1 Когда используются поправки

Вообще не мои изречения все взято из курса прикладной эконометрики на экономическом факультете МГУ, в частности с гитхаба Анны Юрьевны Ставнийчук
Ужасно полезные ссылки: 1, 2, 3

История про множественное тестирование гипотез с примерами. Особенность множественного тестирования гипотез состоит в том, что чем больше гипотез мы проверяем на одних и тех же данных, тем больше будет вероятность допустить как минимум одну ошибку первого рода – эффект множественных сравнений (multiple comparisons/testing).

Как минимум одна ошибка первого рода иначе называется family wise error (FWER):

$$FWER = 1 - (1 - \alpha)^2$$

Источниками множественного тестирования могут быть:

- **Несколько типов воздействия (Multiple treatment arms)** - отберем бесплатную рассрочку для группы 1 и понизим лимит до 20 тысяч для группы 2
- **Гетерогенное воздействие (Heterogeneous treatment effects)** - а как эффект будет только на Digital, Personal, Private, а не на всех в целом.
- **Несколько способов оценки (Multiple estimators)** - давайте накрутим t-test, а потом еще критерием манна-уитни проверим, ну и бутстрапом заодно.
- **Несколько зависимых переменных (Multiple outcomes), эффект на которые мы хотим оценить** - хотим посмотреть, чтобы одна метрика выросла, а другие не просели.

Примеры

В) **(2 балла)** расстроившись из-за статистически незначимого влияния нового алгоритма на удержание пользователя, руководитель предлагает аналитику проверить, нет ли влияния изменения алгоритма на другие показатели, и сравнить на 5% уровне значимости дополнительные результаты. В таблице представлены результаты:

Результирующий показатель	P-значение теста на равенство средних
Показатель 1	0.215
Показатель 2	0.442
Показатель 3	0.002
Показатель 4	0.125
Показатель 5	0.090
Показатель 6	0.200

Объясните с точки зрения статистики, почему аналитик считает, что не нужно внедрять новый алгоритм. Приведите соответствующие расчёты по таблице.

Заметим, что внедрение нового неудачного алгоритма может привести к значительным финансовым потерям.

Руководитель предлагает заняться data-фишингом: провести множественное тестирование. когда заранее не была определена целевая метрика, и ещё и внедрение нового неудачного алгоритма может привести к финансовым потерям, то есть если на самом деле эффекта от нового алгоритма нет, а тесты его найдут, то стоимость ошибки 1 рода будет большой.

Далее самое простое - это сказать, что давайте в такой ситуации проведём корректировку p-значений, из них не требует дополнительных предпосылок и «равноправно» относится к гипотезам коррекция Бонферрони. Достаточно просто поделить $\alpha = 0.05$ на 6, получить 0.0083 и сравнить p-значения из шести гипотез с новым урровнем значимости. (Значимый результат только в 1 гипотезе из 6, поэтому аналитик сделал вывод, что не надо внедрять алгоритм.) Альтернативно можно было пересчитать не уровень значимости, а p-значения. Можно было использовать любую другую коррекцию.

По-хорошему, здесь до расчётов стоило бы ещё обосновать необъодимость корректировки через $FWER = 1-(1-\alpha)^6 = 1-(0.95)^6$, что приближённо равно 0.26. Это верроятность ошибки 1 рода, когда значимость покажет хотя бы один из 6 тестов, хотя реально на самом деле эффекта нет. А можно было бы просто остаться на этом шаге объяснений.

Но только считается FWER не по p-значениям – за это снижали баллы.

Е) **(2 балла)** По результатам эксперимента Кейт не обнаружила значимой на уровне 5% разницы. Менеджер предлагает проверить, нет ли эффекта на какой-нибудь подгруппе пользователей. Исходя из характеристик запросов (город отправления и прибытия, даты поездки) он делит наблюдения на 50 групп и предлагает провести для каждой на 5% уровне значимости такой же эксперимент. Кейт провела все тесты. Среди получиенных p-значений самое маленькое равно 0.0021. Кейт понимает, что внедрение неудачной новой ценовой политики может повлечь за собой финансовые потери компании. Как ей поступить в этой ситуации? Дайте содержательные и формальные пояснения.

В) **(2 балла)** По настоянию руководителя аналитик проверяет ещё 10 вариантов рекламы, сравнивает между собой результаты по каждой паре вариантов на 5% уровне значимости, чтобы выбрать лучший вариант рекламы, и затем показывать его всем пользователям приложения Котобанка. Заметим, что внедрение неудачной рекламы может привести к некоторым финансовым потерям. Какова вероятность того, что по результатам тестов хотя бы один из вариантов рекламы окажется статистически эффективнее остальных, хотя на самом деле разницы между ними нет? Что с точки зрения статистики Вы посоветуете Илье?

Илья и его руководитель столкнулись с проблемой множественного тестирования гипотез, то есть вероятность получить ложно-значений результат (совершить ошибку 1 рода) в хотя бы одном из тестов $= FWER = 1-(1-005)^{12*11/2}$, что значительно больше, чем 0.05, т.е компания рискует, опираясь на ложно-положительный результат, потерять деньги. Что делать? Изложить эти соображения руководителю. А если придёся провести все эти тесты, то использовать более жёсткий уровень значимости, например, применив коррекцию Бонферрони или Хольма (далее краткое описание процедуры).

p-hacking — это практика манипуляции данными или проведением анализа с целью достижения статистически значимого результата

- *Множественное тестирование: исследователь тестирует большое количество гипотез и выбирает только те, которые дали «значимый» результат*

- *Изменение критериев: исследователь может начать с анализа всей выборки, а затем, не получив значимого результата, выделить подгруппы или исключить выбросы*

- *Игра с методами: исследователь может экспериментировать с различными методами анализа, чтобы выбрать тот, который дает желаемый результат.*

- *Применение различных ковариат: добавление или удаление ковариат для изменения результата теста на значимость*

2 Задача про транзитивность

Пусть у нас есть контроль и две тестовые группы. Мы проверили, что контроль статзначимо отличается от тестовой группы №1, и тестовая группа №1 статзначимо отличается от тестовой группы №2. Можно ли утверждать, что контроль статзначимо отличается от тестовой группы №2, не проводя тест *(и не делая поправку на множественное тестирование гипотез)*.

В терминах линейной регрессии это можно записать как:

$$Y_i = \beta_0 + \beta_1 * d_i^1 + \beta_2 * d_i^2 + \varepsilon_i,$$

где

- Y_i — целевая метрика (зависимая переменная) для i-го наблюдения;
- d_i^1 — дамми-переменная того, что наблюдение находит в тестовой группе №1 (равно 1, если клиент в группе №1 и 0 иначе);
- d_i^2 — дамми-переменная того, что наблюдение находит в тестовой группе №2;

$$d_i^j = \begin{cases} 1, & \text{если i-ое наблюдение находится в j тритмент группе} \\ 0, & \text{иначе} \end{cases}$$

Далее с помощью МНК мы получили оценку нашей регрессии:

$$\widehat{Y_i} = \widehat{\beta}_0 + \widehat{\beta}_1 * d_i^1 + \widehat{\beta}_2 * d_i^2$$

Утверждение 1: Оценка эффекта для **контроль vs тестовая группа №1** эквивалентна оценке значимости коэффициента β_1.

$$\begin{array}{ll} H_0: & \beta_1 = 0 \\ H_1: & \beta_1 \neq 0 \end{array} \quad \Rightarrow \quad t_1 = \left| \frac{\widehat{\beta}_1 - 0}{\text{s.e.}(\widehat{\beta}_1)} \right| \quad \text{vs} \quad St^{\alpha}_{n-3}$$

$$\text{Если } t_1 > St^{\alpha}_{n-3}, \text{ то } H_0 \text{ не принимается}$$

Утверждение 2: Оценка эффекта для **тестовая группа №1 vs тестовая группа №2** эквивалентна проверке гипотезы о том, что коэффициенты β_1 и β_2 статзначимо отличаются друг от друга.

$$\begin{array}{llll} H_0: & \beta_1 = \beta_2 & \leftrightarrow & \beta_1 - \beta_2 = 0, \\ H_1: & \beta_1 \neq \beta_2 & \leftrightarrow & \beta_1 - \beta_2 \neq 0 \end{array} \quad \Rightarrow \quad t_{1,2} = \left| \frac{(\widehat{\beta}_1 - \widehat{\beta}_2) - 0}{\text{s.e.}(\widehat{\beta}_1 - \widehat{\beta}_2)} \right| \quad \text{vs} \quad St^{\alpha}_{n-3}$$

$$\text{Если } t_{1,2} > St^{\alpha}_{n-3}, \text{ то } H_0 \text{ не принимается}$$

Утверждение 3: Оценка эффекта для **контроль vs тестовая группа №2** эквивалентна оценке значимости коэффициента β_2.

$$\begin{array}{ll} H_0: & \beta_2 = 0 \\ H_1: & \beta_2 \neq 0 \end{array} \quad \Rightarrow \quad t_2 = \left| \frac{\widehat{\beta}_2 - 0}{\text{s.e.}(\widehat{\beta}_2)} \right| \quad \text{vs} \quad St^{\alpha}_{n-3}$$

$$\text{Если } t_2 > St^{\alpha}_{n-3}, \text{ то } H_0 \text{ не принимается}$$

Следовательно, надо доказать, что

$$t_1 > St^{\alpha}_{n-3} \quad \text{и} \quad t_{1,2} > St^{\alpha}_{n-3}, \quad \Longrightarrow \quad t_2 > St^{\alpha}_{n-3}$$

$$\left|\frac{\widehat{\beta}_1}{\text{s.e.}(\widehat{\beta}_1)}\right| > c \quad \text{и} \quad \left|\frac{\widehat{\beta}_1 - \widehat{\beta}_2}{\text{s.e.}(\widehat{\beta}_1 - \widehat{\beta}_2)}\right| > c \quad \Longrightarrow \quad \left|\frac{\widehat{\beta}_2}{\text{s.e.}(\widehat{\beta}_2)}\right| > c,$$

для некоторого $c > 0$, зависящего от уровня значимости α и числа степеней свободы.

Пусть оценки $\widehat{\beta}_1$ и $\widehat{\beta}_2$ независимы (справедливо, если мы проводим тест) - то есть:

$$\text{Cov}(\widehat{\beta}_1, \widehat{\beta}_2) \ = \ 0$$

$$\text{s.e.}(\hat{\beta}_1) = \text{s.e.}(\overline{Y_1} - \overline{Y_0}) = \sqrt{\hat{\sigma}^2 \left(\frac{1}{n_1} + \frac{1}{n_0}\right)}$$

$$\text{s.e.}(\hat{\beta}_2) = \sqrt{\hat{\sigma}^2 \left(\frac{1}{n_2} + \frac{1}{n_0}\right)}$$

$$\text{s.e.}(\hat{\beta}_1 + \hat{\beta}_2) = \sqrt{\text{s.e.}(\hat{\beta}_1)^2 + \text{s.e.}(\hat{\beta}_2)^2 + 2\ \text{Cov}(\hat{\beta}_1, \hat{\beta}_2)} = \hat{\sigma}\sqrt{\frac{2}{n_0} + \frac{1}{n_1} + \frac{1}{n_2}}$$

$\hat{\sigma}^2$ — оценка дисперсии ошибок (при $N = n_0 + n_1 + n_2$):

$$\hat{\sigma}^2 = \frac{1}{N-3} * \sum_{i=1}^{N} e_i^2 = \frac{1}{N-3} * \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

Таким образом, пытаемся доказать, что:

$$\frac{\left|\widehat{\beta}_1\right|}{\sqrt{\hat{\sigma}^2 \left(\frac{1}{n_1} + \frac{1}{n_0}\right)}} > c \quad \text{и} \quad \frac{\left|\widehat{\beta}_1 - \widehat{\beta}_2\right|}{\hat{\sigma}\sqrt{\frac{2}{n_0} + \frac{1}{n_1} + \frac{1}{n_2}}} > c \quad \Longrightarrow$$

$$\Longrightarrow \quad \frac{\left|\widehat{\beta}_2\right|}{\sqrt{\hat{\sigma}^2 \left(\frac{1}{n_2} + \frac{1}{n_0}\right)}} > c$$

Приведем контрпример, где такая зависимость не соблюдается.
Это история про то, что β_2 статзначимо меньше β_1.

Пусть количество наблюдений в каждой группе одинаково ($n_0 = n_1 = n_2 = n$), а также $\hat{\sigma} = 1, \ \ n = 100, \ \ c = St^{0.05}_{97} = 2$, тогда попытаемся доказать, что:

$$\left|\widehat{\beta}_1\right| > c\hat{\sigma}\sqrt{\frac{2}{n}} \quad \text{и} \quad \left|\widehat{\beta}_1 - \widehat{\beta}_2\right| > 2c\hat{\sigma}\sqrt{\frac{1}{n}} \quad \Longrightarrow \quad \left|\widehat{\beta}_2\right| > c\hat{\sigma}\sqrt{\frac{2}{n}}$$

$$\left|\widehat{\beta}_1\right| > 2\sqrt{0.02} \quad \Longrightarrow \quad \left|\widehat{\beta}_1\right| > 0.2828$$

$$\left|\widehat{\beta}_1-\widehat{\beta}_2\right| > 4\sqrt{\frac{1}{100}} \quad \Longrightarrow \quad \left|\widehat{\beta}_1-\widehat{\beta}_2\right| > 0.4$$

$$\left|\widehat{\beta}_2\right| > 2\sqrt{0.02} \quad \Longrightarrow \quad \left|\widehat{\beta}_2\right| > 0.2828$$

Возьмем $\widehat{\beta}_1 = 1, \ \ \widehat{\beta}_2 = 0.2$:

$$|1| > 0.2828 \quad \text{и} \quad |0.8| > 0.4 \quad \not\Longrightarrow \quad |0.2| > 0.2828 = 0.2\sqrt{2}$$

Если это была история про статзначимо больше, а не про статзначимо различаются, то модули в t-статистике уходят:

$$\begin{array}{ll} H_0: & \beta_1 \leq 0 \\ H_1: & \beta_1 > 0 \end{array} \quad \Rightarrow \quad t_1 = \frac{\widehat{\beta}_1}{\text{s.e.}(\widehat{\beta}_1)} \quad \text{vs} \quad St^{\alpha}_{n-3}$$

Рассмотрим случай, когда $\hat{\beta}_1 > 0, \ \ \hat{\beta}_1 - \hat{\beta}_2 > 0$ и хотим доказать, что возможна ситуация, когда $\hat{\beta}_2 < 0$:

$$\beta_1 > c * \text{s.e.}(\hat{\beta}_1)$$

$$\beta_2 > \beta_1 + c * \text{s.e.}(\hat{\beta}_2 - \hat{\beta}_1)$$

$$\Longrightarrow \beta_2 > c * (\text{s.e.}(\hat{\beta}_1) + \text{s.e.}(\hat{\beta}_2 - \hat{\beta}_1))$$

И хотим:

$$\beta_2 < c * \text{s.e.}(\hat{\beta}_2)$$

То есть

$$c * (\text{s.e.}(\hat{\beta}_1) + \text{s.e.}(\hat{\beta}_2 - \hat{\beta}_1)) < \beta_2 < c * \text{s.e.}(\hat{\beta}_2)$$

$$c\hat{\sigma}\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_0}\right)} + c\hat{\sigma}\sqrt{\frac{2}{n_0}+\frac{1}{n_1}+\frac{1}{n_2}} < \beta_2 < c\hat{\sigma}\sqrt{\left(\frac{1}{n_2}+\frac{1}{n_0}\right)}$$

$$c\hat{\sigma}\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_0}\right)} + c\hat{\sigma}\sqrt{\frac{2}{n_0}+\frac{1}{n_1}+\frac{1}{n_2}} < c\hat{\sigma}\sqrt{\left(\frac{1}{n_2}+\frac{1}{n_0}\right)}$$

В стандартном тесте такое вряд ли возможно (если $\text{Cov}(\widehat{\beta}_1, \widehat{\beta}_2) = 0$ и $n_0 = n_1 = n_2 = n$):

$$c\hat{\sigma}\sqrt{\frac{2}{n}} + c\hat{\sigma}\sqrt{\frac{4}{n}} \quad \not< \quad c\hat{\sigma}\sqrt{\frac{2}{n}}$$

Но существуют случаи, когда:

$$\sqrt{\frac{1}{n_1}+\frac{1}{n_0}} + \sqrt{\frac{2}{n_0}+\frac{1}{n_1}+\frac{1}{n_2}} < \sqrt{\frac{1}{n_2}+\frac{1}{n_0}}$$

- n_0 очень большой — 10^5 наблюдений,
- n_1 достаточно маленький — 50 штук,
- n_2 ещё меньше — 10 штук.

- **s.e.**(β_1) будет в первую очередь определяться

$$\sqrt{\frac{1}{n_1} + \frac{1}{n_0}} \approx \sqrt{\frac{1}{50}} \approx 0.14$$

(поскольку $\frac{1}{n_0}$ совсем крошечное);

- s.e.(β_2) будет ещё крупнее, так как

$$\sqrt{\frac{1}{n_2} + \frac{1}{n_0}} \approx \sqrt{\frac{1}{10}} \approx 0.36$$

- s.e.$(\beta_1 - \beta_2)$ может оказаться относительно небольшой, если оценки β_1 и β_2 сильно отрицательно скоррелированы, тогда

$$\mathrm{Var}(\beta_1 - \beta_2) = \mathrm{Var}(\beta_1) + \mathrm{Var}(\beta_2) - 2\ \mathrm{Cov}(\beta_1, \beta_2)$$

способна заметно «просесть».

Ну тогда можно придумать бредовый пример, который никогда не произойдет.

Короче, Инна права, если тест был даже двухсторонний, и там было показано явно, что $\beta_2 > \beta_1$ (не было лажи в определении того, что β_2 реально больше β_1 - то есть доверительный интервал такой разницы лежит больше нуля), то все ок.

История про поправку. Допустим, что утверждение все же было правильным.

Тогда в 5% случаев мы бы могли ошибиться в утверждении 1 о том, что тестовая группа №1 статзначимо отличается от контроля, то есть приняли бы гипотезу H_1, хотя в реальности гипотеза H_0. Так же мы бы могли ошибиться и при проверки гипотезы из утверждения 2, что тестовая группа №2 статзначимо отличается от тестовой группы №1, в 5% случаев. Таким образом, вероятность, что мы ошибемся хотя бы один раз в одном из двух утверждений равна:

$$1 - (1 - 5\%)^2 = 0.0975$$

Так как из этих двух утверждений следует третье о том, что тестовая группа №2 статзначимо отличается от контроля, то его мы принимаем на другом уровне значимости равном 9.75%, иначе нужно делать поправку на множественное тестирование и делить уровень значимости на два:

$$1 - (1 - \frac{5}{2}\%)^2 = 0.049375$$

Как по мне, надо быть гадалкой, чтобы вытаскивать всегда такой коэф β_1, который статзначимо меньше β_2 и при этом статзначимо лучше контроля, чтобы пользоваться всегда коррекцией на два, а не на три. А если ты уже проверил гипотезу с β_2, то надо пользоваться коррекцией на три, а не закрывать глаза (ну я надеюсь).

Прости Господи, доказательство эквивалетности оценок

Рассмотрим три группы: **Контроль (C)**, **Тритмент №1 (T1)**, **Тритмент №2 (T2)**. Для каждого i-го наблюдения введём дамми-переменные:

$$d_i^j = \begin{cases} 1, & \text{если i-ое наблюдение находится в тритмент группе j} \\ 0, & \text{иначе} \end{cases}$$

Запишем уравнение линейной регрессии:

$$Y_i = \alpha + \beta_1\, d_i^1 + \beta_2\, d_i^2 + \varepsilon_i.$$

Тогда коэффициенты в модели можно интерпретировать следующим образом:

- Для группы **Контроля** ($d_i^1 = 0$, $d_i^2 = 0$): $\mathbb{E}[Y_i \mid \text{Контроль}] = \alpha$.
- Для группы **Тритмента №1** ($d_i^1 = 1$, $d_i^2 = 0$): $\mathbb{E}[Y_i \mid \text{T1}] = \alpha + \beta_1$.
- Для группы **Тритмента №2** ($d_i^1 = 0$, $d_i^2 = 1$): $\mathbb{E}[Y_i \mid \text{T2}] = \alpha + \beta_2$.

Таким образом, β_1 и β_2 показывают, на сколько в среднем отличаются группы T1 и T2 от контрольной.

Пусть:

$$\overline{Y}_C = \frac{1}{n_C}\sum_{i\in C} Y_i, \quad \overline{Y}_{T1} = \frac{1}{n_{T1}}\sum_{i\in T1} Y_i, \quad \overline{Y}_{T2} = \frac{1}{n_{T2}}\sum_{i\in T2} Y_i.$$

МНК-оценка производится путем минимизации суммы квадратов ошибок:

$$L(\alpha, \beta_1, \beta_2) = \sum_{i=1}^{n}\Big(Y_i - \alpha - \beta_1\, d_i^1 - \beta_2\, d_i^2\Big)^2 \to \min_{\alpha,\beta_1,\beta_2}$$

Найдем минимум MSE:

$$\frac{\partial L}{\partial \alpha} = \sum_{i=1}^{n}\Big(Y_i - \alpha - \beta_1\, d_i^1 - \beta_2\, d_i^2\Big) = 0,$$
$$\frac{\partial L}{\partial \beta_1} = \sum_{i=1}^{n}\Big(Y_i - \alpha - \beta_1\, d_i^1 - \beta_2\, d_i^2\Big)d_i^1 = 0,$$
$$\frac{\partial L}{\partial \beta_2} = \sum_{i=1}^{n}\Big(Y_i - \alpha - \beta_1\, d_i^1 - \beta_2\, d_i^2\Big)d_i^2 = 0.$$

Рассмотрим уравнение для β_1:

$$\sum_{i\in T1}\big(Y_i - \alpha - \beta_1\big) = 0 \quad\Longrightarrow\quad T_1(\alpha + \beta_1) = \sum_{i\in T1} Y_i \quad\Longrightarrow\quad \alpha + \beta_1 = \overline{Y}_{T1}.$$

$$\beta_1 = \overline{Y}_{T1} - \alpha.$$

Аналогично для β_2:

$$\sum_{i\in T2}\big(Y_i - \alpha - \beta_2\big) = 0 \quad\Longrightarrow\quad \alpha + \beta_2 = \overline{Y}_{T2},$$

$$\beta_2 = \overline{Y}_{T2} - \alpha.$$

Уравнение для α:

$$\alpha = \overline{Y}_C.$$

Подставляя найденное значение α в уравнения для β_1 и β_2, получаем:

$$\widehat{\alpha} = \overline{Y}_C, \quad \widehat{\beta}_1 = \overline{Y}_{T1} - \overline{Y}_C, \quad \widehat{\beta}_2 = \overline{Y}_{T2} - \overline{Y}_C.$$

Оценка разницы средних $\overline{Y}_{T1} - \overline{Y}_C$ для группы T1 и $\overline{Y}_{T2} - \overline{Y}_C$ для группы T2 в условиях эксперимента интерпретируется как ATE. Таким образом, коэффициенты $\widehat{\beta}_1$ и $\widehat{\beta}_2$ из линейной регрессии являются оценками ATE для соответствующих тритмент-групп относительно контрольной.

3 Задача Ангелины

У нас есть 4 экрана, на каждом из них клиент может нажать на баннер (может и несколько раз нажать на разных экранах). Тритменту определенные баннеры, контролю не эти баннеры. Можно оценить эффект в целом, а можно ли оценить эффект для каждого экрана?

- Ну в идеальном эксперименте надо иметь 4 тритмента для каждого экрана - в таком случае наш t-test может ложно сработать в 19% случаев - хотя бы одна гипотеза подтвердится, а на самом деле там H_1.
- Можно оценить гетерогенный эффект - засунуть в регрессию дамми экранов и их произведения, но в таком случае это просто корреляция, а не причинно-следственная связь — почти наверняка есть всякий selection-bias — чел перешел на этот экран, потому что он разбирается в кредитных продуктах и хороший грейсер, а мы как раз предлагаем ему еще один, в итоге, метрика доход на клиента упадет по сравнению с теми, кто этот баннер не видел — путаем эффект от баннера и эффект от характеристик людей — можно придумать еще кейсы с плохим selection bias.
- Можно почитать про ANOVA: H0 – все средние равны, H1 – хотя бы одно среднее отличается — вроде это просто тест на короткую против длинной регрессию — но это опять упрется в корреляцию или причинно-следственную связь.
- Можно сравнивать все комбинации переходов - но это очень большая корректировка и общая гипотеза: "Влияют ли хотя бы на одной комбинации экранов баннеры" - вроде 16 штук