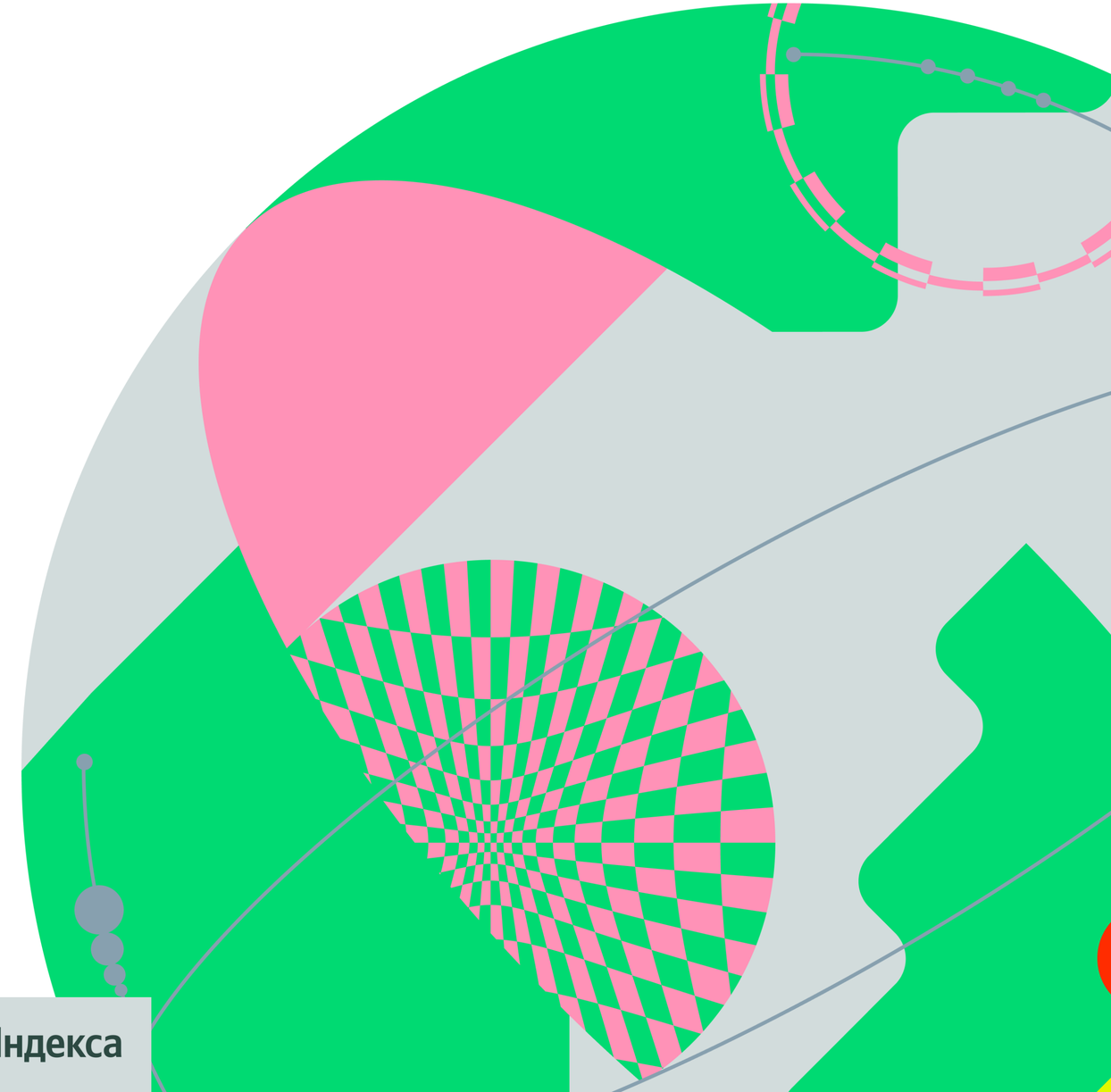
Проверка гипотез



Академия Яндекса

Критерий доверия к выборочным расчётам

С выборками в прошлый раз разобрались. Мы увидели, что для исследований достаточно взять выборку из всей генеральной совокупности. Это облегчает/удешевляет многие исследования.

А что после этого? На сколько можно доверять выводам на основе выборочных данных? Можно ли, анализируя их, распространять выводы на всю генеральную совокупность? Ведь все наши расчёты носят приблизительный характер. Вся статистика – это вероятностные расчёты. Вот мы вычислили в прошлый раз, что уровень корреляции между временем ремонта и типом корабля равен 0.63. Но ведь у нас всего около 70 случаев ремонта было.

Можно ли вывод о связи факторов распространять на все возможные ремонты всех космолётов в будущем?

Т.е. выявленную связь факторов на выборке предположить, и на всей генеральной совокупности

Гипотезы

Предположения о данных называются **гипотезами**.
Важно: подтвердить гипотезу на основе экспериментальных данных нельзя — это фундаментальное ограничение. Всё, что мы можем сделать по итогам проверки, — это отвергнуть гипотезу или нет.

Иными словами, при условии, что гипотеза верна, данные могут лишь не противоречить ей или, наоборот, показывать очень маловероятные результаты. Но и в том, и в другом случае нет оснований утверждать, что выдвинутая гипотеза *доказана*.

Биологи изучают маршруты перелётов птиц редкого вида. Гипотеза состоит в том, что они всегда летят вдоль рек. Учёные случайным образом отловили несколько десятков птиц, прикрепили датчики и отпустили. Все отслеженные маршруты оказались пролегающими вдоль рек.

Доказывает ли это гипотезу? Нет, потому что какие-то птицы, возможно, летают по-другому, а отследить маршруты всех птиц даже редкого вида чаще всего невозможно. Однако уверенность в том, что гипотеза верна, повысилась. Если несколько исследований покажут, что маршруты птиц проходят вдоль рек, а обратное не будет подтверждено экспериментально, это будет признано научным фактом.

Ни в одном факте человечество не уверено 100%. Просто не нашлось экспериментальных данных, которые бы его опровергли. Когда-нибудь могут найтись — и теорию придётся пересмотреть.

Исключение одно: абстрактные математические утверждения — теоремы. Ведь их строго выводят из аксиом — предпосылок, которые мы принимаем за истину.

Чтобы делать такие выводы, опираясь на конкретные числа, а не на интуицию, далее научимся пользоваться алгоритмом работы с численными выборочными данными.

Гипотезы

В общем случае, статистической гипотезой (гипотезой) называется любое утверждение об изучаемом законе распределения или характеристиках случайных величин.

Пример статистических гипотез:
1.Генеральная совокупность распределена по нормальному закону.
2.Дисперсии двух нормально распределенных совокупностей равны между собой.

Проверка гипотез начинается с формулировки **нулевой гипотезы** H0 (от англ. hypothesis — «гипотеза»).

Нулевая гипотеза *(H0)* — предположение о том, что между параметрами генеральных совокупностей нет различий, то есть эти различия носят не систематический, а случайный характер. (часто формулируют именно так, чтобы использовать знак равенства)

Пример1. Нулевая гипотеза записывается следующим образом:
H_0: $\mu_1=\mu_2$ (нулевая гипотеза заключается в том, что генеральное среднее одной совокупности равно генеральному среднему другой совокупности).

Альтернативная гипотеза (H1) – предположение о том, что между параметрами генеральных совокупностей есть достоверные различия. Это утверждение, которое принимается верным, если отвергается H0.

Пример 2. Альтернативные гипотезы записываются следующим образом:
H_1: $\mu_1 \neq \mu_2$ (нулевая гипотеза заключается в том, что генеральное среднее одной совокупности не равно генеральному среднему другой совокупности).

Гипотезы

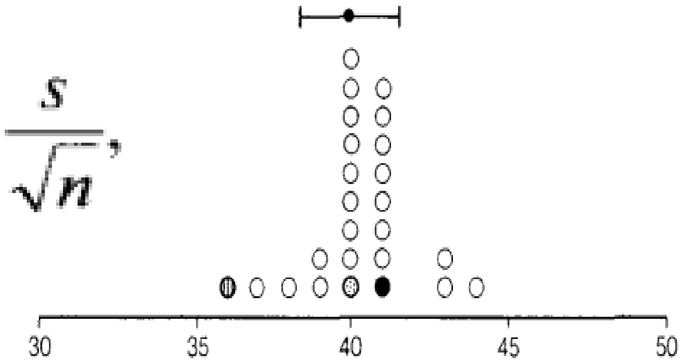
Когда сформулировали нулевую и альтернативную гипотезы, нужно на основе данных сделать выбор: отвергнуть нулевую гипотезу в пользу альтернативной или нет.

Формулировки гипотез касаются какого-либо параметра — чаще всего это среднее (далее мы будем это рассматривать). Параметр оценивается по выборке, и мы получаем наблюдаемое на выборке значение.

Затем нужно решить, отвергнуть нулевую гипотезу в пользу альтернативной или нет. Как это сделать? **Нужно понять, насколько вероятно получить наблюдаемое значение при условии, что нулевая гипотеза верна.**

Если вероятность большая, то отвергать нулевую гипотезу вряд ли стоит. Если маленькая — получается, что при верной нулевой гипотезе наблюдаемое значение маловероятно. Это повод отвергнуть нулевую гипотезу в пользу альтернативной.

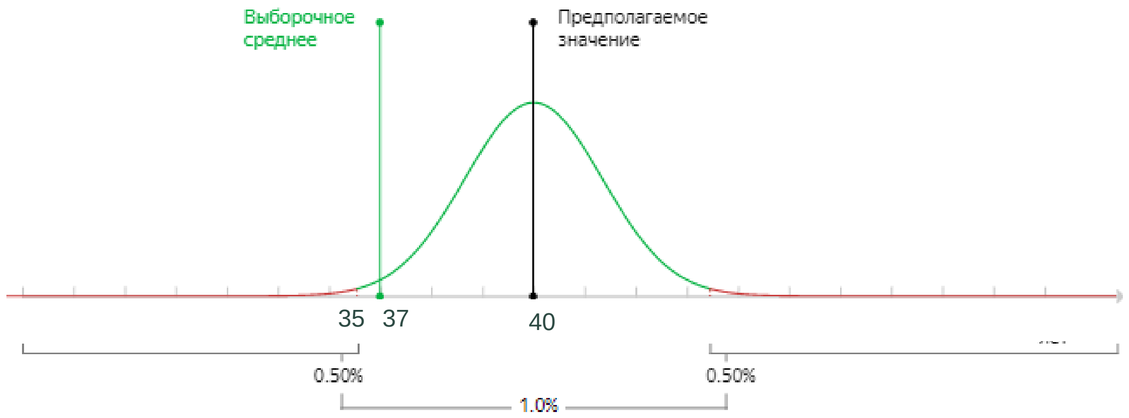
$$s_{\bar{X}} = \frac{s}{\sqrt{n}},$$

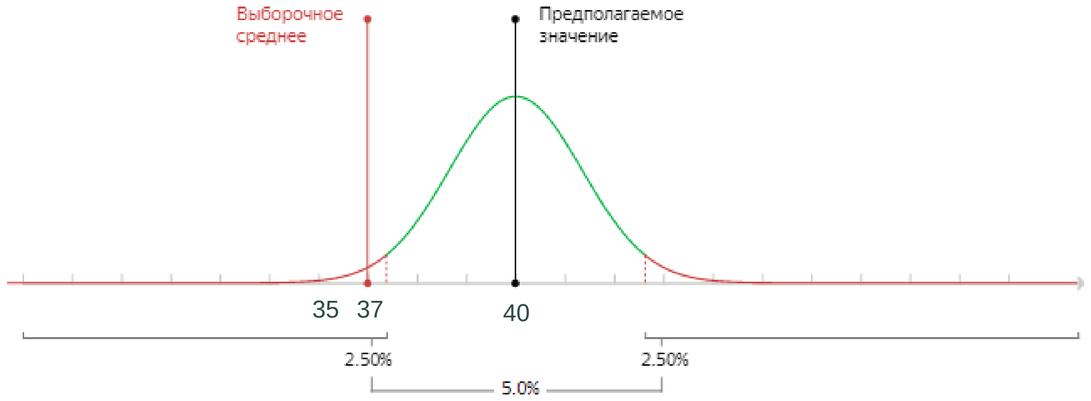


Предположим, выборочное среднее равно 40.
Нулевая гипотеза: средний рост всех жителей равен 40
Альтернативная гипотеза: средний рост всех жителей НЕ равен 40

Как определить, где ещё не стоит отвергать нулевую гипотезу, а где уже пора? Где проходит граница? Чтобы это определить, надо выбрать уровень статистической значимости.

Гипотезы





Нулевая гипотеза отвергается, если вероятность получить среднее, рассчитанное по выборочным данным, *слишком мала*. Уровень значимости численно определяет, где проходит граница.

Уровнем значимости задаётся вероятность попасть слишком далеко от центра распределения.
Вероятность попасть в тот или иной интервал равна площади графика над этим интервалом. Нас интересуют значения, далёкие от предполагаемого, — то есть мы отбрасываем хвосты, площадь которых равна заданному уровню статистической значимости.

P-уровень

Мерой доверия к выводам будет служить так называемый p-уровень. Это число от 0 до 1. Глядя на него и задаваясь уровнем доверия, мы сможем делать выводы.

Давайте обсудим на нашем примере.

Как говорили, все статистические расчеты носят приблизительный/вероятностный характер. Уровень этой приблизительности и определяет «p». Его ещё называют - уровень значимости. Он записывается в виде десятичных дробей, например, 0.023 или 0.965.

Если умножить такое число на 100, то получим показатель p в процентах: 2.3% и 96.5%. Эти проценты отражают вероятность ошибочности нашего предположения о взаимосвязи, например, между «временем ремонта» и «пролётом космолёта».

То есть, коэффициент корреляции 0.63 между «временем ремонта» и «пролётом космолёта» получен при уровне статистической значимости 0.05 или вероятности ошибки 5%.

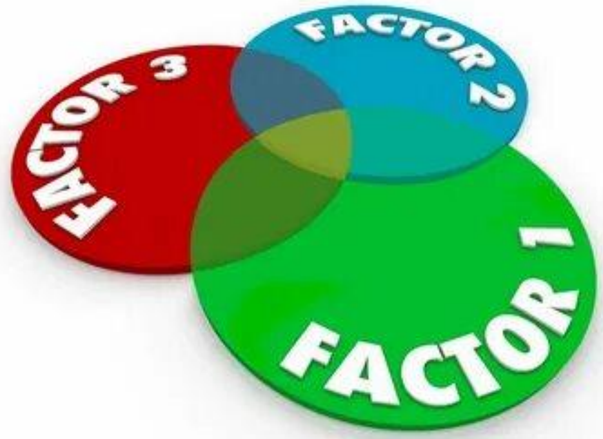
P-уровень

Мы говорили, что выявленная нами корреляция означает, что в нашей выборке наблюдается такая закономерность: чем выше один фактор, тем выше другой. Но так как в статистике все приблизительно, то, утверждая это, мы допускаем, что можем ошибиться, причем вероятность ошибки 5%. То есть, сделав 20 таких сравнений мы можем 1 раз ошибиться с выводом.

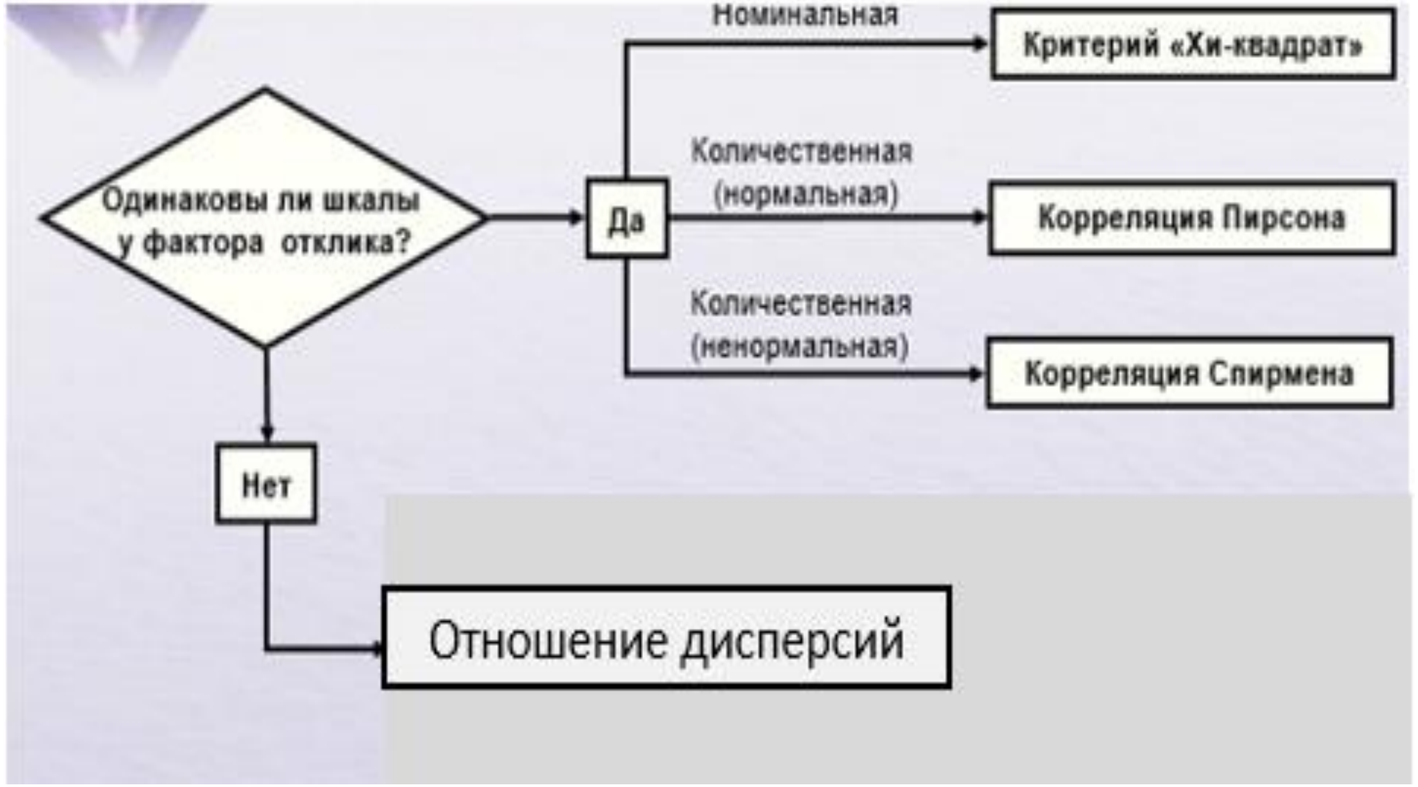
Много это или мало?

Очень интересный вопрос. Всё зависит от конкретной области задач. В промышленной аналитик принято использовать 0,05 (5 %). В медицине (при выводе лекарства, например, такой уровень будет означить, что 1 человеку из 20 лекарства не подойдёт). Поэтому в медицине всё строже – там уровни 0.01, а то и 0.001. В маркетинговых и социологических исследованиях – 0.05. Ну тут жёстких критериев нет.

Взаимосвязь факторов



Мы много говорили, а вы много решали задач на нахождения меры связи между факторами. Причём важно, в какой шкале эти факторы представлены.



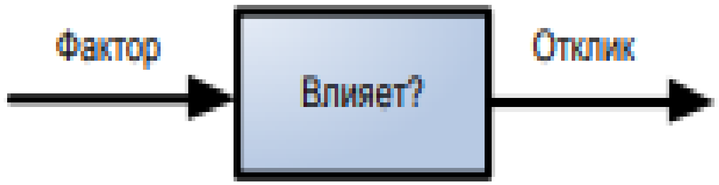
Начало гипотез

Для двух одинаковых количественных шкал (количественная – количественная, например) – как-то легко понять вывод: чем больше один, тем больше другой – всё интуитивно понятно. Вспомните «время ремонта» и «пролёт с последнего ремонта» из темы_3.

А вот если анализируем разные шкалы **«Тип_космолёта» и «кол-во_пробоин». Когда мы изучали его, то сделали вывод, что они не связаны между собой. И что? Как это использовать на практике? Не прозрачно.**

А что если поставить вопрос так: У какого типа космолёта количество пробоин в среднем больше/меньше?

И вот, чтобы отвечать на такого типа вопросы, нам надо окунуться в загадочный, местами сложнопонятный, но интересный мир проверки гипотез. С позиции науки там много нюансов и умопомрачительных названий. Если будет время окунуться туда – не пожалеете. Но это потом. А сегодня мы рассмотрим эту тему с позиции практики. Как бы её не представляли в книгах – тема проста. Суть проблематики можно представить вот так:



Начало гипотез

Мы отвечаем на вопрос: «Влияет ли фактор на отклик?». Как влияет «кол-во пробоин» на «тип космолёта» или лучше наоборот спросить? А не важно. Выбор фактора и отклика осуществляется на основании представления о природе исследуемой проблемы, интуиции специалиста или опыта аналогичных исследований. Поэтому, общая схема может быть изображена следующим образом:

А с помощью p-уровня мы можем говорить о распространении вывода о связи факторов на генеральную совокупность

В результате расчетов по любому из перечисленных методов на выходе получают показатель, который называется **«статистическая значимость»**

Статистическая значимость (p- уровень) – это вероятность того, что фактор не оказывает влияния на зависимую величину

- Статистическая значимость – может принимать значения от 0 до 1
- Чем меньше статистическая значимость, тем больше вероятность того, что фактор оказывает влияние
- Статистическая значимость – это степень уверенности в полученных выводах

**Если статистическая значимость меньше 0,05,
то можно утверждать с высокой степенью уверенности,
что влияние фактора есть**

Нет оснований, чтобы принять нулевую гипотезу, поэтому принимаем альтернативную.

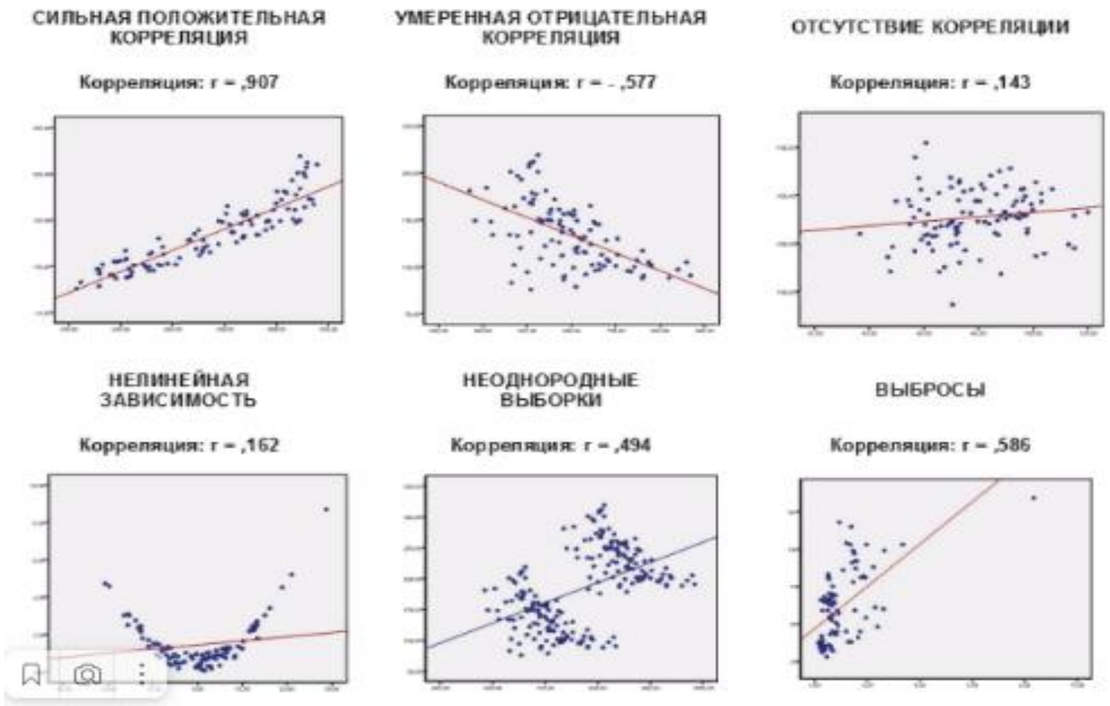
Как решаются задачи

ПОСТАНОВКИ ЗАДАЧИ:

- Влияет ли «кол-во пробоин» на «тип космолёта»
- Повлияла ли проведённая акция на объём продаж
- Одинаково ли относятся сотрудники разных отделов к руководству компании
- Изменился ли уровень квалификации персонала за год работы на предприятии
- ….

Количественная - количественная

Аналитический метод: **Корреляция Пирсона/Спирмена**
Графический: **Диаграмма рассеяния.**



Шкала Чеддока

Количественная мера тесноты связи	Качественная характеристика тесноты связи
$0.1 < r_{xy} < 0.3$	очень слабая
$0.3 < r_{xy} < 0.5$	слабая
$0.5 < r_{xy} < 0.7$	средняя
$0.7 < r_{xy} < 0.9$	высокая
$0.9 < r_{xy} < 1$	очень высокая

Количественная - количественная

Аналитический метод: Корреляция Пирсона/Спирмена
Графический: Диаграмма рассеяния.

Как это можно посмотреть?
Из прошлых встреч давайте возьмём знакомый файл с промышленностью. И посмотрим корреляцию:

	вес	оператор_линии	ингридиент_1	номер_смены	новер_конвейера	ингридиент_2	брак
0	61.7	2	250.9	2	1	6.3	Годен
1	58.6	2	275.7	1	2	6.7	Брак
2	53.6	2	280.6	2	1	7.5	Брак

АНАЛИТИЧЕСКИЙ МЕТОД

ГРАФИЧЕСКИЙ МЕТОД

Количественная - количественная

Аналитический метод: Корреляция Пирсона/Спирмена
Графический: Диаграмма рассеяния.

АНАЛИТИЧЕСКИЙ МЕТОД

```
df1[num].corr(method='pearson').style.background_gradient(cmap='cividis')
```

	вес	ингридиент_1	ингридиент_2
вес	1.000000	-0.087235	0.669274
ингридиент_1	-0.087235	1.000000	0.005579
ингридиент_2	0.669274	0.005579	1.000000

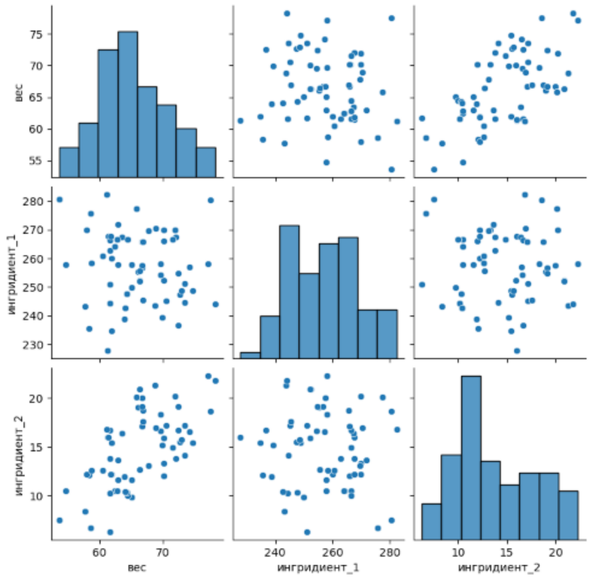
```
from scipy.stats.stats import pearsonr
pearsonr(df1['вес'], df1['ингридиент_2'])
```

```
PearsonRResult(statistic=0.6692739116677272, pvalue=3.6592199820776735e-14)
```

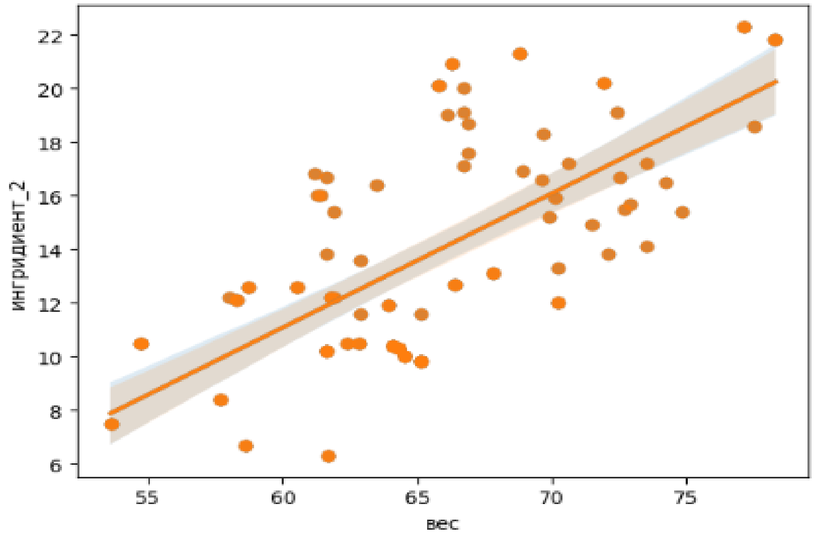
ГРАФИЧЕСКИЙ МЕТОД

```
import seaborn as sns
import matplotlib.pyplot as plt

sns.pairplot(df1[num])
plt.show()
```




```
sns.regplot(data = df1, x = 'вес', y = 'ингридиент_2')
plt.show()
```



Академия Яндекса

Количественная - количественная

Аналитический метод: Корреляция Пирсона/Спирмена
Графический: Диаграмма рассеяния.

	вес	оператор_линии	ингридиент_1	номер_смены	новер_конвейера	ингридиент_2	брак
0	61.7	2	250.9	2	1	6.3	Годен
1	58.6	2	275.7	1	2	6.7	Брак
2	53.6	2	280.6	2	1	7.5	Брак

```
df1[num].corr(method='pearson').style.background_gradient(cmap='cividis')
```

	вес	ингридиент_1	ингридиент_2
вес	1.000000	-0.087235	0.669274
ингридиент_1	-0.087235	1.000000	0.005579
ингридиент_2	0.669274	0.005579	1.000000

```
from scipy.stats.stats import pearsonr
pearsonr(df1['вес'], df1['ингридиент_2'])
```

```
PearsonRResult(statistic=0.6692739116677272, pvalue=3.6592199820776735e-14)
```

Вот что видим по итогу анализа взаимосвязи веса и ингрдиента_2:

- корреляция равна 0.66, доверительный интервал – не большой. (это хорошо!!!)
- этому выводу можно доверять, т.к. p-уровень менее 0.05
- выбросов и иных аномалий, которые бы искажали управленческий вывод, не наблюдается.

Академия Яндекса

номинальная – номинальная

Аналитический метод: хи-квадрат
Графический: круговая диаграмма

Давайте проверим, есть ли связь между браком и номером конвейера. Или по-другому: Влияет ли номер конвейера на уровень брака? Данные такие:

```
df1
```

	брак	новер_конвейера
0	Годен	1
1	Брак	2
2	Брак	1
3	Годен	1
4	Годен	1
...	...	...
94	Годен	1
95	Годен	1

Гипотезы такие:

- H0: категориальные переменные независимы;
- H1: категориальные переменные связаны между собой.

номинальная – номинальная

Аналитический метод: хи-квадрат
Графический: круговая диаграмма

АНАЛИТИЧЕСКИЙ МЕТОД

Помните, при расчёте корреляций отношений, мы подсчитывали так называемые частоты. Т.е. сколько брака/не брака по каждой линии (таблица сопряжённости). В питоне существует команда для этого:

```
contingency_table = pd.crosstab(df1['новер_конвейера'], df1['брак'])
```

```
contingency_table
```

новер_конвейера	брак	Брак	Годен
1		17	55
2		16	11

А потом сам расчёт:

```
: from scipy.stats import chi2_contingency
  chi2_contingency(contingency_table)#[0:3]
```

```
: Chi2ContingencyResult(statistic=9.682291666666668, pvalue=0.0018605242881465018, dof=1, expected_freq=array([[24., 48.],
       [ 9., 18.]]))
```

Что видим?

p-уровень меньше 0.05. Это значит, что мы отвергаем нулевую гипотезу и принимаем альтернативную. Иными словами: с уровнем значимости 0.002 (примерно) может утверждать, что на линиях 1 и 2 уровень брака статистически значимо различается.

номинальная – номинальная

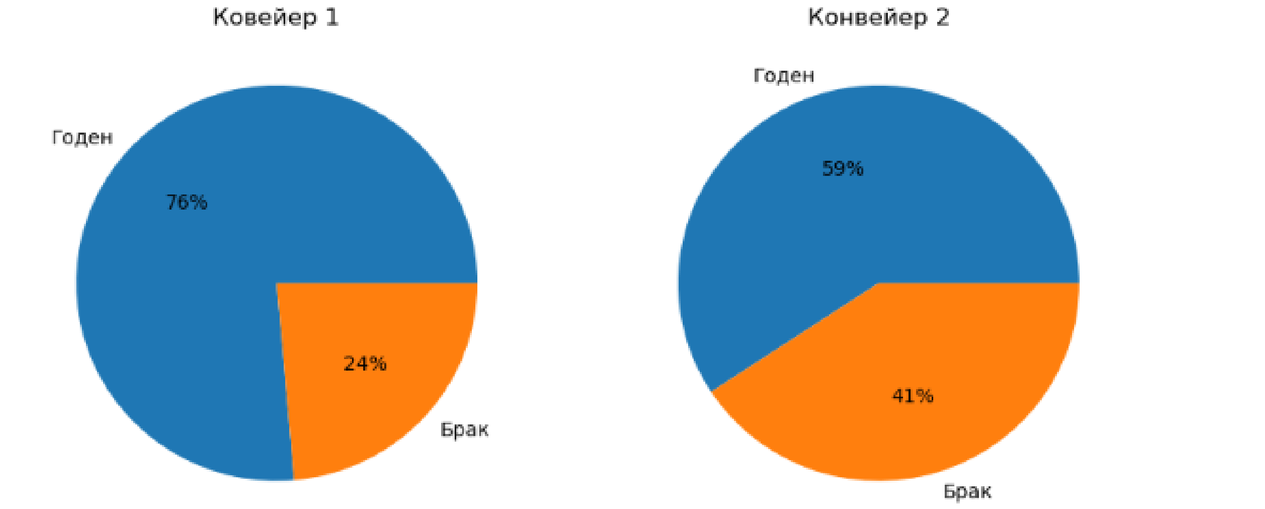
Аналитический метод: хи-квадрат
Графический: круговая диаграмма

ГРАФИЧЕСКИЙ МЕТОД

```
import matplotlib.pyplot as plt

labls = df1['брак'].unique()

fig, axes = plt.subplots(1, 2, figsize=(10, 5))
axes[0].pie(df1.loc[df1['новер_конвейера'] == 1]['брак'].value_counts(), labels = labls, autopct='%.0f%%')
axes[0].set_title('Ковейер 1')
axes[1].pie(df1.loc[df1['новер_конвейера'] == 2]['брак'].value_counts(), labels = labls, autopct='%.0f%%')
axes[1].set_title('Конвейер 2')
plt.show()
```



Ну и теперь графически всё видно, а аналитически доказано: конвейер 2 – существенно больше брака делает, чем конвейер 1.

А раз так – нужны соответствующие управленческие меры! Какие - это уже вопрос менеджмента.

Или дальнейшее исслдедование?

Количественная – номинальная

Аналитический метод: Метод Стьюдента / Манна-Уитни
Графический: ящик с усами (боксплот)

Тут интересно тоже. Поставим себе задачу узнать: какой из факторов – вес, ингридиент_1, ингридиент_2 влияет на брак? Как это сделать?

Надо просто сравнить различие средних в группах брака. Например, вес. Отличается ли средний вес в группе «Брак» от среднего веса в группе «Годен»? Если да – говорим о влиянии одного фактора на другой.

	вес	оператор_линии	ингридиент_1	номер_смены	новер_конвейера	ингридиент_2	брак
0	61.7	2	250.9	2	1	6.3	Годен
1	58.6	2	275.7	1	2	6.7	Брак
2	53.6	2	280.6	2	1	7.5	Брак

Гипотезы такие:

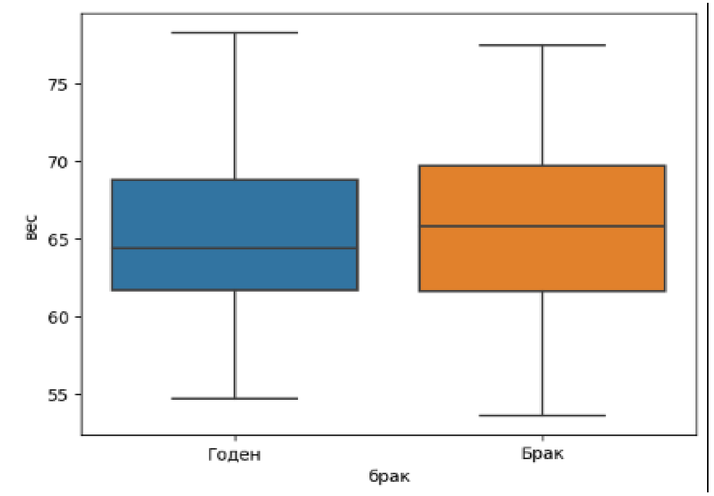
H0: категориальные переменные независимы;
H1: категориальные переменные связаны между собой.

Количественная - номинальная

Аналитический метод: Метод Стьюдента / Манна-Уитни
Графический: ящик с усами (боксплот)

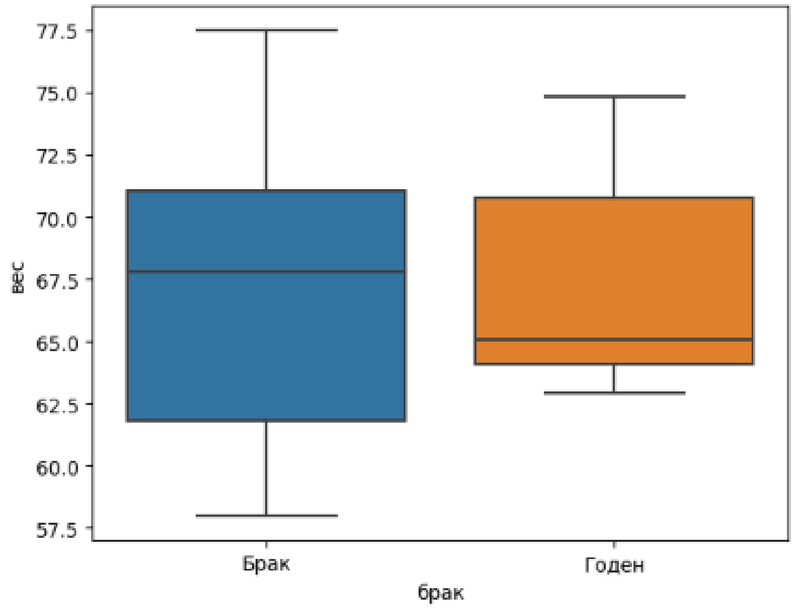
ГРАФИЧЕСИЙ МЕТОД

```
import seaborn as sns
sns.boxplot(x='брак', y='вес', data=df1,)
plt.show()
```



**Но помните, мы же выяснили,
что больше брака даёт конвейер 2.
Давайте только его и посмотрим.**

```
import seaborn as sns
sns.boxplot(x='брак', y='вес', data=df1[df1['новер_конвейера'] == 2],)
plt.show()
```



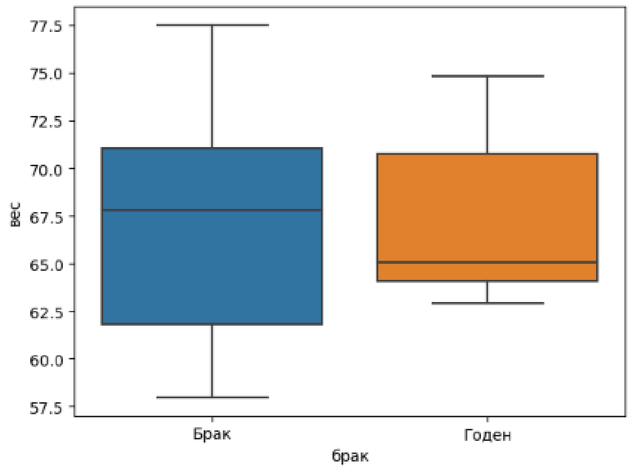
Академия Яндекса

Количественная – номинальная

Аналитический метод: Метод Стьюдента / Манна-Уитни
Графический: ящик с усами (боксплот)

ГРАФИЧЕСИЙ МЕТОД

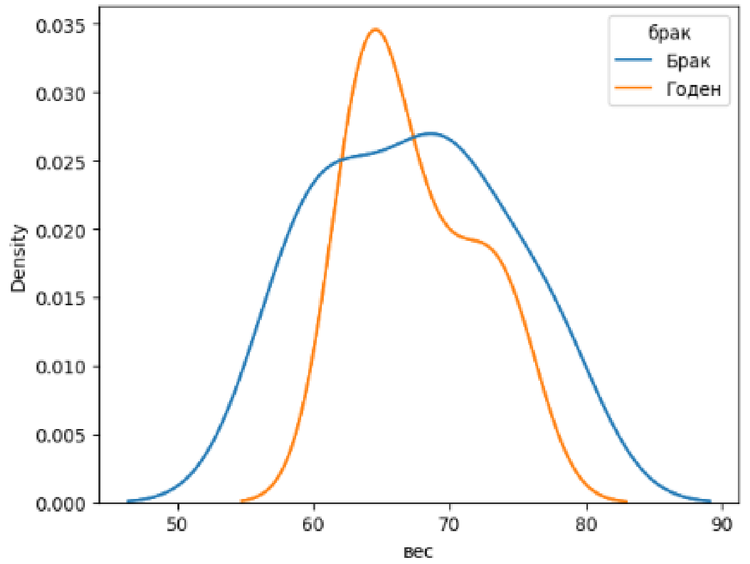
```
import seaborn as sns
sns.boxplot(x='брак', y='вес', data=df1[df1['новер_конвейера'] == 2],)
plt.show()
```



Тут средние кажется различаются. Но также есть и разброс данных (усы показывают) Поэтому графическое различие может быть и случайным. Давайте проверим аналитически. Для этого надо использовать один из двух методов: или метод Стьюдента или метод Манна-Уитни. Выбор зависит от нормальности данных. Если нормальные в каждой группе – Стьюдент. Если хот бы одна группа – не нормально распределена, то Манна-Уитни.

Сначала проверим данные на нормальность. Для этого, чаще всего, достаточно графического способа – построить гистограмму.
(В sns много разных возможностей. Можно histplot, можно kdeplot.)

```
sns.kdeplot(data=df1[df1['новер_конвейера'] == 2], x = 'вес', hue = 'брак')
plt.show()
```



Количественная - номинальная

Аналитический метод: Метод Стьюдента / Манна-Уитни
Графический: ящик с усами (боксплот)

АНАЛИТИЧЕСКИЙ МЕТОД

```
from scipy.stats import mannwhitneyu

br = df1[(df1['новер_конвейера'] == 2) & (df1['брак'] == 'Брак')]['вес']
gd = df1[(df1['новер_конвейера'] == 2) & (df1['брак'] == 'Годен')]['вес']

mannwhitneyu(br, gd, alternative='two-sided')
```

```
MannwhitneyuResult(statistic=85.0, pvalue=0.9016079268389684)
```

Вывод: p-уровень больше 0.05. Это говорит о том, что текущих данных не хватает, чтобы сделать вывод о различии средних.

(если бы нормальное распределение: `stats.ttest_ind`)

КАКОЙ ЖЕ ФАКТОР ВЛИЯЕТ ?

```
# ингридиент_2
br = df1[(df1['новер_конвейера'] == 2) & (df1['брак'] == 'Брак')]['ингридиент_2']
gd = df1[(df1['новер_конвейера'] == 2) & (df1['брак'] == 'Годен')]['ингридиент_2']

mannwhitneyu(br, gd, alternative='two-sided')
```

```
MannwhitneyuResult(statistic=115.0, pvalue=0.18928495025465886)
```

```
# ингридиент_1
br = df1[(df1['новер_конвейера'] == 2) & (df1['брак'] == 'Брак')]['ингридиент_1']
gd = df1[(df1['новер_конвейера'] == 2) & (df1['брак'] == 'Годен')]['ингридиент_1']

mannwhitneyu(br, gd, alternative='two-sided')
```

```
MannwhitneyuResult(statistic=166.0, pvalue=0.0001256183743734838 5)
```

Количественная - номинальная

Аналитический метод: Метод Стьюдента / Манна-Уитни
Графический: ящик с усами (боксплот)

У нас есть основания утверждать, что средние значения «ингридиента_1» в группах брака имеют статистически значимые отличия.

Т.е. различны. Говорит ли это о том, что данный фактор является причиной брака?
Сейчас не известно. Нужно изучать дальше. Мы же рассмотрели только факторы продукта самого. А есть ещё факторы процесса.

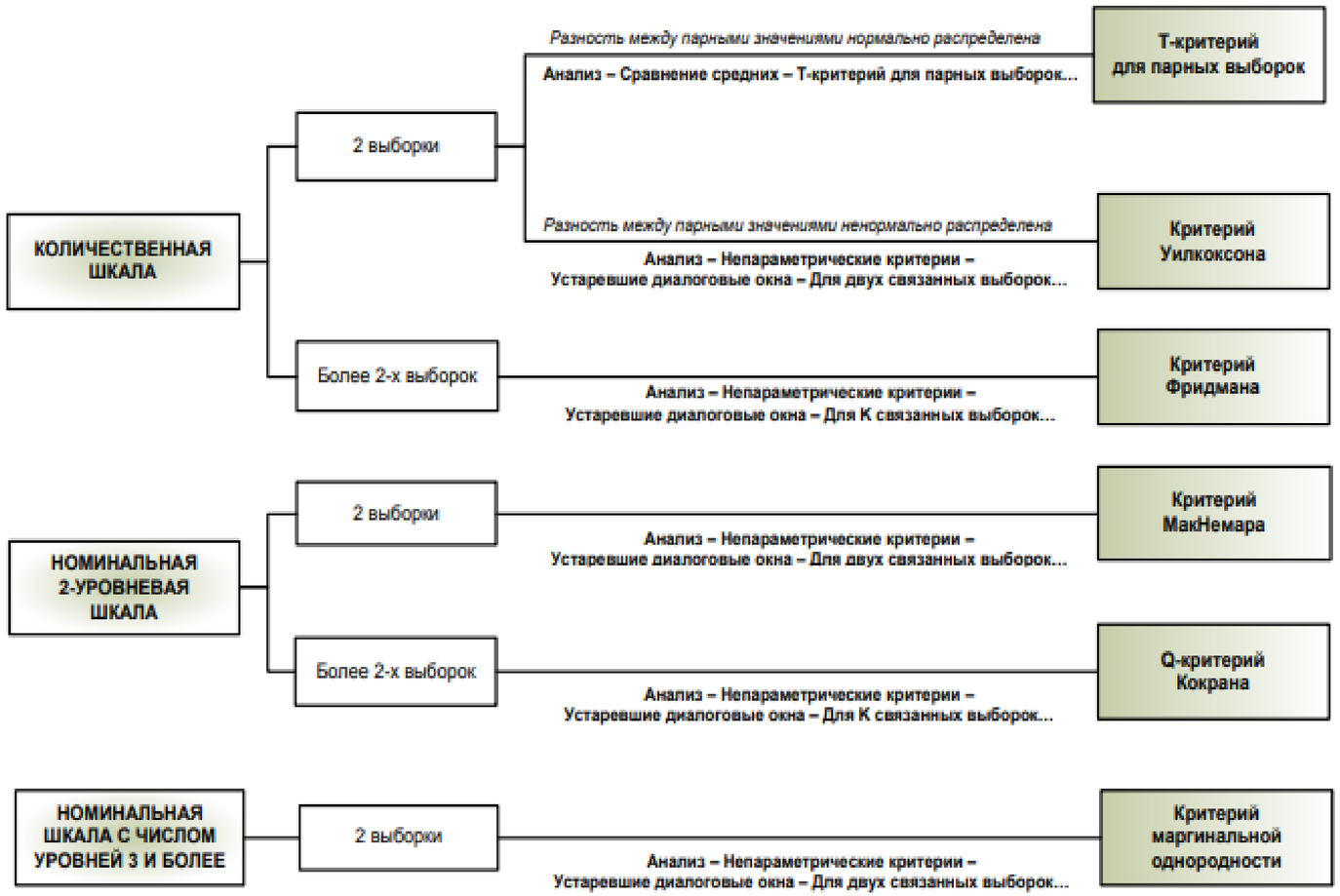
Вывод по методам (не зависимые выборки)

		Номинальная	
		2 уровня	Более 2 уровней
Количественная	Нормальная	Т-критерий Стьюдента	ANOVA
	Ненормальная	U-критерий Манна-Уитни	Критерий Краскела-Уоллеса

Алгоритм действий

1. Определить фактор и отклик
2. Определить шкалы фактора и отклика
3. Собрать данные
4. Выбрать тип графика и провести графический анализ
5. Выбрать метод проверки гипотезы
6. Рассчитать статистическую значимость
7. Сделать окончательный вывод

Другие методы



Что нужно будет для задачек

Проверка гипотез

from scipy.stats import mannwhitneyu, ttest_ind, ПРИМЕР: ttest_ind(g1, g2, equal_var=False)

pd.crosstab()
scipy.stats.chi2_contingency

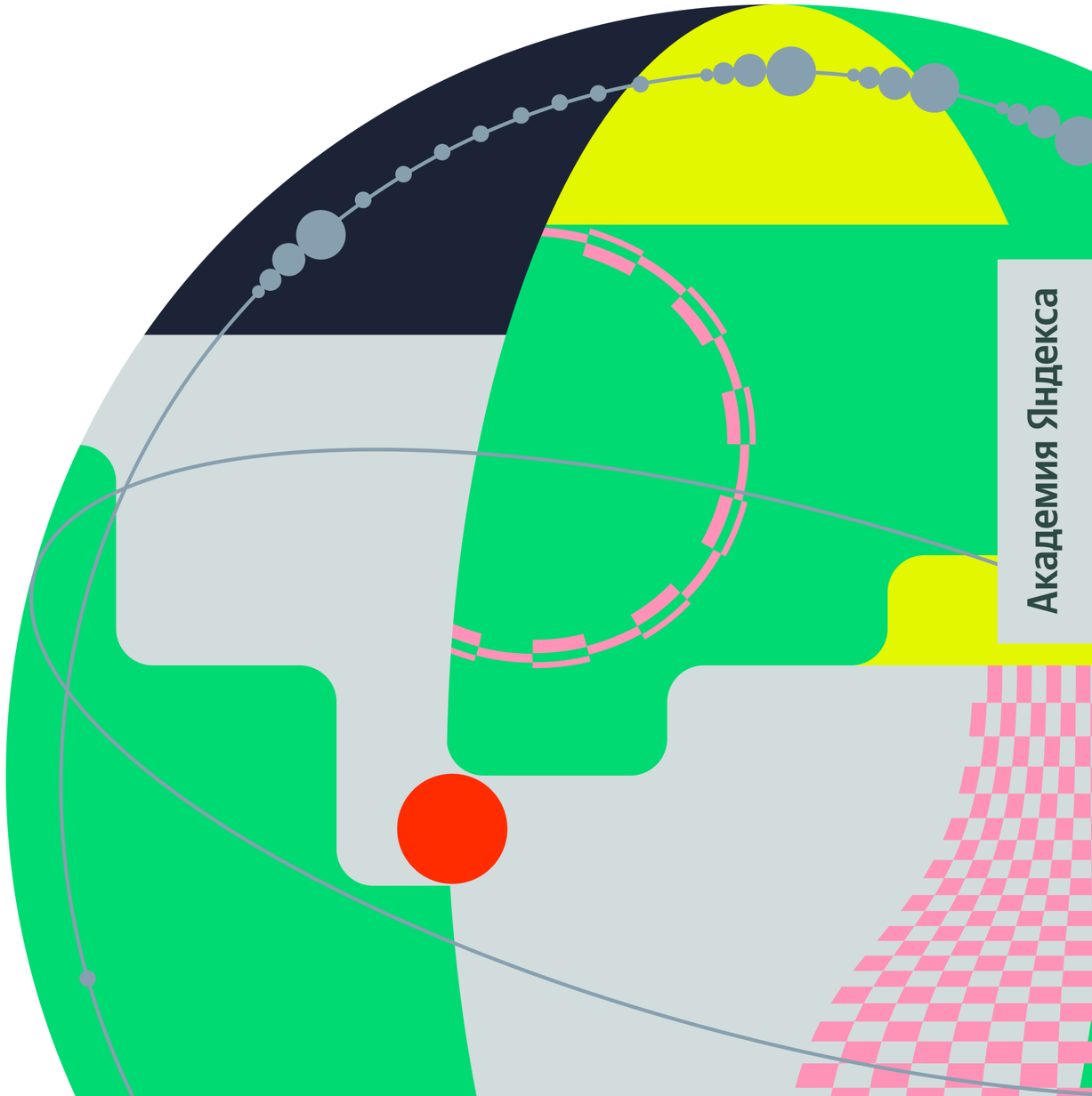
from scipy.stats import pearsonr, spearmanr

Дополнительно

from operator import itemgetter (ДЛЯ СОРТИРОВКИ)

from scipy.stats import shapiro (ПРОВЕРКА НА НОРМАЛЬНОСТЬ) ПРИМЕР: shapiro(g1)[1] >= 0.05

Спасибо за внимание



Академия Яндекса позволяет школьникам
и студентам освоить востребованные IT-профессии
по программам, разработанным экспертами компании