KARPOV.COURSES >>>
КОНСПЕКТ

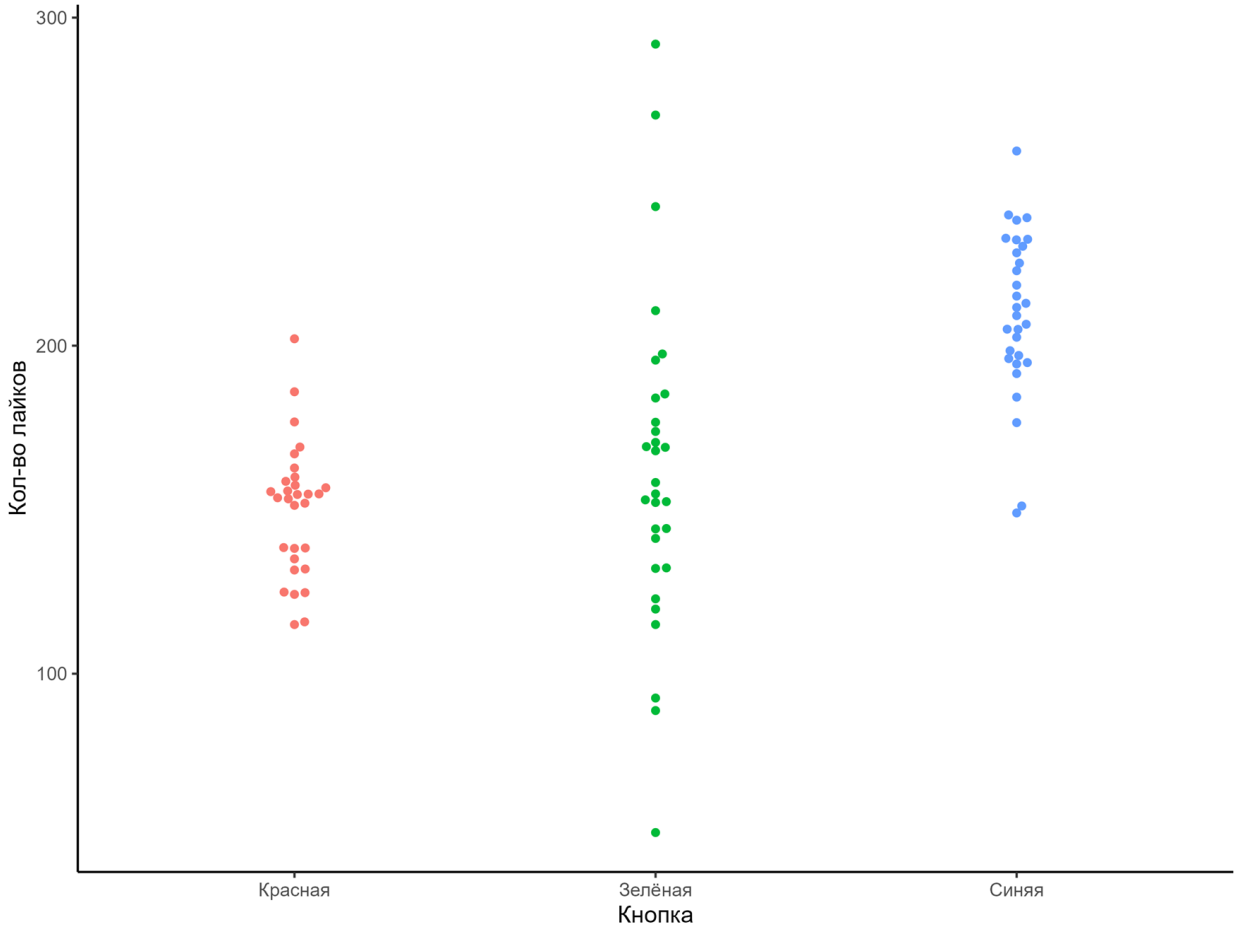
> Конспект > 5 урок > СТАТИСТИКА

> Оглавление

1. Однофакторный дисперсионный анализ
2. Требования к дисперсионному анализу
3. Построение графиков
4. Множественные сравнения
5. Поправка Бонферрони
6. Критерий Тьюки
7. Многофакторный ANOVA
8. Требования к многофакторному дисперсионному анализу
9. Как визуализировать
10. Дополнительно

> Однофакторный дисперсионный анализ

В тех случаях, когда групп становится больше двух, t-критерий перестаёт быть нам полезен. Для таких случаев был создан дисперсионный анализ – он выполняет ту же функцию, но подходит для числа групп больше двух. Технически его можно применять и для двух групп, но это больше вопрос конвенции и удобства.



Как это работает?

Допустим, нам захотелось сравнить между собой аж три выборочных средних X_1, X_2 и X_3.

Гипотезы:

- H_0 – ни одно из выборочных средних не отличается от другого (нет различий)
- H_1 – **хотя бы одно** выборочное среднее отличается от других (есть различия как минимум между двумя группами)

Помимо уже указанных выборочных средних, мы будем работать со средним всех наблюдений – то есть мы рассчитываем среднее не только для каждой группы, но и

среднее по всем группам сразу. Отметим его как $\overline{X}$.

После этого мы считаем так называемую общую сумму квадратов (**SST**, total sum of squares) –мы просто вычитаем из среднего всех наблюдений каждое индивидуальное наблюдение, разницу возводим в квадрат и суммируем все квадраты разниц. Таким образом, у нас получилась мера того, насколько сильно наблюдения отклоняются друг от друга и это можно считать аналогом дисперсии.

$$SST = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(X_{ij} - \overline{\overline{X}})^2$$

Здесь

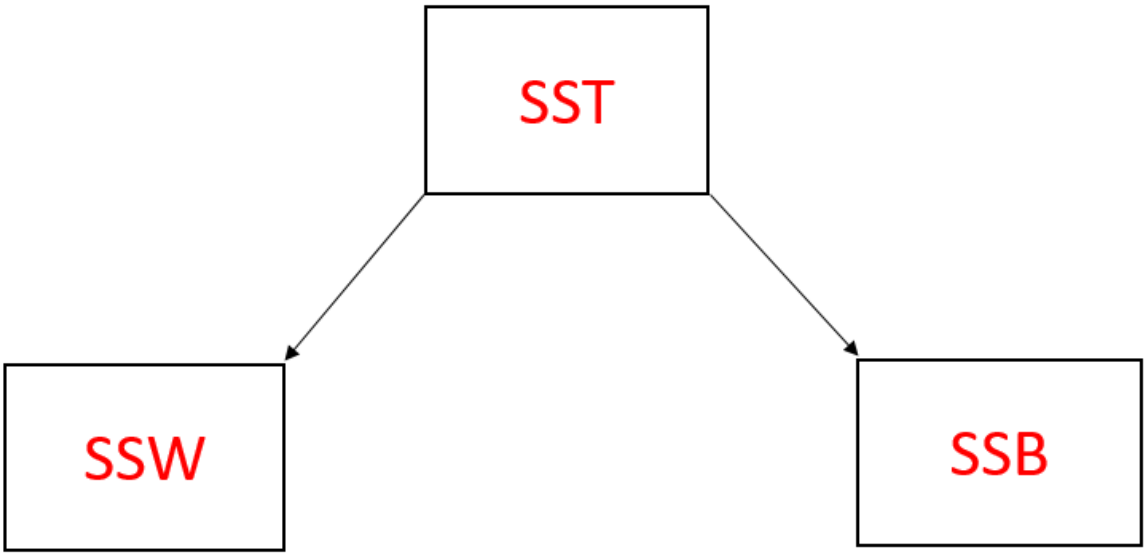
- m - количество групп
- n_i - размер группы под номером i
- X_{ij} - наблюдение под номером j из группы i (индивидуальное наблюдение)
- $\overline{\overline{X}}$ - среднее всех наблюдений
- $\sum$ - знак суммы

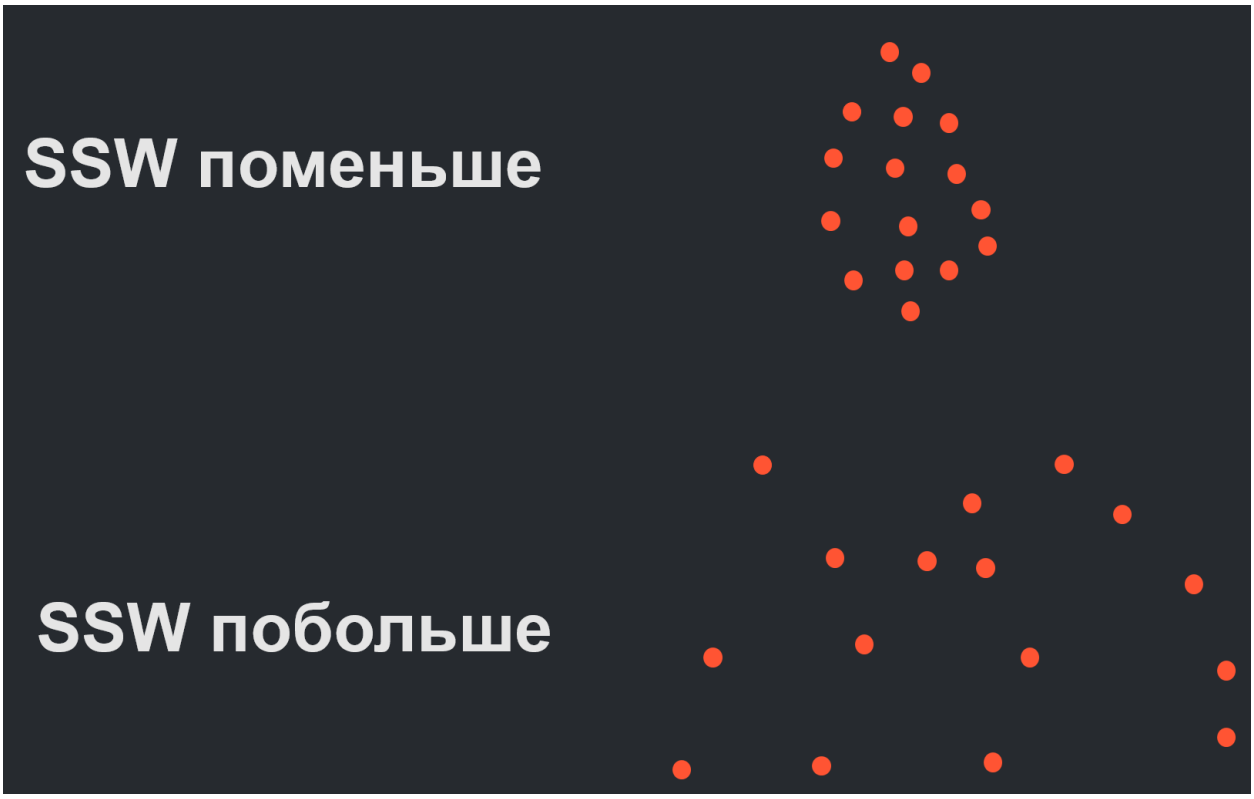
Помимо расчёта этой величины, нам важно посчитать её **степени свободы**. Иначе с ростом выборки SST тоже будет расти, а нам нужно, чтобы это зависело только от изменчивости в самих данных!

Количество степеней свободы для SST рассчитывается как количество наблюдений (N) минус 1, то есть:

$$df_{SST} = N - 1$$

Общая сумма квадратов, по сути, складывается из двух разных источников изменчивости. Во-первых, индивидуальные данные могут различаться между собой в рамках одной группы – внутригрупповая сумма квадратов (**SSW**, sum of squares within groups). Во-вторых, они могут различаться между несколькими группами – межгрупповая сумма квадратов (**SSB**, sum of squares between groups – иногда это называют **SSA**, sum of squares among groups).





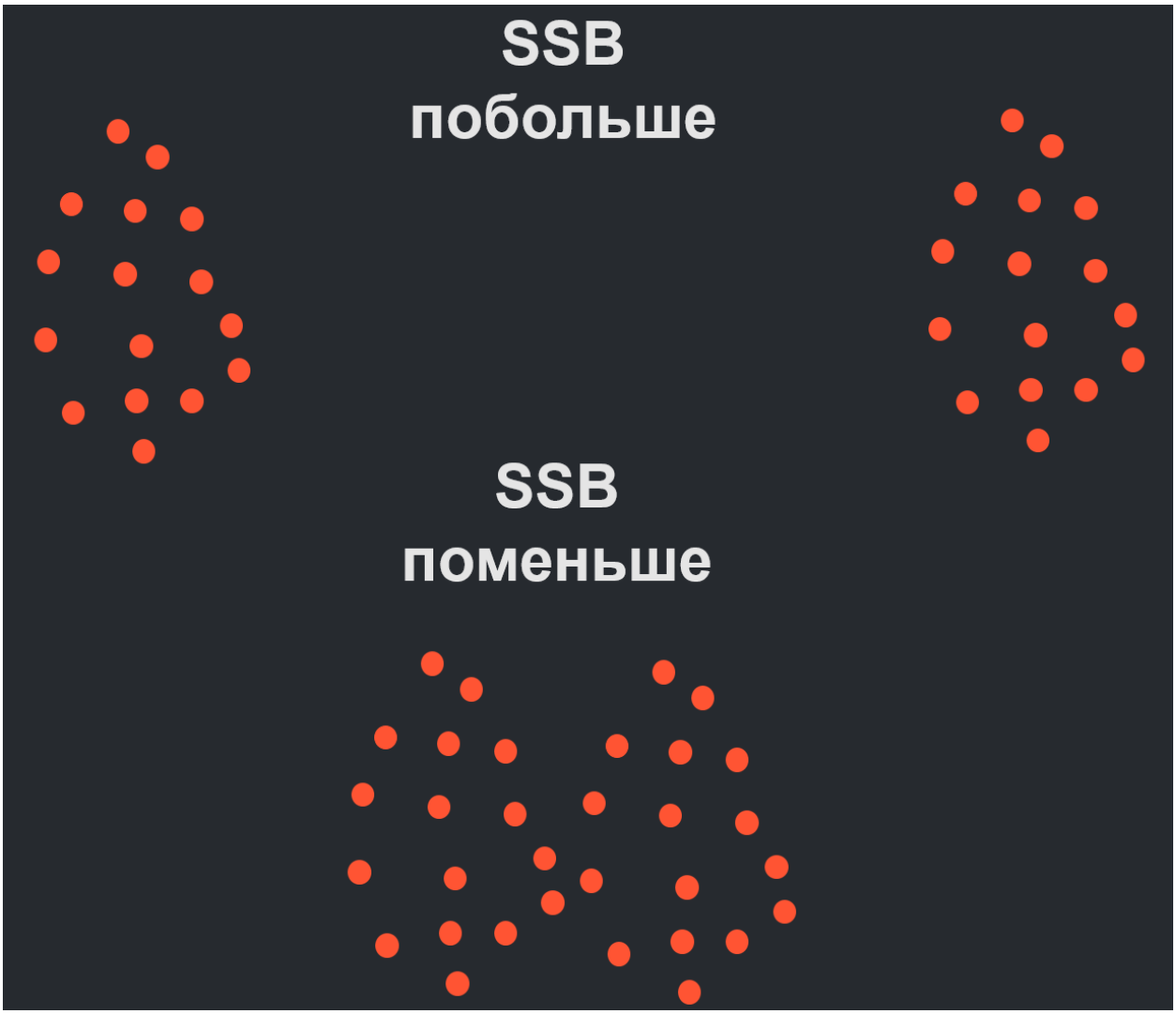
SSW считается на основе выборочных средних – мы просто вычитаем из выборочного среднего одной группы каждое индивидуальное значение, возводим каждую разницу в квадрат и вновь суммируем все квадраты разниц между собой (и так для всех групп).

$$SSW = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$$

Из новых обозначений тут только $\bar{X}_i$ - среднее по группе номер i.

Степени свободы здесь рассчитываются как разница между общим количеством наблюдений (N) и количеством групп (m), то есть:

$$df_{SSW} = N - m$$



В свою очередь, при расчёте SSB мы вычитаем из общего среднего выборочные средние, возводим разницу в квадрат и также суммируем все квадраты разниц.

$$SSB = \sum_{i=1}^{m} n_i (\bar{X}_i - \overline{\overline{X}})^2$$

В некоторых версиях этой формулы n_i нет, это нужно для случая, когда группы неравны по своим размерам.

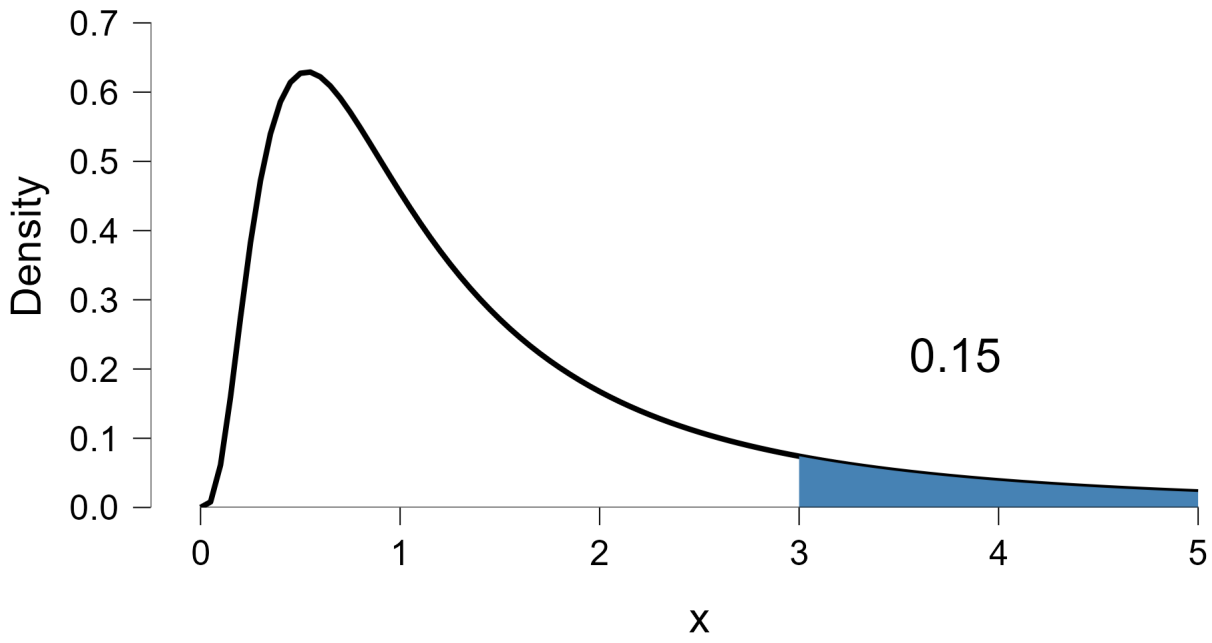
Степени свободы здесь рассчитываются как количество групп (m) минус 1, то есть:

$$df_{SSB} = m - 1$$

Что для нас лучше: чтобы SSB было больше SSW. Это будет означать, что значения между группами различаются больше, чем внутри группы => группы значимо различаются между собой. Формально это выражается через так называемую F-статистику, которая считается по следующей формуле:

$$F = \frac{\frac{SSB}{df_{SSB}}}{\frac{SSW}{df_{SSW}}}$$

Дальше мы традиционно предполагаем, что нулевая гипотеза верна. В таком случае SSW было бы каким-то фиксированным значением, а SSB бы стремилось к нулю => F-статистика бы тоже стремилась к нулю. Соответственно, мы можем проверить, как полученная нами F-статистика соотносится с соответствующим ему F-распределением. Зная F-статистику и соответствующее число степеней свободы, мы можем рассчитать соответствующий p-уровень значимости, который скажет нам какая вероятность получить такое или еще более выраженное отличие между несколькими средними, если на деле верна нулевая гипотеза:

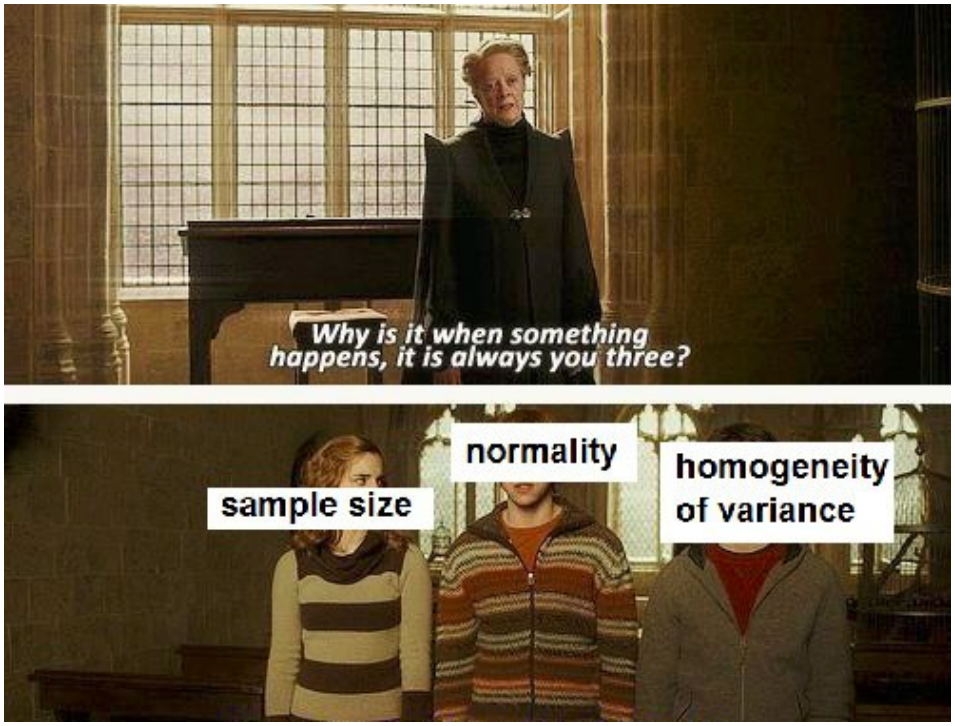


ВНИМАНИЕ: Дисперсионный анализ проверяет только общую гипотезу о том, что где-то есть различия в выборочных средних. Где они именно – нужно проверять отдельно, и об этом позднее.

> Требования к дисперсионному анализу

Как и для t-критерия:

1. Дисперсии внутри наших групп должны быть примерно одинаковы (требование гомогенности дисперсий). Проверить можно с помощью критерия Левена и критерия Бартлетта
2. Если объемы групп недостаточно большие (меньше 30) и не совпадают по размерам, то важно соблюдать требование о **нормальности распределения** выборок.



Как сделать в python?

```
from scipy import stats

stats.f_oneway(a, b, c) # a, b, c - переменные с данными трёх групп
```

Результат выполнения функции может выглядеть так:

```
F_onewayResult(statistic=85.99631112614011, pvalue=3.4370045810218544e-30)
```

Первое число — F-значение, второе - p-значение. Так как p-значение меньше 0.05, то мы отклоняем нулевую гипотезу и делаем вывод, что среднее хотя одной из групп значимо отличается от средних в других группах.

Другой вариант:

```
import statsmodels.formula.api as smf

model = smf.ols(formula = "зависимая_переменная ~ независимая_переменная)", data = данные).fit()
anova_lm(model)
```

	df	sum_sq	mean_sq	F	PR(>F)
C(button)	2.0	201960.286667	100980.143333	85.996311	3.437005e-30
Residual	297.0	348748.710000	1174.238081	NaN	NaN

Здесь p-value отмечено как **PR(>F)**, остальное связано либо с суммами квадратов, либо со степенями свободы.

Третий вариант:

```
import pingouin as pg

pg.anova(data=данные, dv="зависимая_переменная", between="независимая_переменная")
```

	Source	ddof1	ddof2	F	p-unc	np2
0	button	2	297	85.996311	3.437005e-30	0.366728

Важный новый элемент - **размер эффекта**. Он указывает, какой процент всей изменчивости в данных объясняется нашим экспериментальным воздействием.

Ну и если группы не обладают равной дисперсией, то можно сделать **дисперсионный анализ Уэлча**:

```
pg.welch_anova(data=данные, dv="зависимая_переменная", between="независимая_переменная")
```

Эти функции тоже используют научную нотацию.

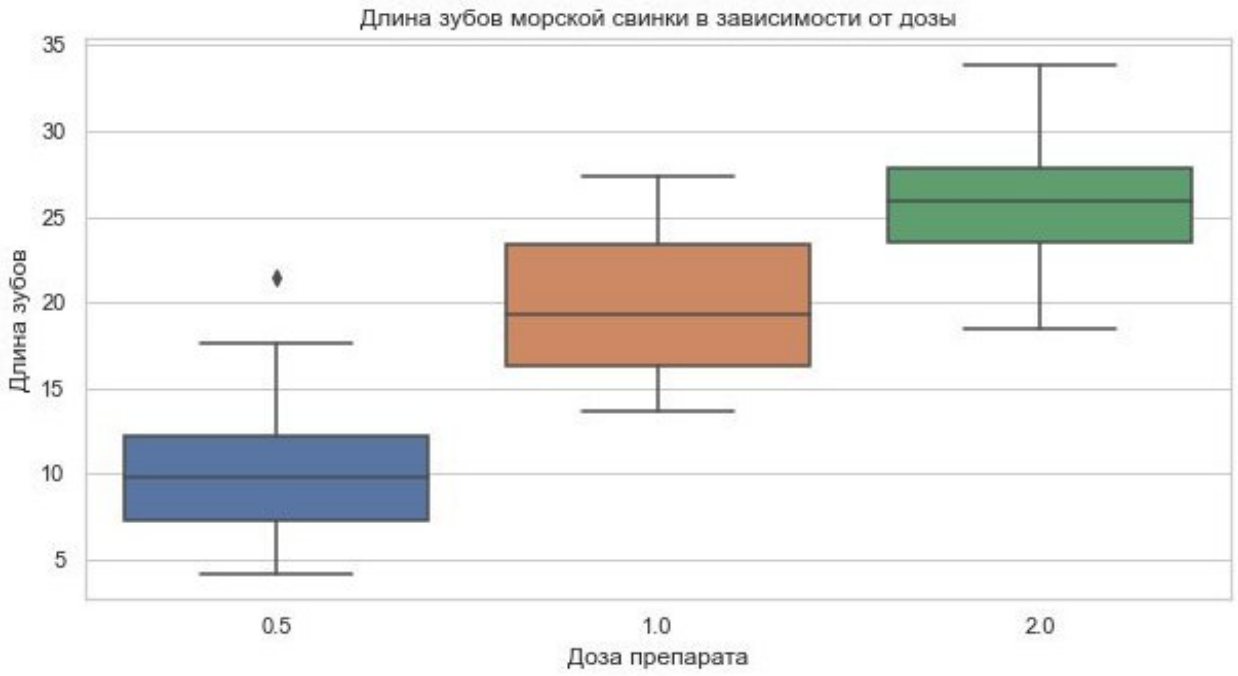
> Построение графиков

Требования к графикам такие же, как и для t-критерия:

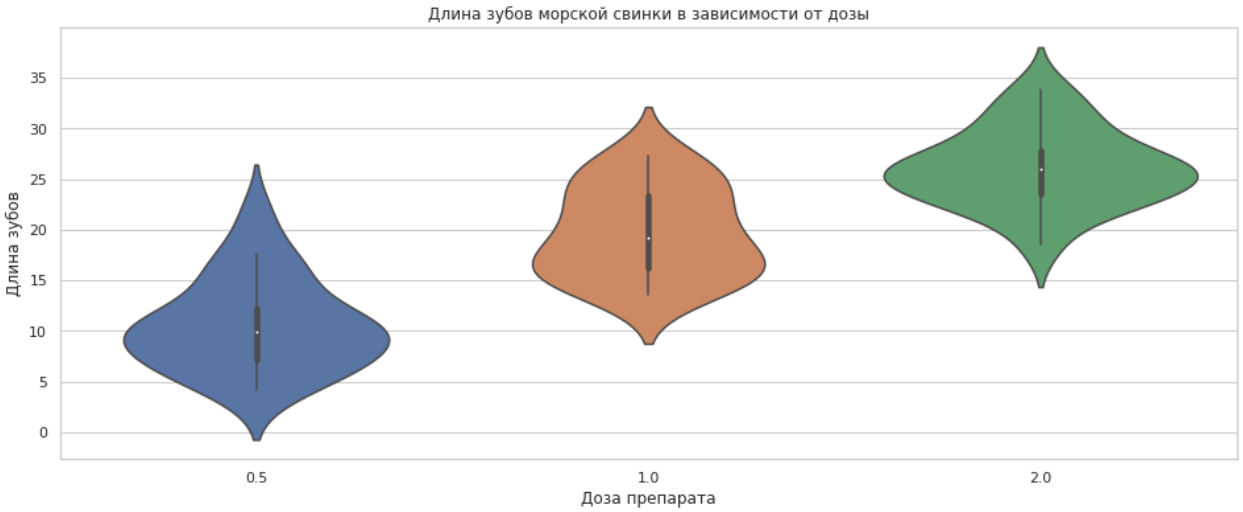
- Указать название графика
- Подписать оси

- Указывать меру изменчивости данных (напр. доверительные интервалы, либо сразу использовать боксплот/скрипку)

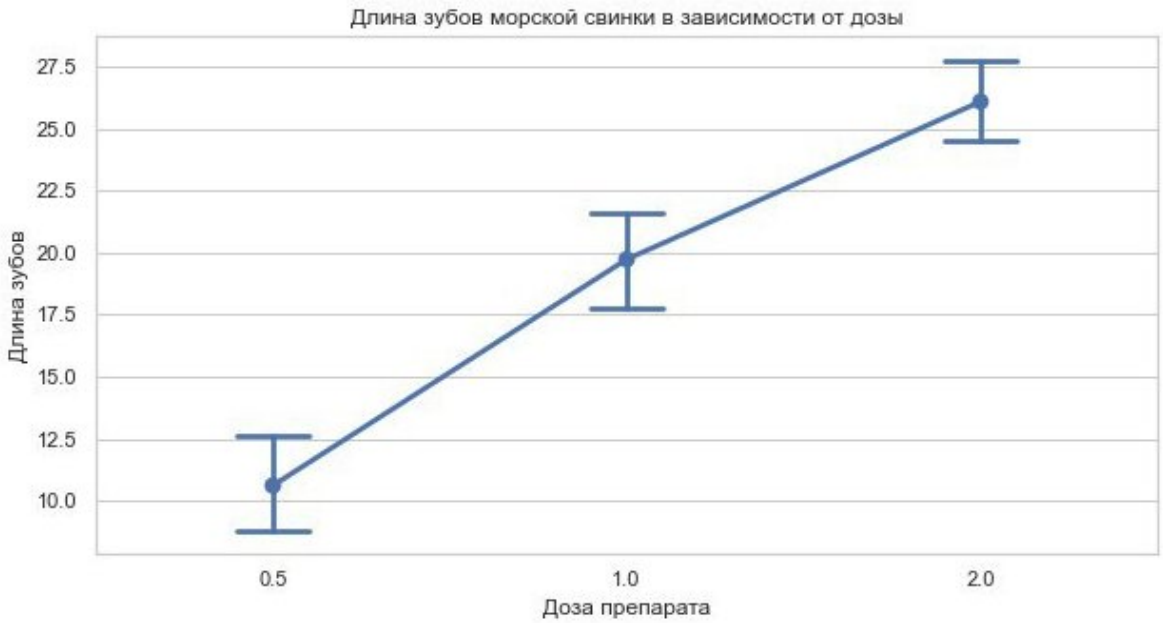
```
sns.boxplot(x = 'dose', y = 'len', data = teeth)
plt.title('Длина зубов морской свинки в зависимости от дозы')
plt.xlabel('Доза препарата')
plt.ylabel('Длина зубов')
```




```
sns.violinplot(x = 'dose', y = 'len', data = teeth)
plt.title('Длина зубов морской свинки в зависимости от дозы')
plt.xlabel('Доза препарата')
plt.ylabel('Длина зубов')
```




```
sns.pointplot(x = 'dose', y = 'len', data = teeth, capsize = .2)
plt.title('Длина зубов морской свинки в зависимости от дозы')
plt.xlabel('Доза препарата')
plt.ylabel('Длина зубов')
```



> Множественные сравнения

Когда нам нужно сравнивать количество групп больше двух, мы неизбежно сталкиваемся с проблемой множественных сравнений. Это неприятный статистический эффект, когда из-за тестирования сразу нескольких гипотез разом вероятность ошибки

I рода искусственно возрастает. Поэтому при увеличении количества групп необходимость сравнивать попарно каждую из них приводит к повышенной вероятности увидеть значимые различия там, где их нет.

Оценить масштаб бедствия можно с помощью следующей формулы:

$$P = 1 - (1 - \alpha)^m$$

Здесь *α* – уровень значимости, m – количество тестируемых гипотез, а P – итоговая вероятность допущения ошибки I рода. Например, при *α* = 0.05 и m = 3 вероятность допустить ошибку I рода уже становится 0.143%. Можете проверить, при каком m эта вероятность приближается к единице.

Посчитать m, в свою очередь, можно по следующей формуле:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Здесь n - количество групп, k - размер комбинации (так как у нас попарные сравнения, то k = 2), ! - факториал числа.

Также можно использовать функцию из модуля SciPy.

> Поправка Бонферрони

В рамках этого метода нам нужно разделить уровень значимости на количество попарных сравнений, и считать различия значимыми только в том случае, если p-значение меньше нового порога. Например, при сравнении 8 групп мы проводим 28 сравнений, и новый уровень значимости становится

$$\frac{0.05}{28} = 0.0018280.05 = 0.0018$$

Проблема метода: при достаточно большом количестве попарных сравнений метод становится слишком консервативным: растет вероятность ошибки II рода, и мы уже не можем отвергнуть нулевую гипотезу даже при выраженных различиях между группами.

Другие варианты поправки на множественные сравнения (с их математической подложкой) можно посмотреть вот тут. Также рекомендуем заглянуть в документацию `pingouin` и посмотреть аргумент `padjust` вот этой функции.

Ещё два варианта мы рассмотрим в следующем степе.

> Критерий Тьюки

Фактически рассчитывается по той же формуле, что и t-критерий, но несколько иначе рассчитывается стандартная ошибка – в результате критерий Тьюки более консервативен, чем обычный t-критерий, но гораздо менее консервативен по сравнению с поправкой Бонферрони.

Критерий Тьюки в Python:

```
from statsmodels.stats.multicomp import (pairwise_tukeyhsd,
                                         MultiComparison)

print(pairwise_tukeyhsd(столбец_с_данными, столбец_с_обозначениями_групп))

#или

MultiComp = MultiComparison(столбец_с_данными, столбец_с_обозначениями_групп)

print(MultiComp.tukeyhsd().summary())
```

В обоих случаях появится похожая табличка:

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
blue	green	-48.01	0.001	-59.4257	-36.5943	True
blue	red	-60.07	0.001	-71.4857	-48.6543	True
green	red	-12.06	0.0356	-23.4757	-0.6443	True

- *group1* и *group2* - названия групп, которые сравниваются в рамках теста
- *meandiff* - разница между значением 2 группы и значением 1 группы
- *p-adj* - скорректированный порог значимости
- *lower* и *upper* - нижняя и верхняя границы доверительного интервала различий в средних
- *reject* - отвергается нулевая гипотеза или нет

Как видно по последней колонке, все нулевые гипотезы были отклонены, поэтому мы делаем вывод, что средние всех трёх групп значимо различаются.

Похожий результат будет, если мы используем функцию из `pingouin`:

```
pg.pairwise_tukey(data=данные, dv="зависимая_переменная", between="независимая_переменная")
```

	A	B	mean(A)	mean(B)	diff	se	T	p-tukey	hedges
0	blue	green	210.03	162.02	48.01	4.846108	9.906919	0.001000	1.395736
1	blue	red	210.03	149.96	60.07	4.846108	12.395514	0.001000	1.746342
2	green	red	162.02	149.96	12.06	4.846108	2.488595	0.035581	0.350606

Дополнительный элемент - стандартизованная разница в средних под названием Hedges G.

Если наши группы имеют разную дисперсию, то применяется **критерий Геймса-Хоувелла**:

```
pg.pairwise_gameshowell(data=данные, dv="зависимая_переменная", between="независимая_переменная")
```

> Многофакторный ANOVA

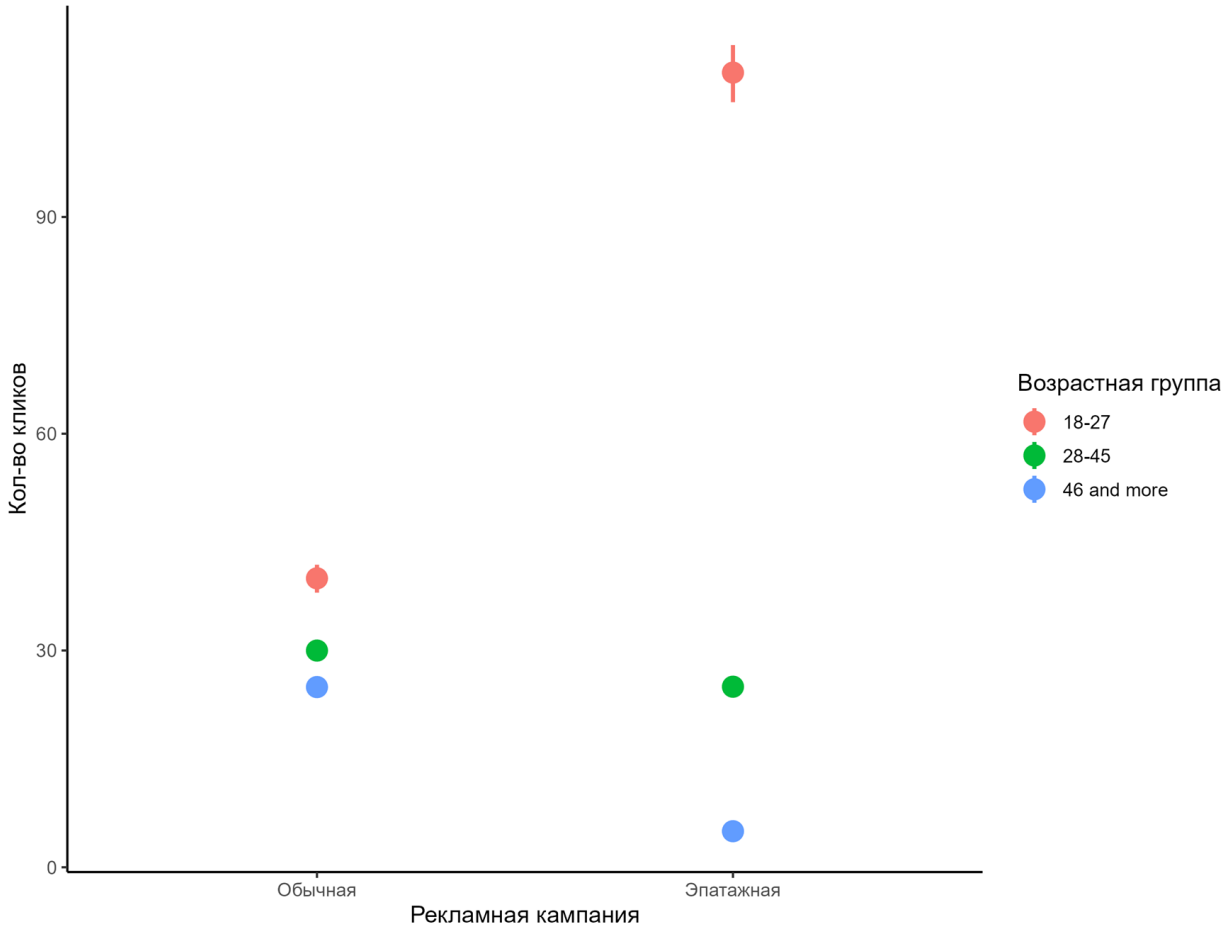
Часто бывает и такое, что одни и те же данные можно разделить на группы по разным основаниям - иначе говоря, мы можем проверять эффект нескольких номинативных переменных на зависимую переменную. Например, двухфакторный дисперсионный анализ проверяет влияние двух номинативных переменных (факторов) – в рамках лекции это было влияние возраста (молодые-пожилые) и дозировки лекарства (низкая-высокая) на экспрессию генов. Можно добавлять и больше факторов.

Таким образом, общая сумма квадратов теперь складывается из большего количества величин:

$$SST = SSW + SSB_A + SSB_B + SSB_A * SSB_B$$

Здесь буквами A и B обозначены межгрупповые суммы квадратов для соответствующих факторов, а $SSB_A * SSB_B$ обозначает их взаимодействие.

Взаимодействие факторов – это случай, когда связь зависимой переменной от одного фактора связана со значениями другого (например, противоположный эффект фактора А для групп, которые сформировались по фактору В).



Внимание:

1. Как и с однофакторным дисперсионным анализом, мы можем констатировать только общую значимость конкретного фактора, но не можем сказать, какие именно группы различаются и каким образом. Для этого нам опять нужны парные сравнения с поправками.
2. Полученные данные, какие бы они ни были, не означают однозначной каузальной связи между переменными – установить её позволяет только правильно организованный эксперимент.

> Требования к многофакторному дисперсионному анализу

Всё те же самые:

1. Дисперсии внутри наших групп должны быть примерно одинаковы (требование гомогенности дисперсий). Проверить можно с помощью критерия Левена и критерия Бартлетта
2. Если объем выборки недостаточно большой (меньше 30), то важно соблюдать требование о нормальности распределения двух выборок.

```
from statsmodels.formula.api import ols
from statsmodels.stats.anova import anova_lm

formula = 'зависимая_переменная ~ фактор1 + фактор2 + фактор1:фактор2'
model = ols(formula, data).fit()
aov_table = anova_lm(model, typ=2)
```

Значком “:” обозначается взаимодействие независимых переменных (НП) - то есть того, что влияет на исследуемую нами величину, зависимую переменную (ЗП).

Результат функции может выглядеть так:

	df	sum_sq	mean_sq	F	PR(>F)
C(ads)	1.0	33735.001667	33735.001667	336.722432	6.446773e-60
C(age_group)	2.0	400495.163333	200247.581667	1998.750536	1.636934e-264
C(ads):C(age_group)	2.0	232685.043333	116342.521667	1161.260853	5.618312e-206
Residual	594.0	59510.710000	100.186380	NaN	NaN

По рядам: первая НП, вторая НП, взаимодействие НП и остатки.

По колонкам:

- *sum_sq* - сумма квадратов
- *df* - степени свободы
- *F* - F-значение
- *PR (>F)* - p-значение

Как видно по последней колонке, значим как эффект обоих факторов, так и их взаимодействие.

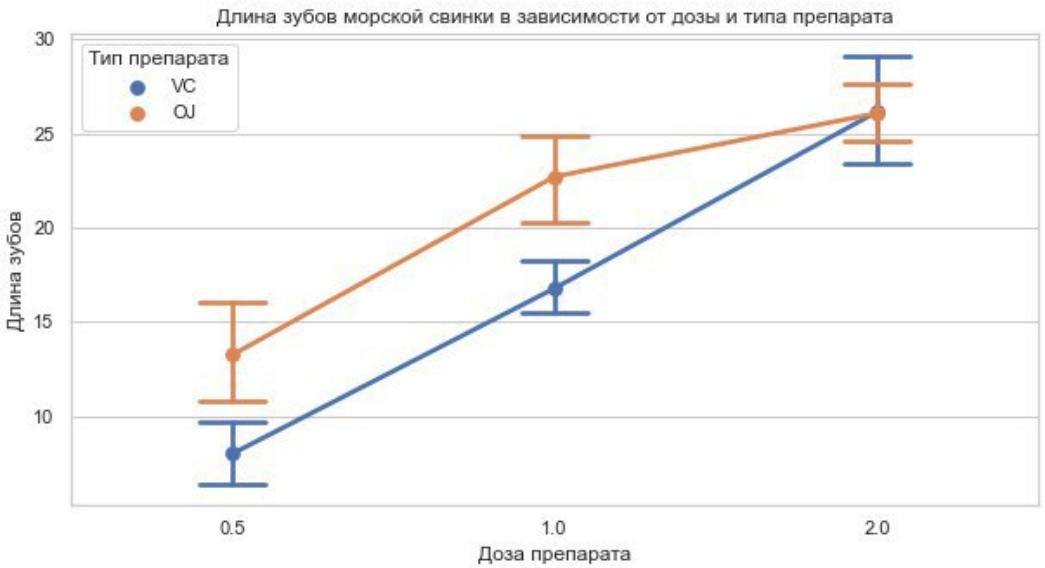
Похожего результата можно добиться и через pingouin:

```
pg.anova(data=данные, dv="зависимая_переменная", between=["фактор1", "фактор2"])
```

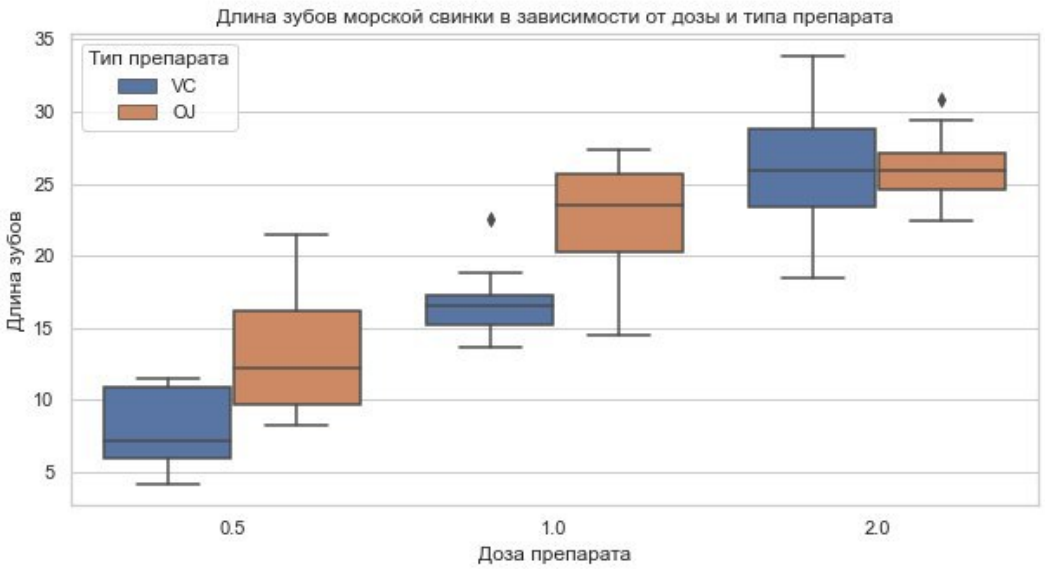
	Source	SS	DF	MS	F	p-unc	np2
0	ads	33735.001667	1	33735.001667	336.722432	6.446773e-60	0.361786
1	age_group	400495.163333	2	200247.581667	1998.750536	1.636934e-264	0.870631
2	ads * age_group	232685.043333	2	116342.521667	1161.260853	5.618312e-206	0.796333
3	Residual	59510.710000	594	100.186380	NaN	NaN	NaN

> Как рисовать

```
sns.pointplot(x = 'dose', y = 'len', hue = 'supp', data = teeth, capsize = .2)
plt.title('Длина зубов морской свинки в зависимости от дозы и типа препарата')
plt.xlabel('Доза препарата')
plt.ylabel('Длина зубов')
plt.legend(title = 'Тип препарата')
```

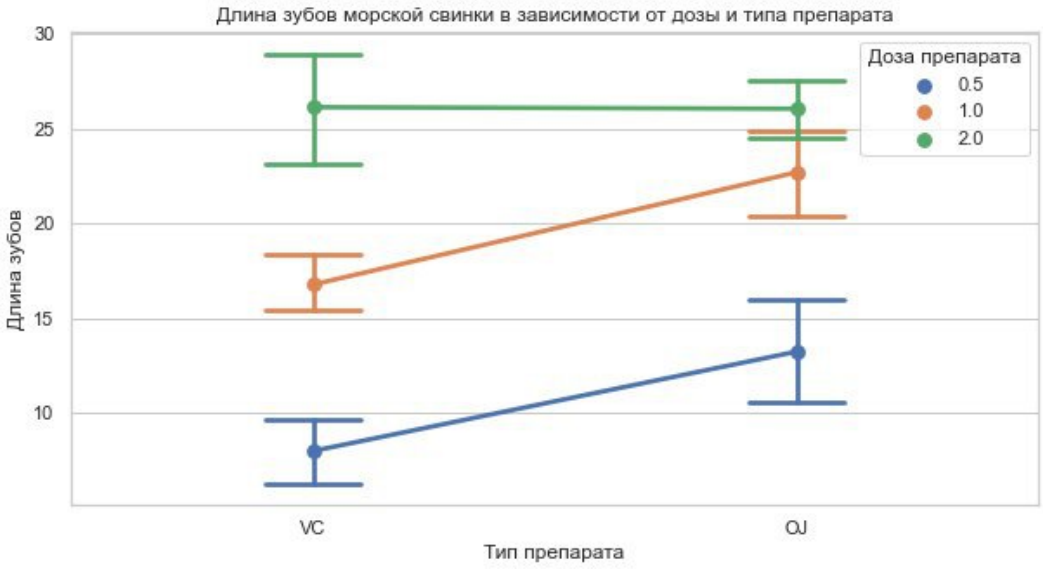


```
sns.boxplot(x = 'dose', y = 'len', hue = 'supp', data = teeth)
plt.title('Длина зубов морской свинки в зависимости от дозы и типа препарата')
plt.xlabel('Доза препарата')
plt.ylabel('Длина зубов')
plt.legend(title = 'Тип препарата')
```



Внимательно выбирайте, какая переменная пойдёт на ось x!

```
sns.pointplot(x = 'supp', y = 'len', hue = 'dose', data = teeth, capsize = .2)
plt.title('Длина зубов морской свинки в зависимости от дозы и типа препарата')
plt.xlabel('Тип препарата')
plt.ylabel('Длина зубов')
plt.legend(title = 'Доза препарата')
```



> Дополнительно

1. Как вы могли видеть, помимо различий на оригинальной шкале измерений (на сколько больше просмотров набрали эти посты?), мы можем смотреть на стандартизированные различия. Например, это **d Коэна** и его близкий родственник **g Хеджеса**, которые по сути отражают различия в средних, скорректированные на стандартные отклонения. Для дисперсионного анализа можно встретить величины **эта-квадрат**, **частичная эта-квадрат** и **омега-квадрат** - все измеряют вклад факторов, учтённых в нашем эксперименте. Подробнее об их расчёте и использовании - вот тут.
2. Вы также могли заметить величину под названием **байес-фактор**. Он часто используется как замена p-value и обозначает степень убеждённости в той или иной гипотезе. Подробнее вот тут (осторожно, R).
3. Также стоит ознакомиться с **типами сумм квадратов** - способами того, как учитывать факторы в многофакторном дисперсионном анализе. Немного подробнее тут.