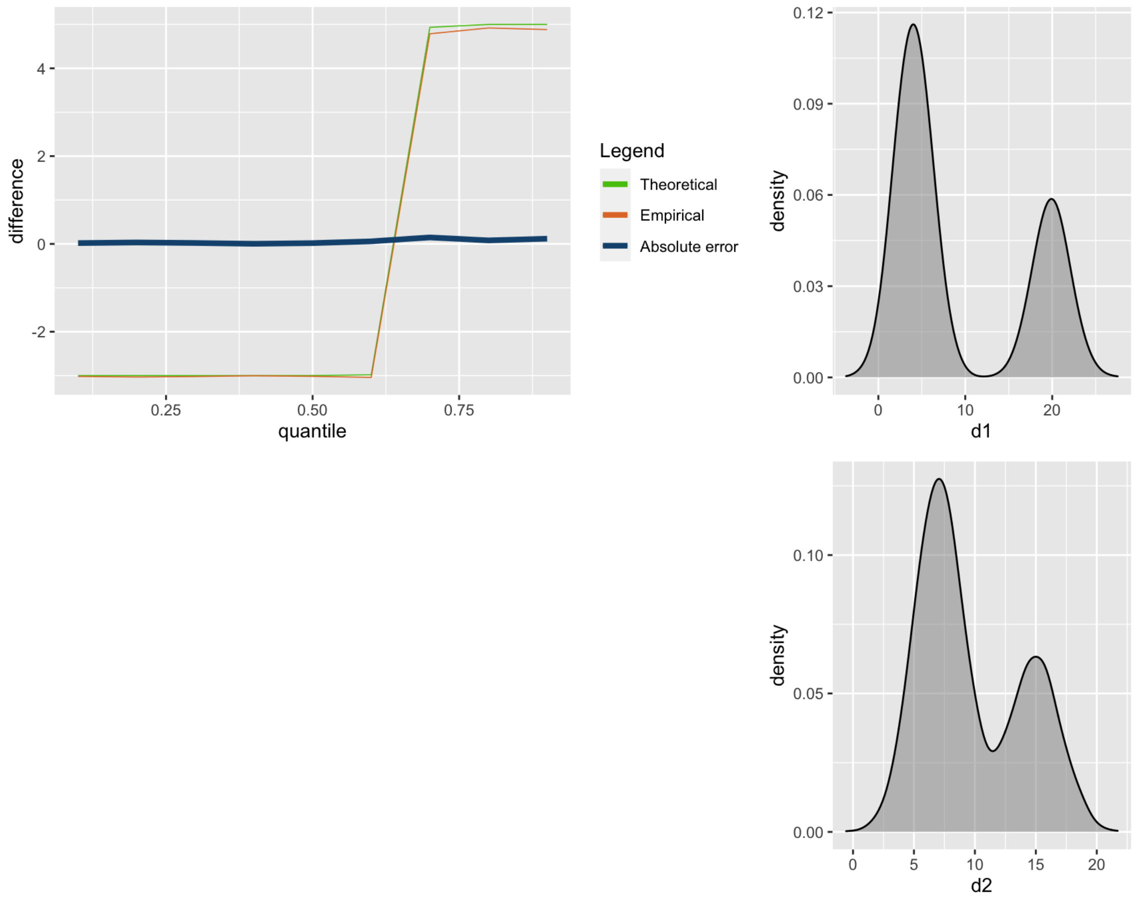

Рис. 5: Плотности смешанных нормальных распределений; функция сдвига

Если бы мы (как неопытные `performance engineer`'ы) увидели такой шифт в распределении (скажем, если бы выкатывали в прод сервис), мы бы незамедлительно начали думать про средние и их свдиг. Однако более опытные коллеги указали бы, что первый график на рисунке 5 показывает следующее: значения большей части квантилей уменьшились на ≈ 3, остальные увеличились на ≈ 5. Таким образом функция сдвига дает сильно больше информации про изменение распределения.

Теперь проанализируем рисунок 6. Для нашей отметки 1% достоверными являются только отрезки квантилей $[0.1, 0.21]$ и $[0.23, 0.55]$.

Если же, как и в прошлых случаях рассматривать значение 0.5%, то достоверным будет только небольшой отрезок возле медианы.

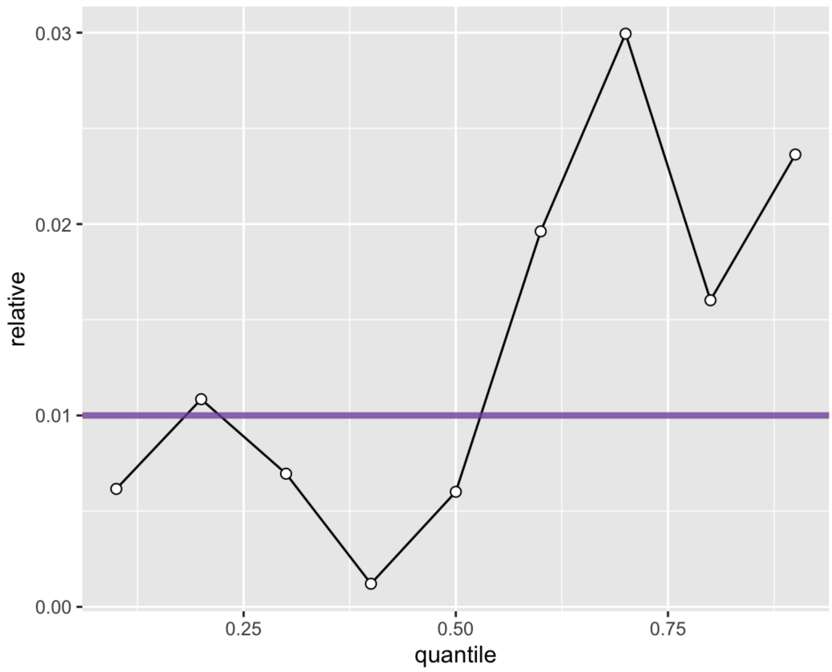

Рис. 6: Сравнение эмпирической и теоретической функции сдвига для коши и бета-распределений

4.4 Случай 4

Здесь в бой пойдут унимодальные распределения, притом еще и `right-skewed`, как любят писать в зарубежной литературе.

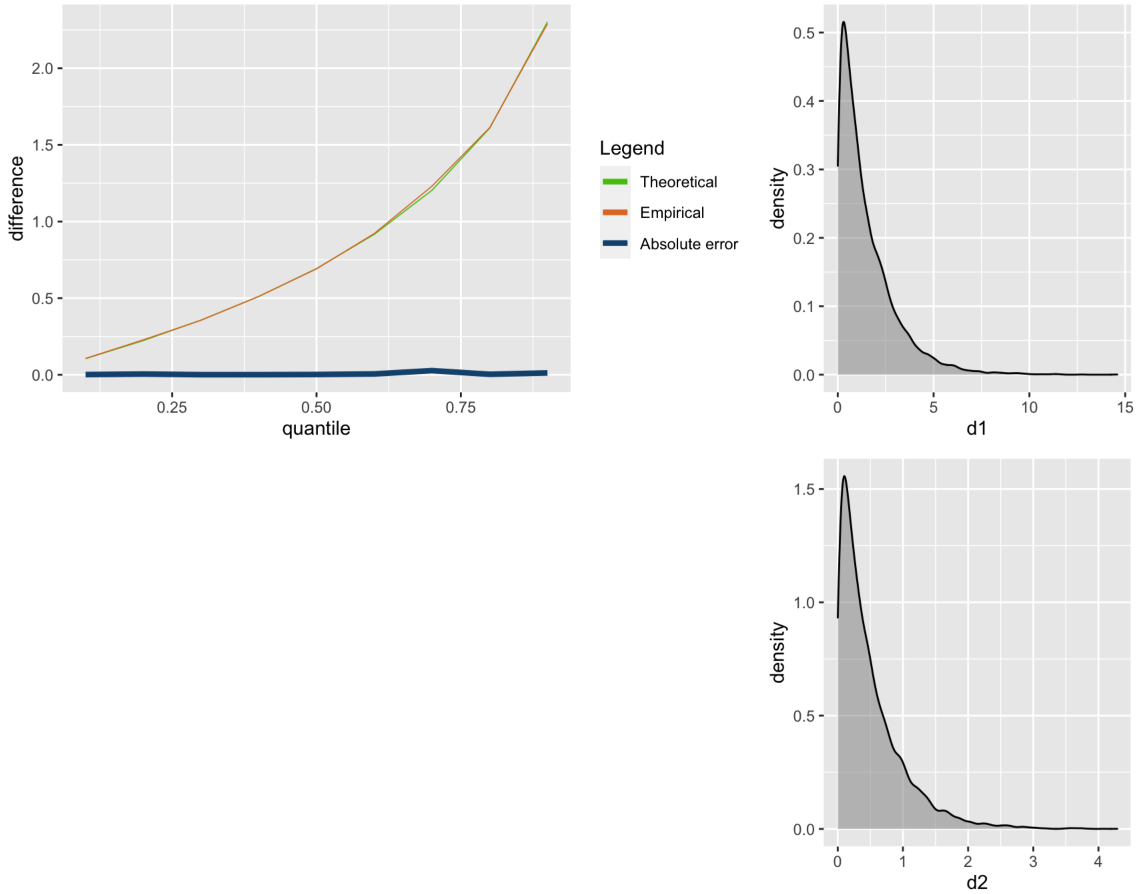

Рис. 7: Плотности распределения Вейбулла и экспоненциального распределения; функция сдвига

Визуально распределения очень похожи. Ощущение, что в реальной жизни мы могли бы такое встретить. Первый график на рисунке 7 можно было бы проинтерпретировать так: значения квантилей увеличивались по мере увеличения значения самого процентиля.

На рисунке 8 мы можем видеть, что в окрестности медианы отрезок квантилей $[0.3, 0.6]$ располагается довольно низко, даже если мы возьмем порог в 0.5%, то этот отрезок все так же будет достовреным.

Однако по краям значения далеки от теоретических.

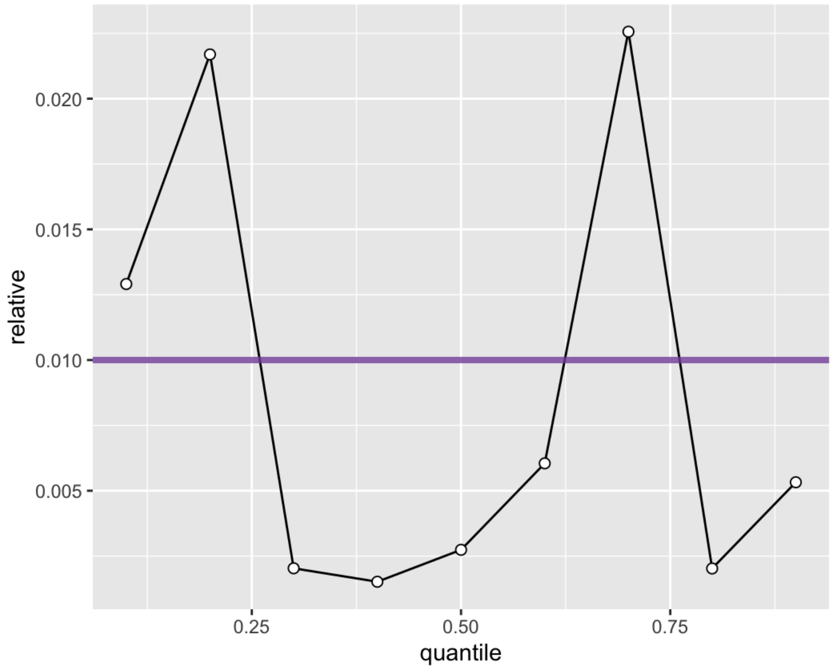

Рис. 8: Сравнение эмпирической и теоретической функции сдвига для распределения Вейбулла и экспоненциального распределения

5 Заключение

В ходе данной работы удалось посмотреть на использование функций сдвига для анализа различий в симуляционных выборках. В некоторых случаях и для некоторых отрезков (в основном в окрестности медианы) значения функции сдвига на симуляционные данных ничуть не уступали теоретическим. Возможно, весомую роль сыграли нетривиальные Harrell-Davis оценки для квантилей.

Предложенный вариант критерия достоверности значений функции сдвига не выглядит исчерпывающим, нуждается в доработке. Также стоит поэкспериментировать с большими выборками: взять $10^5, 10^7, 10^{10} \ldots$ точек.

Мне было приятно поработать над тестовым заданием, я чувстсвую в себе мотивацию и силы разбираться в деталях, анализировать графики, строить выводы. Я бы хотел писать дипломную работу на подобную тему. Еще мне нравится приложение к бизнесу и реальному миру – поработать с тем, как можно применить статистические методы на реальных данных так, чтобы в long run вывести компанию на новый уровень.