Множественное тестирование

История о зомби-лососе

- В 2012 году ряд авторов получил Шнобелевскую премию по нейробиологии
- Надо было протестировать аппарат МРТ
- Для этого обычно в него кладут шарик с маслом и сканируют его

➤ *http://prefrontal.org/files/posters/Bennett-Salmon-2009.pdf*
➤ *https://habr.com/ru/company/ods/blog/325416/*

История о зомби-лососе

- Это скучно, поэтому авторы решили купить
на рынке мёртвого лосося и просканировать его мозг

- Лососю показывали фотографии людей
и проверяли, есть ли у него в мозгу активность

- Оказалось,
что активность есть

➤ *http://prefrontal.org/files/posters/Bennett-Salmon-2009.pdf*
➤ *https://habr.com/ru/company/ods/blog/325416/*

История о зомби-лососе

- Аппарат МРТ возвращает много данных
- Чтобы убедиться, что в мозгу нет реакции, надо проверить много гипотез об отсутствии активности на каждом маленьком участке мозга

Проблема множественного тестирования:
если мы проверяем несколько гипотез подряд,
уровень значимости выходит из-под контроля

Мы начинаем чаще отвергать верные гипотезы,
чем нам хотелось бы

➤ *http://prefrontal.org/files/posters/Bennett-Salmon-2009.pdf*
➤ *https://habr.com/ru/company/ods/blog/325416/*

Множественная проверка гипотез

Проверяем две гипотезы:

Каждую на уровне значимости α

$\mathrm{H}_0\colon \mu_1 = \mu_2 = \mu_3$

Можно ошибиться сразу в двух местах:

$$\mathbb{P}(\text{ошибочно отвергнуть хотя бы одну из } H_0)$$
$$= 1 - \mathbb{P}(\text{не ошибиться ни в одной}) = 1 - (1-\alpha)^2$$
$$= 1 - (1 - 2\alpha + \alpha^2) = 2\alpha - \alpha^2 > \alpha$$

$$\alpha_i = 0.05 \quad \Rightarrow \quad \alpha = 0.1 - 0.025 = 0.075 > 0.05$$

❗ Вероятность ошибки первого рода накапливается и выходит из-под контроля

Множественная проверка гипотез

Пример: показ на странице сервиса нескольких новых элементов

- Изменения взаимосвязаны и их можно протестировать только на одном временном промежутке
- В такой ситуации мы сталкиваемся с множественным тестированием
- С ростом числа гипотез, вероятность получить ошибку растёт экспоненциально: $1 - (1-\alpha)^n$

❗ Нужно взять уровень значимости под контроль

Неравенство Бонферрони

- Нужно как-то скорректировать исходный уровень значимости, в этом помогает неравенство Бонферрони:

$$\mathbb{P}(A + B) \leq \mathbb{P}(A) + \mathbb{P}(B)$$

- То есть каждую гипотезу из двух надо проверять на уровне значимости $\frac{\alpha}{2}$

$$\alpha = \mathbb{P}(\text{ошибочно отвергнуть хотя бы одну из } H_0)$$

$$\leq \mathbb{P}(\text{ош. в } 1) + \mathbb{P}(\text{ош. во } 2) \ = \frac{\alpha_i}{2} + \frac{\alpha_i}{2} = \alpha_i$$

- Если гипотез k, берём уровень значимости $\frac{\alpha}{k}$ для каждой

Неравенство Бонферрони

- Из-за коррекции уровня значимости возникают проблемы с мощностью тестов
- Чем больше гипотез проверяется, тем ниже шансы отклонить неверные гипотезы
- Более того, из-за презумпции нулевой гипотезы для более низкого уровня значимости нам нужно собрать большее число наблюдений, чтобы зафиксировать значимое отклонение от нулевой гипотезы

$\Rightarrow$ процедуру надо улучшить,
чтобы мощность стала выше

Матрица ошибок

Рассмотрим случай, когда мы проверяем n гипотез

	верных H_{0i}	неверных H_{0i}
не отвергнутых H_{0i}	U	T
отвергнутых H_{0i}	V	S

- Неверно отклонили V гипотез, неверно не отклонили T гипотез
- На практике пытаются контролировать обобщения ошибки первого рода, например: FWER и FDR

Family-Wise Error Rate (FWER)

Рассмотрим случай, когда мы проверяем n гипотез

	верных H_{0i}	неверных H_{0i}
не отвергнутых H_{0i}	U	T
отвергнутых H_{0i}	V	S

Групповая вероятность ошибки, FWER (Family-Wise Error Rate) – это вероятность совершить хотя бы одну ошибку первого рода

$$FWER = \mathbb{P}(V > 0)$$

False Discovery Rate (FDR)

Рассмотрим случай, когда мы проверяем n гипотез

	верных H_{0i}	неверных H_{0i}
не отвергнутых H_{0i}	U	T
отвергнутых H_{0i}	V	S

Ожидаемая доля ложны отклонения, FDR (False Discovery Rate) – это математическое ожидание числа ошибок первого рода к общему числу отклонений нулевой гипотезы

$$FDR = \mathbb{E}\left(\frac{V}{V+S}\right)$$

Метод Холма

- Поправка Бонферрони пытается контролировать FWER (вероятность хотя бы одной ошибки 1 рода)

- **Бонферрони:** проверяем k гипотез на уровнях значимости

$$\alpha_1 = \alpha_2 = \cdots = \alpha_k = \frac{\alpha}{k}$$

- **Метод Холма** – улучшение поправки Бонферрони, обладает более высокой мощностью

- Проверяем k гипотез, но уровни значимости пытаемся выбирать разными

Метод Холма

- Отсортируем гипотезы по получившимся P-значениям по возрастанию: $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(k)}$
- Возьмём для них

$$\alpha_{(1)} = \frac{\alpha}{k}, \alpha_{(2)} = \frac{\alpha}{k-1}, \dots, \alpha_{(i)} = \frac{\alpha}{k-i+1}, \dots, \alpha_{(k)} = \alpha$$

- Если $p_{(1)} \geq \alpha_{(1)}$, все нулевые гипотезы не отвергаются, иначе отвергаем первую и продолжаем
- Если $p_{(2)} \geq \alpha_{(2)}$, все оставшиеся нулевые гипотезы не отвергаются, иначе отвергаем вторую и продолжаем
- Идём, пока не кончатся гипотезы

Метод Холма

- Метод Холма обеспечивает контроль $FWER$ на уровне α
- Метод Холма оказывается мощнее корректировки Бонферрони, так как его уровни значимости меньше

(Бонферрони)

Метод Бенджамини-Хохберга

- Отсортируем гипотезы по получившимся P-значениям по возрастанию: $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(k)}$

- Возьмём для них

$$\alpha_{(1)} = \frac{\alpha}{k}, \alpha_{(2)} = \frac{2\alpha}{k}, \dots, \alpha_{(i)} = \frac{i\alpha}{k}, \dots, \alpha_{(k)} = \alpha$$

- Если $p_{(k)} < \alpha_{(k)}$, отвергнуть все гипотезы, иначе не отвергнуть $k-$ую и продолжить

- Если $p_{(k-1)} < \alpha_{(k-1)}$, отвергнуть все оставшиеся гипотезы, иначе не отвергнуть $(k - 1) -$ ую и продолжать

- Идём, пока не кончатся гипотезы

Метод Бенджамини-Хохберга

- Для любой процедуры множественного тестирования гипотез $\mathrm{FDR} \leq FWER$
- Метод Бенджамини-Хохберга обычно оказывается более мощным, чем методы контролирующие $FWER$
- Он отвергает не меньше гипотез с теми же α_i
- Это происходит за счёт того, что метод позволяет допустить большее число ошибок первого рода

Специальные тесты

Альтернатива для процедур множественного тестирования – разработка специальных тестов, которые проверяют гипотезы сразу о нескольких ограничениях

Примеры:

- Тест отношения правдоподобий (обсудим позже)
- ANOVA – равенство сразу же нескольких математических ожиданий
- Тест Бартлета – равенство нескольких дисперсий

Резюме

- Если сделать поправку, мёртвый лосось остаётся мёртвым
- До 2010 около 40% статей по нейробиологии не использовали поправки при множественном тестировании гипотез
- Благодаря работе о лососе и Шнобелевской премии за неё удалось уменьшить число таких статей до 10%
- Корректировка уровня значимости помогает держать под контролем ложно-положительные результаты, это приводит к росту ложно-отрицательных результатов

Сколько надо наблюдений

Ошибки, что мы совершаем

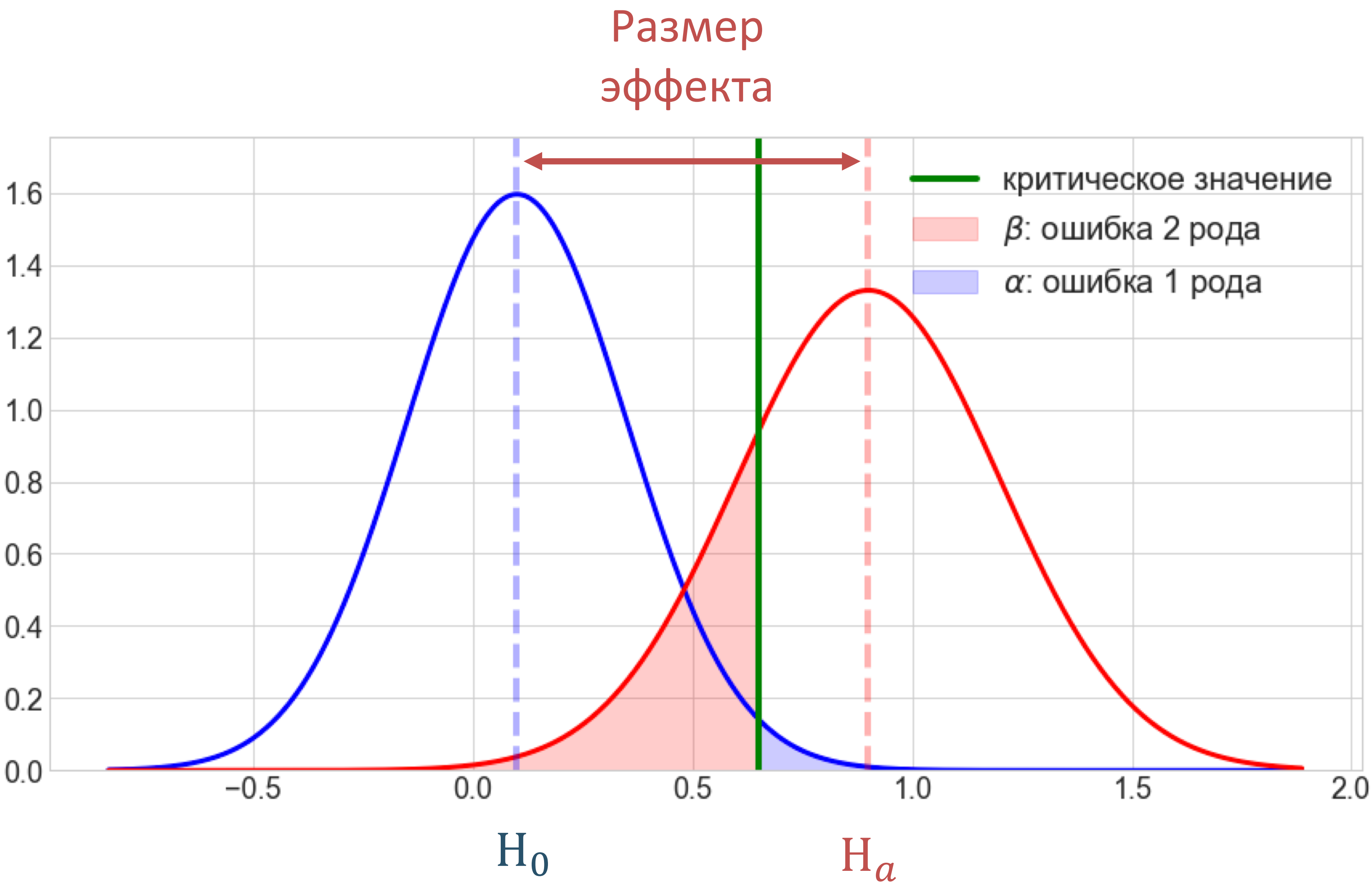
	H_0 верна	H_0 неверна	
H_0 не отвергается	ok	β	**ошибка 2 рода**
H_0 отвергается	α	ok	
	ошибка 1 рода		

$$\alpha = \mathbb{P}(H_0 \text{ отвергнута} \mid H_0 \text{ верна})$$

$$\beta = \mathbb{P}(H_0 \text{ не отвергнута} \mid H_0 \text{ не верна})$$

Величину $1 - \beta$ называют **мощностью** критерия

Размер эффекта



Сколько нужно наблюдений

- Необходимое количество наблюдений зависит от размеров ошибок первого и второго рода, а также от размера эффекта
- Фиксируем уровень значимости (ошибку 1 рода), на которую мы согласны
- Подбираем соотношение между минимальным размером эффекта, желаемой мощностью и объёмом выборки
- В выборе соотношении помогает заказчик эксперимента, у него обычно есть ограничения, с которыми нам придётся работать (количество магазинов, длительность АБ-теста и т.п.)

Таблица эффекта-ошибки

		Ошибка 1/2 рода $\alpha = \beta$			
		0.1%	1%	5%	10%
размер эффекта	1%	много данных			
	1.5%				
	3%				
	5%				
	10%				мало данных

Совокупность этих трёх параметров (ошибка 1/2 рода, размер эффекта) позволяют рассчитать необходимый для эксперимента объём выборки.

Сколько нужно наблюдений

Пример: проверяем равенство конверсий до и после нововведений

$\mathrm{H}_0\colon p_0 = p_a$

$\mathrm{H}_a\colon p_0 \neq p_a$

Используем асимптотически-нормальный тест:

$$z = \frac{p_a - p_0}{\sqrt{P(1-P)\cdot\left(\frac{1}{n}+\frac{1}{n}\right)}} \underset{\mathrm{H}_0}{\overset{asy}{\sim}} N(0,1)$$

размер эффекта

Сколько нужно наблюдений

Ошибка второго рода:

$$\beta = \Phi\left(\frac{\sqrt{p_0(1-p_0)}}{\sqrt{p_a(1-p_a)}} \cdot z_{1-\alpha} + \frac{p_0 - p_a}{\sqrt{\frac{p_a(1-p_a)}{n}}}\right)$$

Число наблюдений:

$$n = \left(\frac{z_{1-\alpha} \cdot \sqrt{p_0(1-p_0)} + z_{1-\beta} \cdot \sqrt{p_a(1-p_a)}}{p_a - p_0}\right)^2$$

размер эффекта

Анализ мощности

До эксперимента:

- Какой нужен объём выборки, чтобы найти различия с разумной степенью уверенности
- Различия какой величины мы можем найти, если известен объём выборки

После эксперимента:

- смогли бы мы найти различия с помощью нашего эксперимента, если бы величина эффекта была равна Δ

Резюме

- Для многих критериев можно вывести формулу для расчёта необходимого числа наблюдений
- Число наблюдений зависит от ошибок ½ рода и минимального размера эффекта, который мы хотим уловить
- Перед экспериментом необходимое число наблюдений определяют исходя из пожеланий заказчика и физических возможностей