**Стандартная ошибка среднего**

Стандартная ошибка среднего , где s – дисперсия, n – кол-во наблюдений

Правило трёх сигм:

где сигма – стандартное отклонение.

Для оценки среднего значения:

, где H – предел погрешности (напр. ), z – доверительный уровень:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| % набл. | 60 | 70 | 80 | 85 | 90 | 95 | 97 | 99 | 99,7 |
| z | 0,84 | 1,03 | 1,29 | 1,44 | 1,65 | 1,96 | 2,18 | 2,58 | 3 |

Для оценки доли:

, где p – уровень разброса, который мы ожидаем (доля опрошенных из всех).

Цель определения размера выборки для описательного анализа:

* Для количественных шкал – оценить среднее значение ()
* Для бинарных шкал – оценить долю признака

Если размер генеральной совокупности известен при определении размера выборки n рекомендуется скорректировать его на известный объём генеральной совокупности N:

**Методы отбора**

* Случайный
* Систематический (каждое i-ое наблюдение, I = N/n)
* Одноступенчатый кластерный (ген. совокупность состоит из кластеров, каждый из которых отражает свойства всей генеральной совокупности, является небольшой моделью ген. совокупность. Из отобранных кластеров в выборку включаются все наблюдения
* Расслоенный (ГС состоит из слоёв, или групп, средние значения изучаемого признака в которых существенно различаются)
  + Равномерный (из каждого слоя случайным образом отбирается одинаковое кол-во наблюдений)
  + Пропорциональный (выборки из каждого слоя пропорциональны размерам слоев ГС)
  + Оптимальный (выборки из каждого слоя пропорциональны разбросу в каждом слое ГС)

Виды сравнения метрик:

1. Покомпонентное
2. Позиционное
3. Повременное
4. Частотное
5. Корреляционное



Виды диаграмм:

1. Круговая (Pie chart)
2. Линейная (Bar chart)
3. Гистограмма (Histogram)
4. График (Line plot)
5. Точечная (Scatter plot)
6. Диаграмма размаха (Box plot)
7. Лепестковая (Radar chart)
8. Тепловая карта (Heat map)

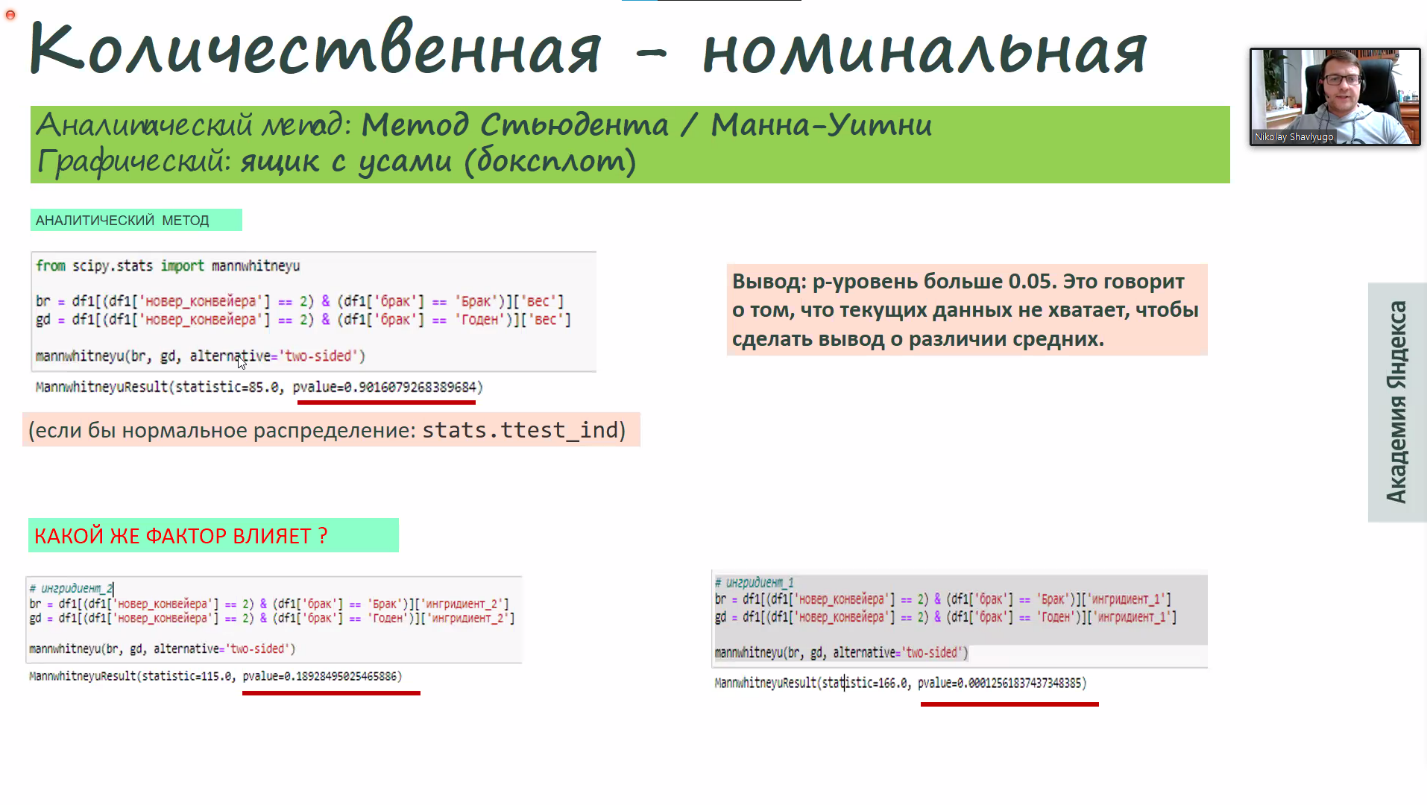
Правила создания диаграмм:

* Подписи графика и осей
* Выделение важного параметра
* Подбор информативной визуализации
* Спокойные тона при выборе цвета

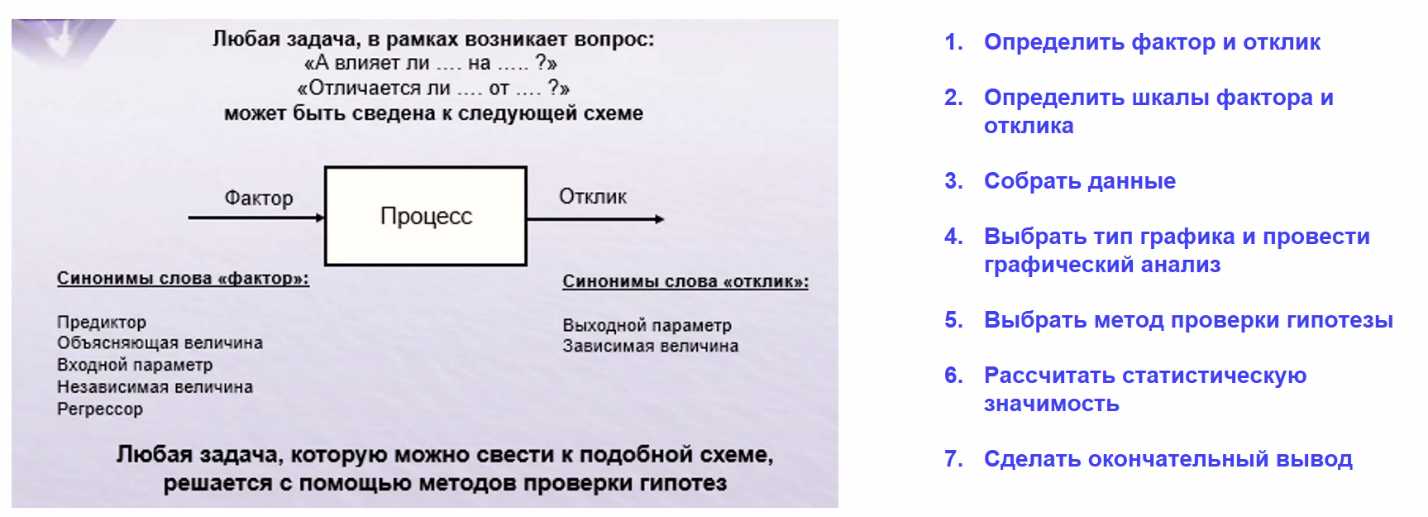
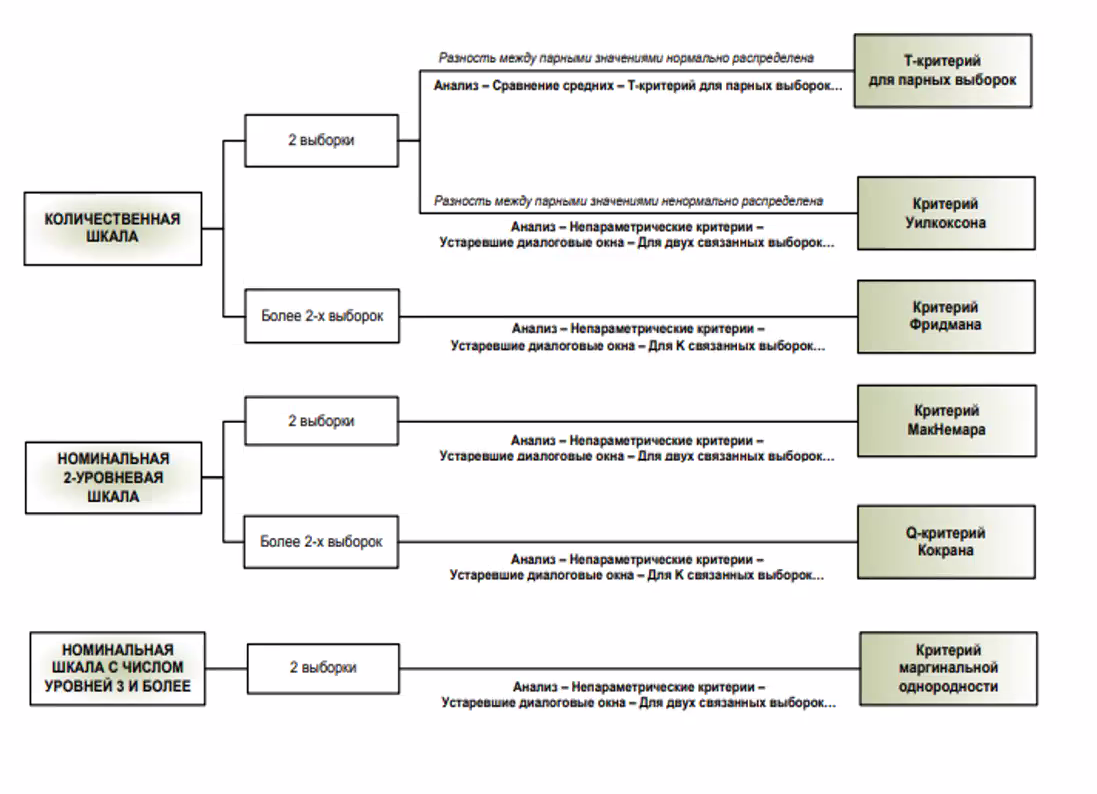
**Проверка гипотез**

Кол-кол: метод Спирмена/Пирсона, диаграмма рассеяния

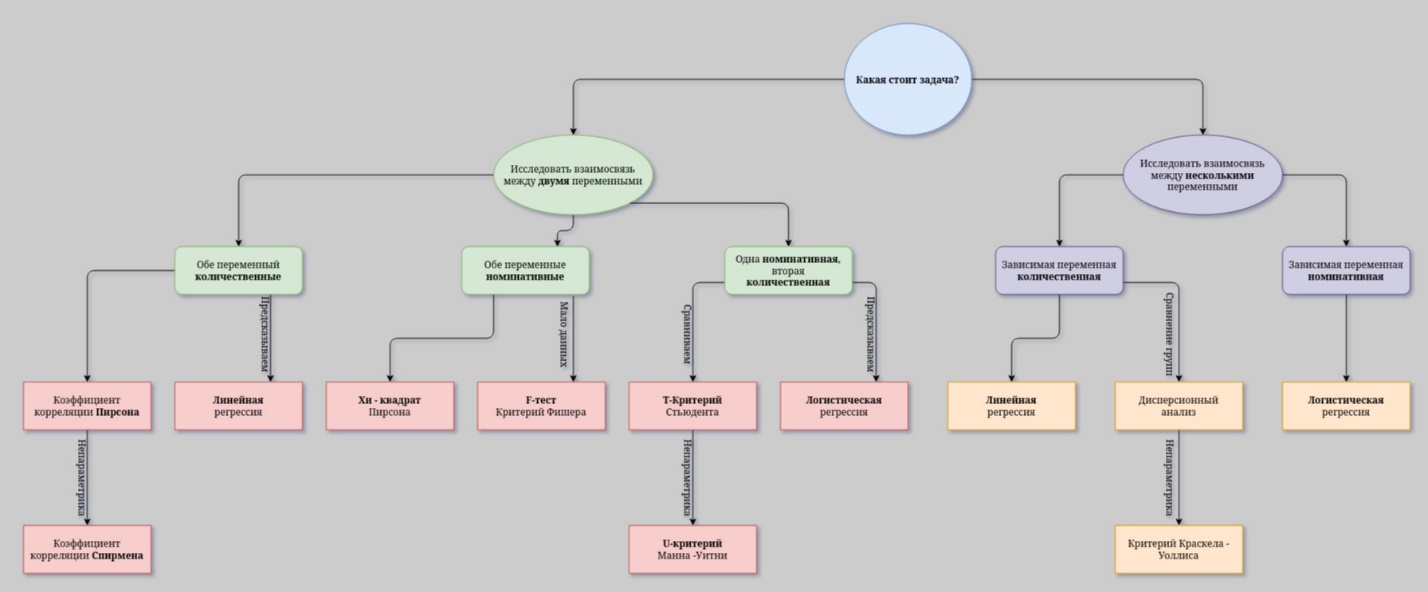
Ном-ном: метод Крамера, пирог

Кол-ном: метод Стьюдента (для норм.)/Манна-Уитни, ящик с усами

Для независимых выборок:

Для зависимых выборок:

df.sort(key=itemgetter())



**Виды метрик**

|  |  |
| --- | --- |
| **Метрика качества** | «Насколько хорошо что-то работает?» |
| **Приёмочная метрика** | Для оценки целесообразности внесения каких-либо изменений |
| **Мониторинговая метрика** | Оперативно сигнализирует о появлении проблемы или сбоя, на которую необходимо обратить внимание |
| **KPI-метрика** | Показатели, которые позволяют оценить степень достижения какой-либо цели |

**Из чего строятся метрики**

* Абсолютные
* Относительные

Метрика-отношение:

* Доля: хотим замерить не численную характеристику объектов
* Среднее: хотим замерить численную характеристику объектов
* И другие

**Свойства метрик**

1. **Сонаправленность.** При выборе прокси-метрики важно, чтобы она была связана с целевой метрикой и чтобы изменения прокси-метрики означало изменение целевой метрики в соответствующем направлении. То есть, положительные изменения прокси-метрики должны приводить к положительным изменениям целевой метрики.
2. **Интерпретируемость.** Метрика должна быть понятна как с точки зрения названия, так и с точки зрения порядка её вычисления.

|  |  |  |
| --- | --- | --- |
| Ответ на аналитический вопрос | Эффект на самом деле есть | Эффекта на самом деле нет |
| Эффект есть |  | *Ошибка первого рода* |
| Недостаточно данных | *Ошибка второго рода* |  |

1. **Точность.**
2. **Чувствительность.** Метрика называется чувствительной, если её использование редко приводит к возникновению ошибок второго рода.