

Домашнее задание №1

Выполнил Макаренко Лев

Тема

Применение метода бутстрап для оценки доверительного интервала для ключевого показателя одной из функциональностей онлайн-сервиса.

Описание потребности

В онлайн-сервисе принято решение изменить функционал для регистрации новых пользователей. Оценка эффективности новой версии функционала выполняется по выбранной ключевой метрике - среднему времени регистрации пользователя. Согласно исходной предпосылке, чем меньше пользователь тратит времени на прохождение пути регистрации - тем лучше для него и для бизнеса в целом., поэтому для дальнейшей проверки гипотез относительно изменений указанной метрики, нужно оценить доверительный интервал для метрики.

Указанная метрика рассчитывается как среднее арифметическое времени регистраций пользователей за выбранный период (месяц) по формуле (1):

$$MART = \frac{Reg\ duration}{Users}$$

Где: *Reg duration* - суммарная продолжительность регистраций всех пользователей (в часах), *Users* - кол-во пользователей, начавших и завершивших процесс регистрации за выбранный период (месяц).

Стоит отметить, что ранее доверительный интервал для имеющихся данных уже был оценен, но используя классический метод для оценки доверительного интервала математического ожидания в случае с неизвестной дисперсией, по формуле (2):

$$P(\bar{X} - t_{1-\frac{\alpha}{2}, n-1} \frac{\hat{\sigma}}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{1-\frac{\alpha}{2}, n-1} \frac{\hat{\sigma}}{\sqrt{n}}) = 1 - \alpha$$

Где: \bar{X} - выборочное среднее, $\hat{\sigma}$ - несмещенная оценка стандартного отклонения, μ - оцениваемое (истинное) мат. ожидание, α - уровень значимости, вероятность того, что μ не войдет в интервал, $P = 1 - \alpha$ - уровень доверия, вероятность того, что μ войдет в интервал, n - кол-во наблюдений, $t_{1-\frac{\alpha}{2}, n-1}$ -

квантиль уровня $1 - \frac{\alpha}{2}$ распределения Стьюдента с $n - 1$ степенями свободы для расчета границ интервала.

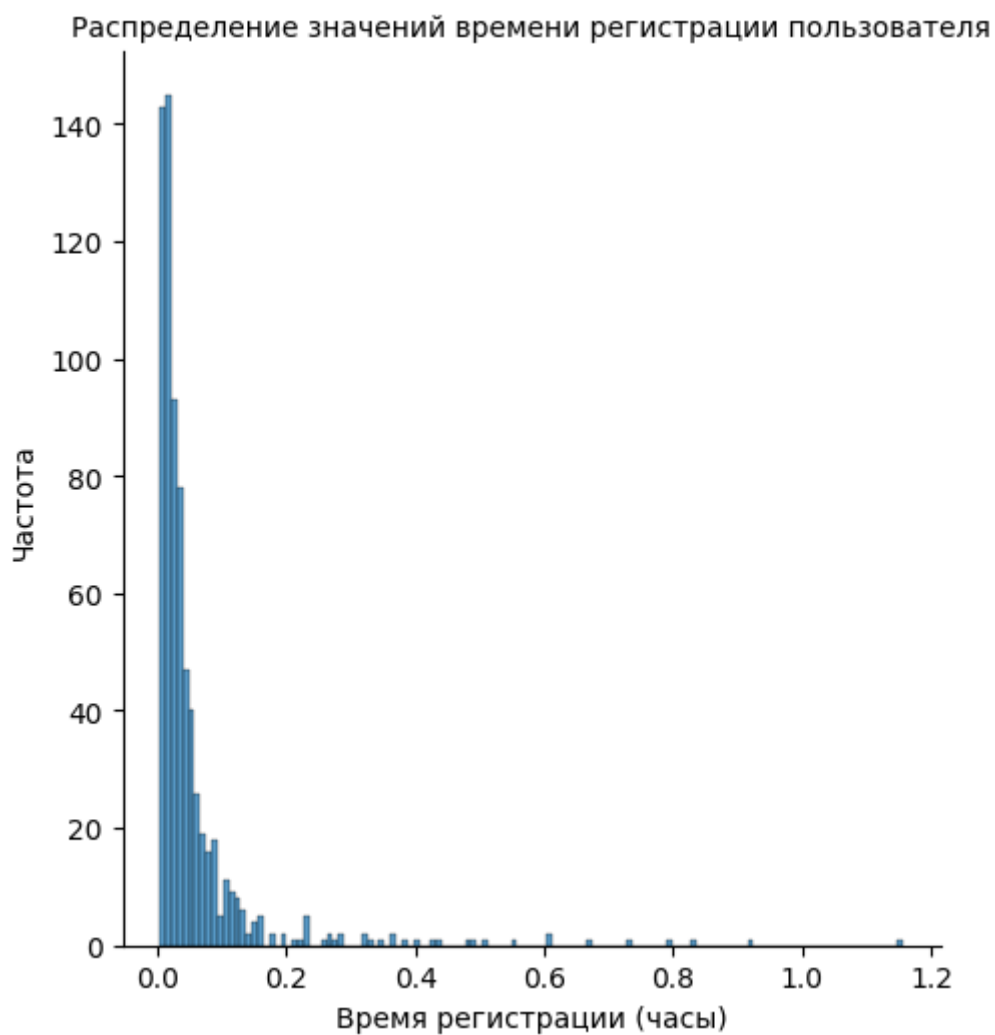
Теперь же, хочется оценить доверительный интервал, используя метод бутстрэпа, т.к. есть мнение, что он более масштабируем - прост в реализации, позволяет решать задачу на порядок быстрее, без существенных потерь в качестве, и что немаловажно - позволит экстраполировать такой подход на оценку доверительных интервалов для множества других метрик, с минимальными временными затратами.

Описание данных

Выборка за месяц для оценки метрики содержит 752 независимых наблюдений, и включает признаки:

Признак	Описание
<i>abonent_id</i>	Уникальный идентификатор пользователя
<i>reg_start_date</i>	Дата и время начала регистрации
<i>reg_end_end</i>	Дата и время окончания регистрации
<i>reg_time_seconds</i>	Продолжительность регистрации (в секундах)
<i>reg_time_minutes</i>	Продолжительность регистрации (в минутах)
<i>reg_time_hours</i>	Продолжительность регистрации (в часах)

Распределение значений времени регистрации (без выбросов):



Оценка доверительного интервала

Результат применения классического метода для оценки 95% доверительного интервала для метрики *MART* по имеющимся данным, по формуле (2):

Дисперсия	Неизвестна
Кол-во наблюдений	752
Уровень доверия ($1 - \alpha$)	0.95
Выборочное среднее	1.3921274443062792
Несмещенная оценка стандартного отклонения	12.815044593835925
Квантиль уровня 0.975 распределения Стьюдента с 751 степенью свободы	1.9631278115521518

Нижняя граница интервала	$0.4747252444139203 \approx 0.47$ часа или 28.5 минут
Верхняя граница интервала	$2.309529644198638 \approx 2.31$ часа

Результат применения метода бутстрэпа для оценки 95% доверительного интервала для метрики *MART* по имеющимся данным:

Кол-во наблюдений	752
Кол-во итераций	1000
Метод	Наивный бутстреп
Функция	Среднее арифметическое
Размер подвыборки	752
Нижняя граница интервала (квантиль уровня 0.025)	$0.591238555999076 \approx 0.59$ часа или 35.47 минут
Верхняя граница интервала (квантиль уровня 0.975)	$2.4481278177061645 \approx 2.45$ часа

Выводы

В целом, результаты¹ оценивания 95% доверительного интервала для среднего времени регистрации пользователя двумя методами похожи, однако оценка бутстрэпа ([35.47 минут; 2.45 часа]) оказалась немного смещена к большим значениям, относительно оценки классическим методом ([28.5 минут; 2.31 часа]).

Есть предположение, что наблюдаемая разница в оценках связана с асимметричностью распределения значений в исходной выборке (наличием выбросов), т.к. исходное распределение значений смещено вправо - имеет длинный правый хвост. Если исходную выборку предварительно очистить от выбросов, тогда результаты оценки двумя методами будут очень близки.

Также возможно стоит учитывать природу распределения исследуемой величины, она может иметь распределение отличное от нормального, в нашем случае оно больше напоминает экспоненциальное распределение, и для него оценка используемым нами классическим методом менее точна, чем оценка методом бутстрэпа, поскольку использованный нами классический метод основан на предположении об асимптотической нормальности распределения (t-статистика).

¹ Оценка интервалов выполнялся в сервисе Google Colab - https://colab.research.google.com/drive/1m-rQjPM3i7_WI55vU9O08kjTnUQrh3Q1

Также можно предварительно преобразовать величину (например логарифмирование), получив распределение больше похожее на нормальное, а уже после применить к ней классический метод для оценки доверительного интервала среднего (t -статистика). При таком подходе оценка преобразованных данных методом бутстрэпа будет близка к оценке классическим методом.