

Интенсив

по анализу А/В-тестирований: день 2

День 2

Проверка статистических гипотез. Виды статистических критериев

e^xperiment fest

Что узнаем?

- Параметрические критерии и непараметрические критерии
- Расчет минимального объема выборки и что такое MDE

Как выбрать статистический критерий?

e^xperiment fest

Как проверяются гипотезы?

Как выбрать статистический критерий?

e^xperiment fest

Для проверки гипотез используются
параметрические и непараметрические
критерии

Как выбрать статистический критерий?

e^xperiment fest

Как выбрать статистический критерий для проверки гипотез?

- Для опровержения/подтверждения гипотез используются статистические критерии.
- Выбор статистического критерия всегда зависит от сформулированной гипотезы
- Например, для сравнения средних используется один класс критериев, для проверки форм распределения – другой

Нулевая гипотеза – принимаемое предположение о том, что не существует связи между наблюдениями в двух (или более) событиях (выборках, феноменах, совокупностях).

Гипотезу отвергают, если данные показывают разницу между выборками (на заданном уровне значимости)

H_0 значит, что средние (если это, например, t-критерий) не отличаются между двумя выборками, а H_1 , соответственно, что это условие не выполняется. Для данного эксперимента положительным итогом является отвержение нулевой гипотезы.

Отвергнуть ее можем при заданной допустимости ошибки первого рода (*p-value*), т.е. чтобы ошибка была меньше 0.05, если нам достаточно 95% уровня значимости

Как выбрать статистический критерий?

e^xperiment fest

Параметрические критерии

Параметрические критерии основаны на том, что известно распределение данных. Их можно использовать, когда главное допущение критерия соблюдается, а именно то, что тип распределения – известен.

Как выбрать статистический критерий?

e^xperiment fest

Аналитик Лёня сравнивает две версии одной формы:
с адресом и без. Цель – оптимизация костов колл-центра. Какая
форма эффективнее?

A

Имя

Александр

email

example@example.com

Телефон

+7(999)9999999

B

Имя

Александр

email

example@example.com

Телефон

+7(999)9999999

Адрес

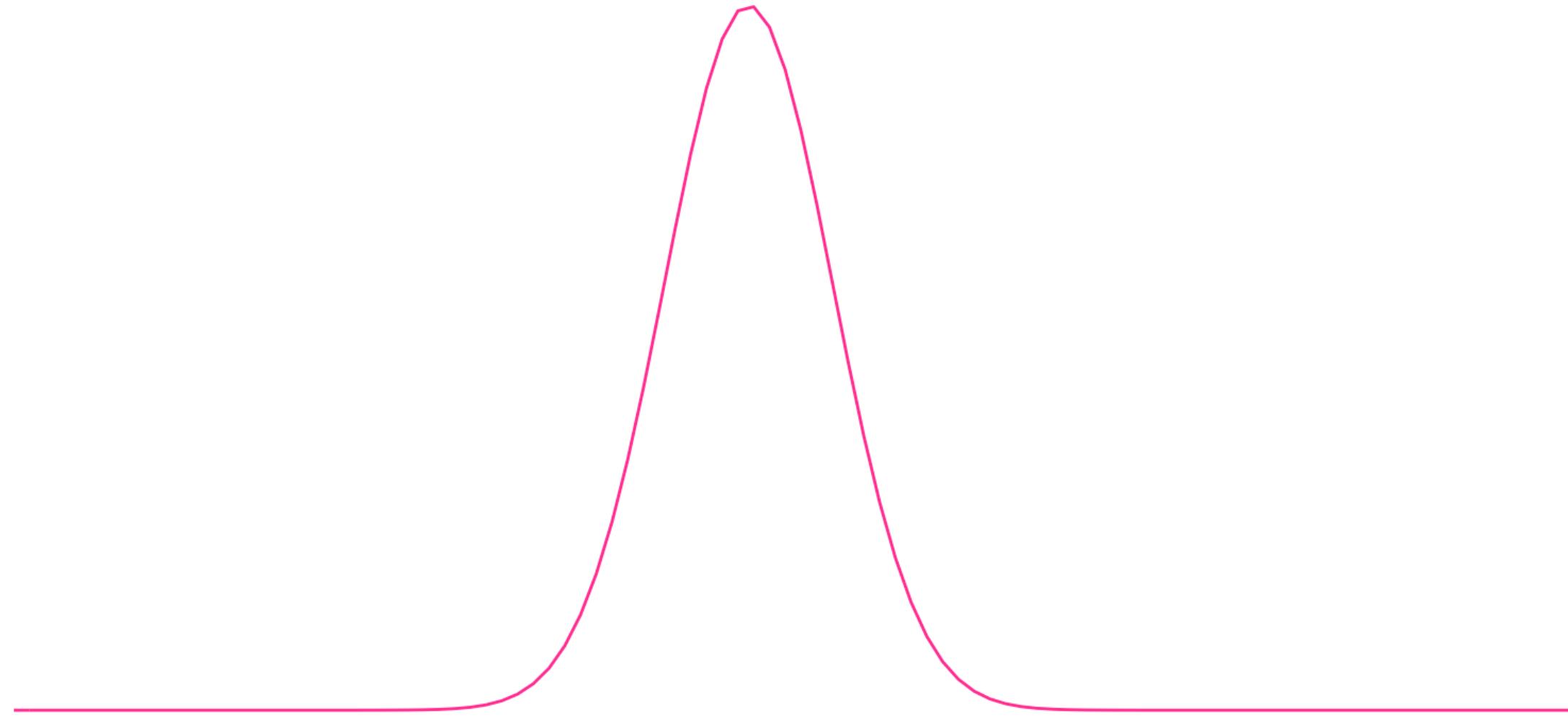
г. Москва, ул. Ленина, д. 1

Как выбрать статистический критерий?

e^xperiment fest

У этой метрики может быть такое распределение

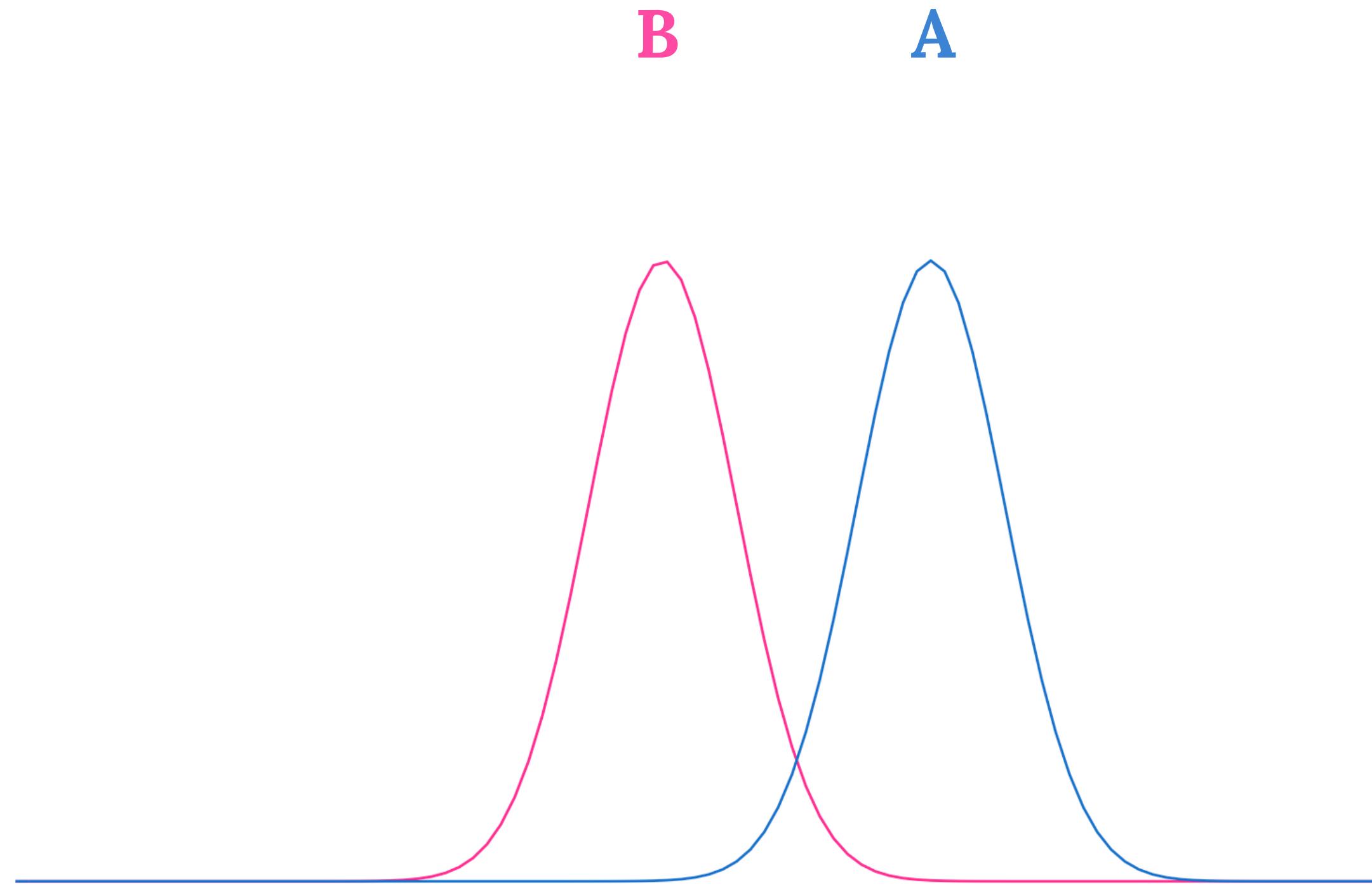
B



Как выбрать статистический критерий?

e^xperiment fest

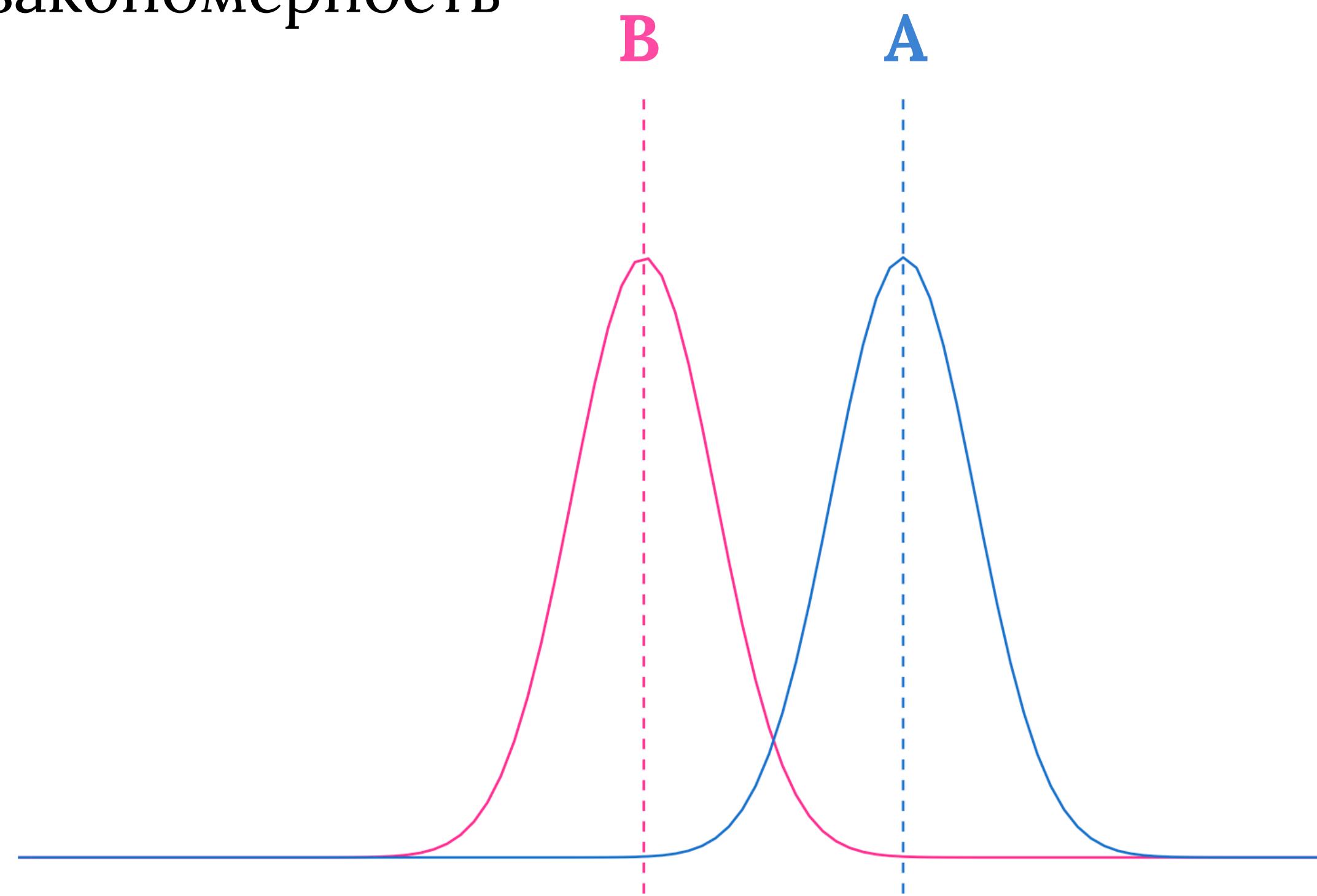
Соответственно, распределение той же метрики должна обладать схожей формой. Как нам их сравнить?



Как выбрать статистический критерий?

e^xperiment fest

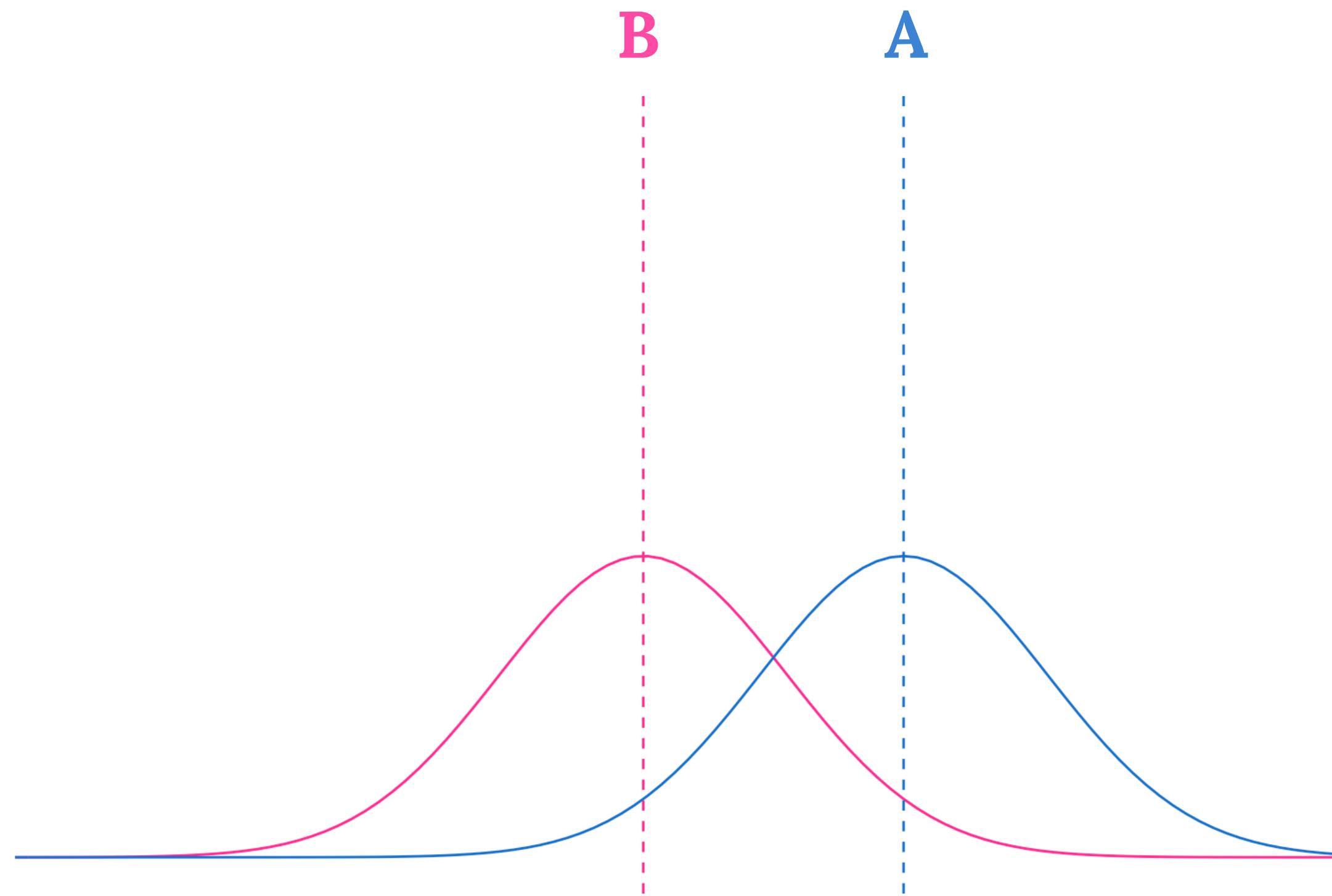
Конечно же, сравнив средние. Средние сильно отдалены, но нам бы хотелось быть уверенными в том, что мы не наблюдаем случайность, а видим закономерность



Как выбрать статистический критерий?

e^xperiment fest

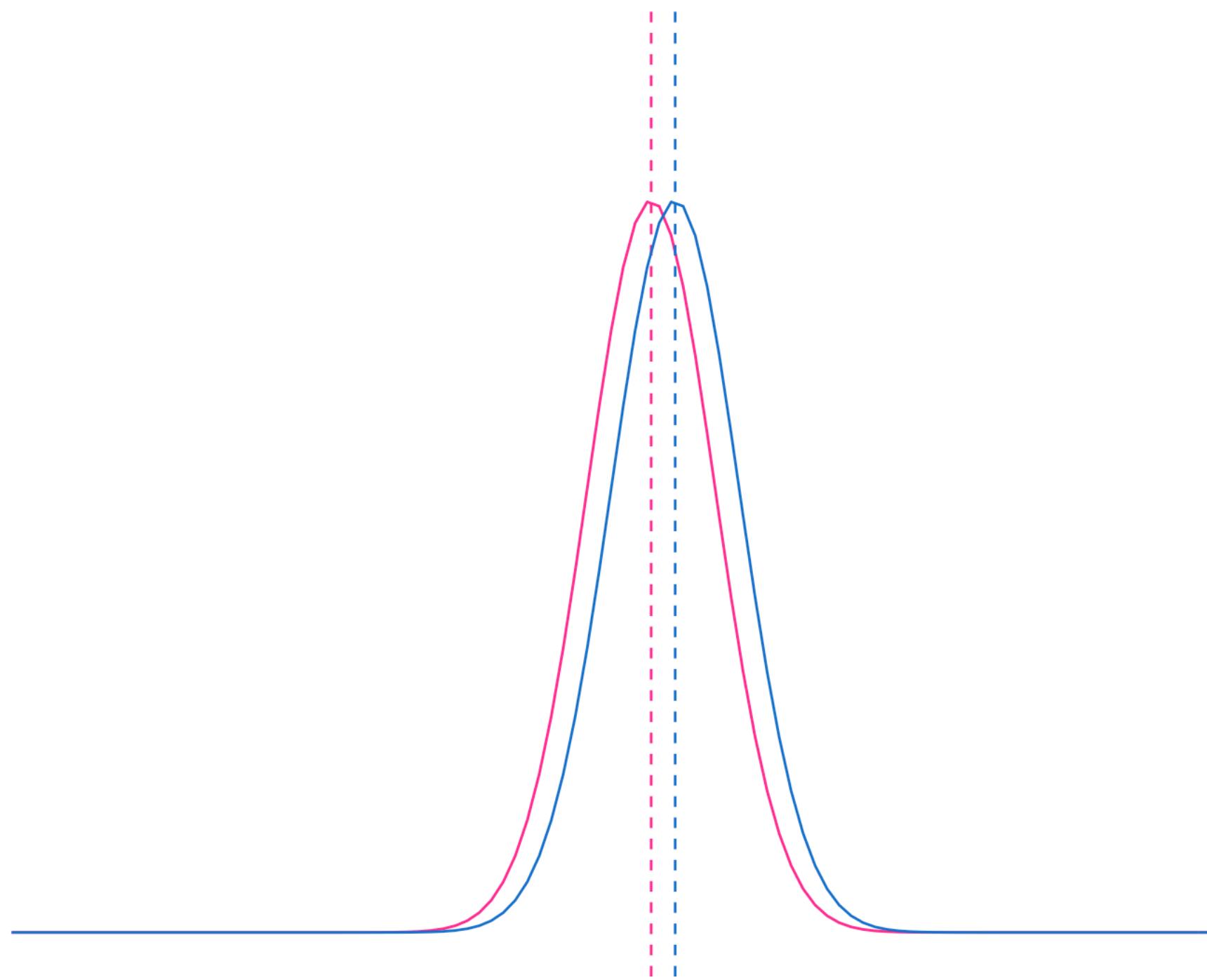
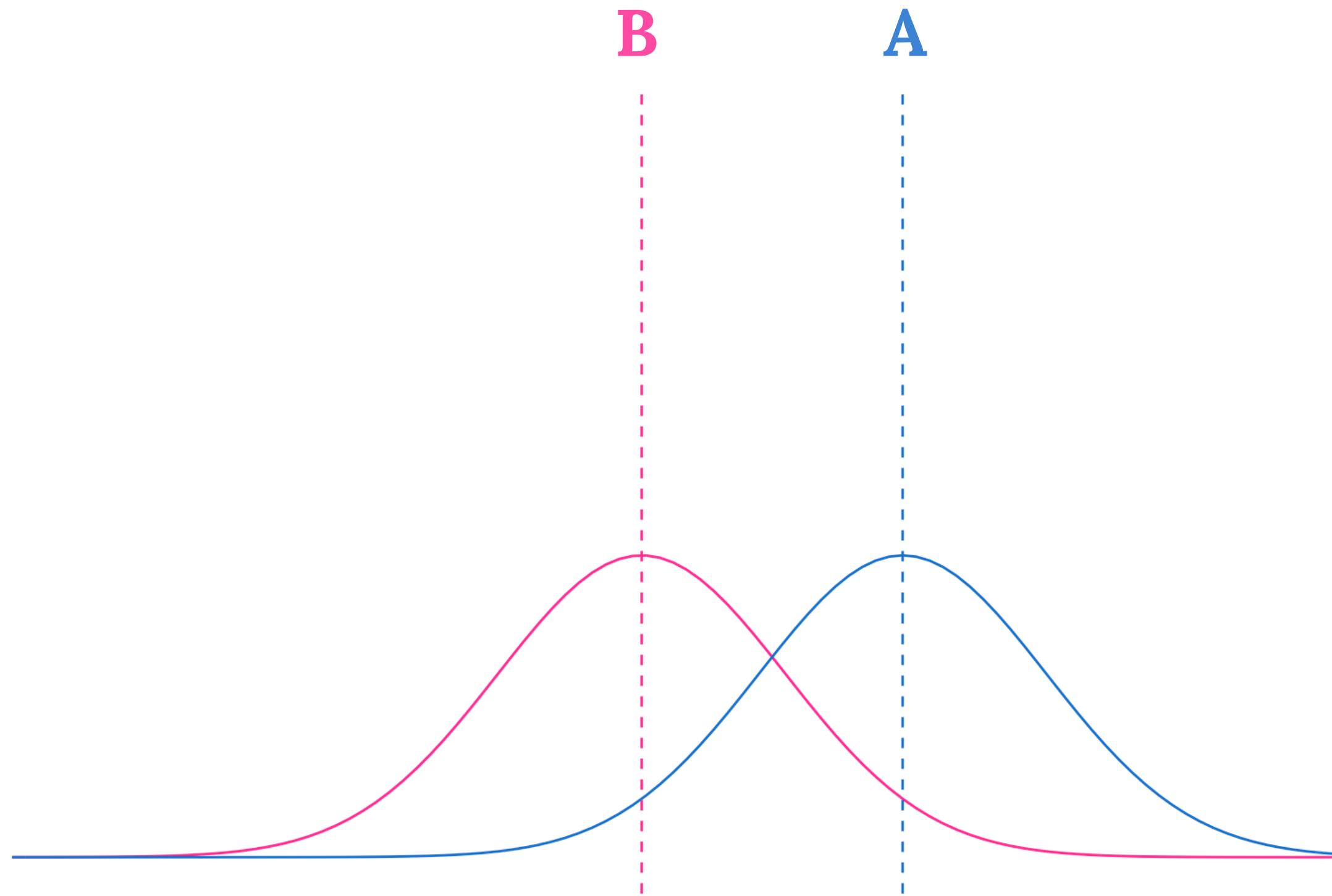
а этому может мешать большая дисперсия у распределений метрики



Как выбрать статистический критерий?

e^xperiment fest

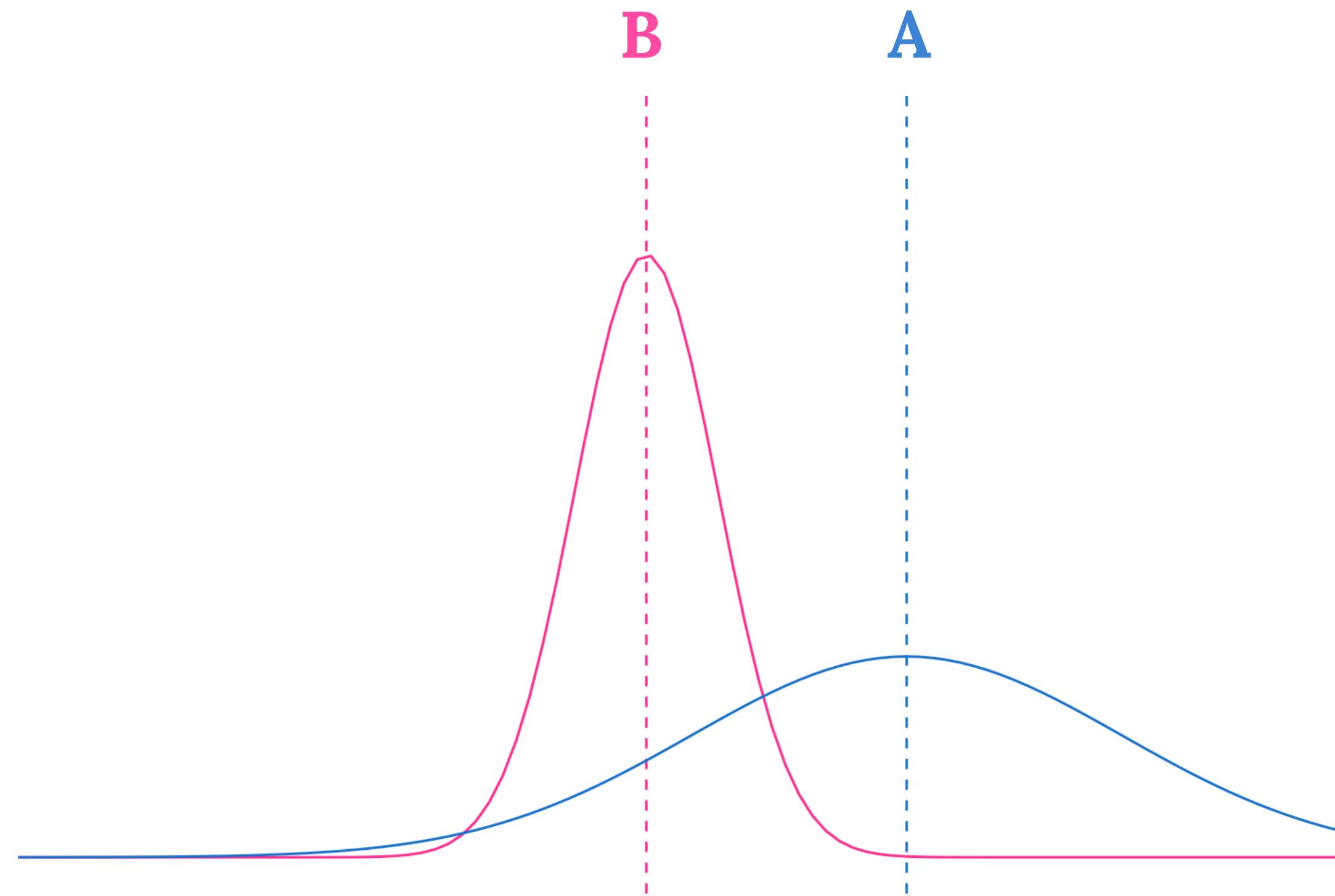
или отсутствие разницы между выборками



Как выбрать статистический критерий?

e^xperiment fest

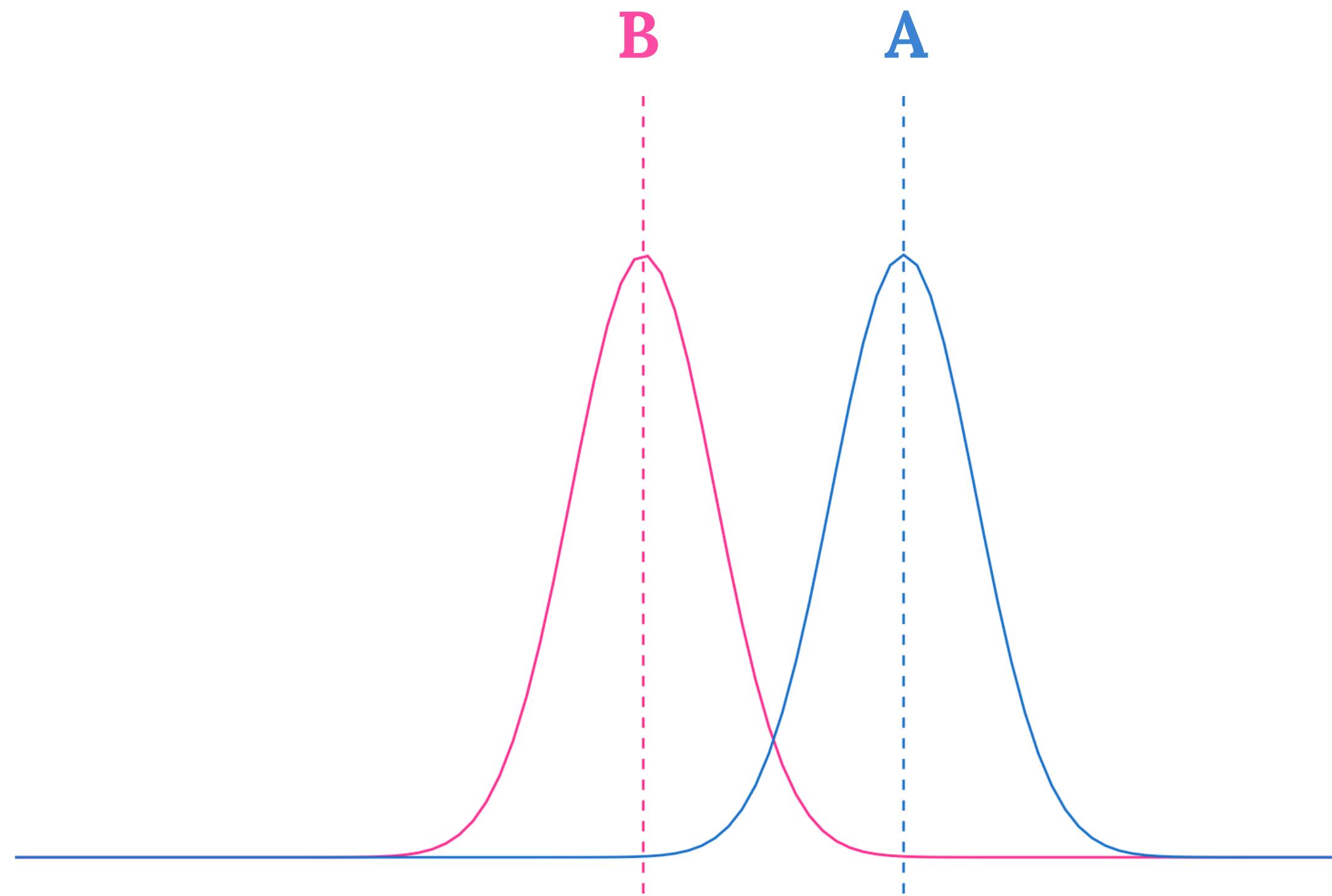
бывает так, что дисперсия у одного распределения больше, чем у другого (напр., из-за дисбаланса выборок)



Как выбрать статистический критерий?

e^xperiment fest

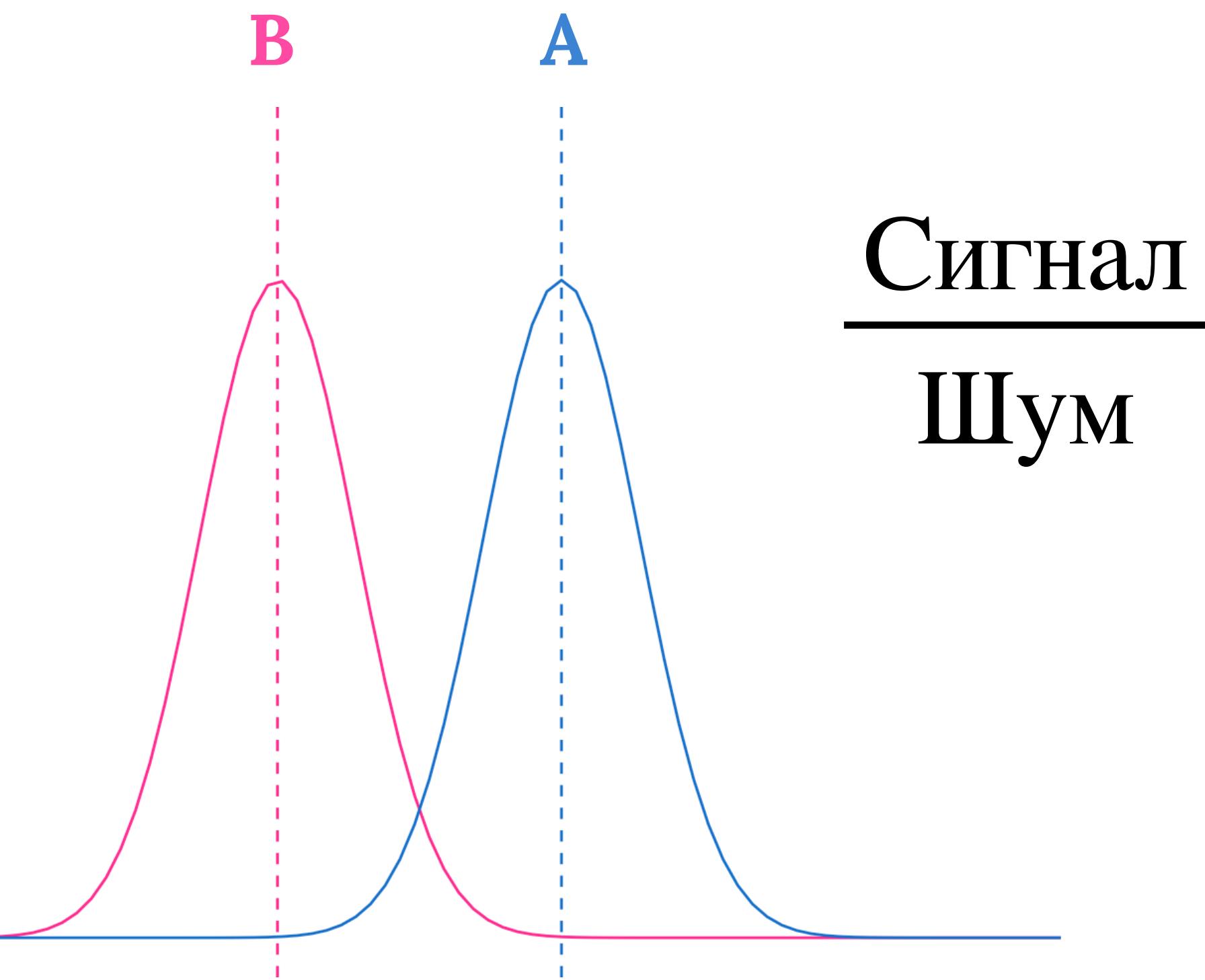
Чтобы сравнить две выборки, нам достаточно рассчитать t -значение:



Как выбрать статистический критерий?

e^xperiment fest

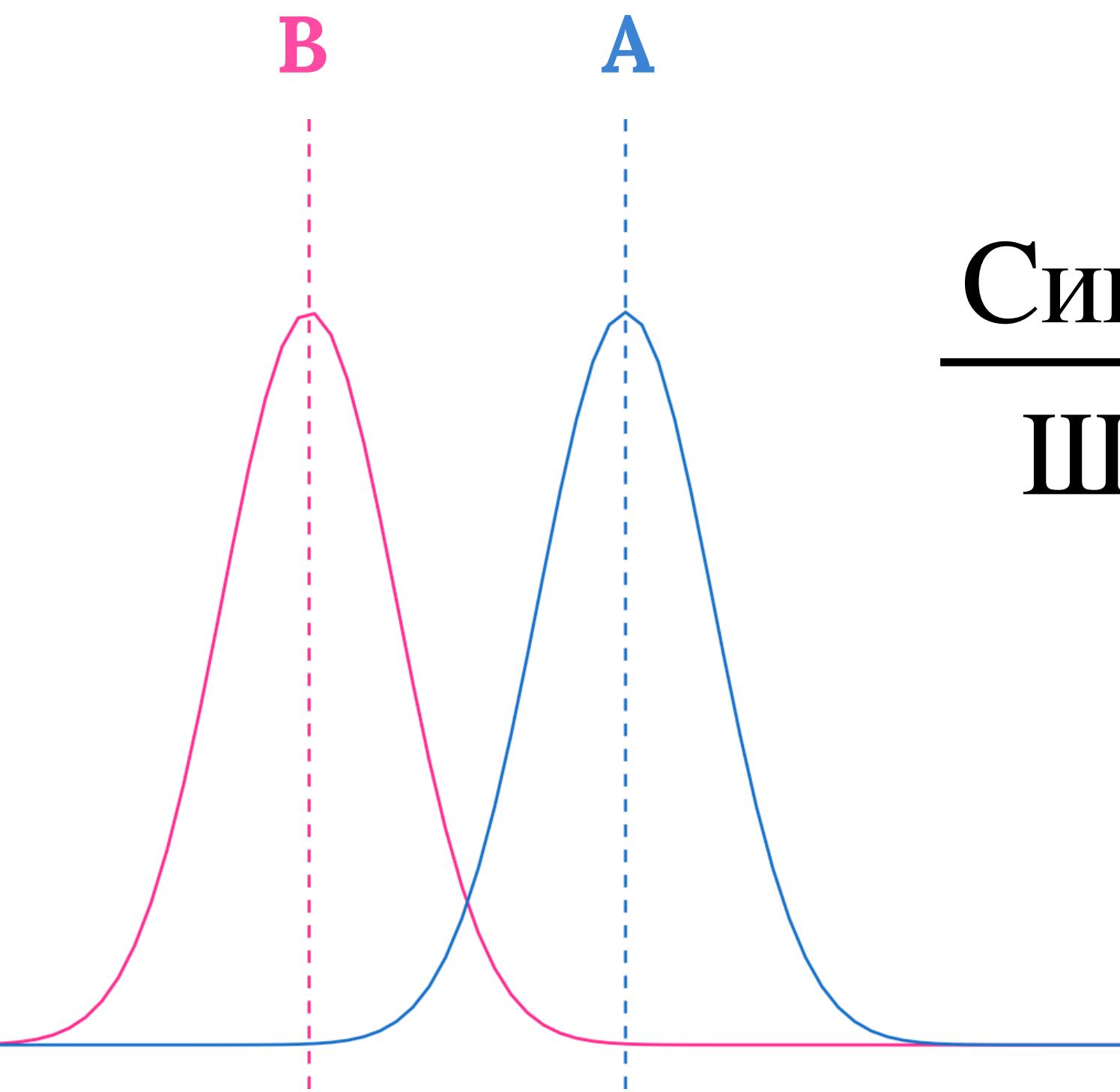
Чтобы сравнить две выборки, нам достаточно рассчитать t -значение:



Как выбрать статистический критерий?

e^xperiment fest

Чтобы сравнить две выборки, нам достаточно рассчитать t -значение:

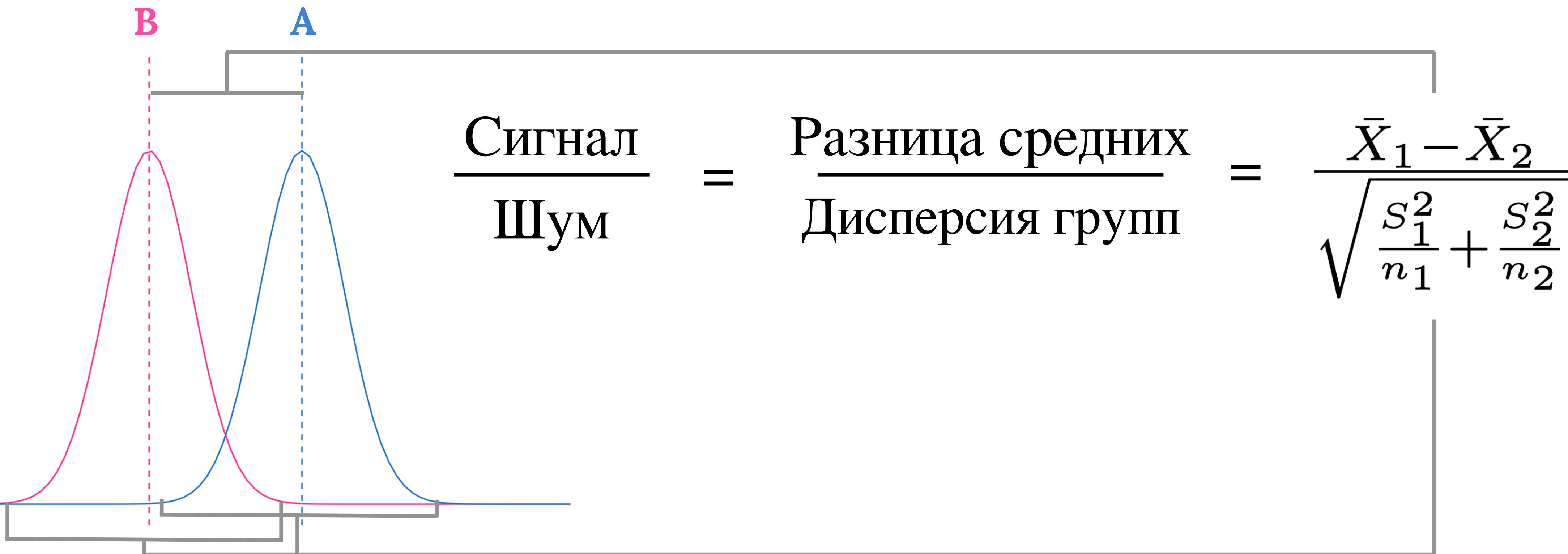


$$\frac{\text{Сигнал}}{\text{Шум}} = \frac{\text{Разница средних}}{\text{Дисперсия групп}}$$

Как выбрать статистический критерий?

e^xperiment fest

Чтобы сравнить две выборки, нам достаточно рассчитать t -значение:



Как выбрать статистический критерий?

e^xperiment fest

Разница средних Дисперсия групп



Если эксперимент направлен на изменение удобства, то разница будет видна в поведенческих метриках (напр., время прохождения сценария). В деньгах разницу мы заметим не сразу (если вообще она есть)

Шум зависит от: размера выборки, сезонности, отсутствии репрезентативности в группах, хаотичном поведении определенных подвыборок пользователей и других внешних факторов

Посмотрим на данные и посчитаем

Форма без адреса

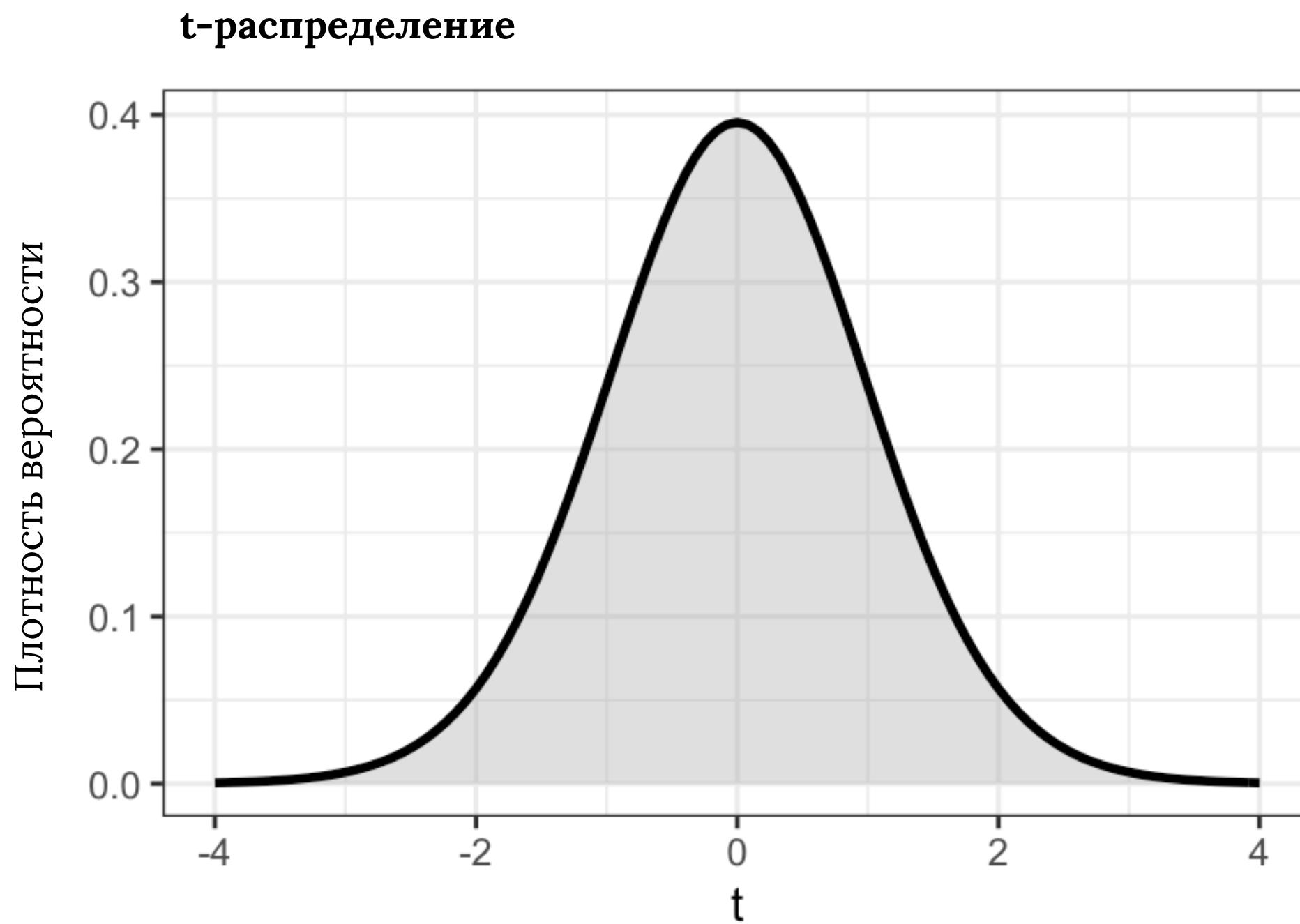
Среднее время звонка = 124 сек

n = 30

Форма с адресом

Среднее время звонка = 108 сек

n = 30



Как выбрать статистический критерий?

e^xperiment fest

Посмотрим на данные и посчитаем

Форма без адреса

Среднее время звонка = 124 сек

n = 30

Форма с адресом

Среднее время звонка = 108 сек

n = 30

Итого

Разница ($\bar{x}_1 - \bar{x}_2$) = 15,93 сек

Стандартная ошибка (SE) = 7,27

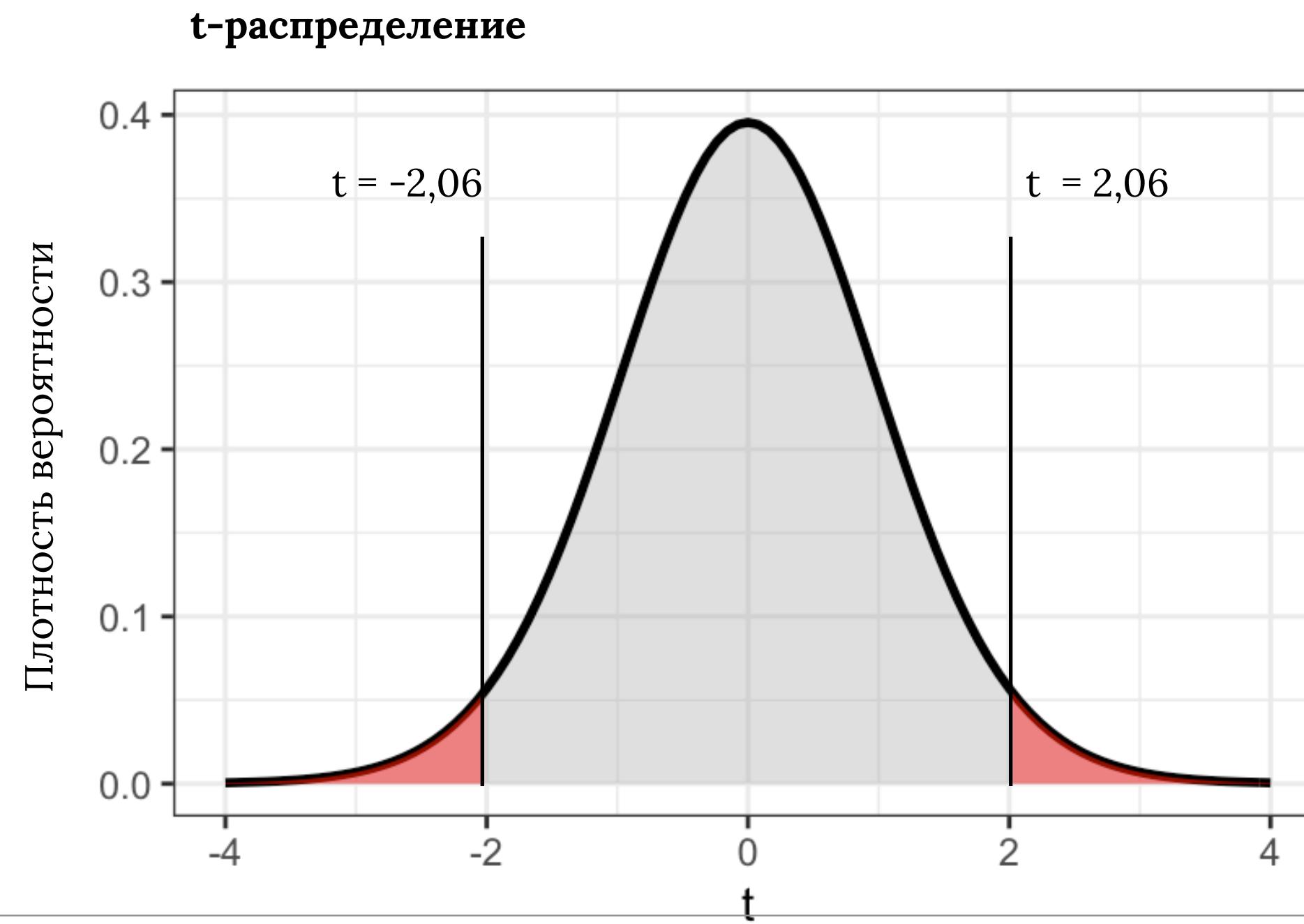
$$t = 15,93 \div 7,27 = 2,06$$

p-val (R) = (1-pt(2.06, df=60-2))

или ищем в таблице критических значений и находим 0,0218

$$p\text{-val} = 0,0218 * 2 = 0,04375$$

Как выбрать статистический критерий?

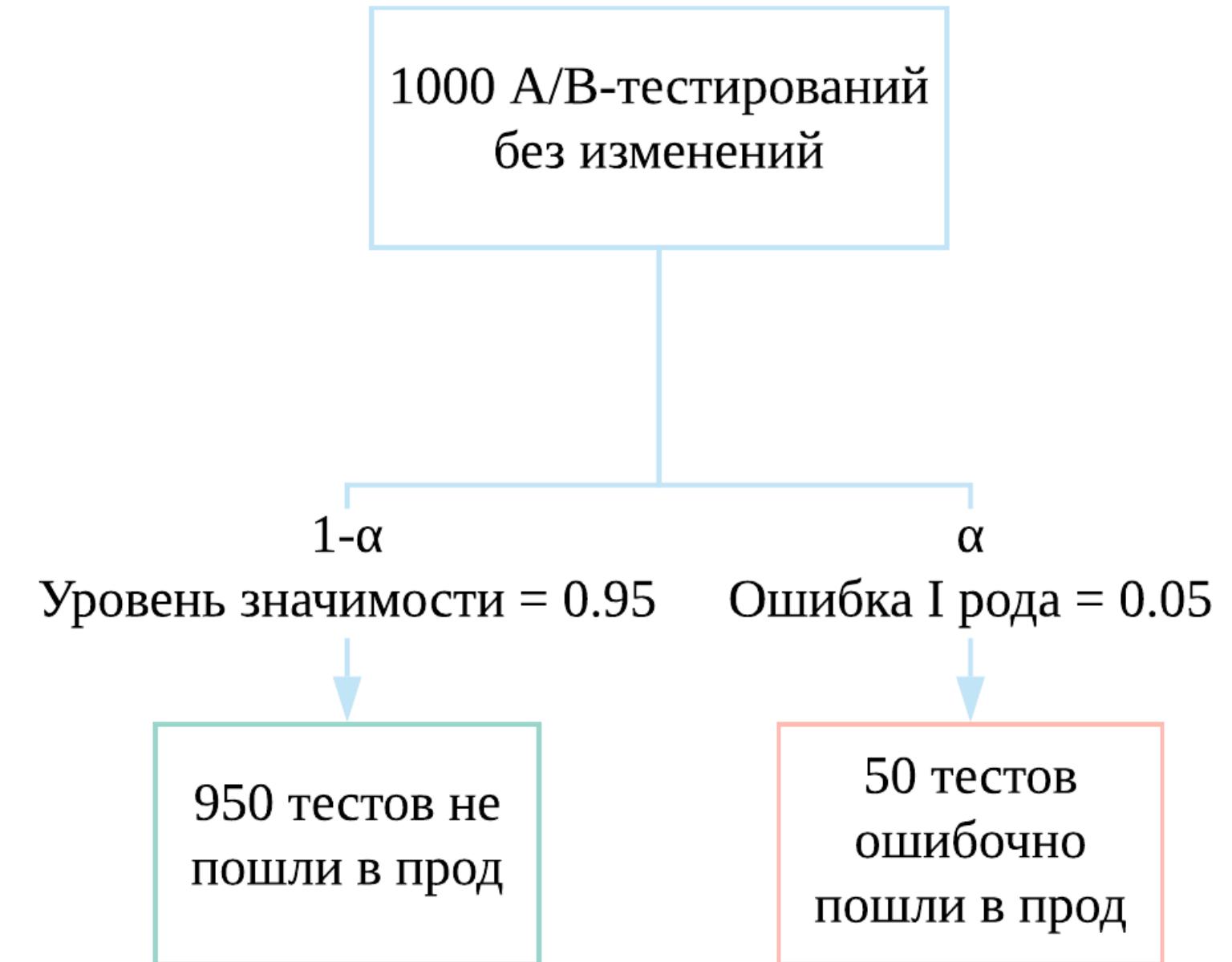


Вероятность наблюдать такие явления (p-value) = 0,04

Наблюдать такие или более экстремальные явления = 0.04. При уровне значимости альфа < 0.05, мы отвергаем нулевую гипотезу и принимаем альтернативу

Мы получили $p\text{-value} = 0.04$. На уровне альфы > 0.05 мы отвергаем нулевую гипотезу, а на 0.01 уже нет. Тогда как быть?

Нет универсальных пороговых значений



Как выбрать статистический критерий?

e^xperiment fest

t-Критерий – целое семейство статистических методов по проверке гипотез:

One Sample T-Test		Two Sample T-Test		
		Independent Sample		
	Paired Sample	Equal Sample Size, not Variance	Equal or Unequal Sample Size, Similar Variance	Equal or Unequal Sample Size, Unequal Variance
$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$	$t = \frac{\bar{X}_D - \mu_0}{s_D/\sqrt{n}}$	$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \sqrt{\frac{2}{n}}}$ <p>where</p> $s_p = \sqrt{\frac{s_{X_1}^2 + s_{X_2}^2}{2}}$	$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ <p>where</p> $s_p = \sqrt{\frac{(n_1 - 1) s_{X_1}^2 + (n_2 - 1) s_{X_2}^2}{n_1 + n_2 - 2}}$	$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{\Delta}}}$ <p>where</p> $s_{\bar{\Delta}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

– здесь перечислены далеко не все вариации t-Критерия. Помимо этого, существует еще проблема отсутствия возможности точно сравнить средние значения в выборках, дисперсии которых неизвестны. Эта проблема названа проблемой Беренса-Фишера. Для этого используют приближение Уэлша и то не во всех случаях

Как выбрать статистический критерий?

e^xperiment fest

Аналитик Лёня сравнивает две версии одной формы: с текстовыми ярлыками для полей и без. Какая форма эффективнее?

Имя

email

Телефон

Имя

Александр

email

example@example.com

Телефон

+7(999)9999999

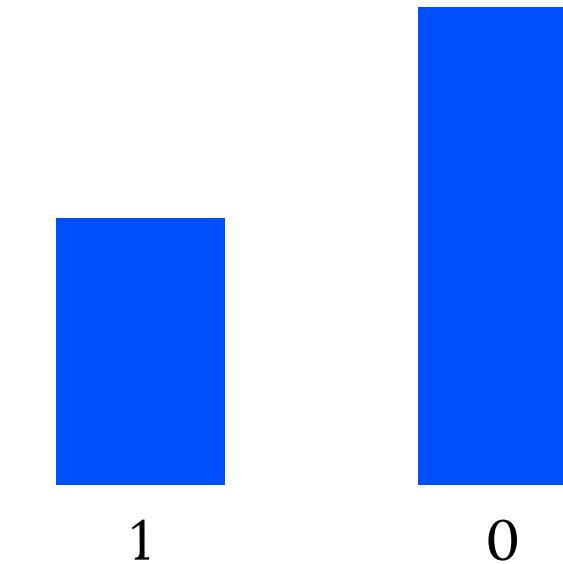
В этом поможет критерий долей

Как выбрать статистический критерий?

e^xperiment fest

Критерии долей – это критерии, которые работают с распределениями Бернулли.

Они принимают на вход выборки из нулей и единиц и проверяют гипотезы о параметрах p этих распределений (вероятность появления единицы в выборке).



Бернулли удобно, т.к. в отличие от нормального распределения, не нужно применять никаких критериев, чтобы узнать взята ли выборка из конкретного распределения

Как выбрать статистический критерий?

e^xperiment fest

Посмотрим на таблицу с результатами и сравним их. Кажется, что тестовая группа (\hat{p}_2) эффективнее, чем контрольная (\hat{p}_1). Но нам необходимо это проверить. Будем использовать *z-тест для проверки пропорций*.

	Увидели форму	Отправили форму	Конверсия (доля отправленных)
Тестовая группа	2100	156	0.0742
Контрольная группа	2100	129	0.0614

Если $\hat{p}_2 - \hat{p}_1 = 0$, тогда разницы между конверсиями нет. Мы посчитаем для этой разницы ДИ (с помощью стандартной ошибки), чтобы узнать, наблюдаем ли мы случайность или все же закономерность

Как выбрать статистический критерий?

e^xperiment fest

Посмотрим на таблицу с результатами и сравним их. Кажется, что тестовая группа (\hat{p}_2) эффективнее, чем контрольная (\hat{p}_1). Но нам необходимо это проверить. Будем использовать *z-тест для проверки пропорций*.

	Увидели форму	Отправили форму	Конверсия (доля отправленных)
Тестовая группа	2100	156	0.0742
Контрольная группа	2100	129	0.0614

Формула расчета
дисперсии для
распределения Бернулли

$$p \times (1 - p)$$

$$\hat{p}_2 - \hat{p}_1 \pm 1.96 \times \sqrt{\hat{p}_1 \times (1 - \hat{p}_1) / n_1 + \hat{p}_2 \times (1 - \hat{p}_2) / n_2}$$

\hat{p}_1 и \hat{p}_2 — измеренные в результате А/Б-теста конверсии,

n_1 и n_2 — количество элементов в каждой группе.

Как выбрать статистический критерий?

e^xperiment fest

Посмотрим на таблицу с результатами и сравним их. Кажется, что тестовая группа (\hat{p}_2) эффективнее, чем контрольная (\hat{p}_1). Но нам необходимо это проверить. Будем использовать *z-тест для проверки пропорций*.

	Увидели форму	Отправили форму	Конверсия (доля отправленных)
Тестовая группа	2100	156	0.0742
Контрольная группа	2100	129	0.0614

Формула расчета дисперсии для распределения Бернулли
 $p \times (1 - p)$

$0.0742 - 0.0614 \pm 1.96 \times \sqrt{0.0742 \times (1 - 0.0742) \div 2100 + 0.0614 \times (1 - 0.0614) \div 2100}$

0.0742 и 0.0614 – измеренные в результате А/Б-теста конверсии,
 2100 и 2100 – количество элементов в каждой группе.

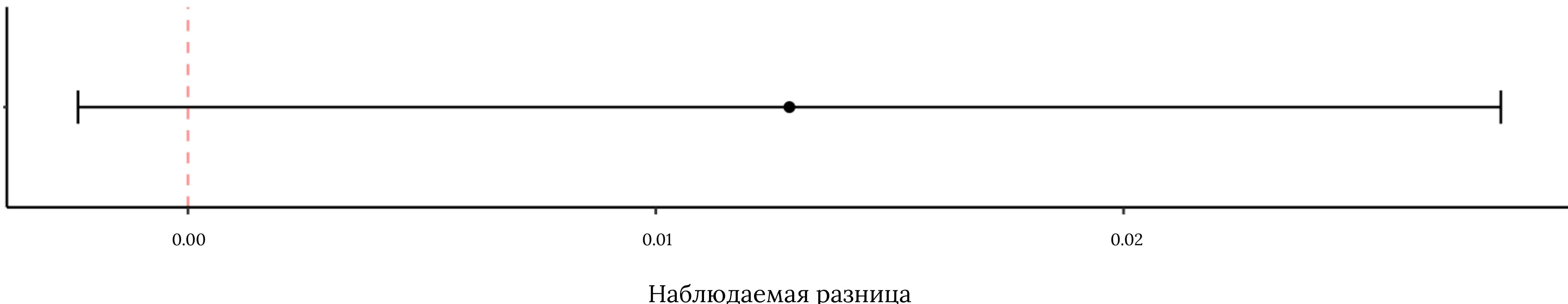
Как выбрать статистический критерий?

e^xperiment fest

Доверительный интервал для разницы = [-0.002350384;0.02806466].

Видим, что разница попадает в ноль. Это значит, что мы не можем отвергнуть нулевую гипотезу на уровне 95%. Обычно в такие моменты ждут дольше для достижения необходимой мощности. Потому как здесь, мощность достаточно низкая (< 80%), а это значит, что верить этим результатам мы не можем

Доверительный интервал для разницы пропорций



Как выбрать статистический критерий?

e^xperiment fest

Непараметрические критерии

В отличие от параметрических критериев, у непараметрических распределение данных неизвестно. При использовании этих критериев часто действия производятся не с самими значениями в выборке и параметрах распределения, а с их рангами

Как выбрать статистический критерий?

e^xperiment fest

Лёне приходится несладко, потому что Петя неистово запускает тесты один за другим. Теперь ему нужно сравнить средние чеки между двумя алгоритмами рекомендательной системы в той же пиццерии!

Bill 1 (Left):

ЗАО МОСКВА-МАКДОНАЛДС СПАСИБО! ЖДЕМ ВАС СНОВА! ПРИЯТНОГО АППЕТИТА.	
Ул. Мясниковская, 30/1/2 стр. 1	тел. 625-42-90
Касса № 08	ИНН: 007710044132
Дата 05.03.2012	ДОК. № 721064
Время 19:00	ФИСК. ЧЕК № 000412
Заказ № 956	
ЛАТТЕ	1 а 73.00 00.83
МАКЧИКЕН	1 а 79.00 00.83
КАРТОФЕЛЬ-ФРИ БОЛ.	1 а 65.00 00.83
СОУС СЫРНЫЙ	1 а 18.00 00.81
ИТОГ	*235.00
В т.ч. Налоги:	35.85
Нал а: 18%	
НАЛИЧНЫЕ РУБ.	500.00
СДАЧА	*265.00
КАССИР ГЕОРГИЕВСКАЯ Ю.	
ФР № 0002634	РЕГ. 000000084794
ЭКПЗ 0699182259	ФП 00158540 #022308

Bill 2 (Right):

ЗАО МОСКВА-МАКДОНАЛДС СПАСИБО! ЖДЕМ ВАС СНОВА! ПРИЯТНОГО АППЕТИТА.	
Ул. Мясниковская, 30/1/2 стр. 1	тел. 625-42-90
Касса № 08	ИНН: 007710044132
Дата 05.03.2012	ДОК. № 721064
Время 19:00	ФИСК. ЧЕК № 000412
Заказ № 956	
ЛАТТЕ	1 а 73.00 00.83
МАКЧИКЕН	1 а 79.00 00.83
КАРТОФЕЛЬ-ФРИ БОЛ.	1 а 65.00 00.83
СОУС СЫРНЫЙ	1 а 18.00 00.81
ИТОГ	*235.00
В т.ч. Налоги:	35.85
Нал а: 18%	
НАЛИЧНЫЕ РУБ.	500.00
СДАЧА	*265.00
КАССИР ГЕОРГИЕВСКАЯ Ю.	
ФР № 0002634	РЕГ. 000000084794
ЭКПЗ 0699182259	ФП 00158540 #022308

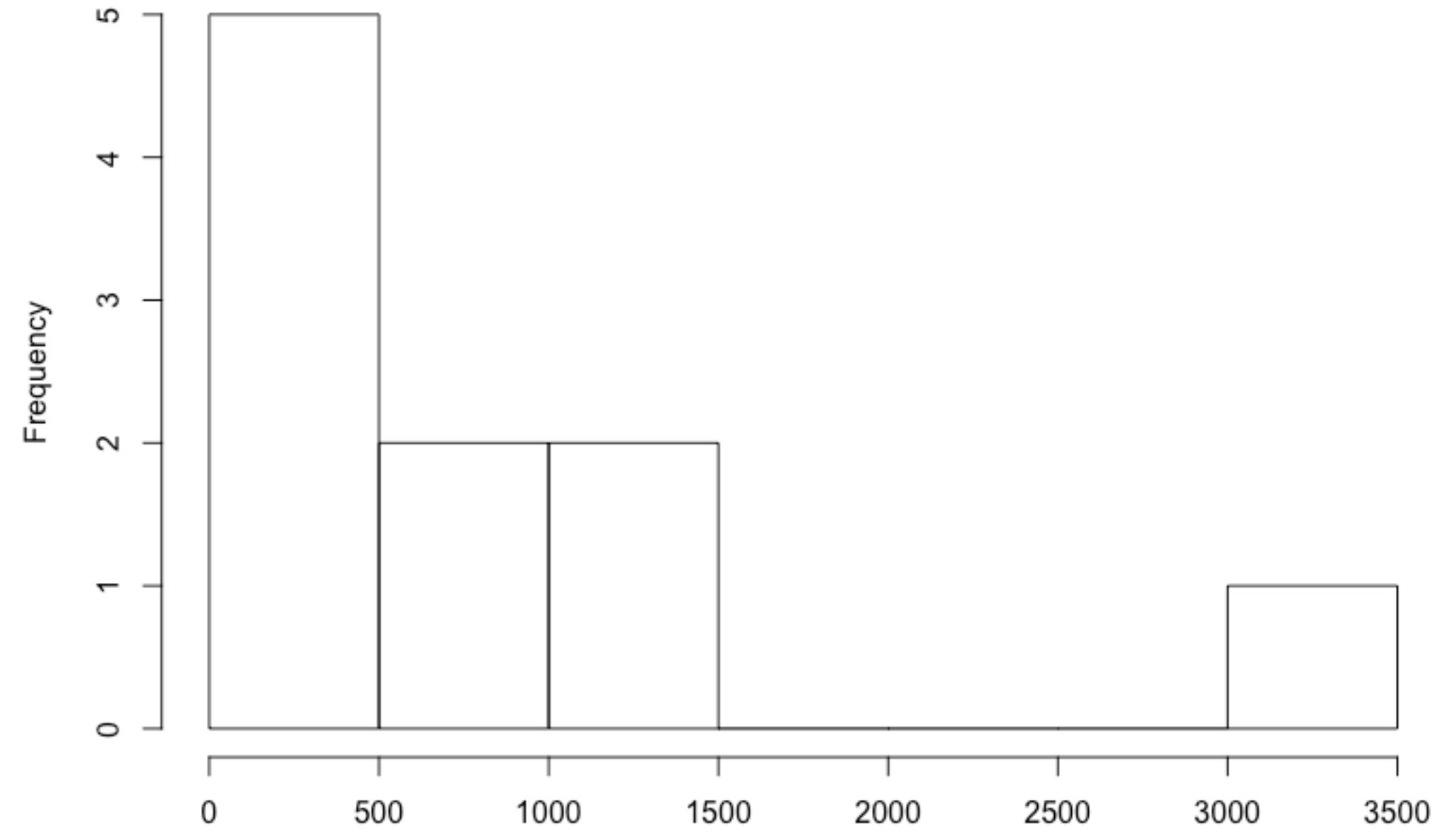
Чтобы решить задачу, нам поможет непараметрический U- критерий Манна-Уитни

Как выбрать статистический критерий?

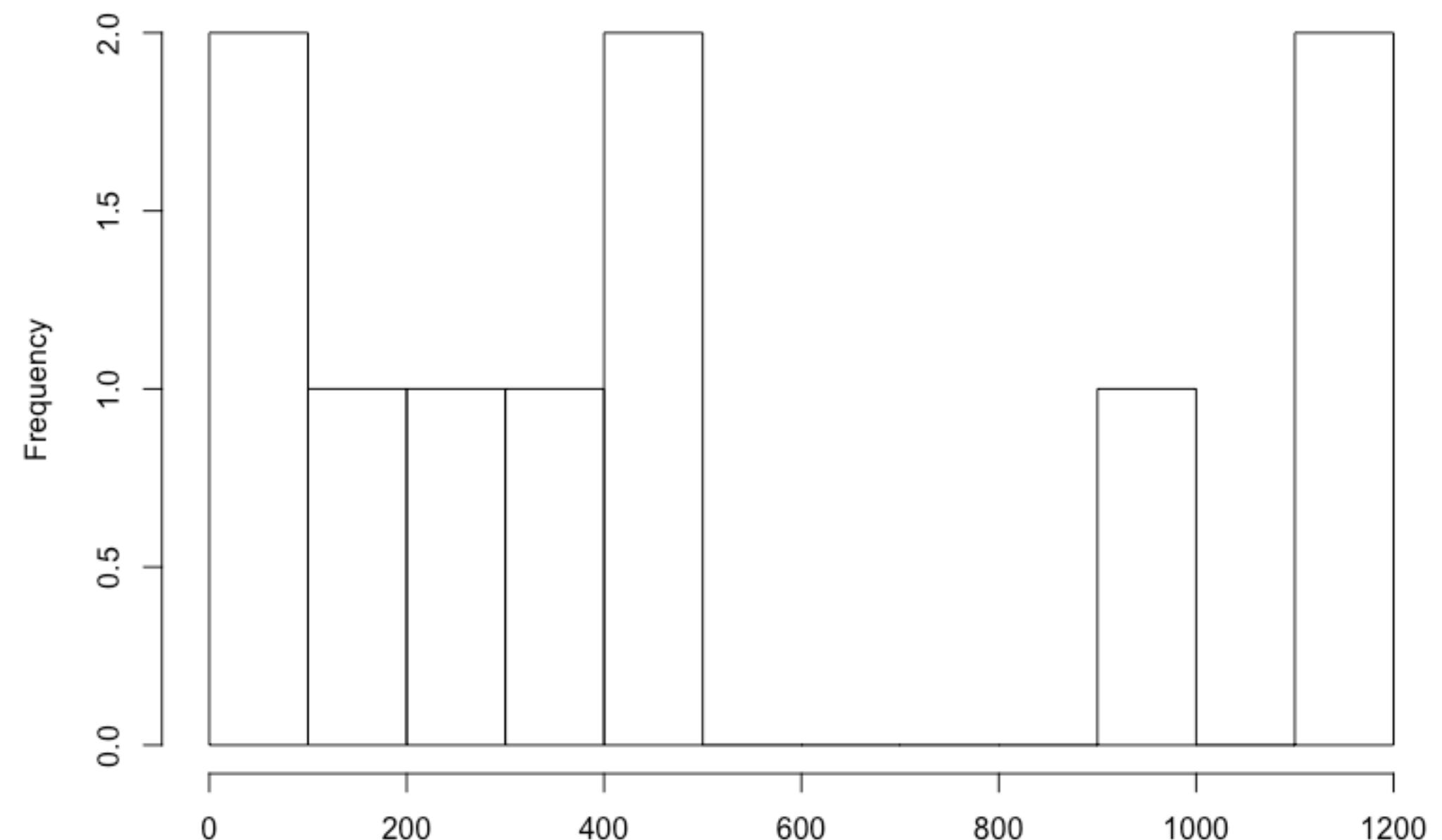
e^xperiment fest

Распределения явно отличаются от нормального

Группа А



Группа Б



Как выбрать статистический критерий?

e^xperiment fest

Статистики придумали хитрое решение таких задач.

Суть заключается в том, что мы преобразуем все значения в их последовательные ранги. Ниже таблица для двух групп

№ чека	1	2	3	4	5	6	7	8	9	10
Чеки А	0	100	0	700	3300	1100	700	500	400	1400
Чеки В	0	300	1000	0	200	500	500	1200	400	1200

Как выбрать статистический критерий?

e^xperiment fest

1. Мы «склеиваем» таблицы для двух групп и отсортируем их по возрастанию

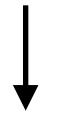
Группа	A	A	B	B	A	A	A	B	A	A	B	B	B	A	B	A	A	B	B	
Чек	0	0	0	0	100	200	300	400	400	500	500	500	700	700	1000	1100	1200	1200	1400	3300

Как выбрать статистический критерий?

e^xperiment fest

1. Мы «склеиваем» таблицы для двух групп и отсортируем их по возрастанию

Группа	A	A	B	B	A	A	A	B	A	A	B	B	B	A	B	A	A	B	B		
Чек	0	0	0	0	100	200	300	400	400	500	500	500	700	700	700	1000	1100	1200	1200	1400	3300



2. Присвоим каждому значению свой ранг. Сначала, пусть это будет порядковый номер

Группа	A	A	B	B	A	A	A	B	A	A	B	B	B	A	B	A	A	B	B		
Чек	0	0	0	0	100	200	300	400	400	500	500	500	700	700	700	1000	1100	1200	1200	1400	3300
Ранг	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	

Как выбрать статистический критерий?

e^xperiment fest

2. Присвоим каждому значению свой ранг. Сначала, пусть это будет порядковый номер

Группа	A	A	B	B	A	A	A	B	A	A	B	B	B	A	B	A	A	B	B	
Чек	0	0	0	0	100	200	300	400	400	500	500	500	700	700	1000	1100	1200	1200	1400	3300
Ранг	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20



3. Но можно заметить, что некоторые значения повторяются, имея при этом разные ранги. Очень важно это поправить, взяв среднее между рангами для повторяющихся значений.

Группа	A	A	B	B	A	A	A	B	A	A	B	B	B	A	B	A	A	B	B	
Чек	0	0	0	0	100	200	300	400	400	500	500	500	700	700	1000	1100	1200	1200	1400	3300
Ранг	2.5	2.5	2.5	2.5	5	6	7	8.5	8.5	11	11	11	13.5	13.5	15	16	17.5	17.5	19	20

Как выбрать статистический критерий?

e^xperiment fest

Группа	A	A	B	B	B	A	A	B	A	A	B	B	B	A	B	A	A	B	B	
Чек	0	0	0	0	100	200	300	400	400	500	500	500	700	700	1000	1100	1200	1200	1400	3300
Ранг	2.5	2.5	2.5	2.5	5	6	7	8.5	8.5	11	11	11	13.5	13.5	15	16	17.5	17.5	19	20

4. А теперь просуммируем ранги погруппно, получаем $R_1 = 98.5$ и $R_2 = 111.5$, где R_1 это вариант A, и $R_2 = B$, соответственно. Подставляем эти значения в формулу

$$U_1 = n_1 \times n_2 + n_1 \times (n_1 + 1) \div 2 - R_1$$

$$U_2 = n_1 \times n_2 + n_2 \times (n_2 + 1) \div 2 - R_2$$

В итоге получаем $U_1 = 56.5$ и $U_2 = 43.5$, берем меньший, т.е. 43.5

Как выбрать статистический критерий?

e^xperiment fest

5. Ищем в таблице критических значений (как мы делали до этого с z и t критериями) соответствующее нашему и видим $U_{\text{критическое}} = 23$.

$U_{\text{критическое}} < U$, отвергаем нулевую гипотезу о равенстве двух распределений

К счастью, в R все это
делается одной командой

wilcox.test()

Таблица

Критические значения критерия У Манна-Уитни для уровней статистической значимости $p \leq 0,05$ и $p \leq 0,01$ (по Гублеру Е.В.,

Генкину А.А., 1973). Различия между двумя выборками можно считать значимыми ($p < 0,05$), если $U_{\text{эмп}} \leq U_{0,05}$, и тем более достоверными ($p < 0,01$), если $U_{\text{эмп}} \leq U_{0,01}$.

n_1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
n_2																			$\rho=0,05$
3	-	0																	
4	-	0	1																
5	0	1	2	4															
6	0	2	3	5	7														
7	0	2	4	6	8	11													
8	1	3	5	8	10	13	15												
9	1	4	6	9	12	15	18	21											
10	1	4	7	11	14	17	20	24	27										
11	1	5	8	12	16	19	23	27	31	34									
12	2	5	9	13	17	21	26	30	34	38	42								
13	2	6	10	15	19	24	28	33	37	42	47	51							
14	3	7	11	16	21	26	31	36	41	46	51	56	61						
15	3	7	12	18	23	28	33	39	44	50	55	61	66	72					
16	3	8	14	19	25	30	36	42	48	54	60	65	71	77	83				
17	3	9	15	20	26	33	39	45	51	57	64	70	77	83	89	96			
18	4	9	16	22	28	35	41	48	55	61	68	75	82	88	95	102	109		
19	4	10	17	23	30	37	44	51	58	65	72	80	87	94	101	109	116	123	
20	4	11	18	25	32	39	47	54	62	69	77	84	92	100	107	115	123	130	137
																			$\rho=0,01$
5	-	-	0	1															
6	-	-	1	2	3														
7	-	0	1	3	4	6													
8	-	0	2	4	6	7	9												
9	-	1	3	5	7	9	11	14											
10	-	1	3	6	8	11	13	16	19										
11	-	1	4	7	9	12	15	18	22	25									
12	-	2	5	8	11	14	17	21	24	28	31								
13	0	2	5	9	12	16	20	23	27	31	35	39							
14	0	2	6	10	13	17	22	26	30	34	38	43	47						
15	0	3	7	11	15	19	24	28	33	37	42	47	51	56					
16	0	3	7	12	16	21	26	31	36	41	46	51	56	61	66				
17	0	4	8	13	18	23	28	33	38	44	49	55	60	66	71	77			
18	0	4	9	14	19	24	30	36	41	47	53	59	65	70	76	82	88		
19	1	4	9	15	20	26	32	38	44	50	56	63	69	75	82	88	94	101	
20	1	5	10	16	22	28	34	40	47	53	60	67	73	80	87	93	100	107	114

Таблица II. Продолжение

n_1	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
n_2	$\rho=0.05$																	
21	19	26	34	41	49	57	65	73	81	89	97	105	113	121	130	138	146	154
22	20	28	36	44	52	60	69	77	85	94	102	111	119	128	136	145	154	162
23	21	29	37	46	55	63	72	81	90	99	107	116	125	134	143	152	161	170
24	22	31	39	48	57	66	75	85	94	103	113	122	131	141	150	160	169	179
25	23	32	41	50	60	69	79	89	98	108	118	128	137	147	157	167	177	187
26	24	33	43	53	62	72	82	93	103	113	123	133	143	154	164	174	185	195
27	25	35	45	55	65	75	86	96	107	118	128	139	150	160	171	182	193	203
28	26	36	47	57	68	79	89	100	111	122	133	144	156	167	178	189	200	212
29	27	38	48	59	70	82	93	104	116	127	139	150	162	173	185	196	208	220
30	28	39	50	62	73	85	96	108	120	132	144	156	168	180	192	204	216	228
31	29	41	52	64	76	88	100	112	124	137	149	161	174	186	199	211	224	236
32	30	42	54	66	78	91	103	116	129	141	154	167	180	193	206	219	232	245
33	31	43	56	68	81	94	107	120	133	146	159	173	186	199	213	226	239	253
34	32	45	58	71	84	97	110	124	137	151	164	178	192	206	219	233	247	261
35	33	46	59	73	86	100	114	128	142	156	170	184	198	212	226	241	255	269
36	35	48	61	75	89	103	117	132	146	160	175	189	204	219	233	248	263	278
37	36	49	63	77	92	106	121	135	150	165	180	195	210	225	240	255	271	286
38	37	51	65	79	94	109	124	139	155	170	185	201	216	232	247	263	278	294
39	38	52	67	82	97	112	128	143	159	175	190	206	222	238	254	270	286	302
40	39	53	69	84	100	115	131	147	163	179	196	212	228	245	261	278	294	311
	$\rho=0.01$																	
21	10	16	22	29	35	42	49	56	63	70	77	84	91	98	105	113	120	127
22	10	17	23	30	37	45	52	59	66	74	81	89	96	104	111	119	127	134
23	11	18	25	32	39	47	55	62	70	78	86	94	102	109	117	125	133	141
24	12	19	26	34	42	49	57	66	74	82	90	98	107	115	123	132	140	149
25	12	20	27	35	44	52	60	69	77	86	95	103	112	121	130	138	147	156
26	13	21	29	37	46	54	63	72	81	90	99	108	117	126	136	145	154	163
27	14	22	30	39	48	57	66	75	85	94	103	113	122	132	142	151	161	171
28	14	23	32	41	50	59	69	78	88	98	108	118	128	138	148	158	168	178
29	15	24	33	42	52	62	72	82	92	102	112	123	133	143	154	164	175	185
30	15	25	34	44	54	64	75	85	95	106	117	127	138	149	160	171	182	192
31	16	26	36	46	56	67	77	88	99	110	121	132	143	155	166	177	188	200
32	17	27	37	47	58	69	80	91	103	114	126	137	149	160	172	184	195	207
33	17	28	38	49	60	72	83	95	106	118	130	142	154	166	178	190	202	214
34	18	29	40	51	62	74	86	98	110	122	134	147	159	172	184	197	209	222
35	19	30	41	53	64	77	89	101	114	126	139	152	164	177	190	203	216	229
36	19	31	42	54	67	79	92	104	117	130	143	156	170	183	196	210	223	236
37	20	32	44	56	69	81	95	108	121	134	148	161	175	189	202	216	230	244
38	21	33	45	58	71	84	97	111	125	138	152	166	180	194	208	223	237	251
39	21	34	46	59	73	86	100	114	128	142	157	171	185	200	214	229	244	258
40	22	35	48	61	75	89	103	117	132	146	161	176	191	206	221	236	251	266

Как выбрать статистический критерий?

Обратите внимание, что нулевая гипотеза у Манна-Уитни **не сравнивает средние с помощью суммы рангов**, а говорит о том, что **выборки взяты из одного и того же распределения**

Главное условие применимости Манн-Уитни – схожесть форм распределений

Для других сравнений существует ряд непараметрических критериев:

- Критерий Краскелла-Уоллиса (от 2-х групп и больше)
- Критерий Данна

...

Как выбрать статистический критерий?

e^xperiment fest

Нужно ли знать все критерии?

Существует огромное количество статистических критериев. И все их знать просто невозможно

На картинке содержание справочника

Прикладная математическая статистика А.И. Кобзарь

Как выбрать статистический критерий?

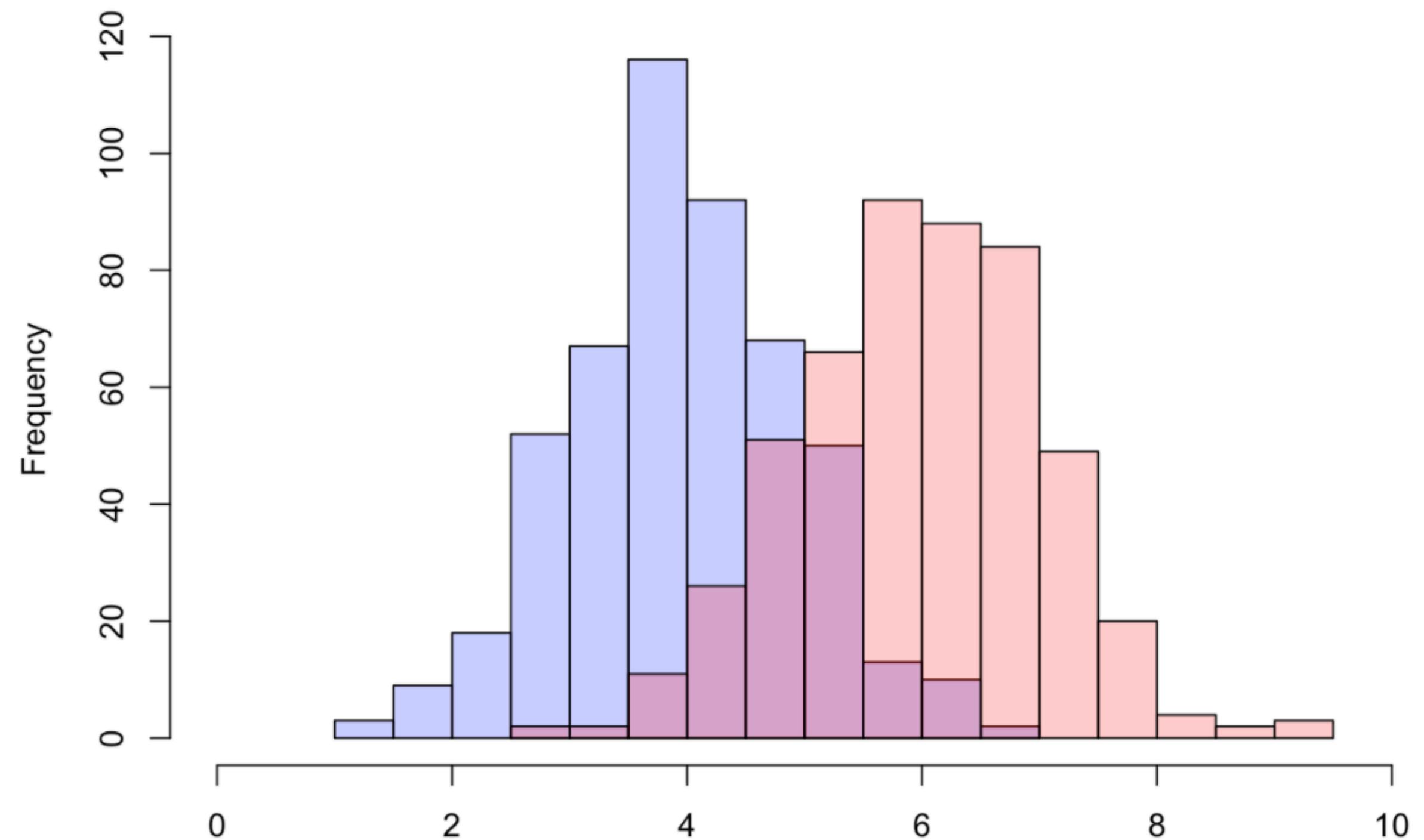
3.3. Критерии проверки экспоненциальности распределения 3.3.1. Критерий Шапиро-Уилка (279). 3.3.2. Критерий типа Колмогорова-Смирнова (282). 3.3.3. Критерий типа Смирнова-Крамера-фон Мизеса для цензурированных данных (286). 3.3.4. Критерий Фрошни (288). 3.3.5. Корреляционный критерий экспоненциальности (288). 3.3.6. Регрессионный критерий Брейна-Шапиро (293). 3.3.7. Критерий Кимбера-Мичела (292). 3.3.8. Критерий Фишера (293). 3.3.9. Критерий Бартлетта-Морана (294). 3.3.10. Критерий Климко-Англа-Радемакера-Рокетта (294). 3.3.11. Критерий Холлендера-Прозана (295). 3.3.12. Критерий Кочара (298). 3.3.13. Критерий Эпса-Палли-Черго-Уэлча (299). 3.3.14. Критерий Бергмана (301). 3.3.15. Критерий Шермана (303). 3.3.16. Критерий наибольшего интервала (304). 3.3.17. Критерий Хартли (305). 3.3.18. Критерий показательных методов (305). 3.3.19. Ранговый критерий независимости интервалов (306). 3.3.20. Критерии, основанные на трансформации экспоненциального распределения в равномерное (308). 3.3.20.1. Критерий \bar{U} (308). 3.3.20.2. Критерий \hat{U} (309). 3.3.20.3. Критерий Гринвуда (309). 3.3.21. Критерий Манн-Фертинга-Шуера для распределения Вейбулла (311). 3.3.22. Критерий Дешпаньде (316). 3.3.23. Критерий Лоулесса (317).	319
3.4. Критерии согласия для равномерного распределения 3.4.1. Критерий Шермана (319). 3.4.2. Критерий Морана (320). 3.4.3. Критерий Ченга-Смирнига (322). 3.4.4. Критерий Саркади-Косика (323). 3.4.5. Энтропийный критерий Дудевича-ван дер Мюлена (324). 3.4.6. Критерий Хегази-Грина (326). 3.4.7. Критерий Янга (328). 3.4.8. Критерий типа Колмогорова-Смирнова (330). 3.4.9. Критерий Фрошни (331). 3.4.10. Критерий Гринвуда-Кэсенбери-Миллера (332). 3.4.11. „Сглаженный“ критерий Неймана-Бартона (333).	336
3.5. Критерии симметрии 3.5.1. „Быстрый“ критерий Кенуя (336). 3.5.2. Критерий симметрии Смирнова (337). 3.5.3. Знаковый критерий симметрии (337). 3.5.4. Одновыборочный критерий Вилкоксона (339). 3.5.5. Критерий Антилла-Керстинга-Цуккини (340). 3.5.6. Критерий Бхатачарья-Гаствира-Райта (модифицированный критерий Вилкоксона) (342). 3.5.7. Критерий Финча (344). 3.5.8. Критерий Босса (345). 3.5.9. Критерий Гупты (348). 3.5.10. Критерий Фрезера (350).	336
3.6. Подбор кривых распределения вероятностей по экспериментальным данным 3.6.1. Кривые распределения Джонсона (352). 3.6.1.1. Семейство распределений S_L Джонсона (353). 3.6.1.2. Семейство распределений S_B Джонсона (355). 3.6.1.3. Семейство распределений S_U Джонсона (357). 3.6.2. Кривые распределений Пирсона (368). 3.6.2.1. Кривые Пирсона типа I (369). 3.6.2.2. Кривые Пирсона типа II (375). 3.6.2.3. Кривые Пирсона типа III (377). 3.6.2.4. Кривые Пирсона типа IV (378). 3.6.2.5. Кривые Пирсона типа V (380). 3.6.2.6. Кривые Пирсона типа VI (381). 3.6.2.7. Кривые Пирсона типа VII (382). 3.6.3. Разложение теоретических распределений (384). 3.6.4. Метод вкладов (385).	352
Глава 4. Проверка гипотез о значениях параметров распределений	
4.1. Сравнение параметров распределений 4.1.1. Сравнение параметров нормальных распределений (389). 4.1.1.1. Сравнение двух средних значений (389). 4.1.1.1.1. Сравнение при известных дисперсиях σ_1^2 и σ_2^2 (389). 4.1.1.1.2. Сравнение при неизвестных равных дисперсиях (390). 4.1.1.1.3. Сравнение при неизвестных неравных дисперсиях (391). 4.1.1.1.4. Критерий Кохрана-Кокса (391). 4.1.1.1.5. Критерий Сатервайта (391). 4.1.1.1.6. Критерий Уэлча (392). 4.1.1.1.7. Критерий Стьюдента (392). 4.1.1.1.8. Модифицированных критерий Стьюдента (393). 4.1.1.1.9. Парный t -критерий сравнения средних (394).	388
4.1.2. Непараметрические (свободные от распределения) критерии однородности статистических данных 4.2.1. Непараметрические критерии сдвига (452). 4.2.1.1. Сравнение параметров сдвига двух совокупностей (452). 4.2.1.1.1. Быстрый (грубый) критерий Кенуя (452). 4.2.1.1.2. Ранговые критерии сдвига (453). 4.2.1.1.2.1. Быстрый (грубый) ранговый критерий (453). 4.2.1.1.2.2. Критерий Манна-Уитни-Вилкоксона (454). 4.2.1.1.2.3. Критерий Фишера-Ййтса-Терри-Гёфдинга (459). 4.2.1.1.2.4. Критерий Ван дер Вардена (460). 4.2.1.1.2.5. Медианный критерий (462). 4.2.1.1.2.6. Критерий Мостеллера (464). 4.2.1.1.2.7. Критерий Розенбаума (464). 4.2.1.1.2.8. Критерий Хаги (464). 4.2.1.1.2.9. Е-критерий Розенбаума (464). 4.2.1.1.2.10. Критерий Хаги (464). 4.2.1.1.2.11. Критерий Хаги (464). 4.2.1.1.2.12. Критерий Хаги (464). 4.2.1.1.2.13. Критерий Хаги (464). 4.2.1.1.2.14. Критерий Хаги (464). 4.2.1.1.2.15. Критерий Хаги (464). 4.2.1.1.2.16. Критерий Хаги (464). 4.2.1.1.2.17. Критерий Хаги (464). 4.2.1.1.2.18. Критерий Хаги (464). 4.2.1.1.2.19. Критерий Хаги (464). 4.2.1.1.2.20. Критерий Хаги (464). 4.2.1.1.2.21. Критерий Хаги (464). 4.2.1.1.2.22. Критерий Хаги (464). 4.2.1.1.2.23. Критерий Хаги (464). 4.2.1.1.2.24. Критерий Хаги (464). 4.2.1.1.2.25. Критерий Хаги (464). 4.2.1.1.2.26. Критерий Хаги (464). 4.2.1.1.2.27. Критерий Хаги (464). 4.2.1.1.2.28. Критерий Хаги (464). 4.2.1.1.2.29. Критерий Хаги (464). 4.2.1.1.2.30. Критерий Хаги (464). 4.2.1.1.2.31. Критерий Хаги (464). 4.2.1.1.2.32. Критерий Хаги (464). 4.2.1.1.2.33. Критерий Хаги (464). 4.2.1.1.2.34. Критерий Хаги (464). 4.2.1.1.2.35. Критерий Хаги (464). 4.2.1.1.2.36. Критерий Хаги (464). 4.2.1.1.2.37. Критерий Хаги (464). 4.2.1.1.2.38. Критерий Хаги (464). 4.2.1.1.2.39. Критерий Хаги (464). 4.2.1.1.2.40. Критерий Хаги (464). 4.2.1.1.2.41. Критерий Хаги (464). 4.2.1.1.2.42. Критерий Хаги (464). 4.2.1.1.2.43. Критерий Хаги (464). 4.2.1.1.2.44. Критерий Хаги (464). 4.2.1.1.2.45. Критерий Хаги (464). 4.2.1.1.2.46. Критерий Хаги (464). 4.2.1.1.2.47. Критерий Хаги (464). 4.2.1.1.2.48. Критерий Хаги (464). 4.2.1.1.2.49. Критерий Хаги (464). 4.2.1.1.2.50. Критерий Хаги (464). 4.2.1.1.2.51. Критерий Хаги (464). 4.2.1.1.2.52. Критерий Хаги (464). 4.2.1.1.2.53. Критерий Хаги (464). 4.2.1.1.2.54. Критерий Хаги (464). 4.2.1.1.2.55. Критерий Хаги (464). 4.2.1.1.2.56. Критерий Хаги (464). 4.2.1.1.2.57. Критерий Хаги (464). 4.2.1.1.2.58. Критерий Хаги (464). 4.2.1.1.2.59. Критерий Хаги (464). 4.2.1.1.2.60. Критерий Хаги (464). 4.2.1.1.2.61. Критерий Хаги (464). 4.2.1.1.2.62. Критерий Хаги (464). 4.2.1.1.2.63. Критерий Хаги (464). 4.2.1.1.2.64. Критерий Хаги (464). 4.2.1.1.2.65. Критерий Хаги (464). 4.2.1.1.2.66. Критерий Хаги (464). 4.2.1.1.2.67. Критерий Хаги (464). 4.2.1.1.2.68. Критерий Хаги (464). 4.2.1.1.2.69. Критерий Хаги (464). 4.2.1.1.2.70. Критерий Хаги (464). 4.2.1.1.2.71. Критерий Хаги (464). 4.2.1.1.2.72. Критерий Хаги (464). 4.2.1.1.2.73. Критерий Хаги (464). 4.2.1.1.2.74. Критерий Хаги (464). 4.2.1.1.2.75. Критерий Хаги (464). 4.2.1.1.2.76. Критерий Хаги (464). 4.2.1.1.2.77. Критерий Хаги (464). 4.2.1.1.2.78. Критерий Хаги (464). 4.2.1.1.2.79. Критерий Хаги (464). 4.2.1.1.2.80. Критерий Хаги (464). 4.2.1.1.2.81. Критерий Хаги (464). 4.2.1.1.2.82. Критерий Хаги (464). 4.2.1.1.2.83. Критерий Хаги (464). 4.2.1.1.2.84. Критерий Хаги (464). 4.2.1.1.2.85. Критерий Хаги (464). 4.2.1.1.2.86. Критерий Хаги (464). 4.2.1.1.2.87. Критерий Хаги (464). 4.2.1.1.2.88. Критерий Хаги (464). 4.2.1.1.2.89. Критерий Хаги (464). 4.2.1.1.2.90. Критерий Хаги (464). 4.2.1.1.2.91. Критерий Хаги (464). 4.2.1.1.2.92. Критерий Хаги (464). 4.2.1.1.2.93. Критерий Хаги (464). 4.2.1.1.2.94. Критерий Хаги (464). 4.2.1.1.2.95. Критерий Хаги (464). 4.2.1.1.2.96. Критерий Хаги (464). 4.2.1.1.2.97. Критерий Хаги (464). 4.2.1.1.2.98. Критерий Хаги (464). 4.2.1.1.2.99. Критерий Хаги (464). 4.2.1.1.2.100. Критерий Хаги (464). 4.2.1.1.2.101. Критерий Хаги (464). 4.2.1.1.2.102. Критерий Хаги (464). 4.2.1.1.2.103. Критерий Хаги (464). 4.2.1.1.2.104. Критерий Хаги (464). 4.2.1.1.2.105. Критерий Хаги (464). 4.2.1.1.2.106. Критерий Хаги (464). 4.2.1.1.2.107. Критерий Хаги (464). 4.2.1.1.2.108. Критерий Хаги (464). 4.2.1.1.2.109. Критерий Хаги (464). 4.2.1.1.2.110. Критерий Хаги (464). 4.2.1.1.2.111. Критерий Хаги (464). 4.2.1.1.2.112. Критерий Хаги (464). 4.2.1.1.2.113. Критерий Хаги (464). 4.2.1.1.2.114. Критерий Хаги (464). 4.2.1.1.2.115. Критерий Хаги (464). 4.2.1.1.2.116. Критерий Хаги (464). 4.2.1.1.2.117. Критерий Хаги (464). 4.2.1.1.2.118. Критерий Хаги (464). 4.2.1.1.2.119. Критерий Хаги (464). 4.2.1.1.2.120. Критерий Хаги (464). 4.2.1.1.2.121. Критерий Хаги (464). 4.2.1.1.2.122. Критерий Хаги (464). 4.2.1.1.2.123. Критерий Хаги (464). 4.2.1.1.2.124. Критерий Хаги (464). 4.2.1.1.2.125. Критерий Хаги (464). 4.2.1.1.2.126. Критерий Хаги (464). 4.2.1.1.2.127. Критерий Хаги (464). 4.2.1.1.2.128. Критерий Хаги (464). 4.2.1.1.2.129. Критерий Хаги (464). 4.2.1.1.2.130. Критерий Хаги (464). 4.2.1.1.2.131. Критерий Хаги (464). 4.2.1.1.2.132. Критерий Хаги (464). 4.2.1.1.2.133. Критерий Хаги (464). 4.2.1.1.2.134. Критерий Хаги (464). 4.2.1.1.2.135. Критерий Хаги (464). 4.2.1.1.2.136. Критерий Хаги (464). 4.2.1.1.2.137. Критерий Хаги (464). 4.2.1.1.2.138. Критерий Хаги (464). 4.2.1.1.2.139. Критерий Хаги (464). 4.2.1.1.2.140. Критерий Хаги (464). 4.2.1.1.2.141. Критерий Хаги (464). 4.2.1.1.2.142. Критерий Хаги (464). 4.2.1.1.2.143. Критерий Хаги (464). 4.2.1.1.2.144. Критерий Хаги (464). 4.2.1.1.2.145. Критерий Хаги (464). 4.2.1.1.2.146. Критерий Хаги (464). 4.2.1.1.2.147. Критерий Хаги (464). 4.2.1.1.2.148. Критерий Хаги (464). 4.2.1.1.2.149. Критерий Хаги (464). 4.2.1.1.2.150. Критерий Хаги (464). 4.2.1.1.2.151. Критерий Хаги (464). 4.2.1.1.2.152. Критерий Хаги (464). 4.2.1.1.2.153. Критерий Хаги (464). 4.2.1.1.2.154. Критерий Хаги (464). 4.2.1.1.2.155. Критерий Хаги (464). 4.2.1.1.2.156. Критерий Хаги (464). 4.2.1.1.2.157. Критерий Хаги (464). 4.2.1.1.2.158. Критерий Хаги (464). 4.2.1.1.2.159. Критерий Хаги (464). 4.2.1.1.2.160. Критерий Хаги (464). 4.2.1.1.2.161. Критерий Хаги (464). 4.2.1.1.2.162. Критерий Хаги (464). 4.2.1.1.2.163. Критерий Хаги (464). 4.2.1.1.2.164. Критерий Хаги (464). 4.2.1.1.2.165. Критерий Хаги (464). 4.2.1.1.2.166. Критерий Хаги (464). 4.2.1.1.2.167. Критерий Хаги (464). 4.2.1.1.2.168. Критерий Хаги (464). 4.2.1.1.2.169. Критерий Хаги (464). 4.2.1.1.2.170. Критерий Хаги (464). 4.2.1.1.2.171. Критерий Хаги (464). 4.2.1.1.2.172. Критерий Хаги (464). 4.2.1.1.2.173. Критерий Хаги (464). 4.2.1.1.2.174. Критерий Хаги (464). 4.2.1.1.2.175. Критерий Хаги (464). 4.2.1.1.2.176. Критерий Хаги (464). 4.2.1.1.2.177. Критерий Хаги (464). 4.2.1.1.2.178. Критерий Хаги (464). 4.2.1.1.2.179. Критерий Хаги (464). 4.2.1.1.2.180. Критерий Хаги (464). 4.2.1.1.2.181. Критерий Хаги (464). 4.2.1.1.2.182. Критерий Хаги (464). 4.2.1.1.2.183. Критерий Хаги (464). 4.2.1.1.2.184. Критерий Хаги (464). 4.2.1.1.2.185. Критерий Хаги (464). 4.2.1.1.2.186. Критерий Хаги (464). 4.2.1.1.2.187. Критерий Хаги (464). 4.2.1.1.2.188. Критерий Хаги (464). 4.2.1.1.2.189. Критерий Хаги (464).	

Квиз

Как выбрать статистический критерий?

e^xperiment fest

1. С помощью какого критерия здесь можно проверять гипотезы?



Как выбрать статистический критерий?

e^xperiment fest

День 2

Способы определения объема выборки для А/В-тестов. Мощность эксперимента

e^xperiment fest

**Как вы думаете, если за год
провести 1000 А/В-тестов,
сколько из них будут с
ошибочными результатами?**

e^xperiment fest

Как вы думаете, если за год
провести 1000 А/В-тестов,
сколько из них будут с
ошибочными результатами?

$$\leq \alpha \text{ и } \leq 1-\beta$$

Мы выделяем 2 способа определения необходимого времени на эксперимент:

Fixed Horizon

Простая реализация

Высокая ошибка ранней остановки

- Фиксированный временной горизонт проведения эксперимента
- Один раз рассчитали, один раз подсмотрели и приняли решение

Sequential testing

Оптимизация «проблемы поглядывания»

Сложная реализация

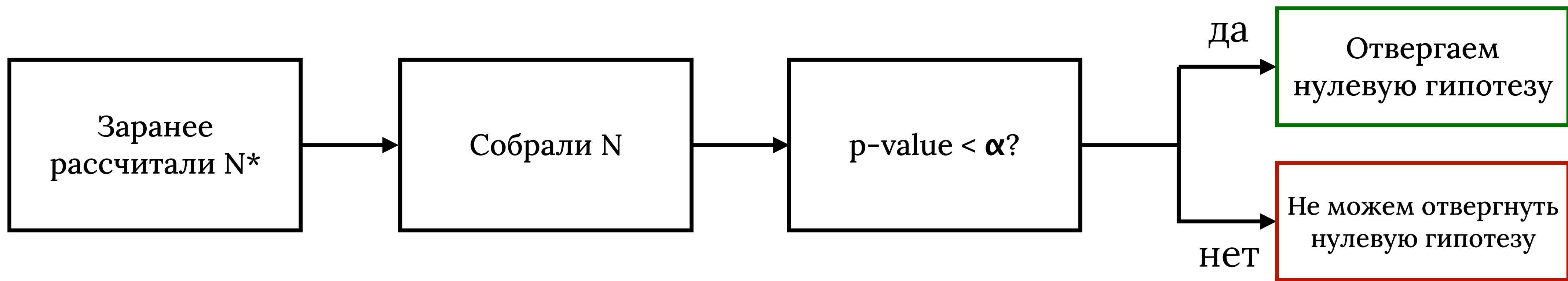
- Границы остановки эксперимента рассчитываются real-time
- Значимость накопительно рассчитывается по дополнительным порогам принятия решения

Как рассчитать нужный объем выборки?

e^xperiment fest

Fixed horizon

Этапы

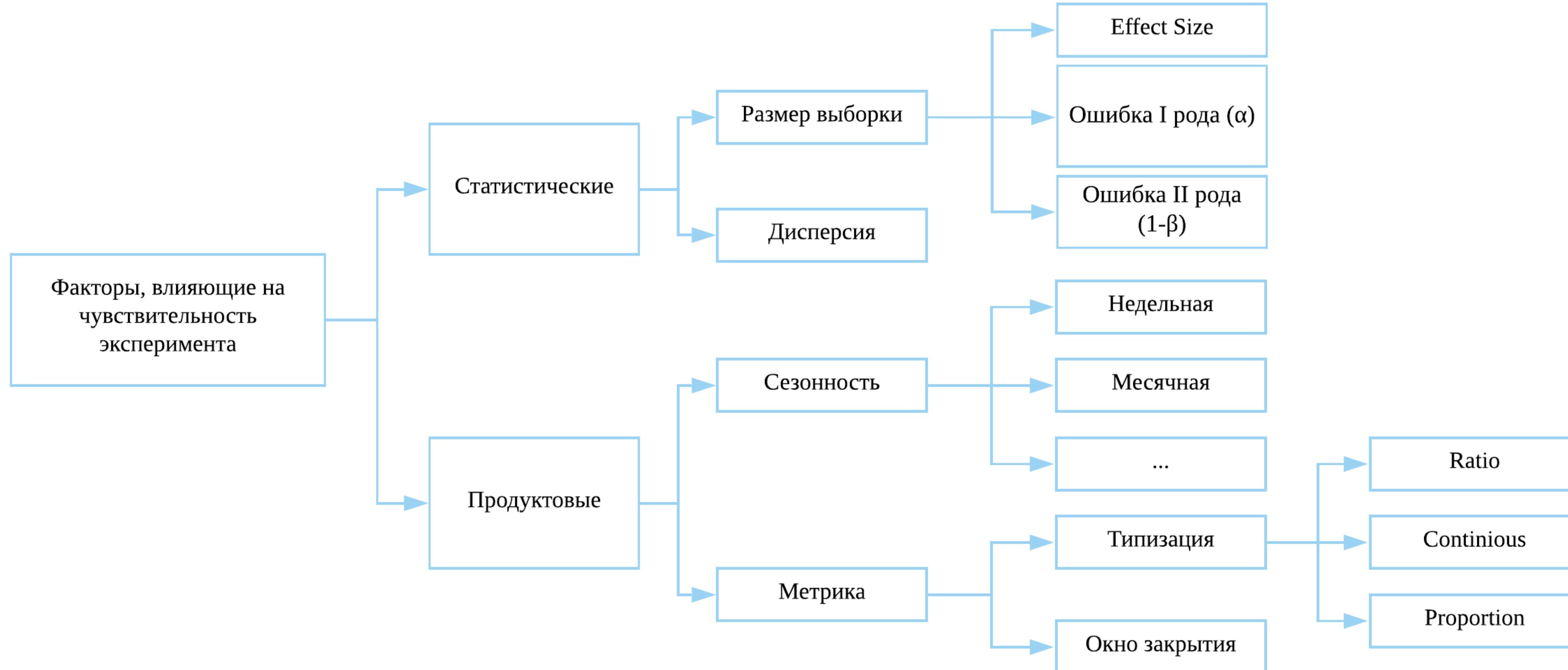


*согласно параметрам ТР (мощность), TN (уровень значимости), MDE (minimum detectable effect)

Как рассчитать нужный объем выборки?

e^xperiment fest

Факторы, влияющие на время



Как рассчитать нужный объем выборки?

e^xperiment fest

Чувствительность

Способность увидеть значимые различия в метрике там, где они на самом деле должны быть называется **чувствительностью**

Высокая чувствительность метрики позволяет:

- видеть достаточно маленькие изменения
- или использовать меньшее количество пользователей

Как рассчитать нужный объем выборки?

e^xperiment fest

Чувствительность зависит от метрики и пользовательских циклов

Чувствительность эксперимента зависит от набора метрик и отрасли, в которой эти метрики оцениваются:

- Клик по баннеру чувствительнее, чем факт покупки в интернет-магазине
- С1 чувствительнее, чем С2, С3 Сn
- Транзакционная активность в free-to-play играх чувствительнее, чем в travel отрасли

Как рассчитать нужный объем выборки?

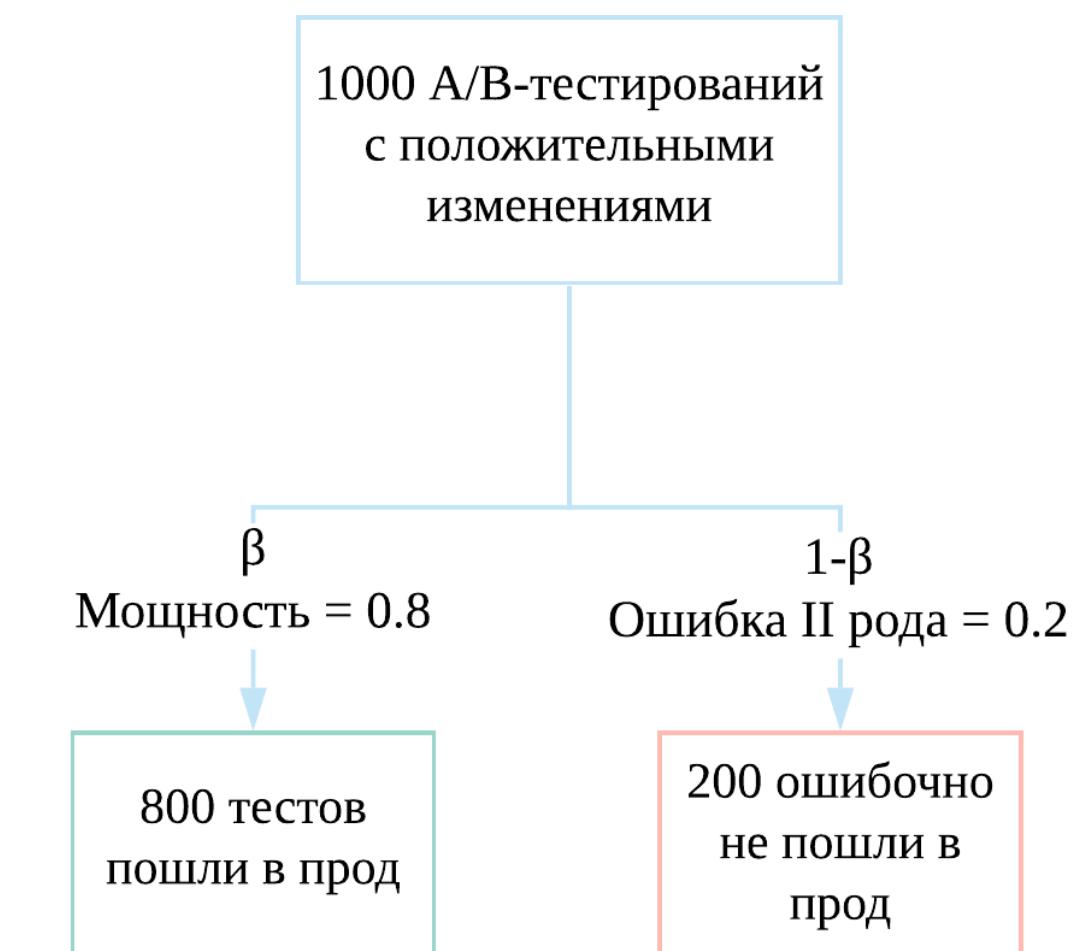
e^xperiment fest

Мощность и ошибка II рода

Мощность для продуктовых реалий является важным параметром, потому что никому не хотелось бы выкидывать эксперименты с реальными эффектами.

Пример:

Допустим, мы берем уровень мощности в 80% как минимальный допустимый порог для экспериментов, то из 1000 экспериментов с реальным приростом в метрику, в 800 мы были бы уверены, что прирост есть. Остаются 200, которые будут выкинуты в мусорку зря, потому что остается False Negative (ошибка II) = 0.2.



*prod – Production – публичная версия продукта

Как рассчитать нужный объем выборки?

e^xperiment fest

Мощность и ошибка II рода

Получается, что...

Мощность нужно максимизировать на столько, на сколько позволяют возможности продукта. В первую очередь это зависит от MAU и, в целом, от трафика.

Вывод

Чем выше уровень мощности, тем меньше хороших экспериментов будут ошибочно забыты

Как рассчитать нужный объем выборки?

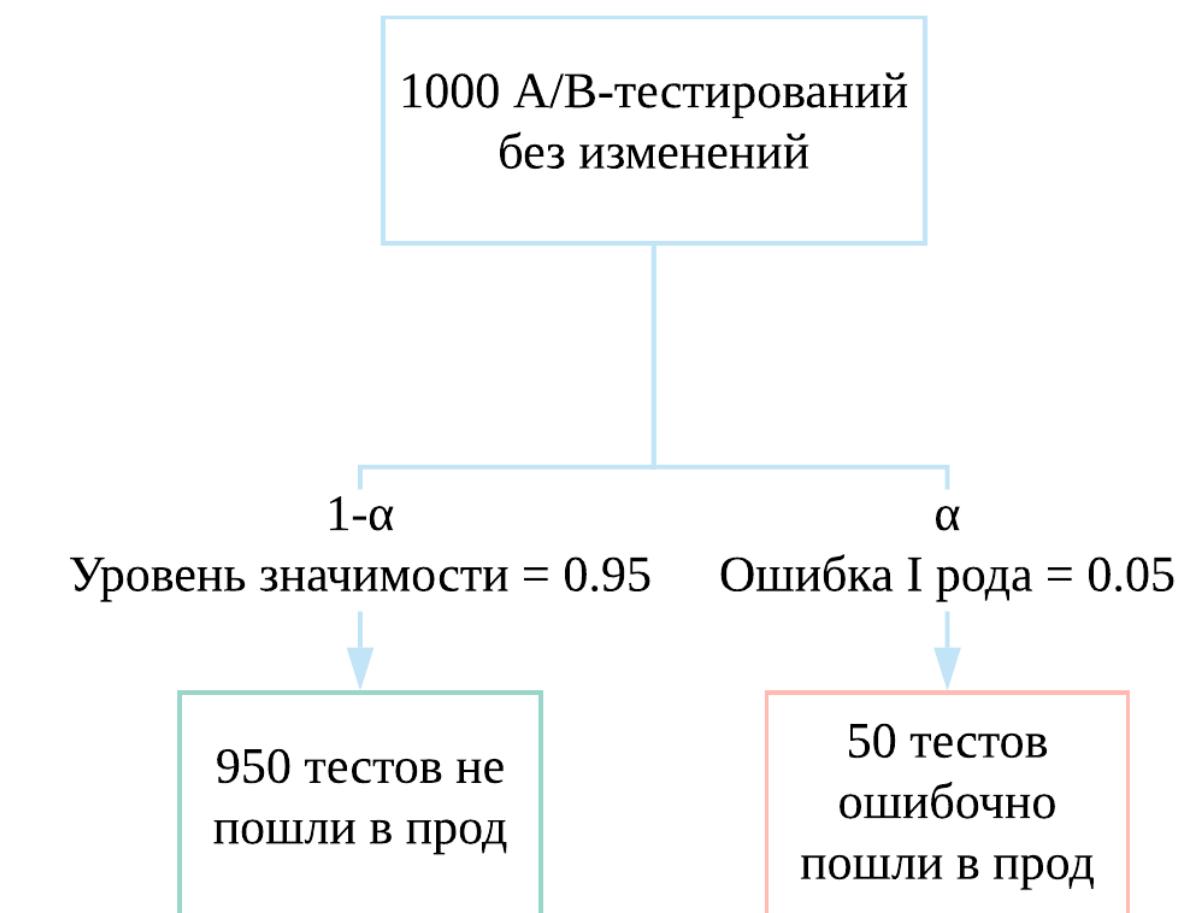
e^xperiment fest

Уровень значимости и ошибка I рода

Когда принимается решение о результатах эксперимента, считается, что в первую очередь нужно смотреть на p-value (ошибка I рода или false positive):
положительное изменение, отрицательное изменение или отсутствие изменения

Пример:

Допустим, мы берем уровень значимости в 95% как минимальный допустимый порог для экспериментов, то из 1000 безуспешных экспериментов (нет эффекта) в 950 мы были бы уверены, что эффекта реально нет. Но 50 ошибочно выкатили бы в продакшен, хотя в этом нет никакого смысла, потому что в них мы наблюдаем случайность, а не реальную закономерность.



Как рассчитать нужный объем выборки?

e^xperiment fest

Уровень значимости и ошибка I рода

Получается, что...

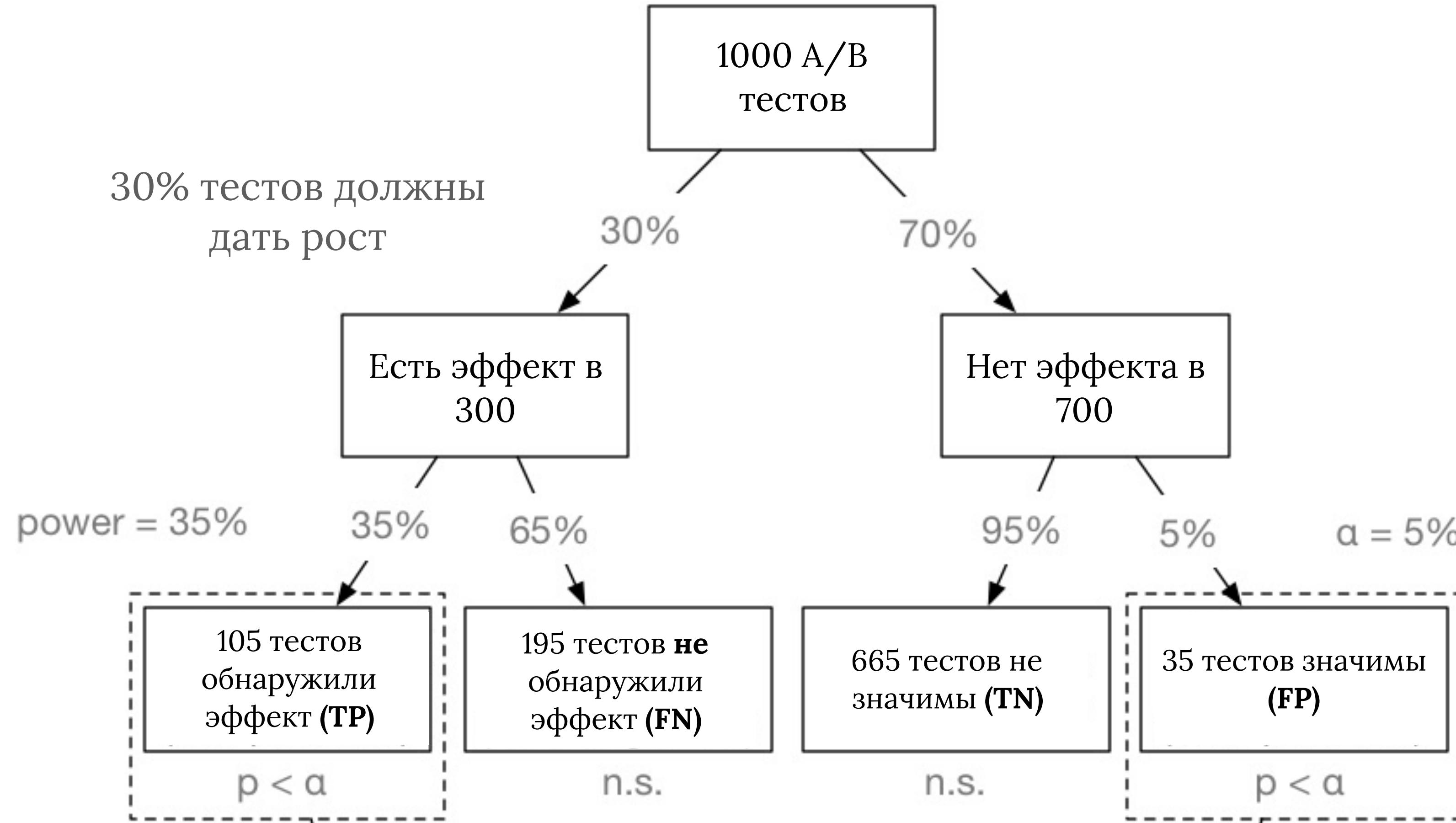
Так же как и мощность, уровень значимости необходимо увеличивать по мере возможностей. Классические 95% оторваны от современных реалий с бигдатами и требований бизнеса к точности. Следовательно, 5% кажутся не позволительной ошибкой для больших компаний

Вывод

Чем выше уровень значимости, тем меньше бесполезных экспериментов будут выкапываться в продакшн.

Как рассчитать нужный объем выборки?

e^xperiment fest



Как выбрать статистический критерий?

e^xperiment fest

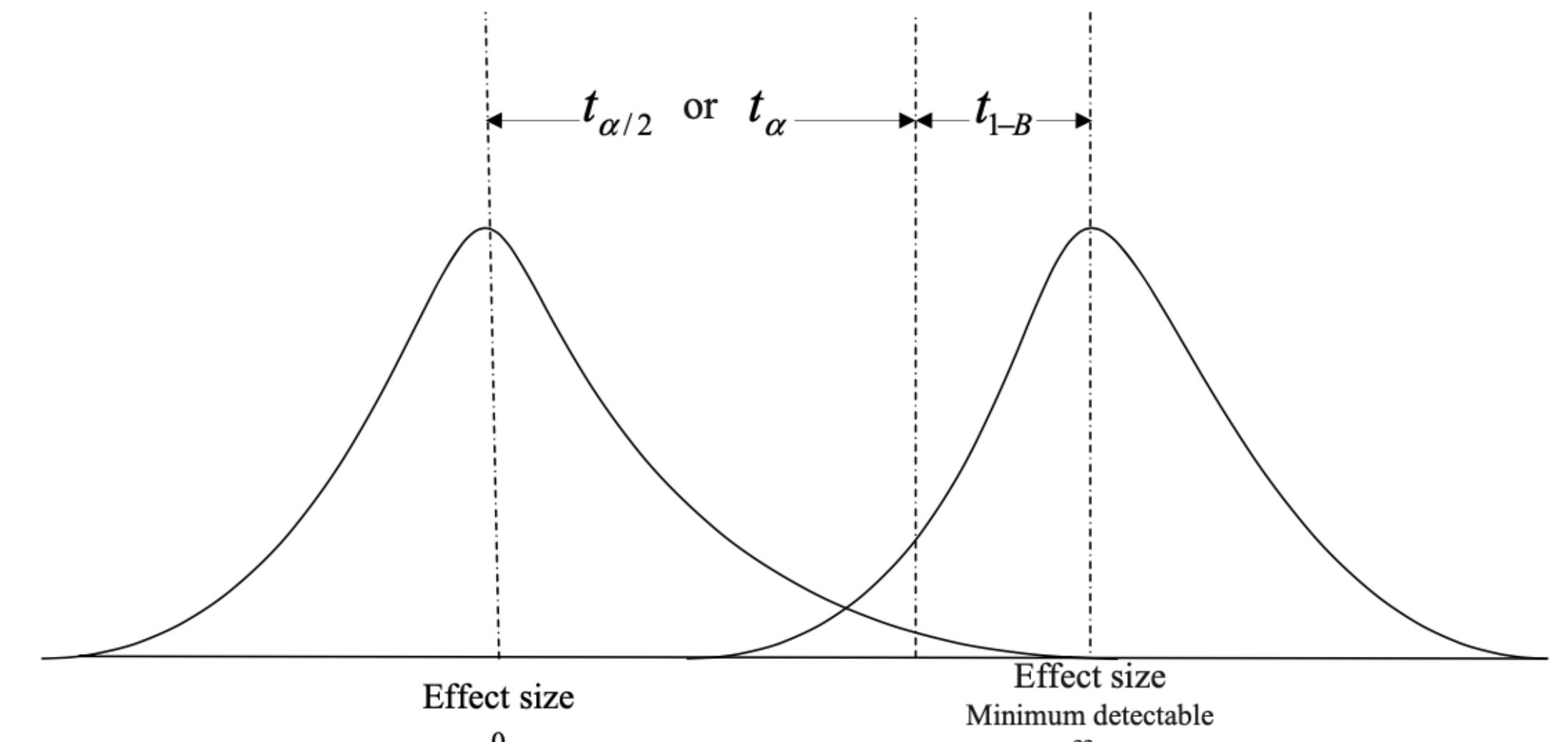
Минимальный ожидаемый эффект (MDE)

Что такое MDE?

Изменения, которые мы хотели бы увидеть с заданными порогами ошибок I и II типов рассчитываются с помощью минимального ожидаемого эффекта (minimum detectable effect или MDE).

Для этого параметра не хватит короткого описания. Разберем основные вопросы далее

Как рассчитать нужный объем выборки?



$$\text{One-tail multiplier} = t_{\alpha} + t_{1-B}$$

$$\text{Two-tail multiplier} = t_{\alpha/2} + t_{1-B}$$

MDE, Lift, ES (effect size)

Начнем с того, что люди часто путают MDE и lift (или uplift). Типично, в обсуждениях имеют ввиду одно и то же, когда говорят про эффект, прирост, значимый прирост и прочие подобные. Но на самом деле это не так. В этих определениях семантика играет важную роль, когда заходит речь про оценку результатов А/В.

(вымысленный диалог)

– Ну что? Выросла метрика?

Варианты чем ответить:

a) Lift

b) Effect Size

c) Minimum Detectable Effect

Как рассчитать нужный объем выборки?

e^xperiment fest

Lift

$$\text{Lift} = \frac{\bar{Y}_t - \bar{Y}_c}{\bar{Y}_c}, \text{ где } \bar{Y}_t, \bar{Y}_c - \text{метрики теста и контроля}$$

Например, RPV (revenue per visit) в контроле 118, в тесте 121. Тогда прирост = $(121-118)/118 = 0.0254 = 2.54\%$.

Lift – это отличие метрики теста от метрики контроля. Считаем их разницу и делим на контроль. Умножаем на 100 и получаем %-ное изменение.

Мы по факту на данных эксперимента рассчитываем прирост и говорим какая у метрик относительная разница. Тут пока ни о каких ошибках I и II типов речи не идет. Это обычный прокси параметр для квази-оценки эксперимента.

Как рассчитать нужный объем выборки?

e^xperiment fest

Effect size

$$\text{Effect Size} = \frac{\bar{Y}_t - \bar{Y}_c}{\sigma}, \text{ где } \sigma = \frac{\sigma_t + \sigma_c}{2}$$

Например, RPV в контроле 118, в teste 121, сигма 2. Тогда Effect size = $(121-118)/2 = 1.5$

Effect size считается по разному для разных типов метрик, но смысл остается (выше показана формула для [cohens' d](#)).

Чем больше полученное число, тем сильнее размер эффекта. Это справедливо, если дисперсия устойчива и почти не изменяется.

Effect size измеряет степень отклонения наблюдаемых значений от тех, которые можно было бы ожидать при отсутствии эффекта. Т.е. считая разницу к дисперсии можно понять насколько сильно отклонился наблюдаемый эффект от эффекта, который равен нулю.

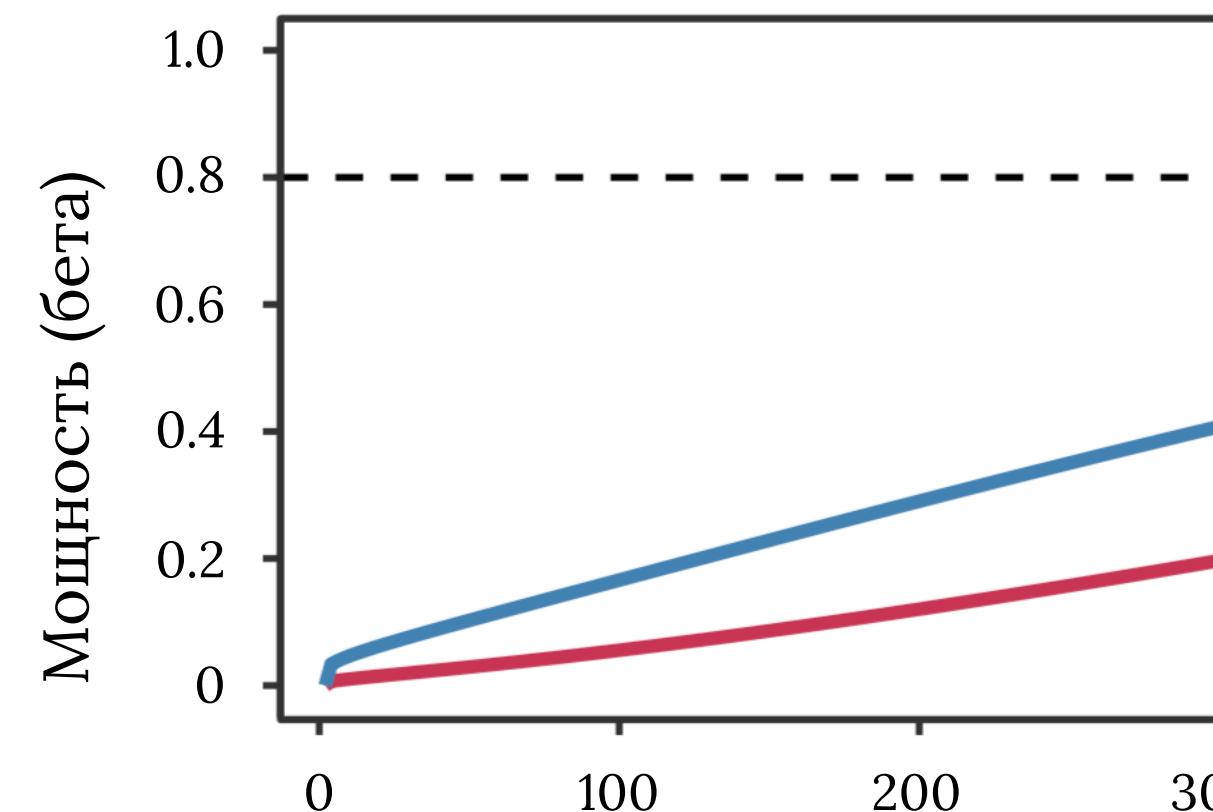
Как рассчитать нужный объем выборки?

e^xperiment fest

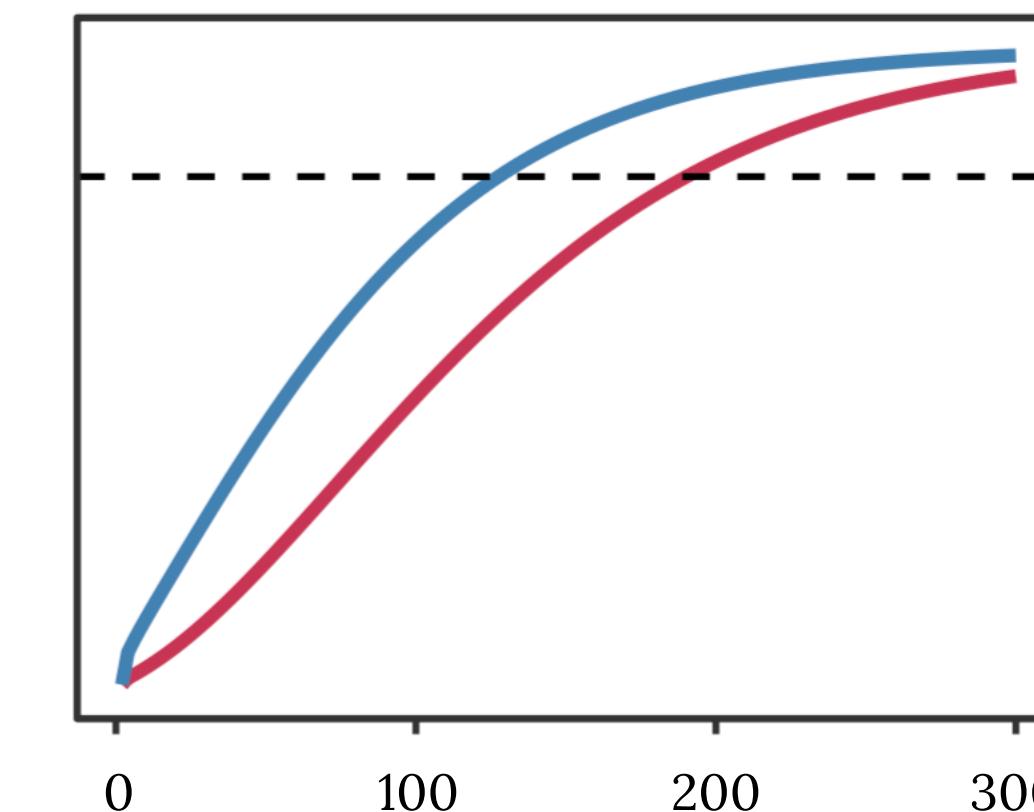
Концепция MDE

График расчета MDE

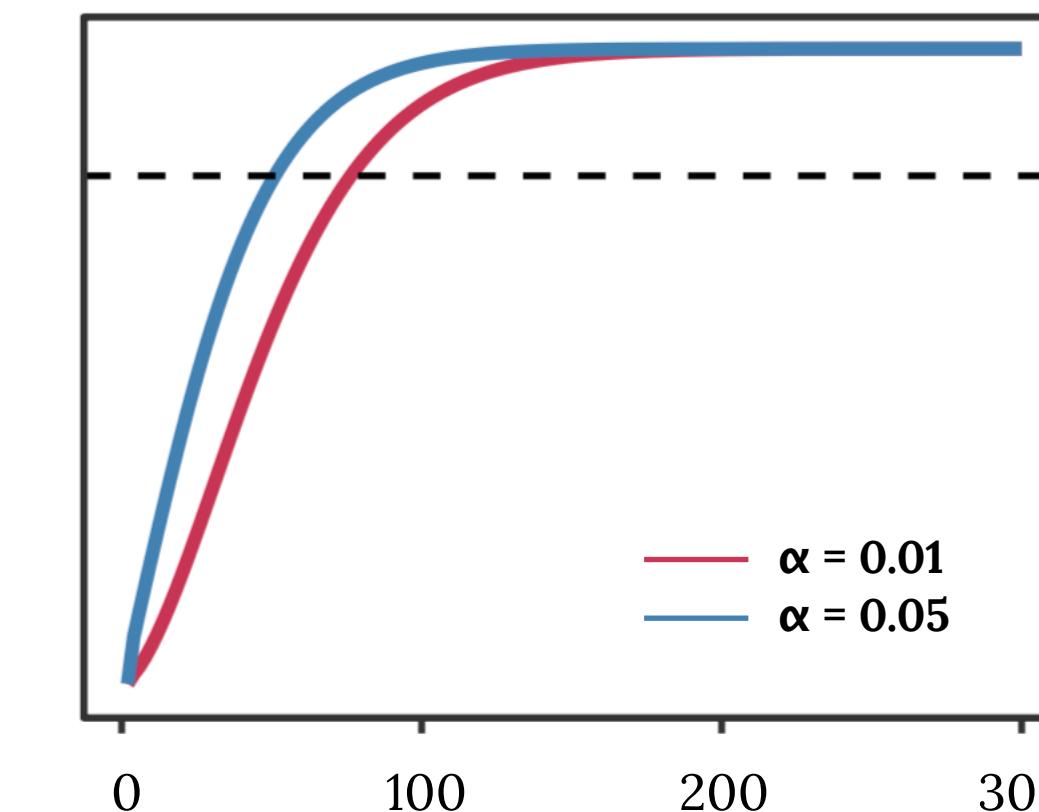
**True MDE = 0.1
(низкий эффект)**



**True MDE = 1
(средний эффект)**



**True MDE = 5
(сильный эффект)**



Как рассчитать нужный объем выборки?

e^xperiment fest

MDE

Минимальный ожидаемый эффект – это наименьший истинный эффект полученный от изменений, который с уверенностью сможет обнаружить статистический критерий.

Более формально: это наименьший истинный эффект полученный от изменений, который имеет определенный уровень статистической мощности для определенного уровня статистической значимости, учитывая конкретный статистический тест.

Как рассчитать нужный объем выборки?

e^xperiment fest

MDE

Минимальный ожидаемый эффект – это наименьший **истинный эффект** полученный от изменений, который с уверенностью сможет обнаружить статистический критерий.

Более формально: это наименьший **истинный эффект** полученный от изменений, который имеет определенный уровень статистической мощности для определенного уровня статистической значимости, учитывая конкретный статистический тест.

... **истинный эффект**... – здесь имеется ввиду, что если значимого изменения при $n = 10000$ у такого эффекта нет (например, 2%), то достаточно уверенно можно сказать лишь то, что возможный истинный эффект не больше, чем MDE. Если выборка была бы больше, скажем $n = 20000$, то может эффект мы бы и увидели, но точно меньше (т.е. < 2%).

Как рассчитать нужный объем выборки?

e^xperiment fest

MDE

Минимальный ожидаемый эффект – это наименьший истинный эффект полученный от изменений, который с уверенностью сможет обнаружить статистический критерий.

Более формально: это наименьший истинный эффект полученный от изменений, который имеет определенный уровень **статистической мощности** для определенного уровня **статистической значимости**, учитывая конкретный статистический тест.

...определенный уровень статистической мощности для определенного уровня статистической значимости... – в расчетах MDE участвуют альфа, бета и размер выборки. Исходя из заданных уровней мощности и значимости, MDE будет варьироваться при одном и том же объеме выборки. Это можно понять на Графике расчета MDE: чем ниже альфа и бета, тем больше потребуется пользователей для одного и того же MDE

Как рассчитать нужный объем выборки?

e^xperiment fest

MDE

Минимальный ожидаемый эффект – это наименьший истинный эффект полученный от изменений, который с уверенностью сможет обнаружить статистический критерий.

Более формально: это наименьший истинный эффект полученный от изменений, который имеет определенный уровень статистической мощности для определенного уровня статистической значимости, учитывая конкретный **статистический тест**.

...*статистический тест* – в формуле расчета MDE учитывается распределение из которого взята метрика. MDE по разному рассчитывается для t-критерия, хи-квадрата и других статистических критериев.

Как рассчитать нужный объем выборки?

e^xperiment fest

Как считать MDE?

$$\text{Minimum Detectable Effect} = M \frac{S}{\bar{Y}_c}, \text{ где } M = t_{\alpha/2} + t_{\beta}, S = \sqrt{\frac{\sigma_t^2}{n_t} + \frac{\sigma_c^2}{n_c}}$$

В M считаем t -значения для параметров альфа и бета. Для двусторонней проверки альфу делим на 2. Далее когда получим MDE , мы будем уверены в результатах, согласно выбранным порогам значимости и мощности.

Если нас интересует уровень значимости 99%, то соответственно берем 0.01 и получаем $t_{\alpha/2} = 2.58$. Расчет для беты: $t_{\beta} = 0.842$ для 80% мощности. Тогда $M = 2.58 + 0.842 = 3.422$.

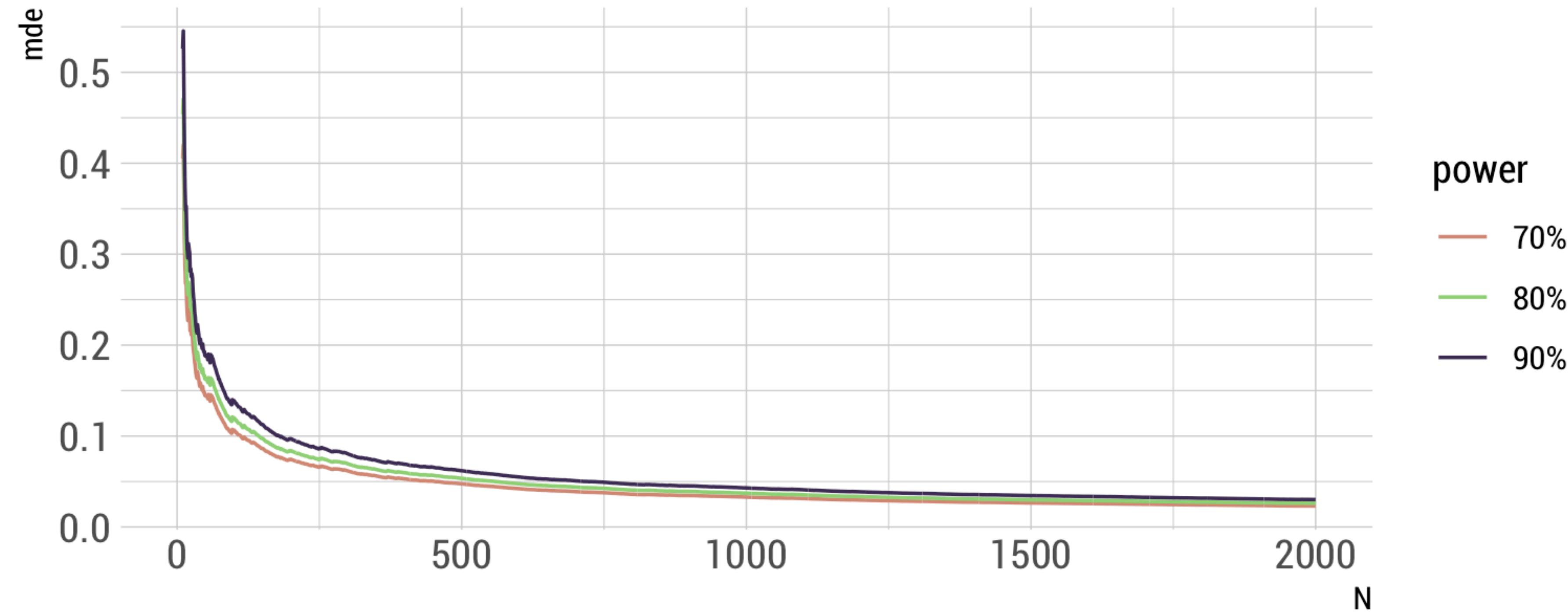
Далее умножаем на стандартную ошибку и получаем MDE !

Отношение к среднему по контролю \bar{Y}_c здесь для получения процента.

Как рассчитать нужный объем выборки?

e^xperiment fest

Чем меньший MDE мы хотели бы получить, тем больше наблюдений потребуется для его обнаружения



Как рассчитать нужный объем выборки?

e^xperiment fest

Как принимать решение на основе MDE?

Если p -value выше уровня α , то это не означает отсутствие эффекта. Эффект может и есть, но он точно не больше, чем MDE для α , β и дисперсии.

Например, вы наблюдаете 14-й день эксперимента, MDE на уровне 1%, p -value выше уровня α . Это означает, что если вы продолжите эксперимент и увидите значимые результаты эксперимента, то только для эффекта равный или меньший 1%.

Принимать решение продолжать эксперимент или нет, можно принимая в расчет MDE.

Как рассчитать нужный объем выборки?

e^xperiment fest

Демо в R и калькуляторе

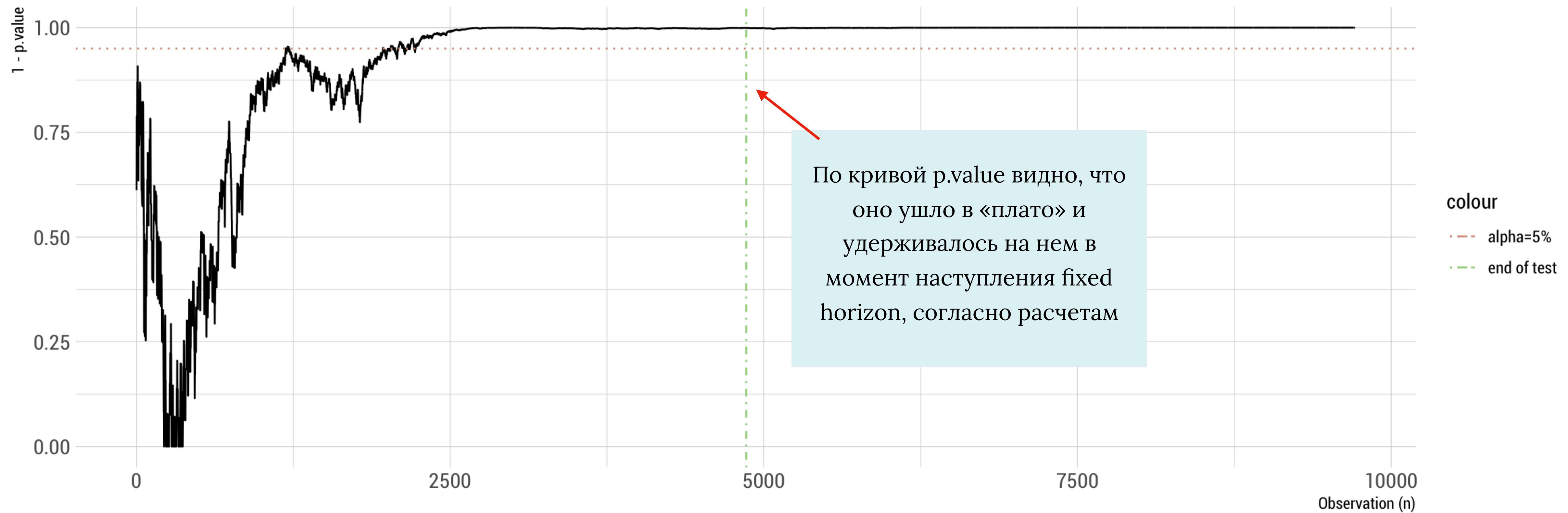
Как рассчитать нужный объем выборки?

e^xperiment fest

Симуляция А/В

CR A: 0,25 Power: 0,80
CR B: 0,2875 Alpha = 0,05

70



Как рассчитать нужный объем выборки?

e^xperiment fest

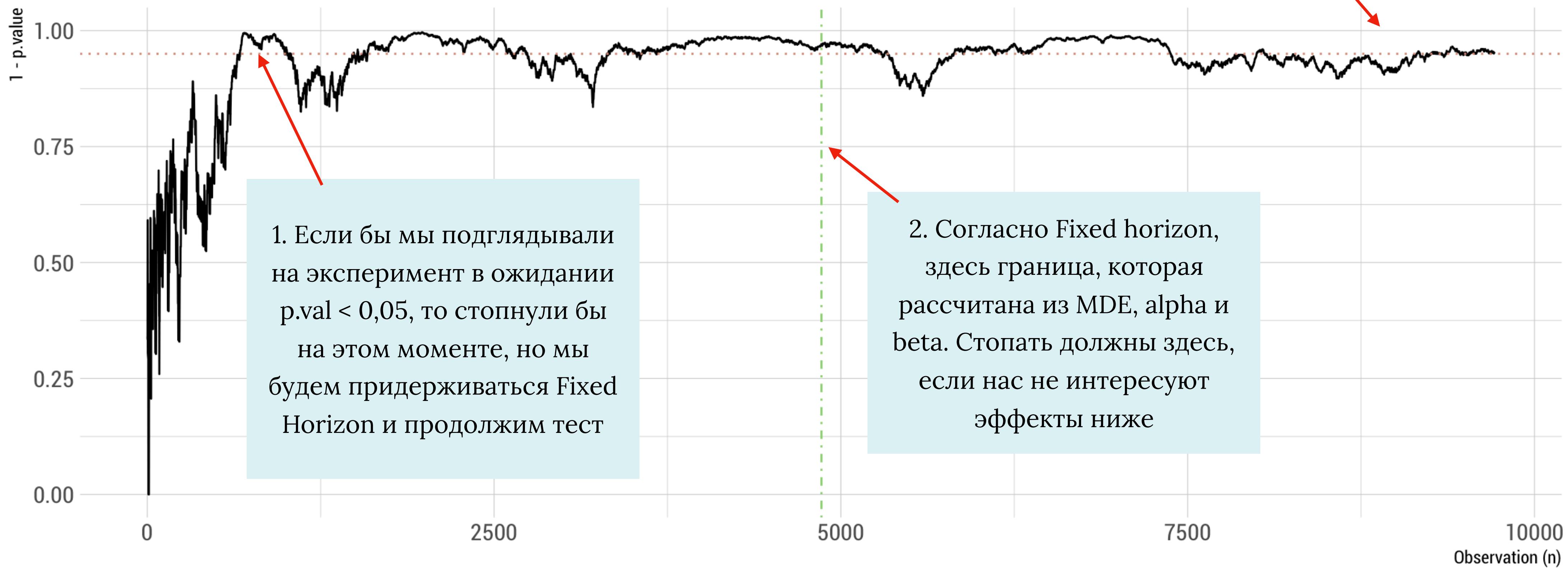
Симуляция А/В

CR A: 0,25

Power: 0,80

CR B: 0,26

Alpha = 0,05

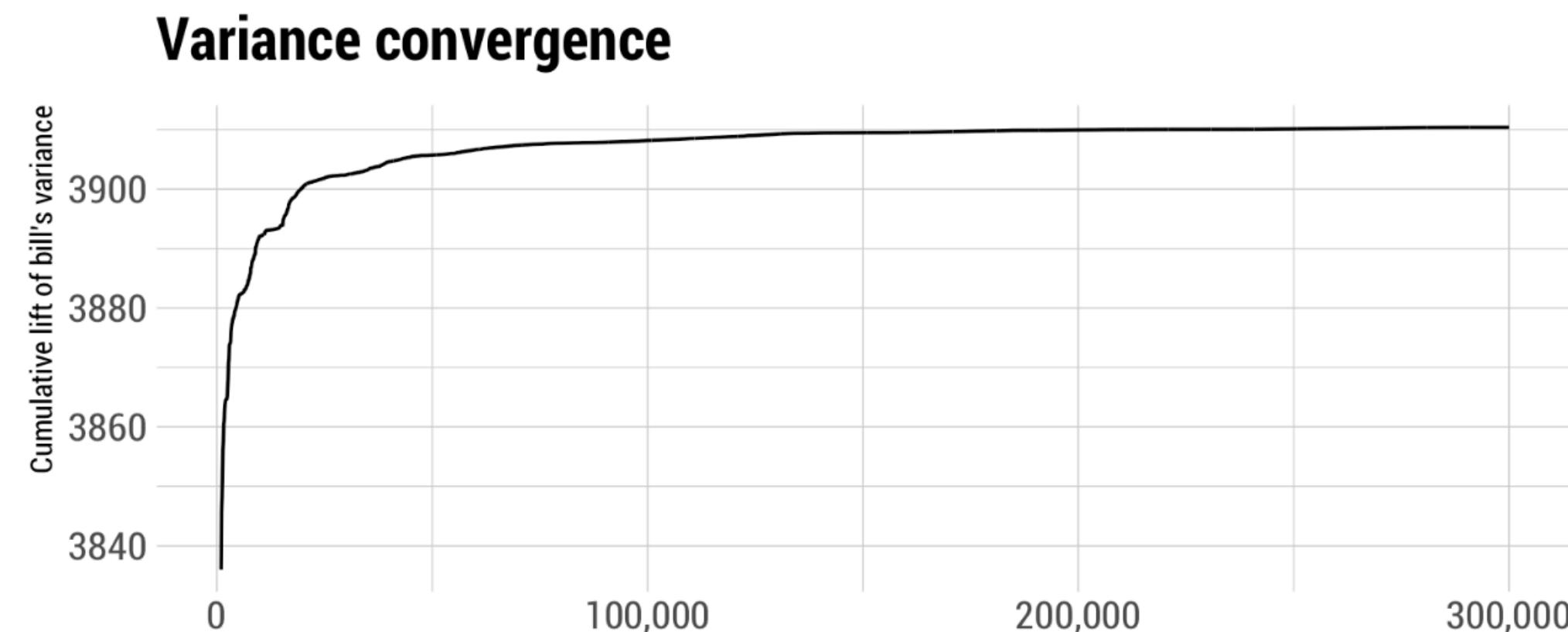


Как рассчитать нужный объем выборки?

e^xperiment fest

Еще один дополнительный ориентир

Минимальный размер выборки, который нужен при любых обстоятельствах можно посчитать по сходимости прироста дисперсии. Неважно на какой период планируется запускать эксперимент, минимальный порог по размеру выборки на конкретную метрику достаточно выбрать по схождению дисперсии в плато:



Например: перед запуском нового эксперимента требуется определить минимальный размер выборки. На какой либо из эффект рассчитывать нет цели, но нужно понять, в какой момент паттерн пользователя не изменяется

Как рассчитать нужный объем выборки?

e^xperiment fest

Конец второго дня

e^xperiment fest

Мирмахмадов Искандер

Черемисинов Виталий

07/2020

experiment-fest.ru