

ИНТЕНСИВ

по анализу A/B-тестирований: день 4

День 4

Ускорение А/Б тестов и прочие способы
уменьшить дисперсию: Стратификация

e^xperiment *f*est

Влияние внешних факторов

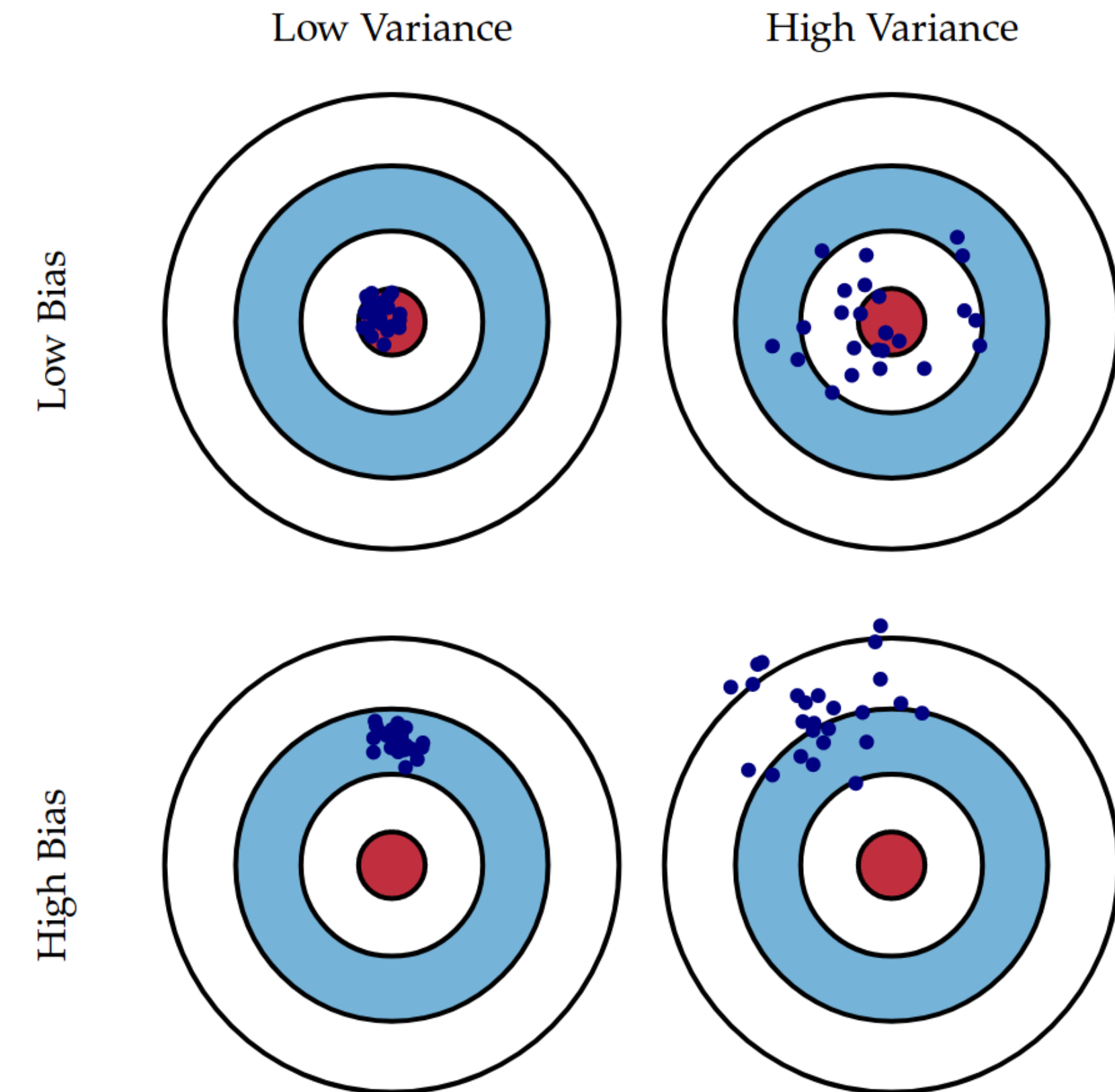
Пример 1: маленькая выборка -> низкая репрезентативность

Пример 2: влияние недельной/месячной сезонности

Влияние оценщика

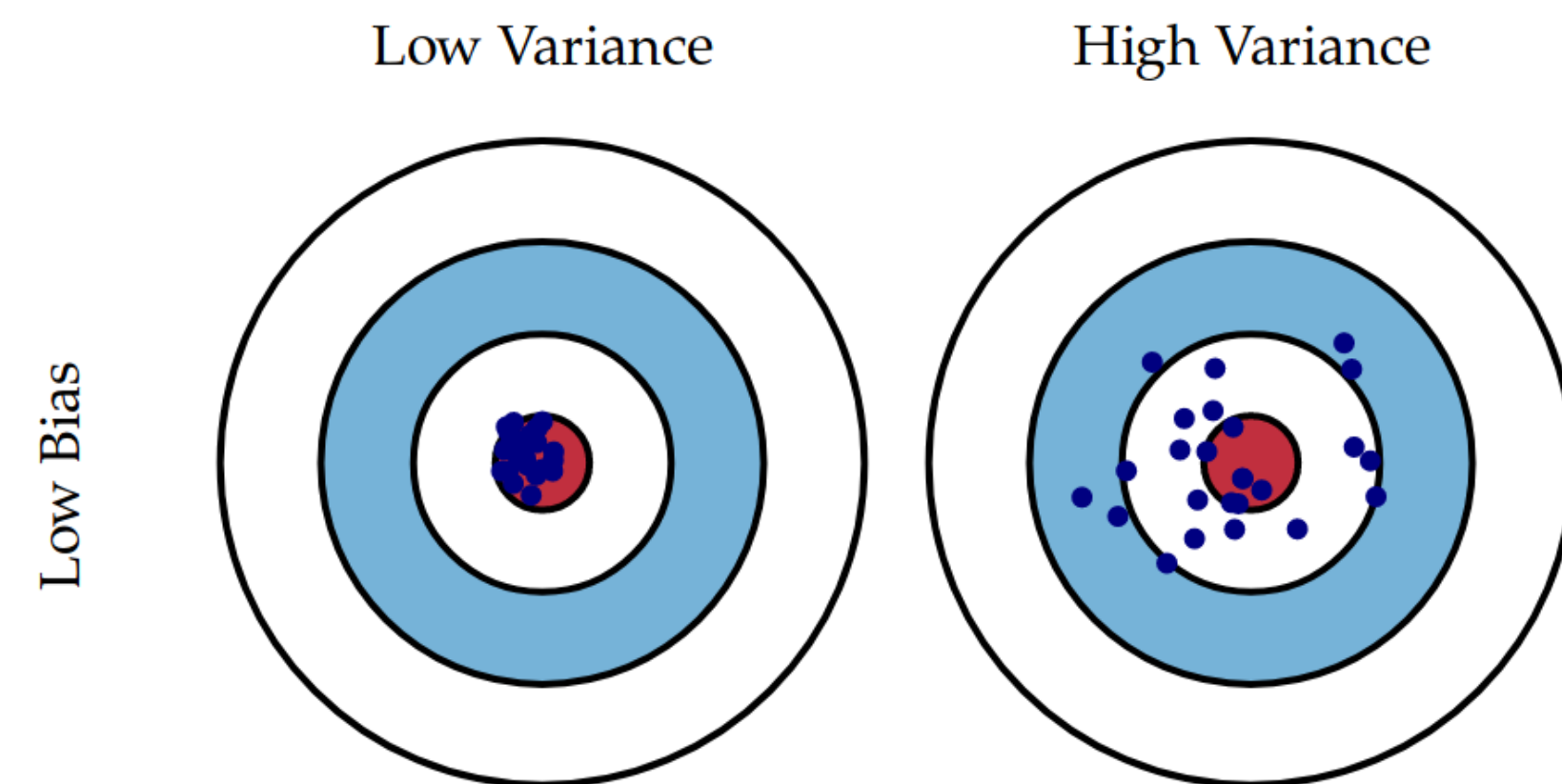
Пример 1: высокий bias возможен при проверке гипотез t-Критерием на распределении с «жирными» хвостами

Пример 2: высокий bias возможен при (непараметрическом) бутстрэпе медиан -> средняя оценки boot-медиан может не совпадать с медианой изначального распределения



Увеличение чувствительности метрики

Для непрерывных метрик можно использовать методы повышения чувствительности путем сокращения дисперсии



Стратификация

e^xperiment *f*est

Идея

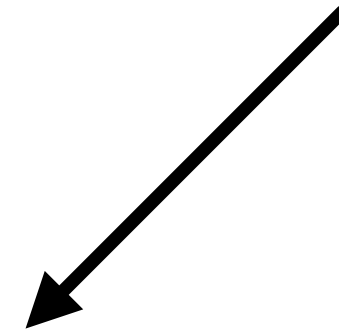
Сокращение дисперсии → Повышение чувствительности → Ускорение эксперимента

Стратификация

e^xperiment *f*est

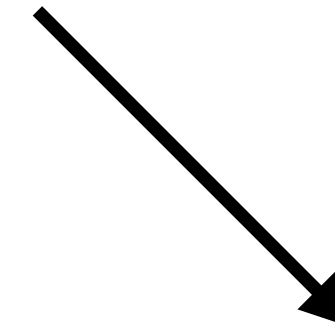
Стратификация

Стратификация



Стратификация (stratified sample)

На этапе дизайна эксперимента,
разбиваем пользователей на k
групп (страт), определяется
вероятность попадания в страту



Пост-стратификация (post-stratification)

Получив случайную выборку,
применяется стратификация на
основе заранее известных
вероятностей попадания в страту

Среднее взвешенное

Какая средняя позиция у баннера?

Баннер показывался в списке на 1,2,3,4 и 5 позициях. Но просто взять среднее по всем позициям – неправильно. Чтобы узнать реальное среднее, будем считать среднее взвешенное

Способ №1

Позиция	Показы
1	300
2	150
3	200
4	50
5	30

Стратификация

e^xperiment fest

Среднее взвешенное

Какая средняя позиция у баннера?

Баннер показывался в списке на 1,2,3,4 и 5 позициях. Но просто взять среднее по всем позициям – неправильно. Чтобы узнать реальное среднее, будем считать среднее взвешенное

Способ №1

						Показы *	
Позиция	Показы		Позиция		Показы		Позиция
1	300	Умножаем позицию на показы →	1	*	300	=	300
2	150		2	*	150	=	300
3	200		3	*	200	=	600
4	50		4	*	50	=	200
5	30		5	*	30	=	150
			Сумма = 730			Сумма = 1550	

Среднее взвешенное

Какая средняя позиция у баннера?

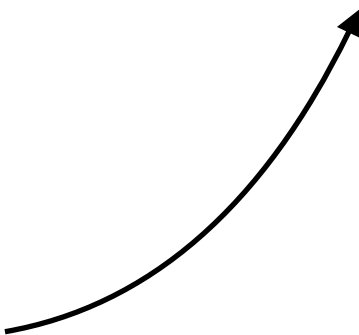
Баннер показывался в списке на 1,2,3,4 и 5 позициях. Но просто взять среднее по всем позициям – неправильно. Чтобы узнать реальное среднее, будем считать среднее взвешенное

Способ №1

						Показы *	
Позиция	Показы		Позиция	Показы		Позиция	
1	300	Умножаем позицию на показы →	1	*	300	=	300
2	150		2	*	150	=	300
3	200		3	*	200	=	600
4	50		4	*	50	=	200
5	30		5	*	30	=	150
				Сумма = 730		Сумма = 1550	

Средняя взвешенная позиция

= $\frac{\text{Сумма (Показы * Позиция)}}{\text{Сумма (Позиция)}}$

 $= \frac{1550}{730} = 2,12$

Среднее взвешенное

Какая средняя позиция у баннера?

Баннер показывался в списке на 1,2,3,4 и 5 позициях. Но просто взять среднее по всем позициям – неправильно. Чтобы узнать реальное среднее, будем считать среднее взвешенное

Способ №2

Позиция	Показы
1	300
2	150
3	200
4	50
5	30

Стратификация

e^xperiment fest

Среднее взвешенное

Какая средняя позиция у баннера?

Баннер показывался в списке на 1,2,3,4 и 5 позициях. Но просто взять среднее по всем позициям – неправильно. Чтобы узнать реальное среднее, будем считать среднее взвешенное

Способ №2

		Показы / Сумма			
Позиция	Показы	показов			Доля
1	300	Переводим показы в доли ————→	300 / 730	=	0,411
2	150		150 / 730	=	0,205
3	200		200 / 730	=	0,274
4	50		50 / 730	=	0,068
5	30		30 / 730	=	0,041
Сумма = 730			Сумма = 1		

Среднее взвешенное

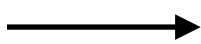
Какая средняя позиция у баннера?

Баннер показывался в списке на 1,2,3,4 и 5 позициях. Но просто взять среднее по всем позициям – неправильно. Чтобы узнать реальное среднее, будем считать среднее взвешенное

Способ №2

Позиция	Показы	Показы / Сумма показов		Доля
1	300	300 / 730	=	0,411
2	150	150 / 730	=	0,205
3	200	200 / 730	=	0,274
4	50	50 / 730	=	0,068
5	30	30 / 730	=	0,041
		Сумма = 730	Сумма = 1	

Переводим показы
в доли



Умножаем позицию
на долю (вес)

Позиция		Доля		Взвешенная позиция
1	*	0,411	=	0,411
2	*	0,205	=	0,411
3	*	0,274	=	0,822
4	*	0,068	=	0,274
5	*	0,041	=	0,205

Среднее взвешенное

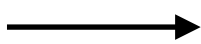
Какая средняя позиция у баннера?

Баннер показывался в списке на 1,2,3,4 и 5 позициях. Но просто взять среднее по всем позициям – неправильно. Чтобы узнать реальное среднее, будем считать среднее взвешенное

Способ №2

Позиция	Показы	Показы / Сумма показов		Доля
1	300	300 / 730	=	0,411
2	150	150 / 730	=	0,205
3	200	200 / 730	=	0,274
4	50	50 / 730	=	0,068
5	30	30 / 730	=	0,041
Сумма = 730		Сумма = 1		

Переводим показы в доли



Умножаем позицию на долю (вес)

Позиция		Доля		Взвешенная позиция
1	*	0,411	=	0,411
2	*	0,205	=	0,411
3	*	0,274	=	0,822
4	*	0,068	=	0,274
5	*	0,041	=	0,205
Суммируем взвешенную позицию				
= 0,411 + 0,411 + 0,822 + 0,274 + 0,205 = 2,12				

Пост-стратификация

Способ расчета повторяет логику расчета любой взвешенной оценки:

1. Выбираем переменные, наиболее влияющие на метрику. Страта (или группа) определяется k -градацией этой переменной.
2. Считаем вероятность попадания пользователя в группу по аналогии с тем, как мы считали доли до этого в примере
3. Считаем дисперсию var внутри группы (страты)
4. Считаем взвешенную оценку var_{strat}
5. Добиваемся сокращения дисперсии

отличие от пре-стратификации от пост в том, что веса определяются на этапе конфига

Вероятность попадания в страту, попросту – вес

Дисперсия по страте

$$var_{strat}(\hat{Y}_{strat}) = \frac{1}{n} \sum_{k=1}^K p_k \sigma_k^2.$$

для нормирования по выборке (по аналогии в формуле стандартной ошибки – это $\frac{\sigma}{\sqrt{n}}$, или

это дисперсия среднего)

Стратификация

e^xperiment *f*est

Среднее стратифицированное по выборке будет такое же как и само среднее по ГС, т.к. считается среднее взвешенное по стратам. Можно особо не углубляться:

$$E_{strat}(\hat{Y}_{strat}) = \sum_{k=1}^K p_k E_{strat}(\bar{Y}_k) = \sum_{k=1}^K p_k \mu_k = \mu.$$

Y : метрика

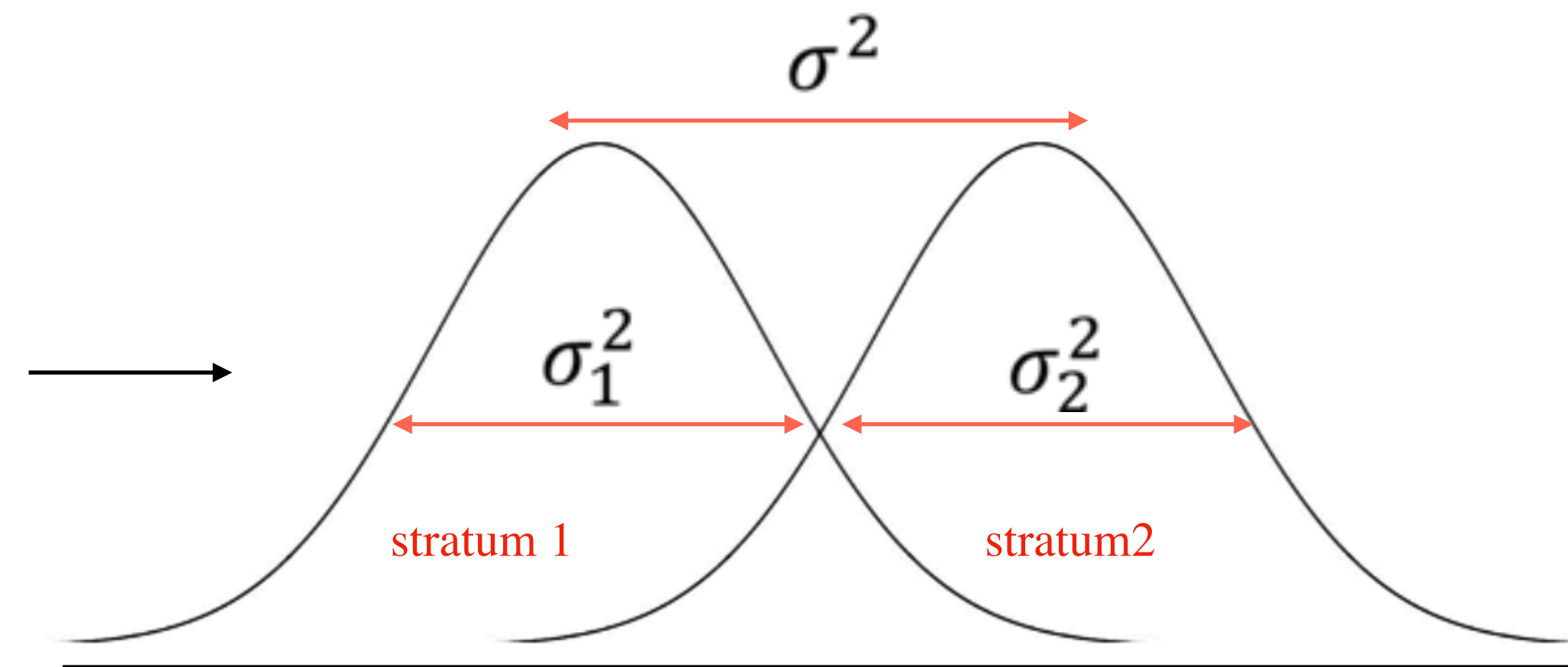
$\mu = E(Y)$: среднее по метрике в ГС

p_k : доля k страты в ГС = вероятность попадания в страту

Стратифицированная дисперсия по выборке считается взвешенно по каждой страте. Вес определяется вероятностью попадания пользователя в страту.

Например, если за страты мы берем регионы и их у нас 3 с аудиторией разбитые по долям 30%, 50%, 20%, то веса для них $p_{k1} = 0.3$, $p_{k2} = 0.5$, $p_{k3} = 0.2$

$$var_{strat}(\hat{Y}_{strat}) = \frac{1}{n} \sum_{k=1}^K p_k \sigma_k^2.$$



9, 7, 8, 6, 9, 6, 7, 5, 4

$$\text{var}(\overline{Y}) = \sum_{k=1}^K \frac{w_k}{n} \sigma_k^2 + \sum_{k=1}^K \frac{w_k}{n} (\mu_k - \mu)^2$$

Выборочное среднее
и дисперсия

$$\bar{x} = 6.777778$$

$$\sigma^2 = 2.617284$$

9, 7, 8, 6, 9, 6, 7, 5, 4

$$\text{var}(\bar{Y}) = \sum_{k=1}^K \frac{w_k}{n} \sigma_k^2 + \sum_{k=1}^K \frac{w_k}{n} (\mu_k - \mu)^2$$

Стратификация

e^xperiment fest

Выборочное среднее
и дисперсия

$$\bar{x} = 6.777778$$

$$\sigma^2 = 2.617284$$

9, 7, 8, 6, 9, 6, 7, 5, 4

$$\text{var}(\bar{Y}) = \sum_{k=1}^K \frac{w_k}{n} \sigma_k^2 + \sum_{k=1}^K \frac{w_k}{n} (\mu_k - \mu)^2$$

Выборочное среднее
и дисперсия

$$\bar{x} = 6.777778$$

$$\sigma^2 = 2.617284$$

9, 7, 8, 6, 9, 6, 7, 5, 4

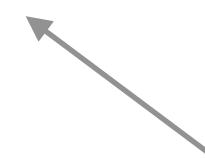
$$\text{var}(\bar{Y}) = \sum_{k=1}^K \frac{w_k}{n} \sigma_k^2 + \sum_{k=1}^K \frac{w_k}{n} (\mu_k - \mu)^2$$

Выборочное среднее
и дисперсия

$$\bar{x} = 6.777778$$

$$\sigma^2 = 2.617284$$

$$\begin{aligned}\bar{x} &= 8 \\ \sigma^2 &= 1 \\ w &= 0.22\end{aligned}$$



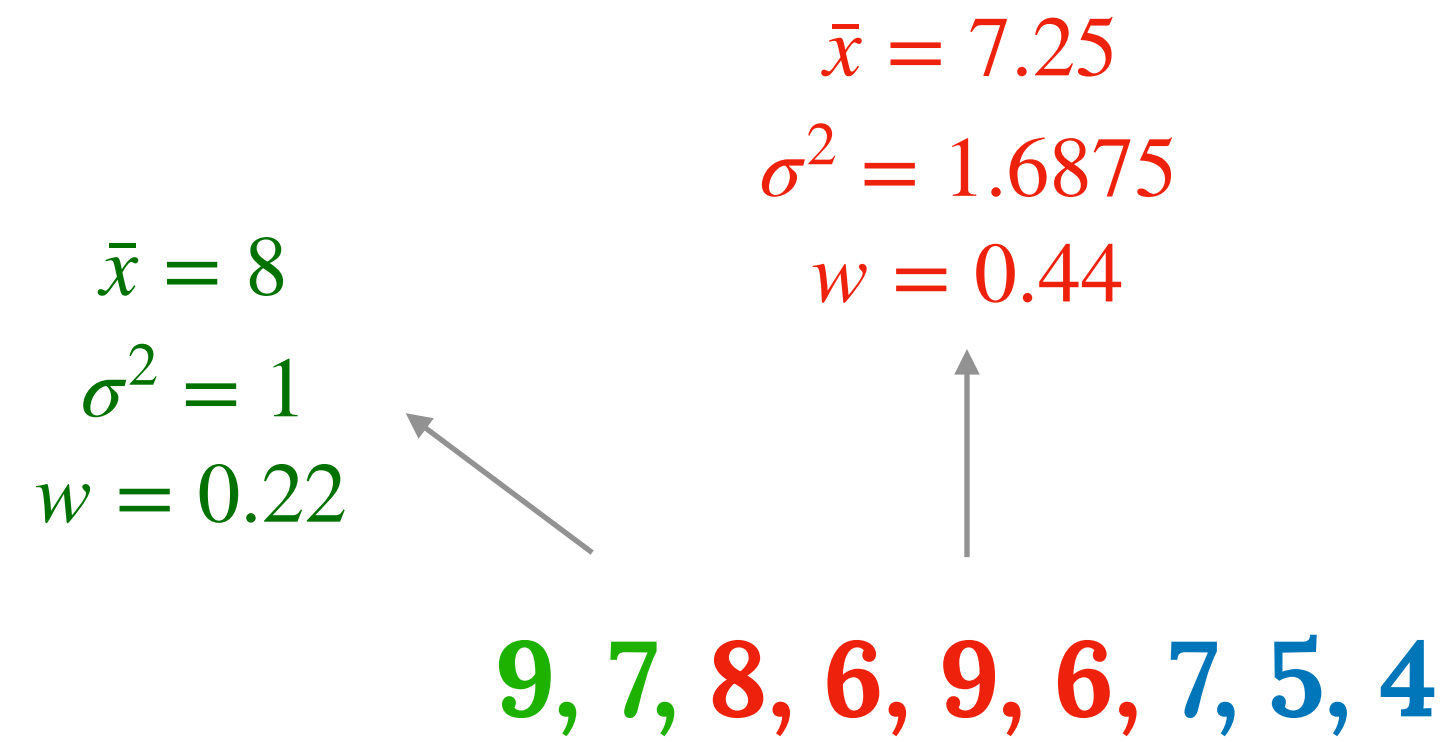
9, 7, 8, 6, 9, 6, 7, 5, 4

$$\text{var}(\bar{Y}) = \sum_{k=1}^K \frac{w_k}{n} \sigma_k^2 + \sum_{k=1}^K \frac{w_k}{n} (\mu_k - \mu)^2$$

Выборочное среднее
и дисперсия

$$\bar{x} = 6.777778$$

$$\sigma^2 = 2.617284$$

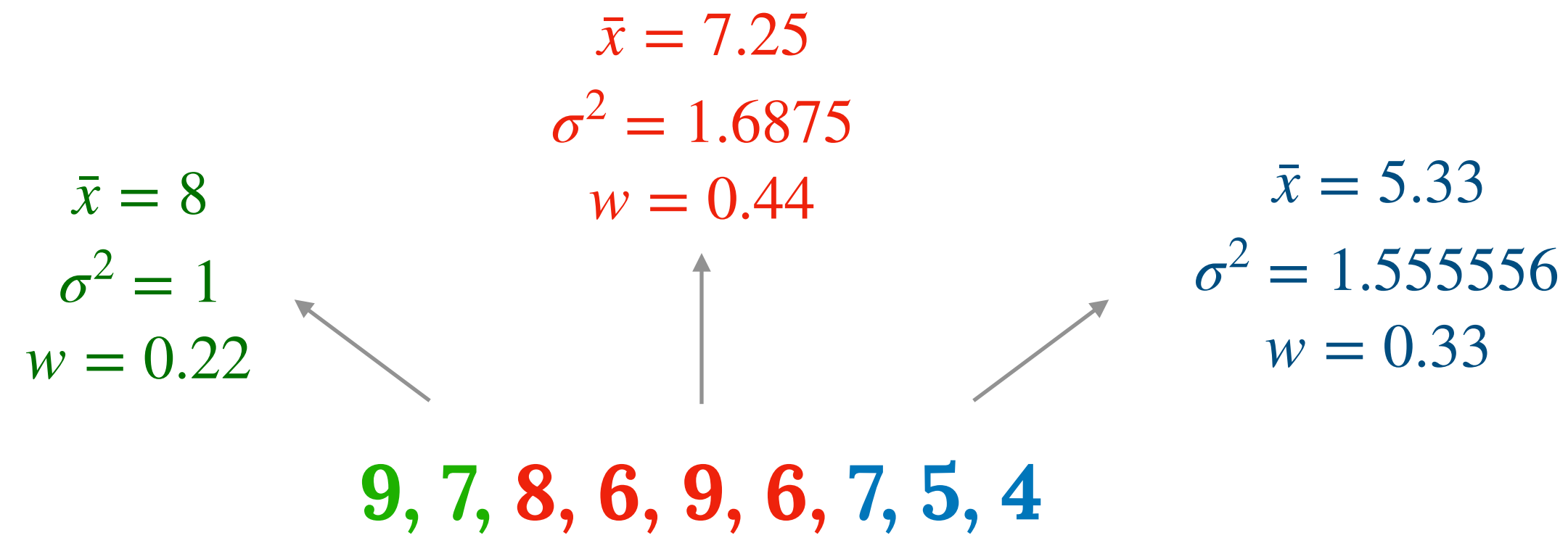


$$\text{var}(\bar{Y}) = \sum_{k=1}^K \frac{w_k}{n} \sigma_k^2 + \sum_{k=1}^K \frac{w_k}{n} (\mu_k - \mu)^2$$

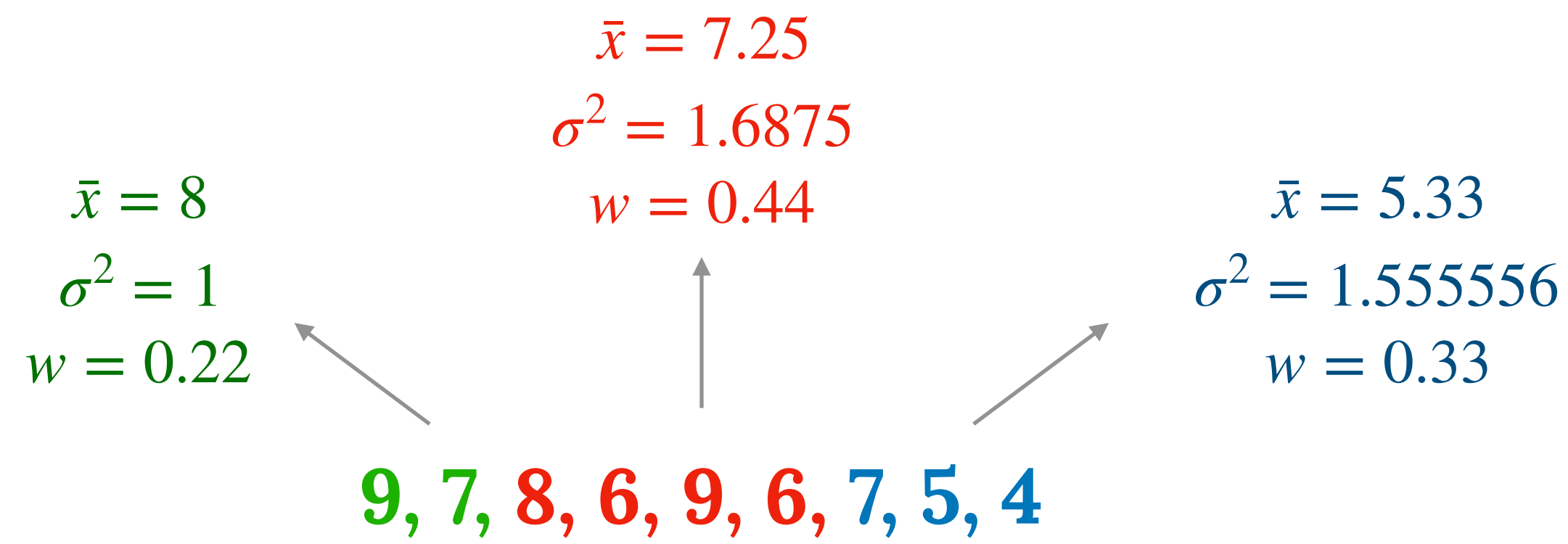
Выборочное среднее
и дисперсия

$$\bar{x} = 6.777778$$

$$\sigma^2 = 2.617284$$



$$\text{var}(\bar{Y}) = \sum_{k=1}^K \frac{w_k}{n} \sigma_k^2 + \sum_{k=1}^K \frac{w_k}{n} (\mu_k - \mu)^2$$

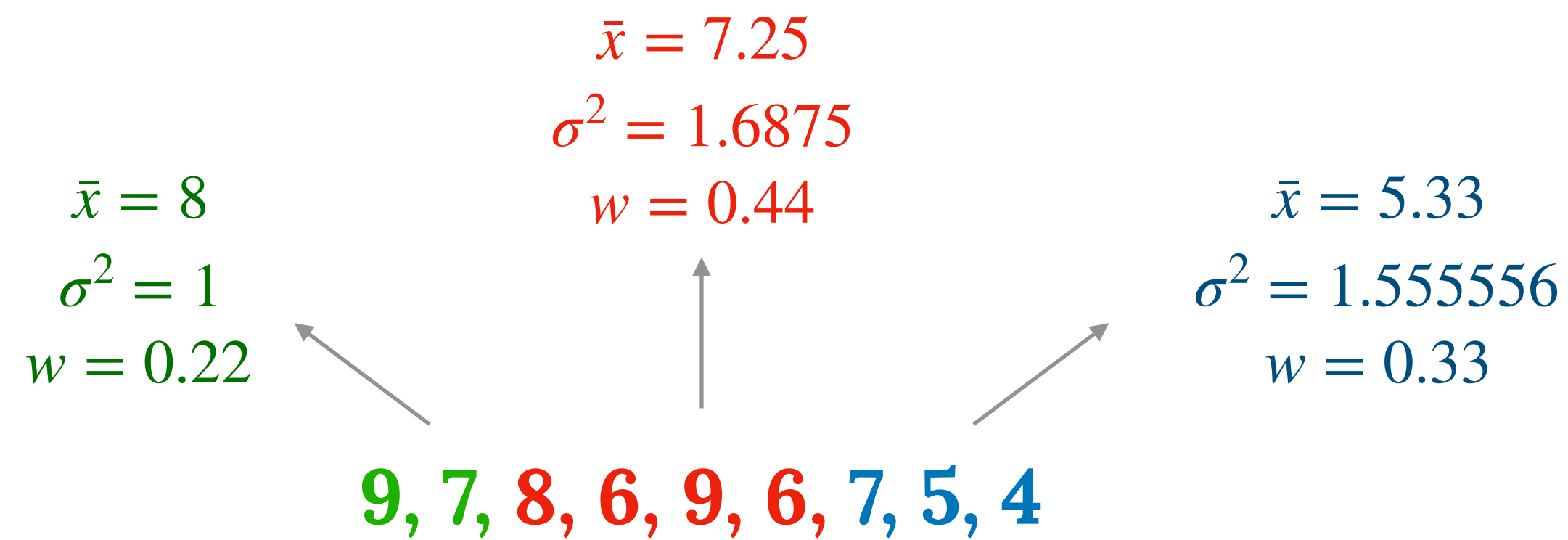


Выборочное среднее
и дисперсия

$$\bar{x} = 6.777778$$

$$\sigma^2 = 2.617284$$

$$\text{var}(\bar{Y}) = \sum_{k=1}^K \frac{w_k}{n} \sigma_k^2 + \sum_{k=1}^K \frac{w_k}{n} (\mu_k - \mu)^2$$



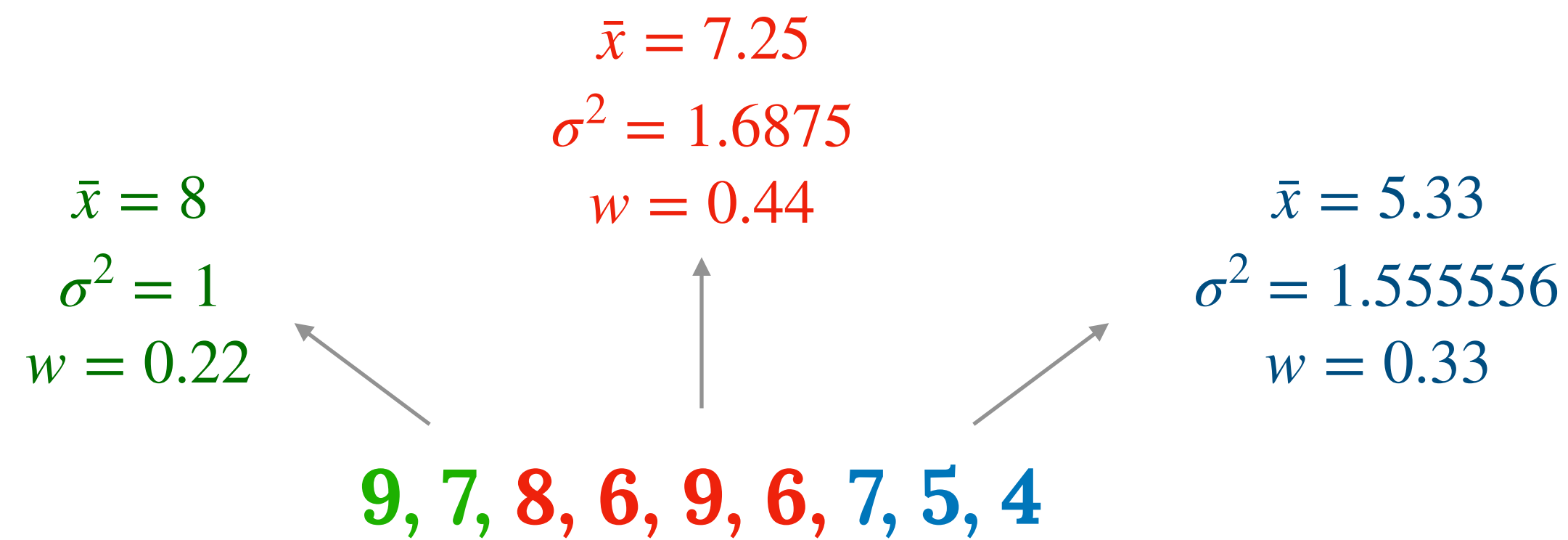
$$\text{var}(\bar{Y}) = \sum_{k=1}^K \frac{w_k}{n} \sigma_k^2 + \sum_{k=1}^K \frac{w_k}{n} (\mu_k - \mu)^2$$

Внутригрупповая дисперсия

$$\begin{aligned}
 &0.22 * 1 \\
 &+ 0.44 * 1.6875 \\
 &+ 0.33 * 1.555556 \\
 &= 1.490741
 \end{aligned}$$

Стратификация

e^xperiment *f*est



$$\text{var}(\bar{Y}) = \sum_{k=1}^K \frac{w_k}{n} \sigma_k^2 + \sum_{k=1}^K \frac{w_k}{n} (\mu_k - \mu)^2$$

Внутригрупповая дисперсия

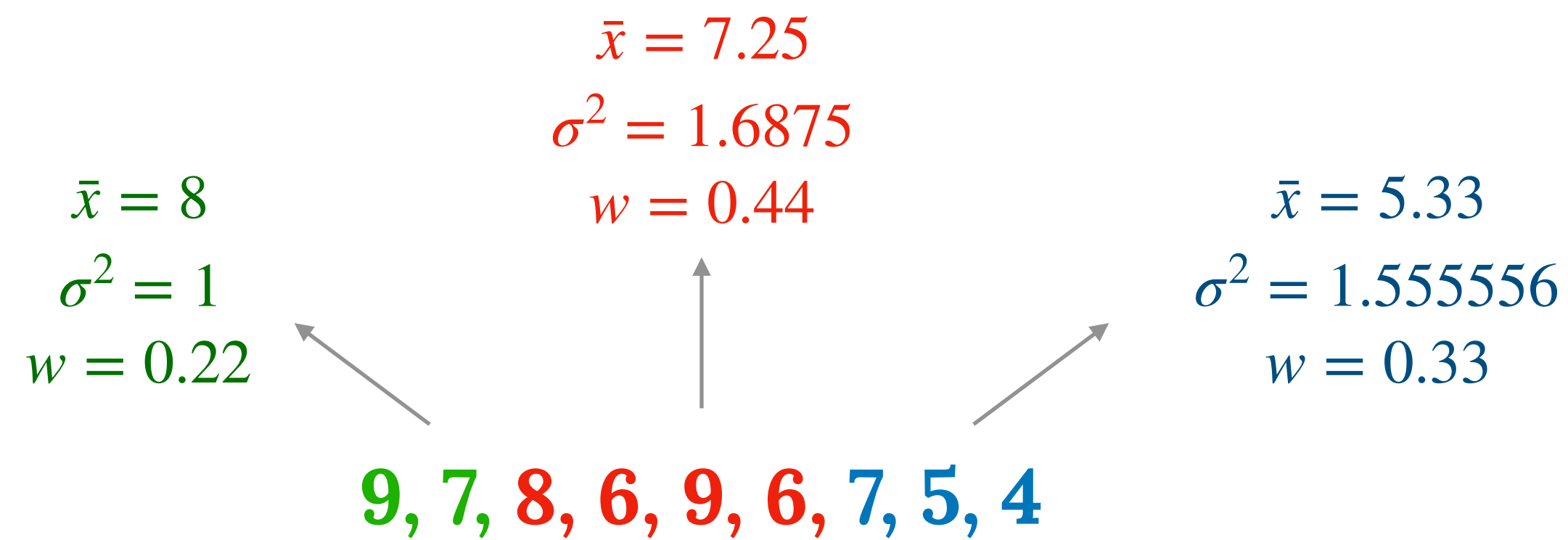
$$\begin{aligned}
 &0.22 * 1 \\
 &+ 0.44 * 1.6875 \\
 &+ 0.33 * 1.555556 \\
 &= 1.490741
 \end{aligned}$$

Межгрупповая дисперсия

$$\begin{aligned}
 &0.22 * (8 - 6.777778)^2 \\
 &+ 0.44 * (7.25 - 6.777778)^2 \\
 &+ 0.33 * (5.33 - 6.777778)^2 \\
 &= 1.126543
 \end{aligned}$$

Стратификация

e^xperiment *f*est



Выборочное среднее
и дисперсия

$$\bar{x} = 6.777778$$

$$\sigma^2 = 2.617284$$

$$\text{var}(\bar{Y}) = \sum_{k=1}^K \frac{w_k}{n} \sigma_k^2 + \sum_{k=1}^K \frac{w_k}{n} (\mu_k - \mu)^2 = 1.490741 + 1.126543 = 2.617284$$

Внутригрупповая дисперсия

$$\begin{aligned}
 &0.22 * 1 \\
 &+ 0.44 * 1.6875 \\
 &+ 0.33 * 1.555556 \\
 &= 1.490741
 \end{aligned}$$

Межгрупповая дисперсия

$$\begin{aligned}
 &0.22 * (8 - 6.777778)^2 \\
 &+ 0.44 * (7.25 - 6.777778)^2 \\
 &+ 0.33 * (5.33 - 6.777778)^2 \\
 &= 1.126543
 \end{aligned}$$

Стратификация

e^xperiment *f*est

Если разобрать дисперсию на составные компоненты, то получим внутригрупповую и межгрупповую дисперсии:

$$\begin{aligned}\text{var}(\bar{Y}) &= \sum_{k=1}^K \frac{w_k}{n} \sigma_k^2 + \sum_{k=1}^K \frac{w_k}{n} (\mu_k - \mu)^2 \\ &\geq \sum_{k=1}^K \frac{w_k}{n} \sigma_k^2 = \text{var}(\hat{Y}_{strat})\end{aligned}$$

Дисперсия по случайной выборке будет либо такой же, как в стратифицированной выборке, либо больше, т.к. в стратифицированной выборке мы избавляемся от межгрупповой дисперсии.

Демонстрация R

Стратификация

e^x periment fest

Выводы

- Страты должны быть стабильными и не меняться во времени — иначе вес страты будет нестабильным и это приведет к непредсказуемому изменению дисперсии
- Плохо подобранные переменные для стратификации влекут за собой увеличение дисперсии
- Стратификацию можно использовать, когда есть проблемы с балансом выборок. Например, аномальный всплеск покупок из одной страты с большой дисперсией можно приглушить стратификацией
- Разные исследования показывают, что стратификация может сократить дисперсию от 1% до 40%

День 4

Ускорение А/Б тестов и прочие способы
уменьшить дисперсию: CUPED

Типичная картина

На маленьких эффектах чаще всего мы видим низкую мощность на ранних этапах эксперимента. Причиной этому является высокая дисперсия:

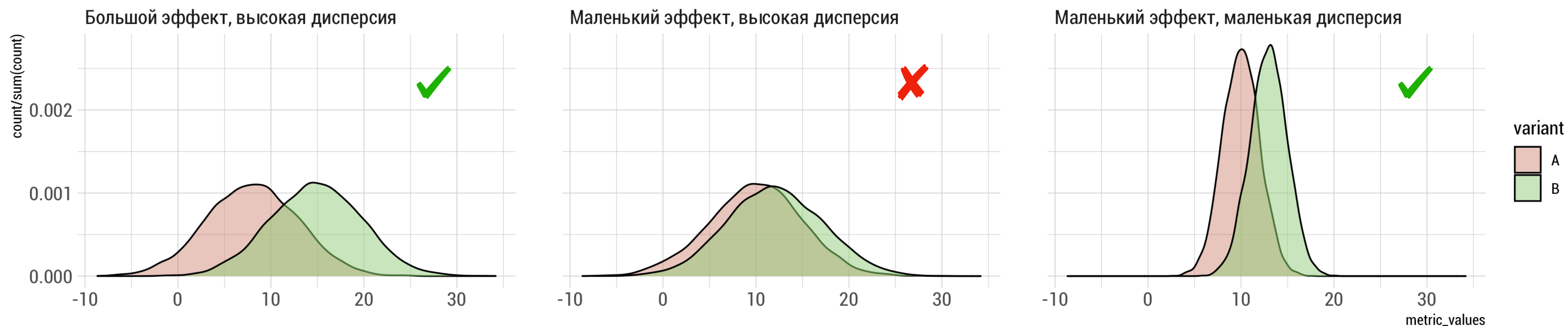


Cuped

e^x periment fest

Проблема: высокая дисперсия на маленьких эффектах

Чтобы добиться сокращения дисперсии необязательно ждать окончания эксперимента для прокраса



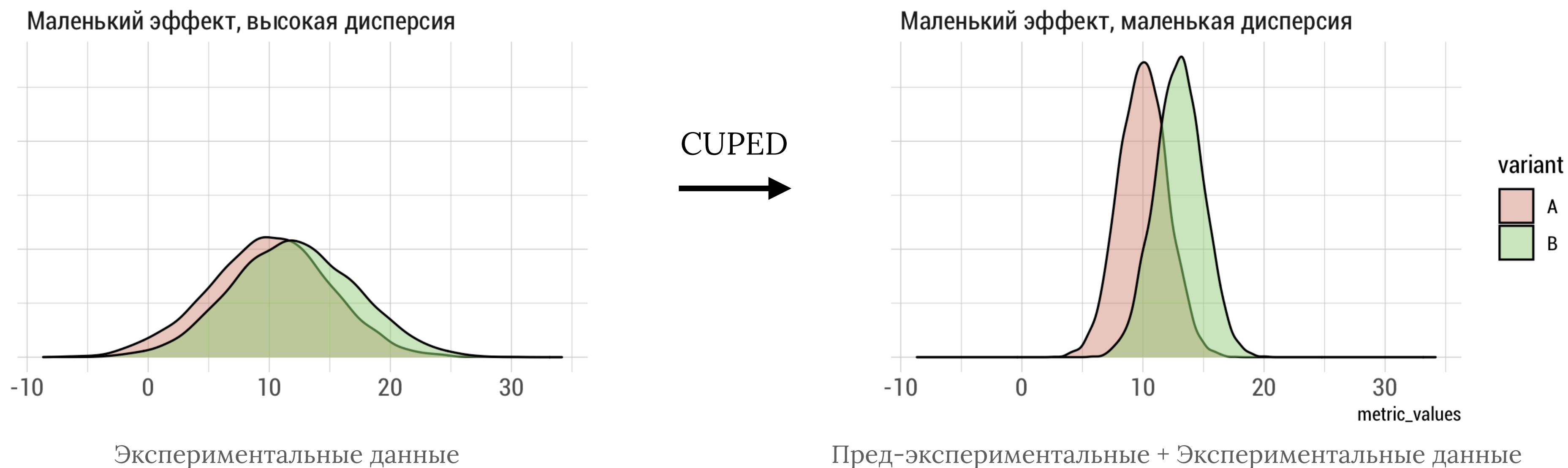
CupEd

e^x periment fest

Решение: Сокращение за счет истории

Сокращение достигается путем использования пред-экспериментальных данных:

- а) использование стратификации
- б) использование ковариат



Cuped

e^xperiment fest

$$Y_{CUPED} = Y - \theta X$$

, где

Y – экспериментальная метрика

X – метрика до проведения эксперимента.

Для сохранения среднего используется

$$X - \hat{X}$$

$$\theta = \frac{cov(X, Y)}{var(X)}$$

в CUPED мы смотрим не просто на бизнес-метрику Y , а на измененную Y_{CUPED} , которая чувствительнее, благодаря ее связи с пред-экспериментальным периодом

$$Y_{CUPED} = Y - \theta X$$

Идея заключается в том, что дисперсия Y обусловлена двумя компонентами: дисперсией ковариаты X и дисперсией неизвестных переменных. После коррекции Y_{CUPED} , мы избавляемся от дисперсии X и оставляем влияние неизвестных переменных

CUPED: Принцип работы 1/3

При использовании CUPED'a учитывается поведение пользователя до эксперимента (ковариата X) и во время эксперимента (фактическая метрика Y).

После нахождения между этими метриками зависимости уменьшается дисперсия → увеличивается чувствительность

Таким образом, мы узнаем как изменилось поведение пользователя от среднего.

CUPED: Принцип работы 2/3

CUPED имеет две модификации: использование пред-стратификации или ковариаты:

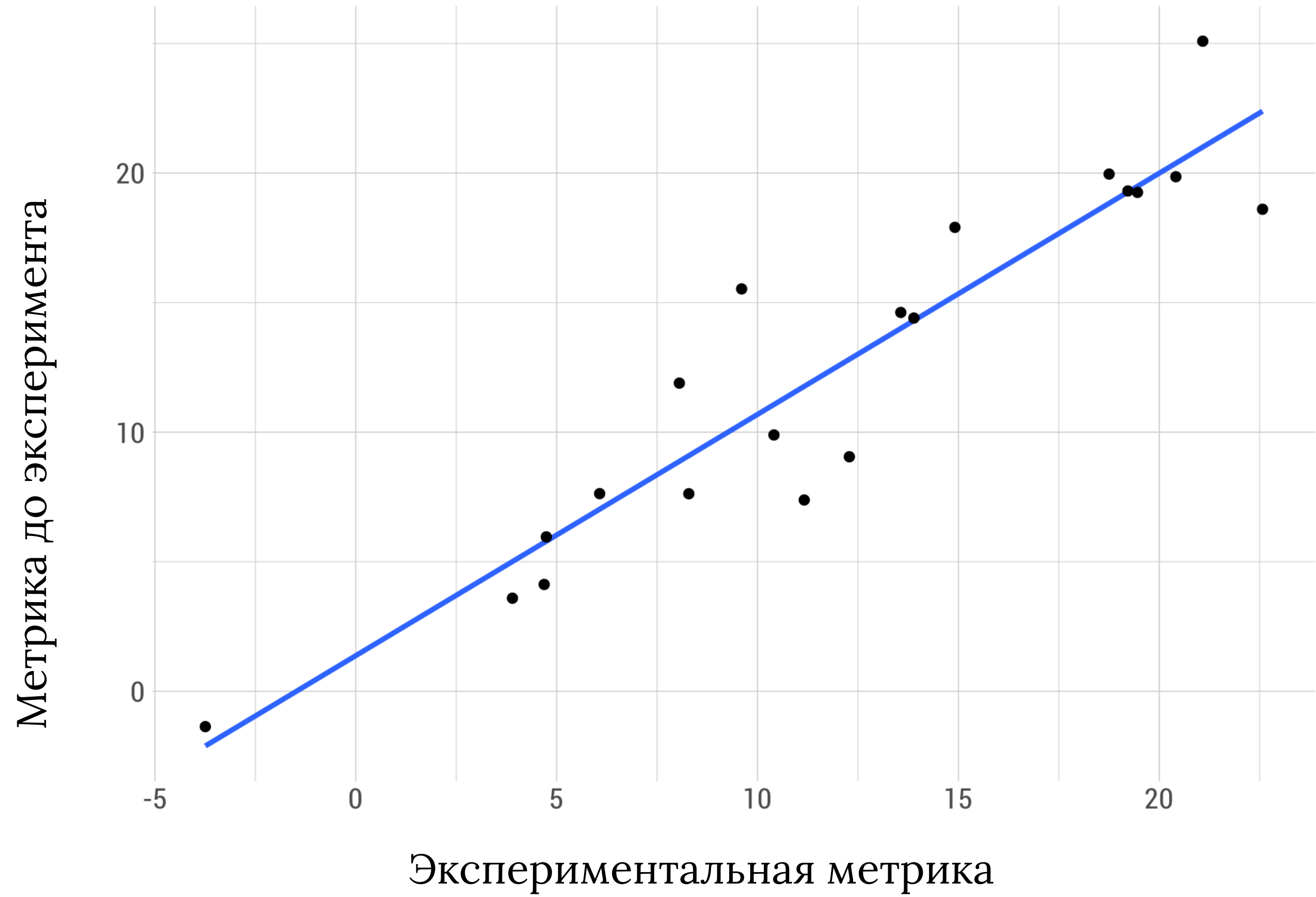
- Пред-стратификация использует категориальные данные типа региона, пакетный тариф и т.п.
- Ковариата – числовая непрерывная метрика (средний чек) взятая из истории пользователя

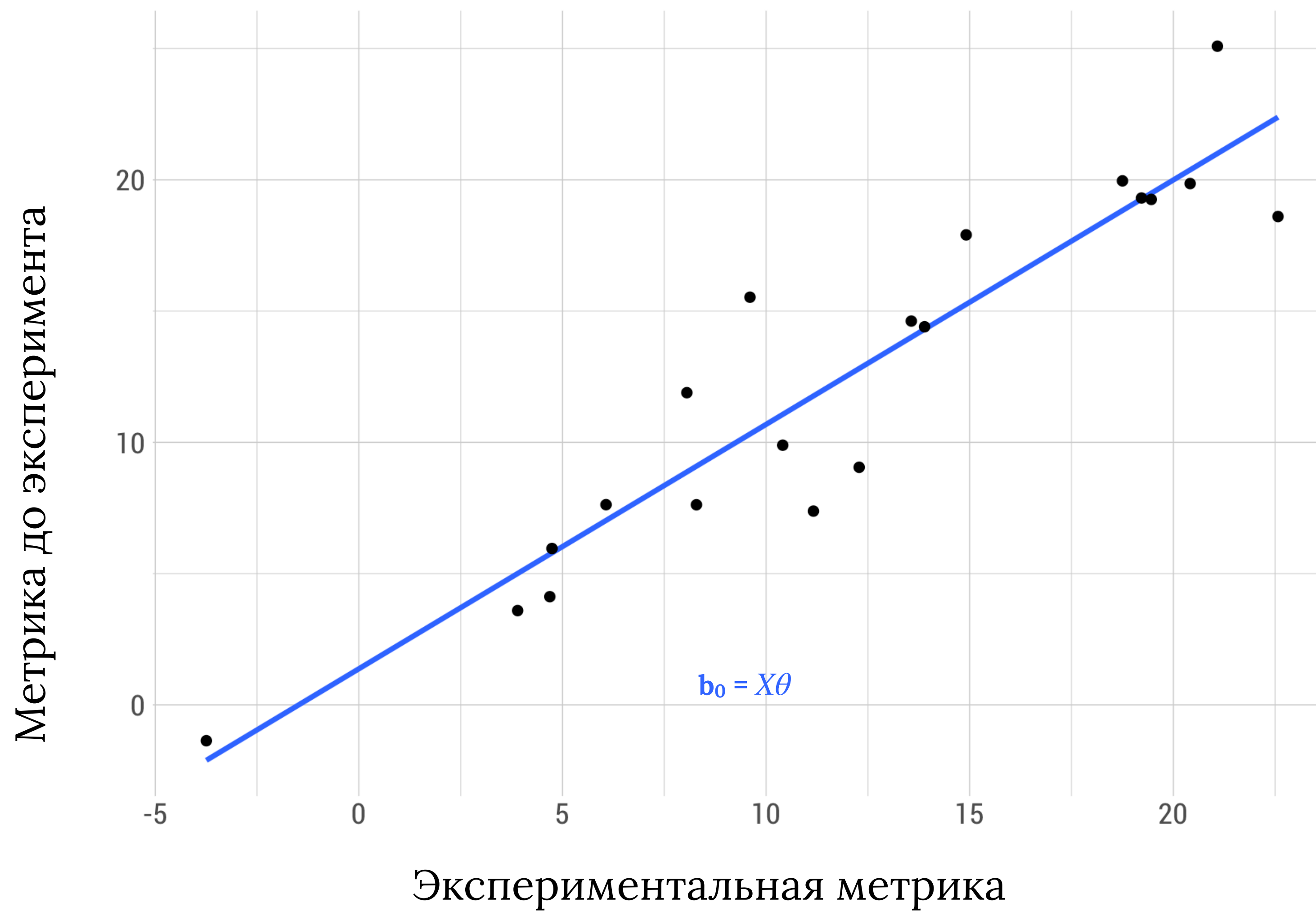
На практике чаще используется ковариата, т.к. она позволяет сильнее сократить дисперсию, чем стратификация

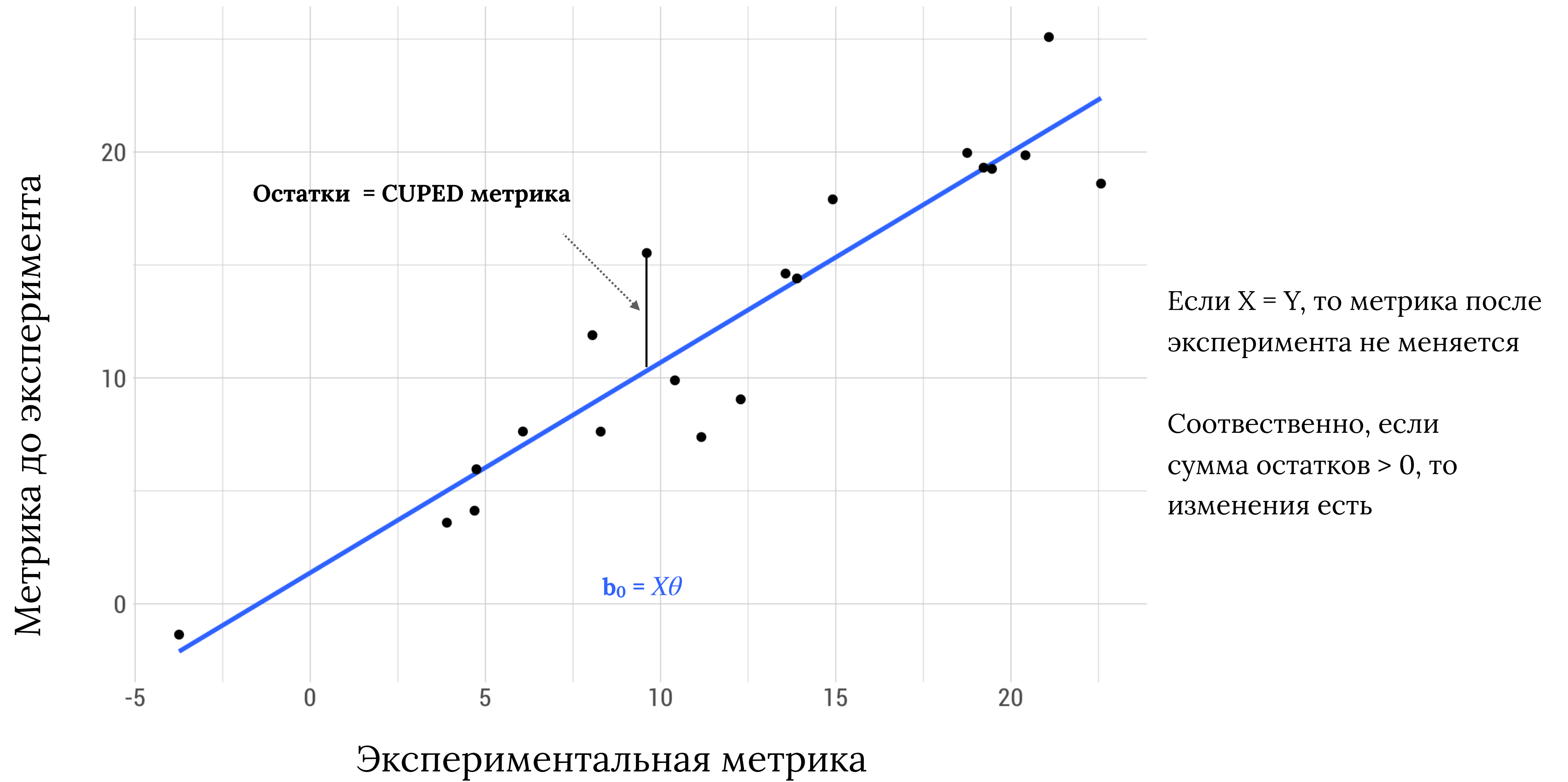
CUPED: Принцип работы 3/3

uid	exp_variant	sessionDuration (метрика)	sessionDuration (пред-эксп. ковариата)	CUPED метрика
1	control	31	30	$31 - (30 - \text{средняя ковариата}) * \theta$
2	control	23	20	$23 - (20 - \text{средняя ковариата}) * \theta$
3	test	25	22	$25 - (22 - \text{средняя ковариата}) * \theta$
4	test	41	46	$41 - (46 - \text{средняя ковариата}) * \theta$

*среднее по ковариате и тета подбираются по всем группам









Демонстрация R

Ограничения и требования

- Необходимы исторические данные о пользователе
- Неочевидно какой пред-экспериментальный период выбрать для ковариаты
- Подходит для работы с разницей средних
- Для дискретных (бинарных) величин не подходит

День 4

Ratio-метрика и линеаризация

e^x periment fest

ВЫБОРОЧНАЯ МЕТРИКА

$$\bar{X} = \frac{\sum_i X_i}{n}$$

- СРЕДНИЙ ЧЕК
- ARPU

ПОЮЗЕРНАЯ МЕТРИКА

$$\bar{R} = \frac{\bar{Y}}{\bar{Z}} = \frac{\sum_i Y_i}{\sum_i Z_i}$$

- CTR
- Action Per Session

Пользователь	Действие на странице	Просмотры страницы
1	10	200
2	3	259
4	90	139
5	10	20

$ratioUser1 = \frac{10}{200} = 0.05$

$ratioUser1 = \frac{3}{259} = 0.011$

$ratioUser1 = \frac{90}{139} = 0.647$

$ratioUser1 = \frac{10}{20} = 0.5$

$naiveAVG = \frac{0.05 + 0.011 + 0.647 + 0.5}{4} = 0.302$

Пользователь	Действие на странице	Просмотры страницы
1	10	200
2	3	259
4	90	139
5	10	20

$$globalRatio = \frac{10 + 3 + 90 + 10}{200 + 259 + 139 + 20} = 0,182$$

$$naiveAVG = \frac{0.05 + 0.011 + 0.647 + 0.5}{4} = 0,302$$

$$bias = 0,302 - 0,182 = 0.12$$

Линеаризация

e^xperiment *f*est

ratio метрику мы можем посчитать как минимум тремя способами

naiveRatio

- Имеет сильный сдвиг относительно globalRatio
- За счет сдвига pvalue может «краситься» случайно

globalRatio

- Требуется считать дисперсию для отношения двух случайных величин
- Невозможно сохранить «поюзерную» направленность
- Для расчета дисперсии требуется использовать bootstrap или разложение ряда для второго момента (читай дополнительные материалы)

L-метрика (линеаризация)

- Сохраняет «поюзерную» направленность
- Можно использовать базовые статистические оценщики
- Можно использовать для дальнейшей оптимизации дисперсии
- Просто вычисляется
- Не сохраняется global ratio

Принцип работы линеаризации

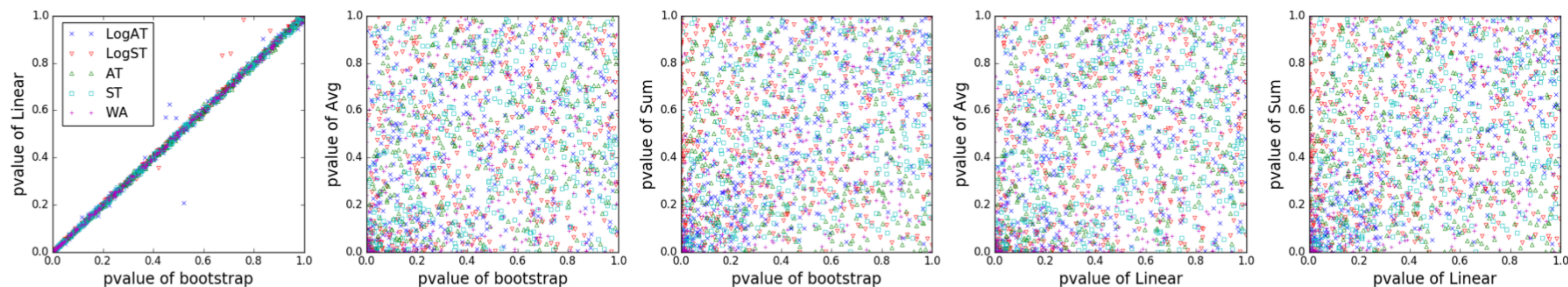
$$Lx, y, k(U) = X(u) - kY(u)$$

- K – ratio по контролю
- X – клики пользователя
- Y – просмотры пользователя

мы хотим понять отклонение метрики в эксперименте относительно ratio в контроле.

иначе говоря – смотрим, что изменилось в эксперименте относительно ситуации в контроле

Сравнение способов анализа ratio-метрик



- pvalue по L метрики сонаправлен с pvalue для ratio bootstrap
- Среднее всех отношений сильно отличается от bootstrap и L метрики за счет сдвига
- Линеаризация **не позволяет ускорять a/b теста**

Плюсы L метрики

- Сохраняется поюзерная направленность
- На L метрику можно использовать методы оптимизации дисперсии вроде CUPED
- Если перед вами не стоит задача сохранять поюзерную направленность или упрощать статистические операции — то можно использовать bootstrap для оценки ratio метрик. Результат будет сонаправлен L метрики

e^x periment fest

Мирмахмадов Искандер
Черемисинов Виталий

07/2020

experiment-fest.ru