ТПОЭ - 23/24

Лекция 7

Нерсес Багиян

Смотрим на три типа метрик

Две из них уже свели к нормальному распределению

Средняя

- User Average Metrics (ARPU/ARPPU/etc)

Ratio

- User-level Conversion Metrics (Retention / etc)
- Page-level Conversion Metrics (Global CTR / etc)

Квантиль

- Ну тут просто квантиль (.99 latency / перцентиль чека)

Абсолюты

- Метрики (GMV / Выручка / Просмотры)

Зависимые данные ломают t-test

При наличии зависимых данных оценка дисперсии в тесте Стьюдента ломается. Это приводит к повышению вероятности ошибки первого рода

Пример:

Мы в нашем магазине мыла проводим эксперимент, чтобы понять увеличивают ли наши рекомендации средней чек. У нас есть разные юзеры, некоторые из них могут купить несколько раз за эксперимент. **Данные в выборке в таком случае являются зависимыми**

Среднее средних это не путь

Сонаправленность не соблюдается

При использовании среднего средних мы не обеспечиваем сонаправленность изменений. По сути, это означает, что в любом тесте мы можем сказать, что изменение есть, хотя его нет или наоборот

А что если есть способ подобрать такую агрегацию, чтобы все работало?

Линеаризация

Линеаризация - поюзерная прокси-метрика, показывающая отклонение от поведения "усредненного" поведения контрольной группы

Пример:

Мы в нашем магазине мыла проводим эксперимент, чтобы понять увеличивают ли наши рекомендации средней чек. У нас есть разные юзеры, некоторые из них могут купить несколько раз за эксперимент. **Как бы работала линеаризация в данном случае?**

Линеаризация на примере

Как выглядит среднее поведения юзеров в контроля?

Группа А

Пользователь 1

2 покупки по 1000 рублей

Пользователь 2

1 покупка по 400 рублей

Средний чек: 800 рублей

Группа Б

Пользователь 1

2 покупки по 400 рублей

Пользователь 2

1 покупка по 1300 рублей

Средний чек: 700 рублей

А как выглядит каждый пользователь относительного этого поведения?

Вопрос аудитории

Группа А

Пользователь 1

2 покупки по 1000 рублей

Пользователь 2

1 покупка по 400 рублей

Средний чек: 800 рублей

Группа Б

Было	Стало
Пользователь 1 2 покупки по 400 рублей	Пользователь 1
Пользователь 2 1 покупка по 1300 рублей	Пользователь 2

А как выглядит каждый пользователь относительного этого поведения?

Посчитали по группе Б

Группа А

Пользователь 1

2 покупки по 1000 рублей

Пользователь 2

1 покупка по 400 рублей

Средний чек: 800 рублей

Группа Б

Было	Стало
Пользователь 1 2 покупки по 400 рублей	Пользователь 1 800 - 2 * 800 = -800
Пользователь 2 1 покупка по 1300 рублей	Пользователь 2 1300 - 1 * 800 = 500

А как выглядит каждый пользователь относительного этого поведения?

Теперь по группе А

Группа А

Было	Стало
Пользователь 1 2 покупки по 1000 рублей	Пользователь 1 2000 - 2 * 800 = 400
Пользователь 2 1 покупка по 400 рублей	Пользователь 2 400 - 1 * 800 = -400

Группа Б

Пользователь 1

800 - 2 * 800 = -800

Пользователь 2

1300 - 1 * 800 = 500

Что мы получили?

Вопрос аудитории

Группа А

Пользователь 1

2000 - 2 * 800 = 400

Пользователь 2

400 - 1 * 800 = -400

Среднее по юзерам: 0

Группа Б

Пользователь 1

800 - 2 * 800 = -800

Пользователь 2

1300 - 1 * 800 = 500

Среднее по юзерам: -150

Линеаризация

Линеаризация - поюзерная прокси-метрика, показывающая отклонение от поведения "усредненного" пользователя контрольной группы

Группа А

Было $$R_C = \frac{\sum_{u \in C} X(u)}{\sum_{u \in C} Y(u)}$$

Стало $$L_C = \frac{1}{|\{u : u \in C\}|} \sum_u l(u)$$

$$l(u) = X(u) - kY(U) \qquad k = R_C$$

Группа Б

Было $$R_C = \frac{\sum_{u \in T} X(u)}{\sum_{u \in T} Y(u)}$$

Стало $$L_T = \frac{1}{|\{u : u \in T\}|} \sum_u l(u)$$

$$l(u) = X(u) - kY(U) \qquad k = R_C$$

Линеаризация

Линеаризация - поюзерная прокси-метрика, показывающая отклонение от поведения “усредненного” пользователя контрольной группы

Группа А

Было $$R_C = \frac{\sum_{u \in C} X(u)}{\sum_{u \in C} Y(u)}$$

Стало $$L_C = \frac{1}{U_C} \sum_u l(u)$$

$$l(u) = X(u) - kY(U) \qquad k = R_C$$

Группа Б

Было $$R_T = \frac{\sum_{u \in T} X(u)}{\sum_{u \in T} Y(u)}$$

Стало $$L_T = \frac{1}{U_T} \sum_u l(u)$$

$$l(u) = X(u) - kY(U) \qquad k = R_C$$

Reminder: свойства прокси метрик

Прокси метрика - косвенная метрика целевой метрики, с которой она сильно коррелирует и с которой есть причинно-следственная связь.

Корреляция

Высоко скоррелирована с основной метрикой

Каузальная связь

Имеет причинно-следственную связь с целевой метрикой

Интерпретируемая

Метрику может объяснить каждый сотрудник

Можно заменить на “Сонаправленность”

Чувствительная

Насколько долго нужно ждать, чтобы увидеть изменения в метрике

Достоверная

Можно ли получить точное подтверждение из данных?

Является ли метрика сонаправленной и чувствительной?

Прокси метрика - косвенная метрика целевой метрики, с которой она сильно коррелирует и с которой есть причинно-следственная связь.

Сонаправленная

Изменение прокси-метрики имеет такой же знак и эффект, как и основная метрика

Чувствительная

Насколько долго нужно ждать, чтобы увидеть изменения в метрике

Интерпретируемая

Метрику может объяснить каждый сотрудник

Достоверная

Можно ли получить точное подтверждение из данных?

Что значит сонаправленность?

Пусть у нас есть поюзерные метрики $X(u)$ и $Y(u)$, R наша ratio метрика и L линеаризованная метрика, а параметр $k = R_C$. Тогда $\Delta(L) = Y_T\Delta(R)$, где:

- $Y_T = \dfrac{1}{|\{u : u \in T\}|}\sum_{u \in T} Y(u)$
- $\Delta(L) = L_C - L_T$
- $\Delta(R) = R_C - R_T$

Сонаправленность линеаризации (1/4)

Чтобы это доказать, нам нужно показать, что:

$$\Delta(L) = Y_T \Delta(R)$$

Давайте распишем $\Delta(L)$ по определению:

$$\Delta(L) = L_C - L_T = \frac{1}{U_C} \sum_{u \in C} X(u) - R_C Y(u) - \frac{1}{U_T} \sum_{u \in T} X(u) - R_C Y(u)$$

Вынесем $Y(u)$ из под знаков суммирования:

$$\Delta(L) = \frac{1}{U_C} \sum_{u \in C} X(u) - R_C \frac{1}{U_C} \sum_{u \in C} Y(u) - \frac{1}{U_T} \sum_{u \in T} X(u) - R_C \frac{1}{U_T} \sum_{u \in T} Y(u)$$

Сонаправленность линеаризации (2/4)

Вынесем $Y(u)$ из под знаков суммирования:

$$\Delta(L) = \frac{1}{U_C}\sum_{u\in C} X(u) - R_C\frac{1}{U_C}\sum_{u\in C} Y(u) - \frac{1}{U_T}\sum_{u\in T} X(u) - R_C\frac{1}{U_T}\sum_{u\in T} Y(u)$$

Введем обозначения, чтобы не тащить формулу дальше:

- $X_C = \frac{1}{U_C}\sum_{u\in C} X(u)$
- $X_T = \frac{1}{U_T}\sum_{u\in T} X(u)$
- $Y_C = \frac{1}{U_C}\sum_{u\in C} Y(u)$
- $Y_T = \frac{1}{U_T}\sum_{u\in T} Y(u)$

Сонаправленность линеаризации (3/4)

Получаем формулу более приятную:

$$\Delta(L) = X_C - R_C * Y_C - X_T + R_C * Y_T$$

Сгруппируем слагаемые:

$$\Delta(L) = (X_C - X_T) - R_C(Y_C - Y_T)$$

Разложим $R_C = \frac{X_C}{Y_C}$**:**

$$\Delta(L) = (X_C - X_T) - \frac{X_C}{Y_C}(Y_C - Y_T)$$

Сонаправленность линеаризации (4/4)

Разложим $R_C = \frac{X_C}{Y_C}$**:**

$$\Delta(L) = (X_C - X_T) - \frac{X_C}{Y_C}(Y_C - Y_T)$$

Раскрываем скобки:

$$\Delta(L) = X_C - X_T - X_C + \frac{X_C Y_T}{Y_C} = -X_T + \frac{X_C Y_T}{Y_C}$$

Выносим Y_T**:**

$$\Delta(L) = Y_T(\frac{X_C}{Y_C} - \frac{X_T}{Y_T}) = Y_T \Delta(R)$$

От эффекта линеаризации к эффекту ratio метрики

Для того, чтобы получить эффект метрики, за которой мы наблюдаем надо воспользоваться следующей формулой:

$$\Delta(R) = \frac{1}{Y_T}\Delta(L)$$

От эффекта линеаризации к эффекту ratio метрики

Для того, чтобы получить эффект метрики, за которой мы наблюдаем надо воспользоваться следующей формулой:

$$\Delta(R) = \frac{1}{Y_T}\Delta(L)$$

Пример:

Мы в нашем магазине мыла проводим эксперимент, чтобы понять увеличивают ли наши рекомендации средней чек. У нас есть разные юзеры, некоторые из них могут купить несколько раз за эксперимент. **Как бы работала линеаризация в данном случае?**

Корректируем эффект линеаризованной метрики

Группа А

Пользователь 1

2000 - 2 * 800 = 400

Пользователь 2

400 - 1 * 800 = -400

Среднее по юзерам: 0

Группа Б

Пользователь 1

800 - 2 * 800 = -800

Пользователь 2

1300 - 1 * 800 = 500

Среднее по юзерам: -150

Как придти к эффекту?

Вопрос аудитории

Корректируем эффект линеаризованной метрики

Делим на среднее Y по группе Б и эффекты совпадут

Группа А

Пользователь 1

2000 - 2 * 800 = 400

Пользователь 2

400 - 1 * 800 = -400

Среднее по юзерам: 0

Группа Б

Пользователь 1

800 - 2 * 800 = -800

Пользователь 2

1300 - 1 * 800 = 500

Среднее по юзерам: -150 / 1.5 = - 100

Чем будем тестировать линеаризованную метрику?

Вопрос аудитории

T-test для линеаризованной метрики

Формулируем нулевую гипотезу

Сформулировали в прошлый раз:

$$H_0 : \theta = R_C - R_T = 0$$
$$H_1 : \theta = R_C - R_T \neq 0$$

В нашем случае:

$$H_0 : \theta = L_C - L_T = 0$$
$$H_1 : \theta = L_C - L_T \neq 0$$

T-test для линеаризованной метрики

Записываем тест статистику

Запишем тест статистику T нашего теста в случае ratio метрик:

$$T = \frac{L_C - L_T}{\sqrt{\frac{\sigma^2(L_C)}{N_C} + \frac{\sigma^2(L_T)}{N_T}}} \sim N(0,1)$$

Алгоритм применения линеаризация

Пишем пошагово

Пусть у нас есть метрики $X(u)$, $Y(u)$ и наша ratio метрики R для каждой из групп. Тогда алгоритм применения линеаризации выглядит следующим образом:

1. Для каждого юзера считаем $l(u) = X(u) - kY(U)$, $k = R_C$
2. Считаем линеаризованную метрики L_C и L_T
3. Считаем t-статистику и соответствующий ей p-value
4. Принимаем решение на основе p-value и значению α

Как думаете какую формулу для расчета выборки можем применять тут?

Вопрос аудитории

Мы можем пользоваться той же формулой

Потому что свели все к t-test и нормальному распределению

$$n \geq \frac{2(F^{-1}(1 - \frac{\alpha}{2}) - F^{-1}(\beta))^2 s^2}{MDE^2}$$

Переходим к квантилям

Осталось разобраться только с ним

Средняя

- User Average Metrics (ARPU/ARPPU/etc)

Ratio

- User-level Conversion Metrics (Retention / etc)
- Page-level Conversion Metrics (Global CTR / etc)

Квантиль

- Ну тут просто квантиль (.99 latency / перцентиль чека)

Абсолюты

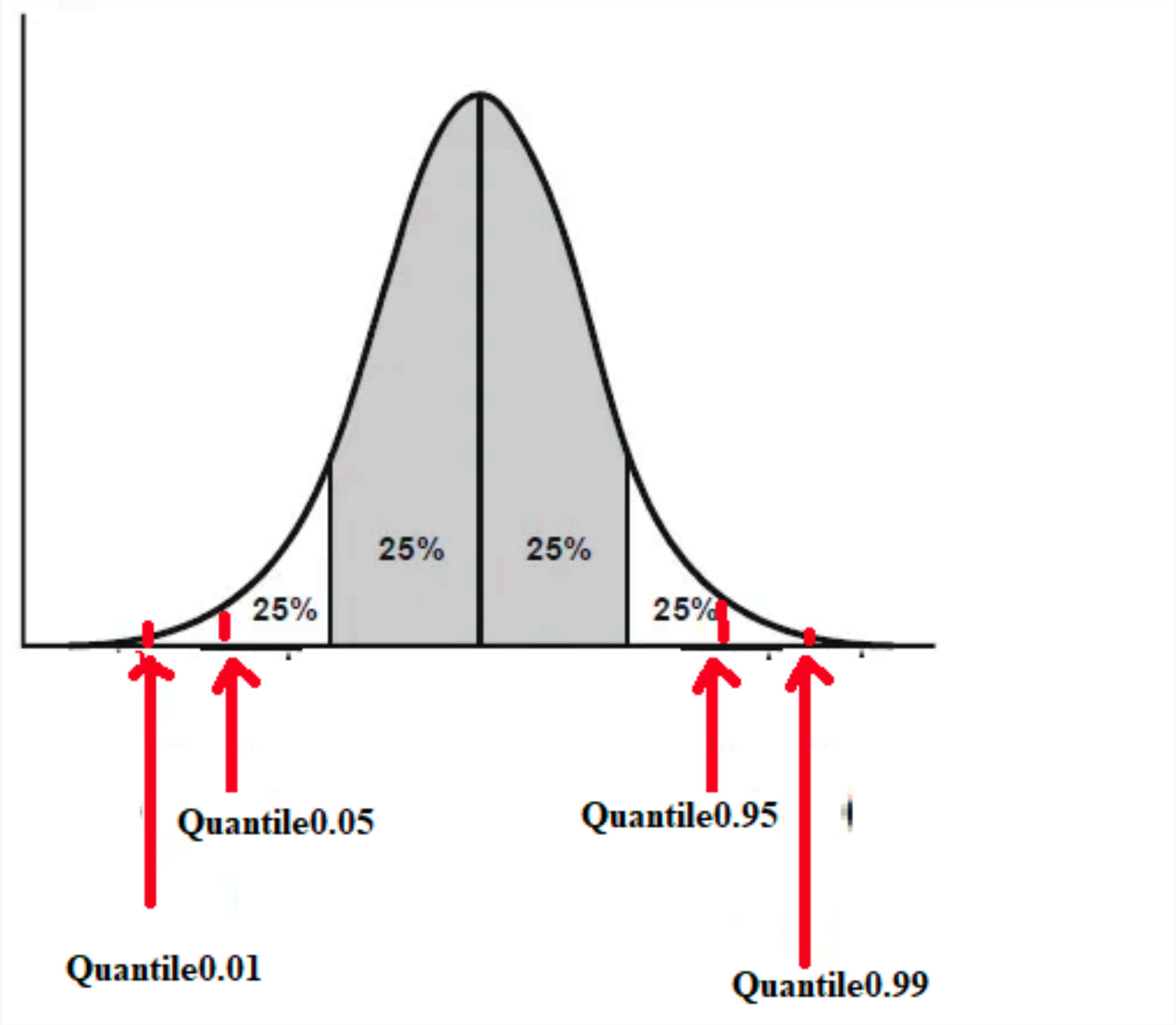
- Метрики (GMV / Выручка / Просмотры)

Reminder: что такое квантиль?

Определение из математической статистики

Квантиль - α квантилем называют такую величину x, что:

$$P(X \leq x) = \alpha \text{ или } F(x) = \alpha$$



Зачем нам его тестировать в продукте?

Определение из математической статистики

Есть несколько причин почему это делают:

- Измерение перформанса продукта - мы хотим убедиться, что большинство страниц загружаются в приемлемое время
- Мы хотим следить не только за средними, но и за распределением в целом

Эксклюзив для самых лояльных

Хорошо сформулированная гипотеза:

1. Предпосылка

   Опросы лояльных клиентов выявили интерес к эксклюзивным продуктам и персонализированным предложениям.Возможность Создать более персонализированные и визуально привлекательные рекламные баннеры, которые будут соответствовать интересам и предпочтениям целевой аудитории, основываясь на их предыдущем поведении на сайте и покупках.

2. Кого коснется

   Программа коснется лояльных клиентов магазина, которые совершают покупки регулярно и имеют историю покупок выше среднего.

3. Мотивация

   Клиенты получат доступ к эксклюзивным продуктам и персонализированным предложениям, что сделает их покупательский опыт более индивидуализированным и премиальным.

4. Эффект, который мы ожидаем

   Ожидается, что внедрение таких рекомендаций улучшит показатель суммы заказа на 10 процентов

Как убедиться, что поменялось не только среднее?

Вопрос аудитории

Растим ARPU с помощью рекомендаций аксессуаров

1. Формулировка гипотезы с сформулированным ожидаемым размером эффекта

   Предложение мыла с подборкой популярных аксессуаров на основе анализа предпочтений покупателей и данных о самых продаваемых товарах увеличит средний доход на пользователя (ARPU) на 20%.

2. Описание аудитории

   Покупатели онлайн-магазина мыла, включая как новых, так и возвращающихся пользователей.

3. Описание вариантов с размером каждой группы

   Контрольная группа (A): Покупателям предлагается стандартный ассортимент без акцентов на комплекты.
   Экспериментальная группа (B): Покупателям активно предлагаются комплекты мыла с популярными ароматами на главной странице и в разделе рекомендаций.
   Размер каждой группы составляет 50% от общего числа посетителей в период эксперимента.

4. Ожидаемые исходы и метрики

   Основная метрика: Увеличение **среднего чека, .75 квантиля, .1 квантиля**
   Контрметрика: Не падение **конверсии в покупку**

5. Продолжительность

   Эксперимент продлится 4 недели, чтобы собрать достаточно данных для статистически значимых результатов, учитывая недельные колебания трафика и поведения покупателей.

6. Результаты

   TBD

Пример дизайн дока: https://docs.google.com/document/d/1y1uawhMsQ3tyZ8dl8W9jDmOWMICPnZhJnsnayt9UX5c/edit?usp=sharing

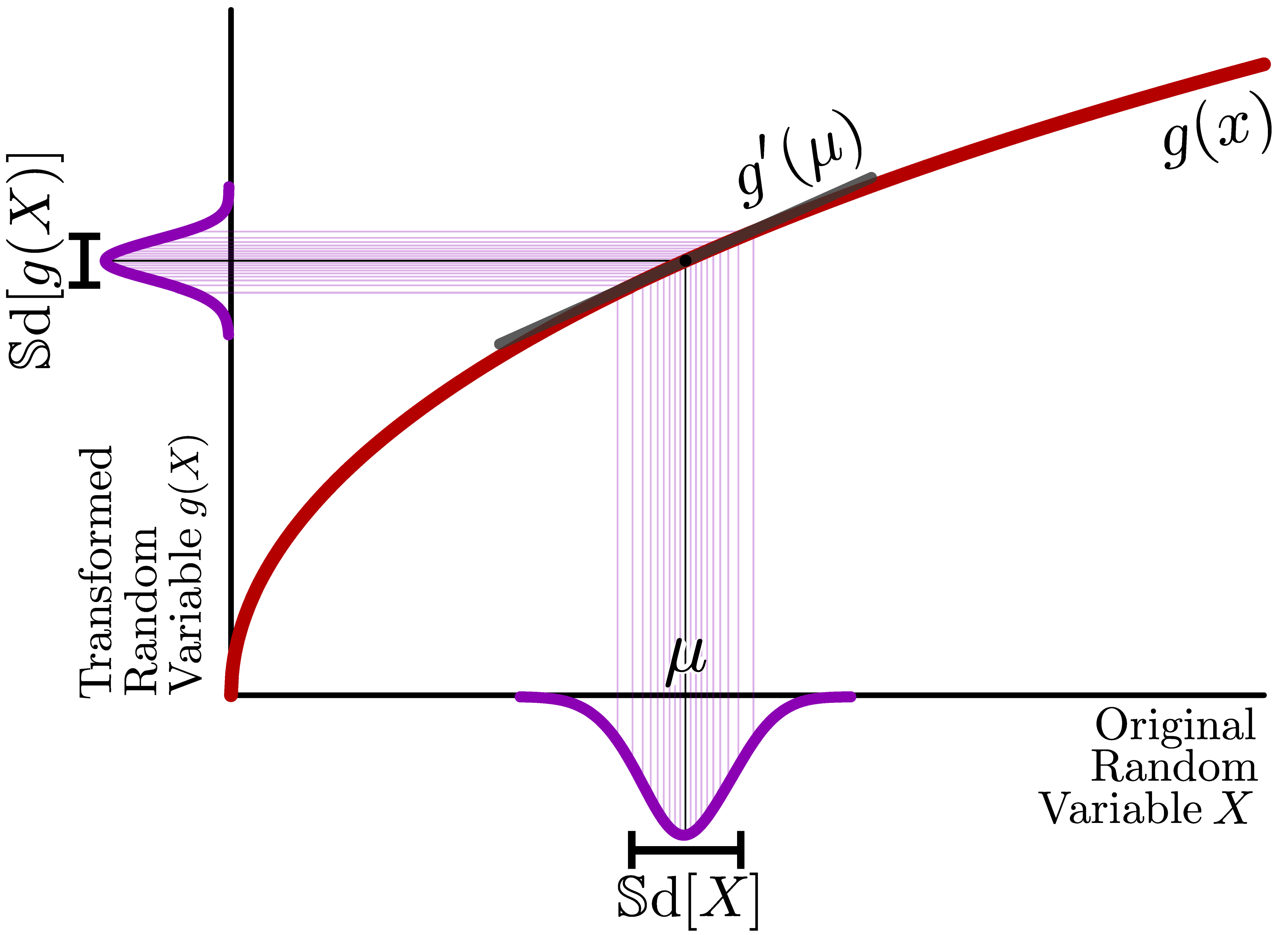
Как будем его тестировать?
Напоминание про дельта-метод

Если существует последовательность случайных величин X_n, удовлетворяющая:

$$\sqrt{n}(X_n - \theta) \to^D N(0,\sigma^2)$$

где σ^2 и θ - конечные константы,
а D обозначает сходимость по распределению, то верно:

$$\sqrt{n}(g(X_n) - g(\theta)) \to^D N(0,\sigma^2[g'(\theta)]^2)$$



Как будем его тестировать?
Можем им воспользоваться и здесь?

Пусть у нас есть квантиль уровня p выборки размером n, есть истинное, ненаблюдаемое значение квантиля $F^{-1}(p)$, тогда согласно дельта - методу:

$$\sqrt{n}(X_{\lfloor np \rfloor} - F^{-1}(p)) \to^{D} N(0, \frac{\sigma^2}{f(F^{-1}(p))^2})$$

Как будем его тестировать?

Тогда статистика будет выглядеть классическим для дельта метода образом

Получается можем пользоваться нашей гипотезой?

$$H_0 : \theta = Q_C - Q_T = 0$$
$$H_1 : \theta = Q_C - Q_T \neq 0$$

И записать статистику:

$$T = \frac{Q_C - Q_T}{\sqrt{\frac{\sigma_C^2}{f_C^2(F_C^{-1}(p))N_C} + \frac{\sigma_T^2}{f_T^2(F_T^{-1}(p))N_T}}} \sim N(0,1)$$

Как думаете, в чем проблема?

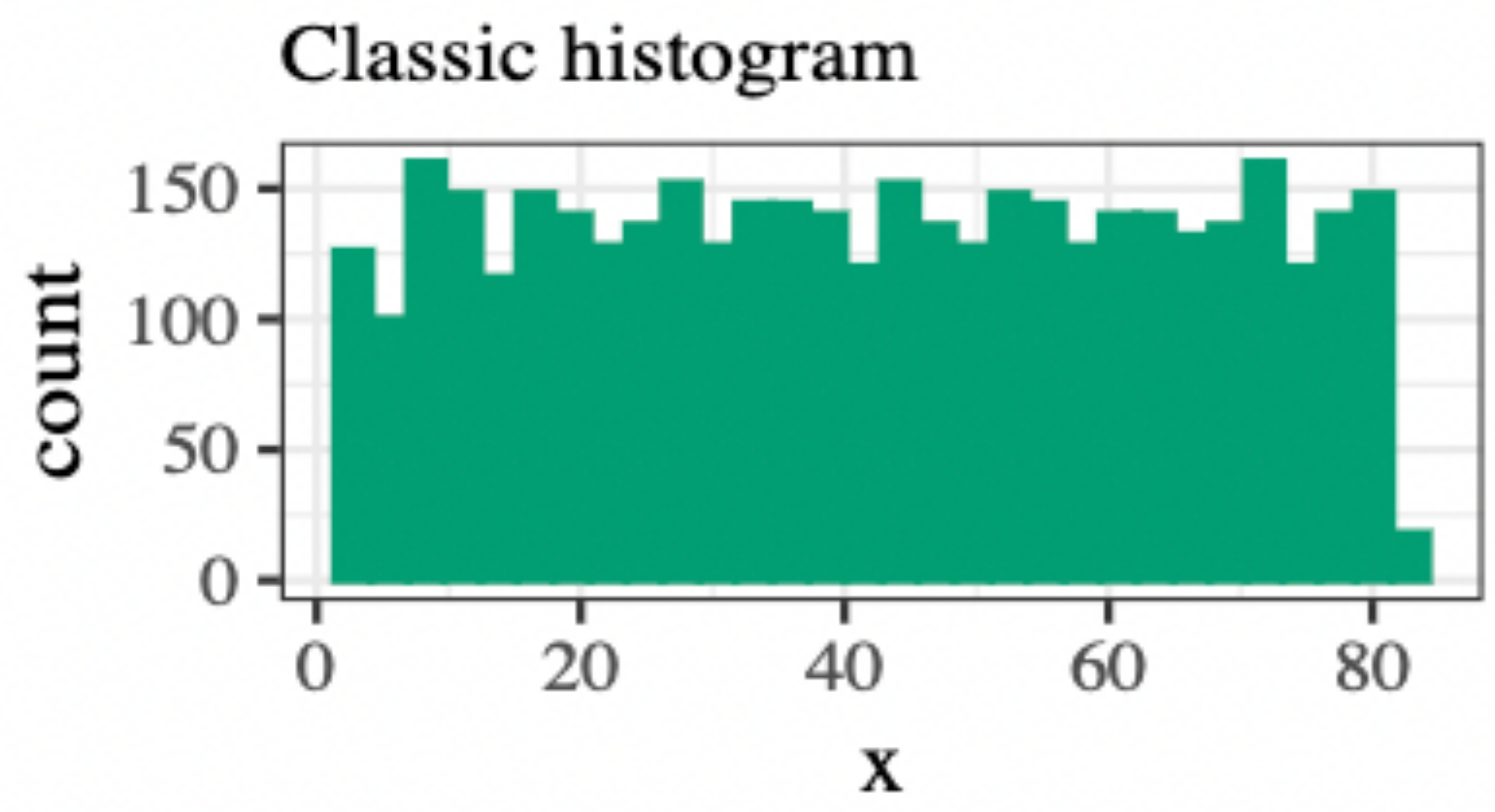
Вопрос аудитории

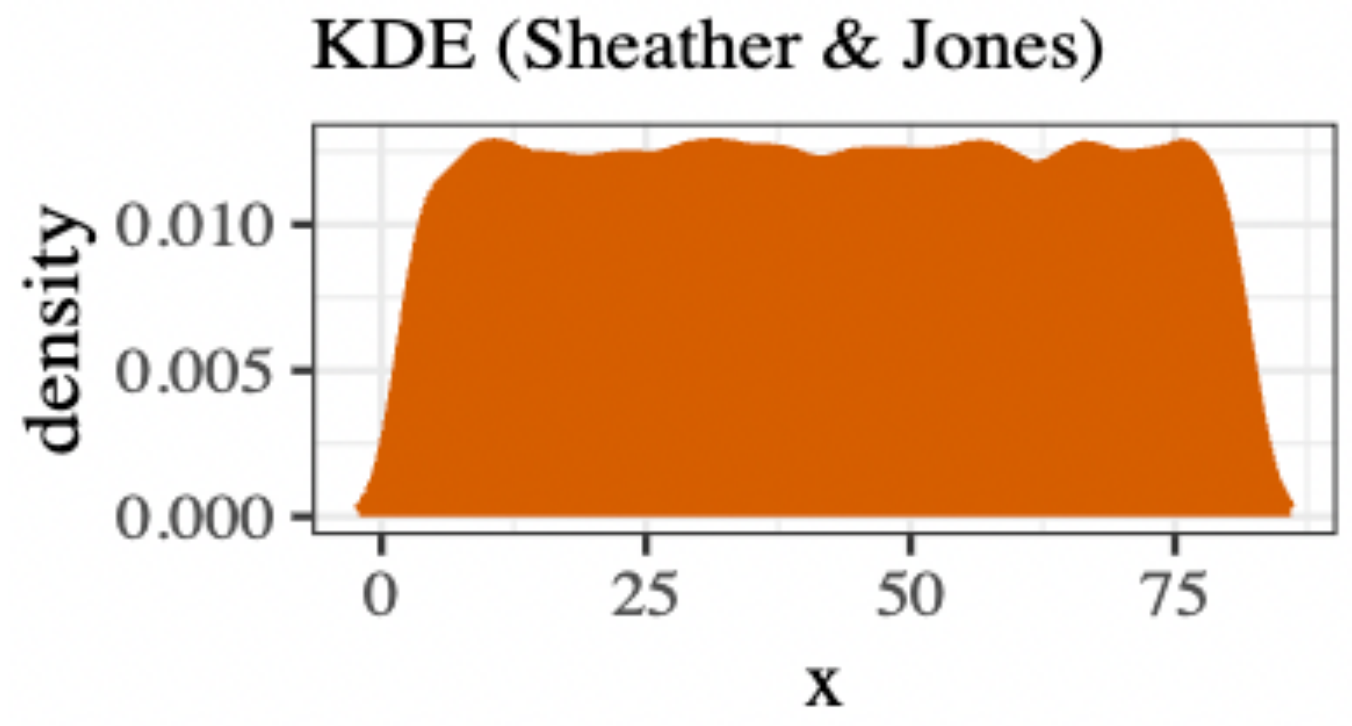
Как будем оценивать плотность?

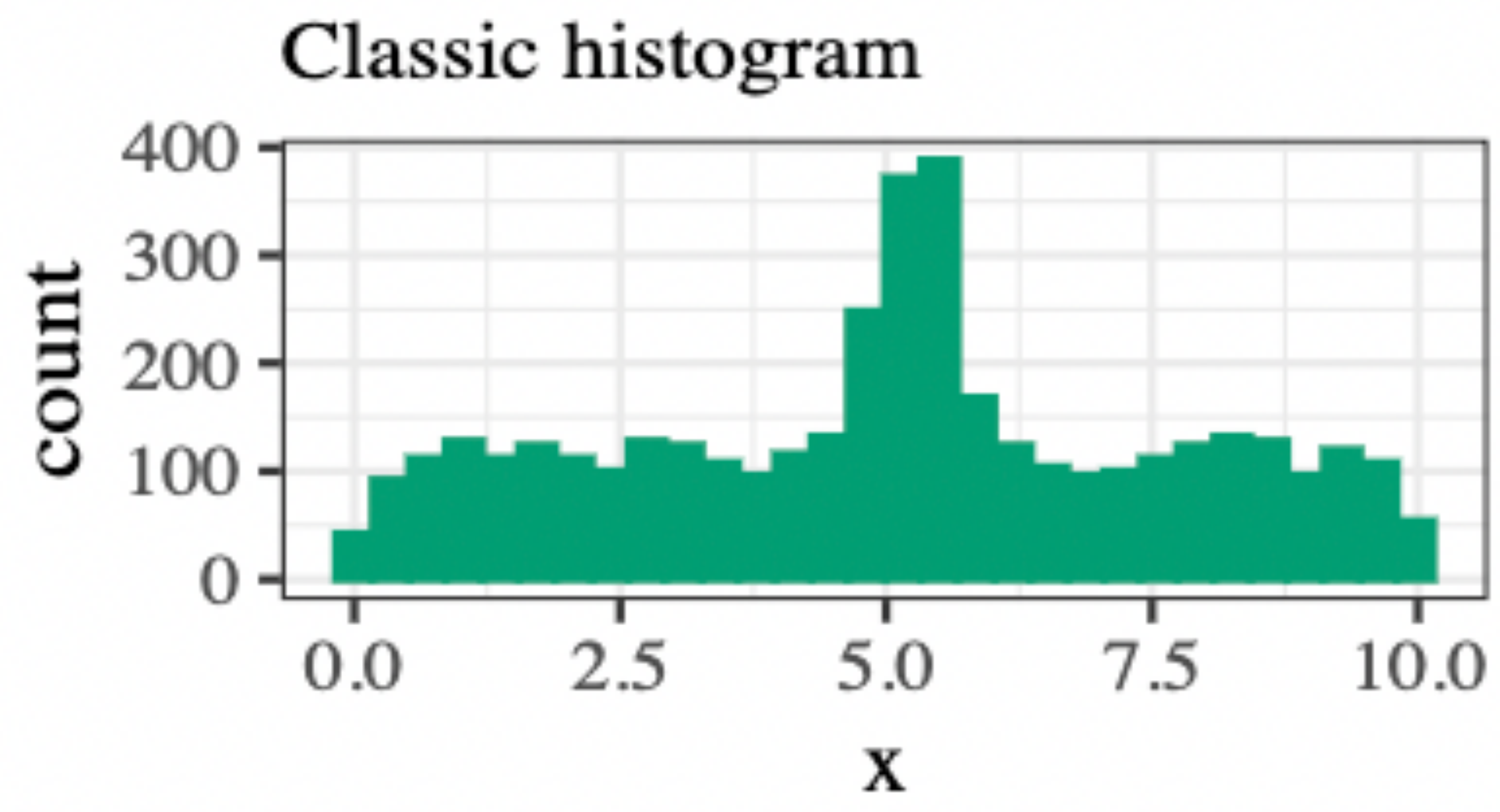
Вопрос аудитории

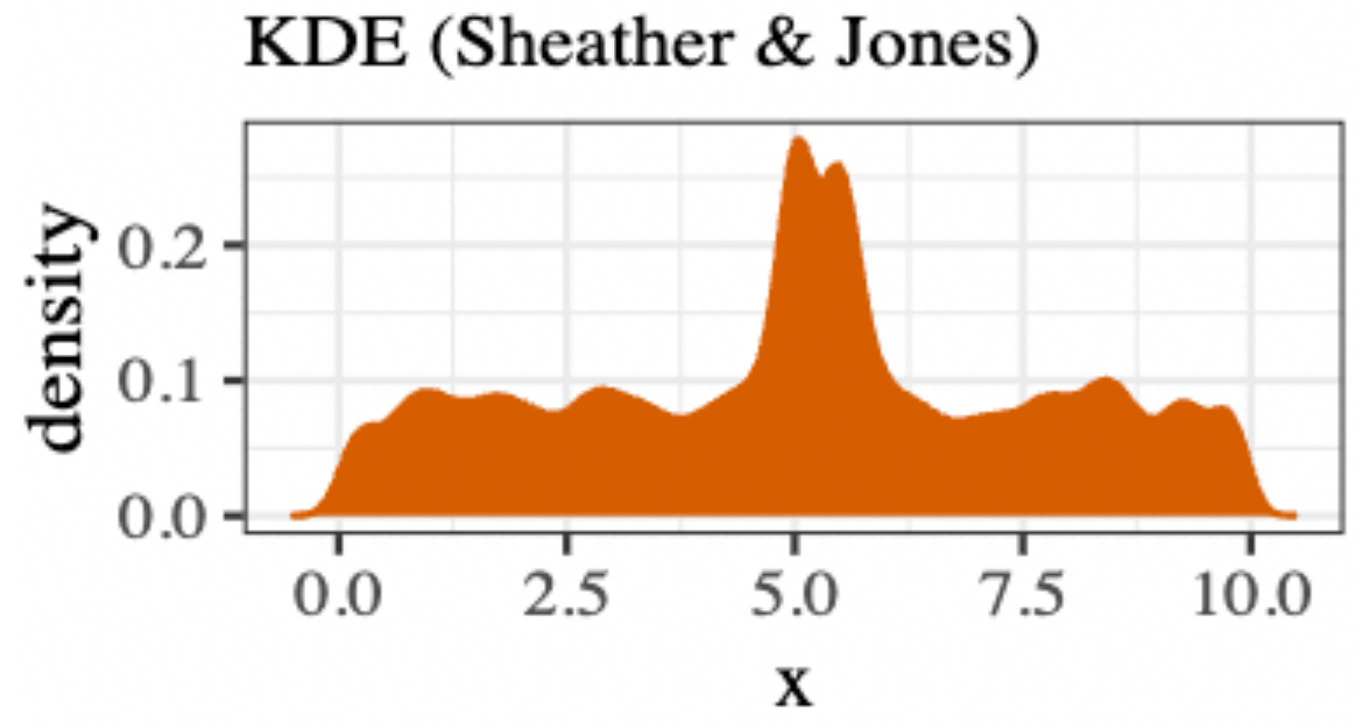
А как оценить плотность?

Можно воспользоваться KDE, но это приводит к bias









А как еще можно?
Можно воспользоваться бутстрапом

Подход к оценке через доверительный интервал
Строим оценку для квантиля

Пусть у нас есть выборка X размером n с неизвестным распределением F, квантиль уровня p, тогда мы можем записать квантиль следующим образом:

$$p = \frac{\sum_{i=0}^{n} I[X_i \leq F^{-1}(p)]}{n} = \bar{I}$$

Как распределен $\bar{I}$? Ответ CLT:

$$\sqrt{n}(\bar{I} - p) \sim N(0, p(1-p))$$

Подход к оценке через доверительный интервал

Строим оценку для квантиля

Зная распределение $\bar{I}$:

$$\sqrt{n}(\bar{I} - p) \sim N(0, \sigma^2)$$

Мы можем вывести доверительный интервал:

$$p \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Это значит, что настоящий перцентиль лежит в выборке между элементами L и U:

$$L = n(p - z_{\alpha/2}\sigma/\sqrt{n})$$

$$U = n(p + z_{\alpha/2}\sigma/\sqrt{n}) + 1$$

Можем ли мы просто посчитать дисперсию в нашем примере?

Вопрос аудитории

Эксклюзив для самых лояльных

Хорошо сформулированная гипотеза:

1. Предпосылка

   Опросы лояльных клиентов выявили интерес к эксклюзивным продуктам и персонализированным предложениям.Возможность
   Создать более персонализированные и визуально привлекательные рекламные баннеры, которые будут соответствовать интересам и предпочтениям целевой аудитории, основываясь на их предыдущем поведении на сайте и покупках.

2. Кого коснется

   Программа коснется лояльных клиентов магазина, которые совершают покупки регулярно и имеют историю покупок выше среднего.

3. Мотивация

   Клиенты получат доступ к эксклюзивным продуктам и персонализированным предложениям, что сделает их покупательский опыт более индивидуализированным и премиальным.

4. Эффект, который мы ожидаем

   Ожидается, что внедрение таких рекомендаций улучшит показатель суммы заказа на 10 процентов

А если вспомнить, что это ratio метрика?

Вопрос аудитории

Алгоритм расчета интервала для квантиля
Пишем пошагово

Пусть у нас есть выборка X размером n с неизвестным распределением F, квантиль уровня p, тогда:

1. Считаем выборочный квантиль $X_{\lfloor np \rfloor}$
2. Считаем индикаторы $I_i = I[X_i \leq X_{\lfloor np \rfloor}]$
3. Применяем дельта-метод для оценки дисперсии $\bar{I}$
4. Считаем $L, U = n(p \pm z_{\alpha/2}\sigma/\sqrt{n})$, где $\sigma/\sqrt{n}$ получена из шага 3
5. Получаем доверительный интервал для квантиля X_L и X_R

И с квантилями разобрались

Дальше можно не делать

Средняя

- User Average Metrics (ARPU/ARPPU/etc)

Ratio

- User-level Conversion Metrics (Retention / etc)
- Page-level Conversion Metrics (Global CTR / etc)

Квантиль

- Ну тут просто квантиль (.99 latency / перцентиль чека)

Абсолюты

- Метрики (GMV / Выручка / Просмотры)