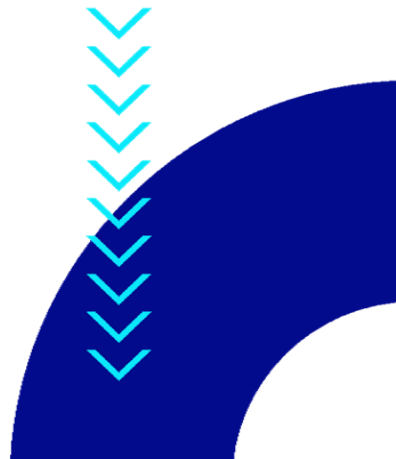
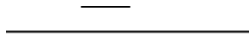


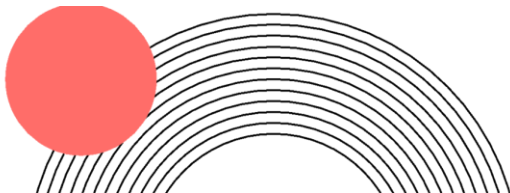
# Занятие №1



# В ходе первого занятия:

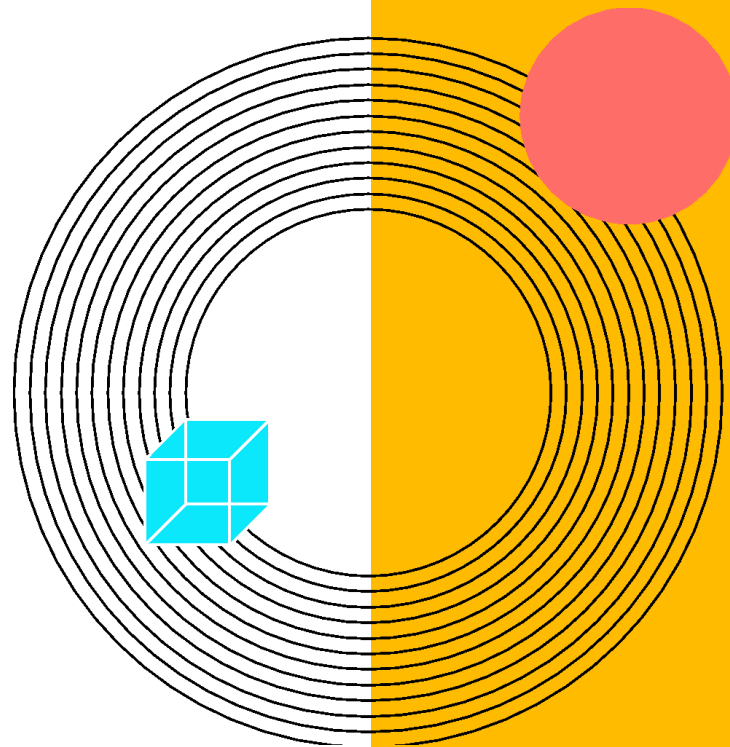
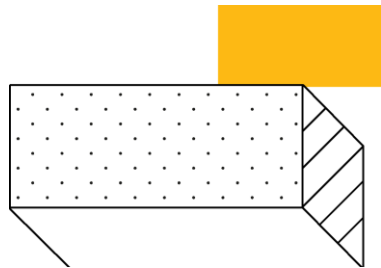


- ▲ поговорим, а зачем вообще нужны АБ - тесты;
- ▲ вспомним, какими бывают распределения и почему это важно при выборе теста;
- ▲ обсудим, что такое статистическая значимость и как она связана с p-value;
- ▲ поймем разницу между параметрическими и непараметрическими тестами, а также посмотрим на реализацию в библиотеках;
- ▲ поговорим о том, как выбрать правильный тест для конкретного кейса;
- ▲ обсудим понятие “нормальности”, когда это важно (и если важно, как приблизить распределение к нормальному), какие проверки существуют;
- ▲ проговорим, как используется Bootstrap.





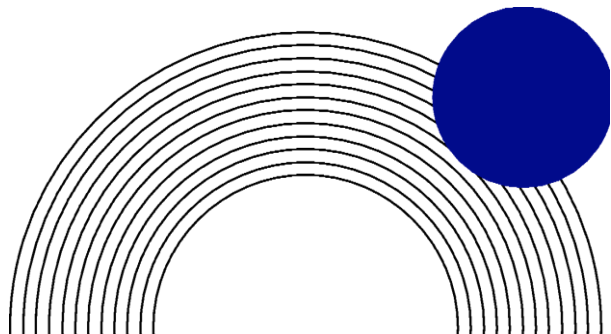
# Примеры использования А/В-тестирования



# Зачем нужно A/B-тестирование?



**Для проверки любых бизнес-гипотез!**



# Зачем нужно A/B-тестирование?



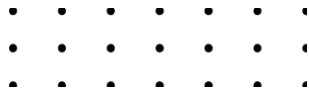
## Ритейл



**Гипотеза:** Если предложить скидку на любимый продукт, то клиенты будут больше покупать и средний чек в месяц вырастет.



**Как проверяем:** Небольшой части клиентов предлагаем скидку на любимый продукт, а другой части не предлагаем. Через месяц сравниваем две группы по приросту среднего чека. Если у группы со скидкой средний чек выше (то есть гипотеза подтвердилась), то кампанию со скидкой ставим на регламент.



# Зачем нужно A/B-тестирование?



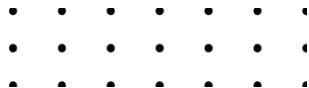
## Телеком



**Гипотеза:** У нас есть новая рекомендательная модель, которая, по мнению дата сайнтиста, лучше подбирает дополнительные услуги для абонентов, то есть конверсия в покупку будет выше, чем со старой моделью.



**Как проверяем:** Небольшой части абонентов подбираем предложения новой моделью и сопоставляем с теми абонентами, которым предложения подобраны старой моделью. Если конверсия у абонентов с новой моделью выше, то заменяем старую модель новой.



# Зачем нужно A/B-тестирование?



## Страховая



**Гипотеза:** Бизнес предполагает, что есть сегмент клиентов, которым мы отказываем в страховке на основе стандартного скоринга, хотя на самом деле мы могли бы подобрать для них выгодные предложения.



**Как проверяем:** Проанализировав данный сегмент клиентов, мы формулируем для них предложения и запускаем тест: выделяем их в отдельную группу и сравниваем их доля страховых случаев с другими клиентами. Если уровень дефолта приемлемый, то расширяем наши предложения для данного сегмента.



# Зачем нужно A/B-тестирование?



## Золотодобывающее производство



**Гипотеза:** Есть технология обогащения руды при помощи бактерий. У этого подхода нет стройного теоретического обоснования. Есть лишь гипотезы о том, каким образом среда и состав бактерий влияет на обогащение. Чтобы найти оптимальное сочетание среды и бактерий проводят эксперименты, то есть A/B тесты.

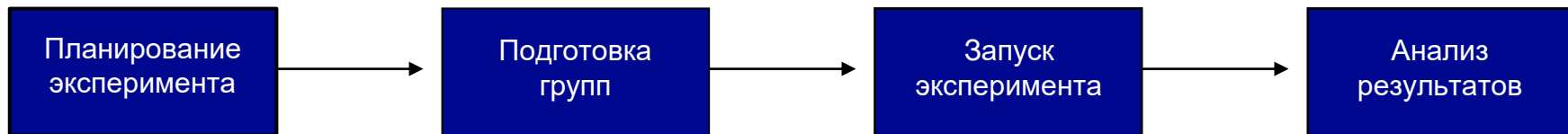


**Как проверяем:** В лабораторных условиях создается разная среда (влажность, температурный режим и т.п.) и разный состав бактерий. При помощи тестов выясняется, какое сочетание оптимально.



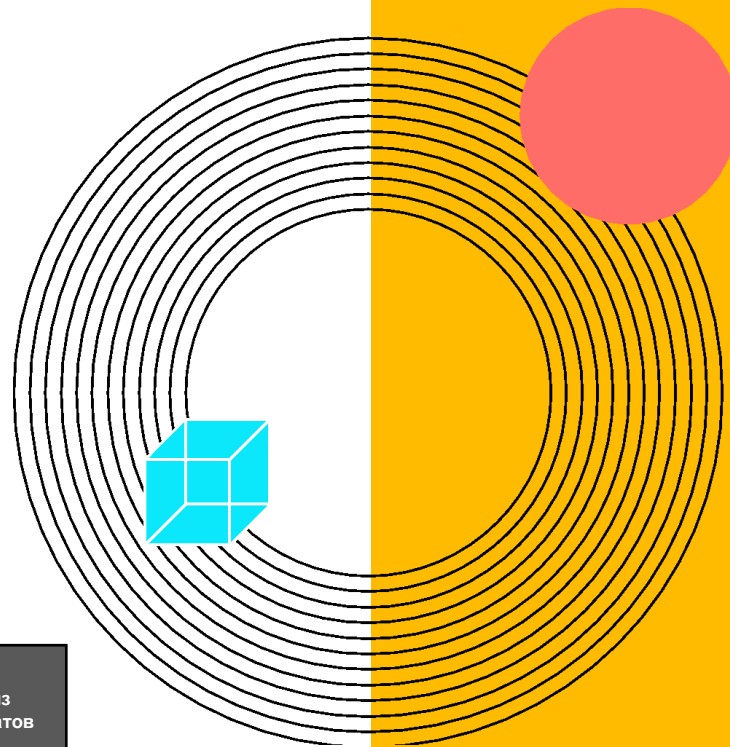
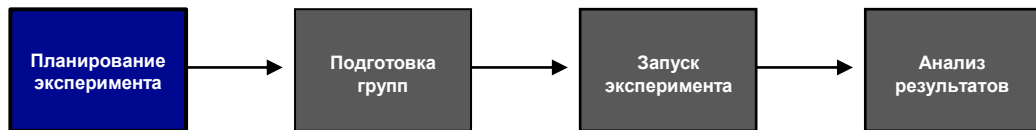


# A/B - тест

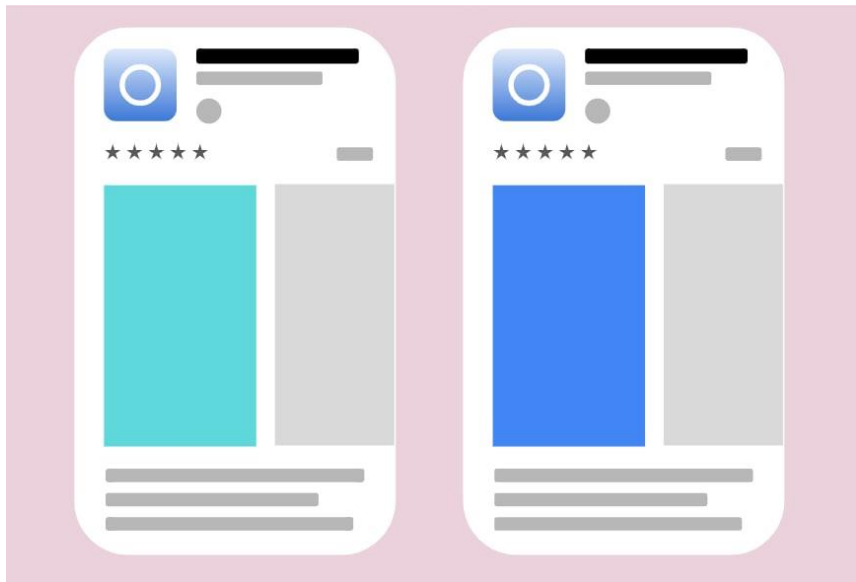




# Основные определения



# Вводный пример



50к

51к

Тест двух видов дизайна

- ▲ Отклик на первый дизайн: 50к
- ▲ Отклик на второй дизайн: 51к

Выбираем второй?



# Понятие частоты



Числовой ряд:

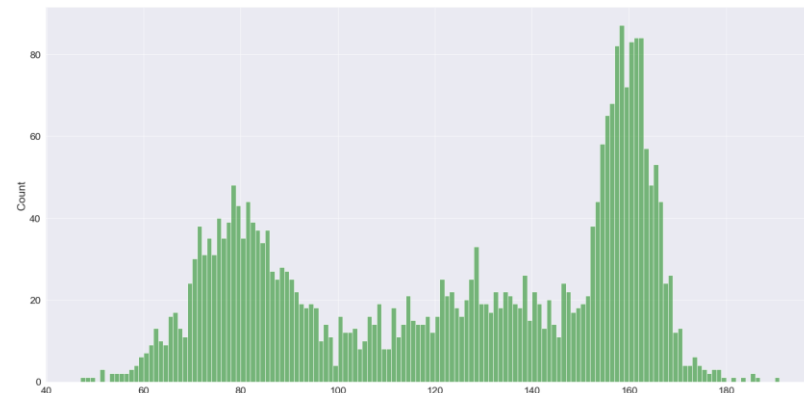
89, 122, 90, 159, 89, 157, 145, 67, 157, 162, 157...

Как его можно охарактеризовать?

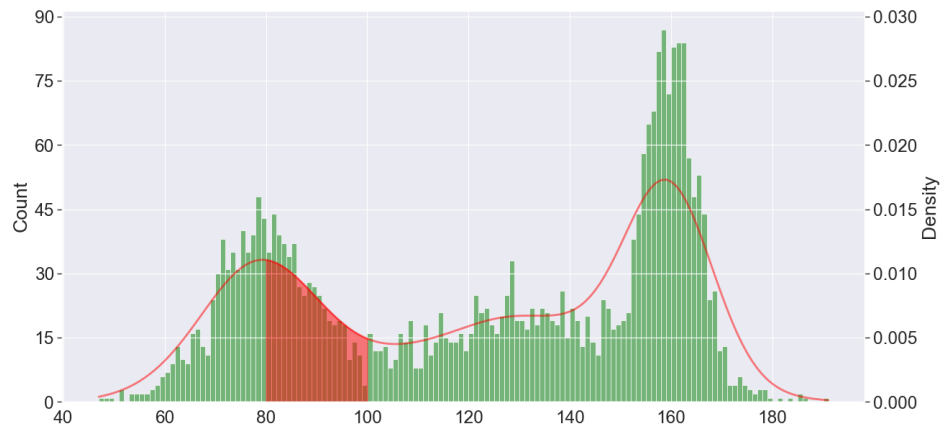
**А если значение не повторяется?**

- ▲ Упорядочиваем ряд в порядке возрастания
- ▲ Бьем весь диапазон значений на N частей
- ▲ Строим гистограмму

Значение		Сколько раз встретилось
89	→	2
157	→	3
...		



# Плотность распределения, вероятность, правдоподобие



- **Вероятность -**

площадь под графиком функции для определенного диапазона значений метрики

- **Плотность распределения -**

Делим частоту конкретного значения - на общее число наблюдений (на общую площадь под графиком распределения)

- **Правдоподобие -**

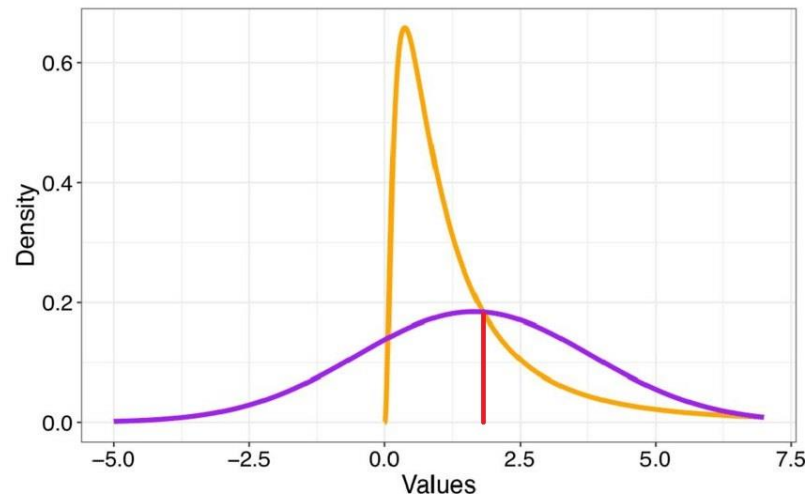
значение функции распределения плотности вероятности для конкретного значения метрики



# Правдоподобие



- На сколько конкретное значение является редким по сравнению с другими значениями метрики?
- Правдоподобие выбранного значения в обоих распределениях одинаково
- Однако, в одном распределении это значение является одним из наиболее правдоподобных
- А в другом - одним из наименее правдоподобных
- Как оценить, насколько выбранное значение правдоподобно - по сравнению с другими?



**“Значимость” значения в выборке** - доля значений столь же или менее правдоподобных, чем выбранное значение

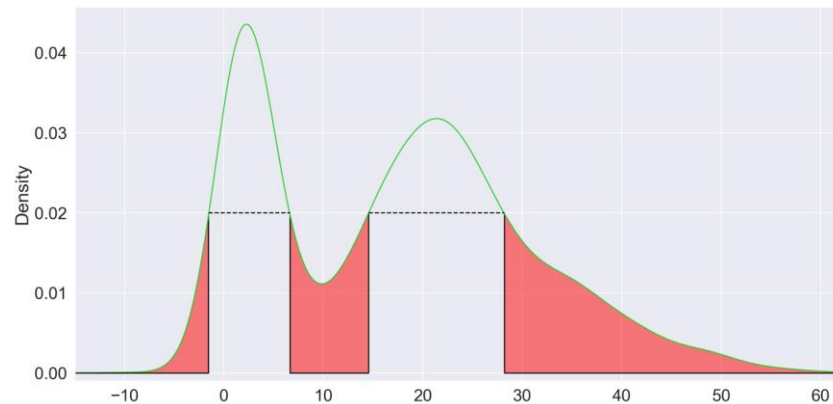
# “Значимость” значения в выборке



- ▲ Какова доля значений метрики - **столь же или более редких** (столь же или менее правдоподобных), чем выбранное значение метрики?

▼ Где может пригодиться:

разметка клиентов с точки зрения их  
характерности для бизнеса

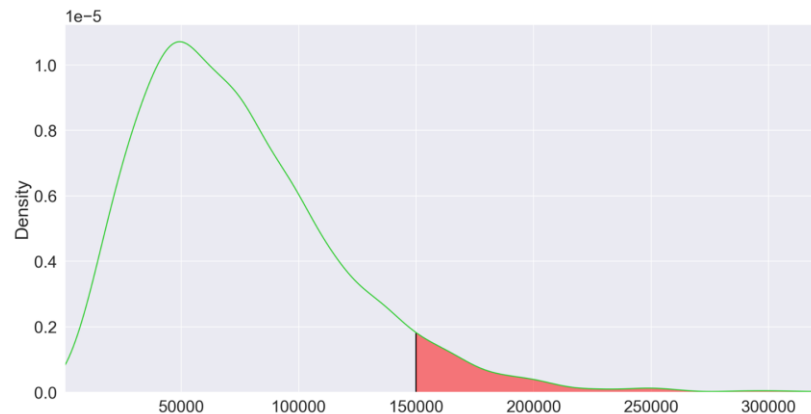


# “Значимость” значения в выборке



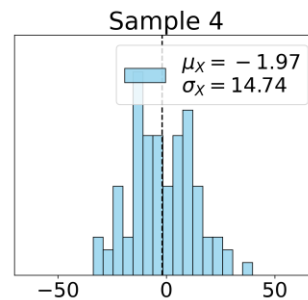
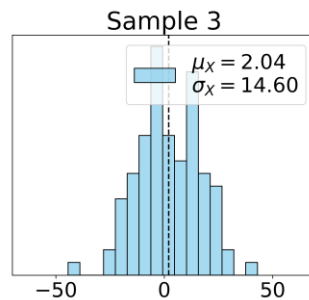
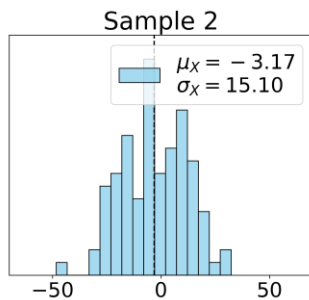
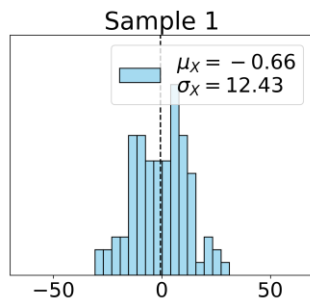
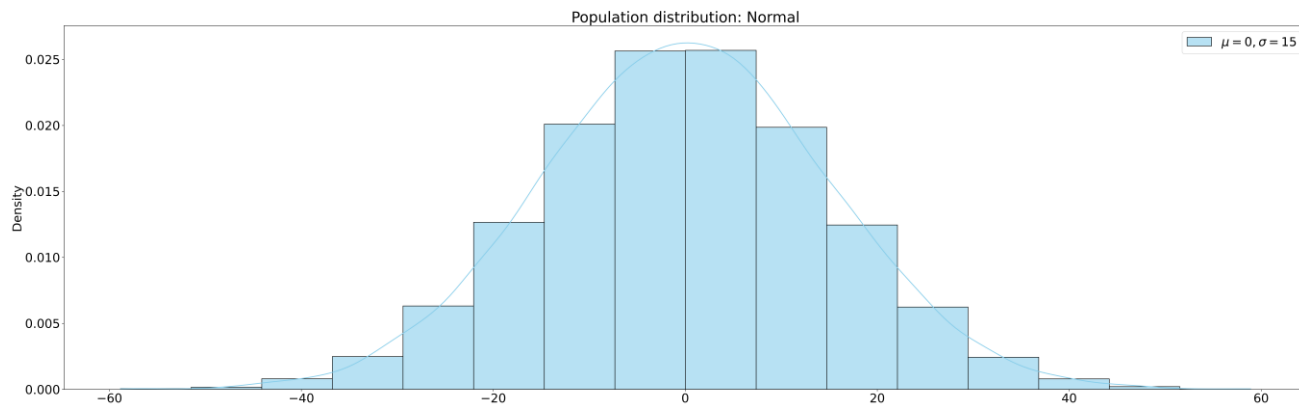
- ▲ Какова доля значений метрики - **таких же или более экстремальных**, чем выбранное значение метрики?

▼ *Где может пригодиться:*  
оценка потенциала к росту

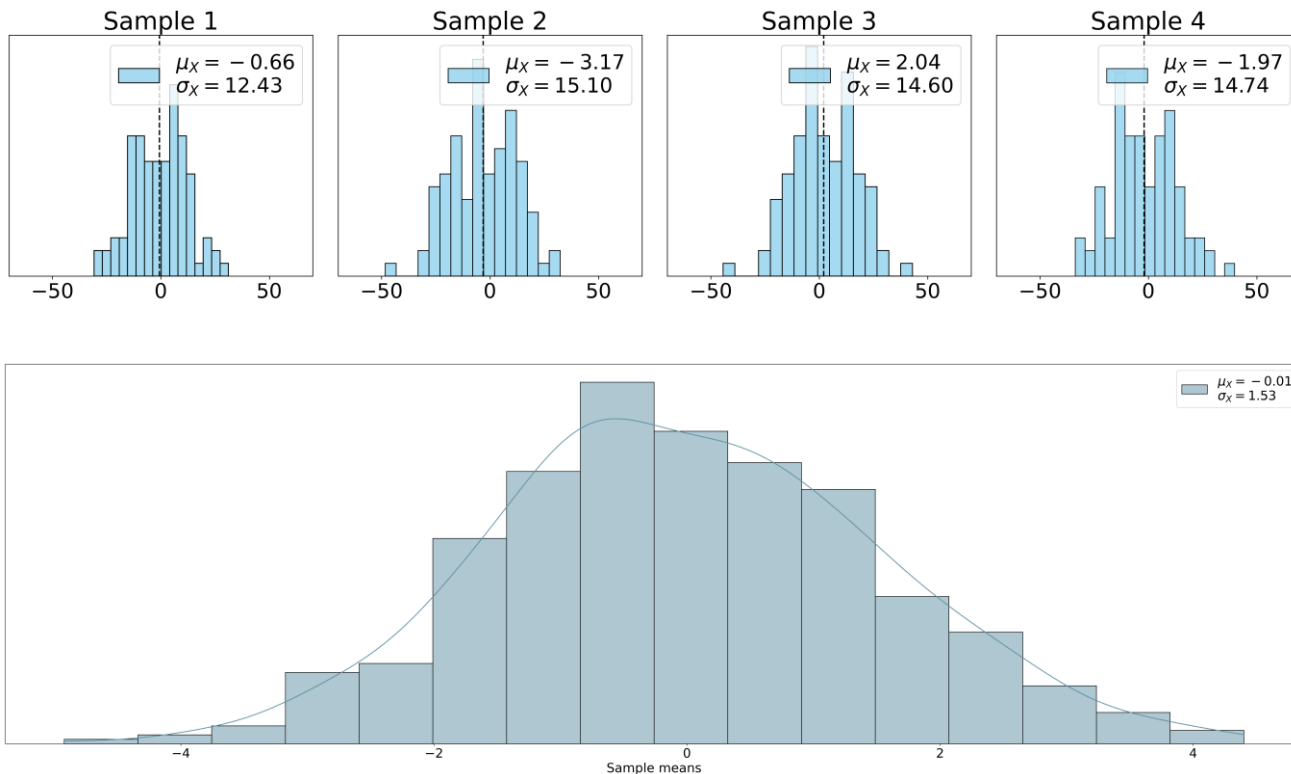




# Центральная предельная теорема



# Центральная предельная теорема



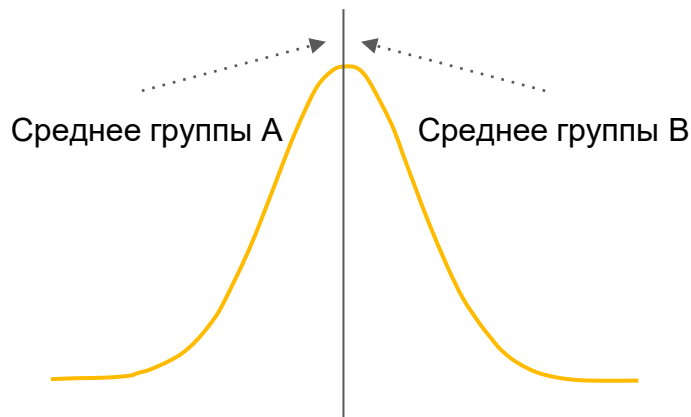
# Статистическая значимость. Тестирование гипотез



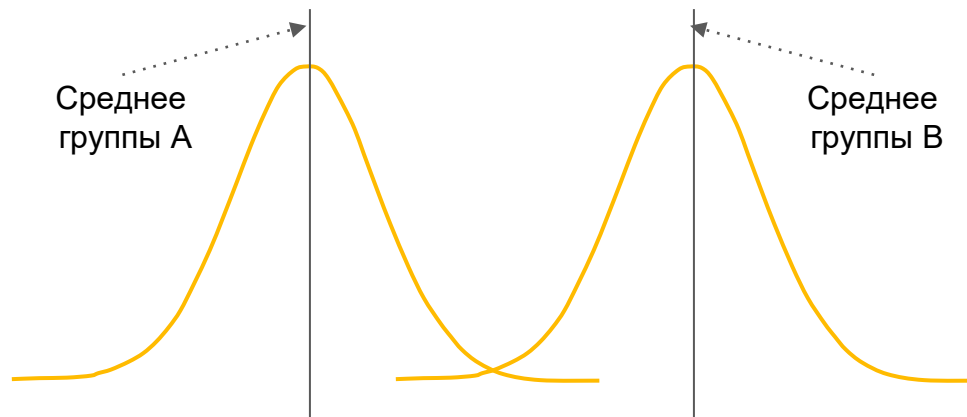
**A:** группа без воздействия

**B:** группа с воздействием

$H_0$ : средние значения для групп A и B совпадают



$H_1$ : средние значения для групп A и B не совпадают



## Теория!

# Статистическая значимость. Тестирование гипотез



## Практика

Известно:

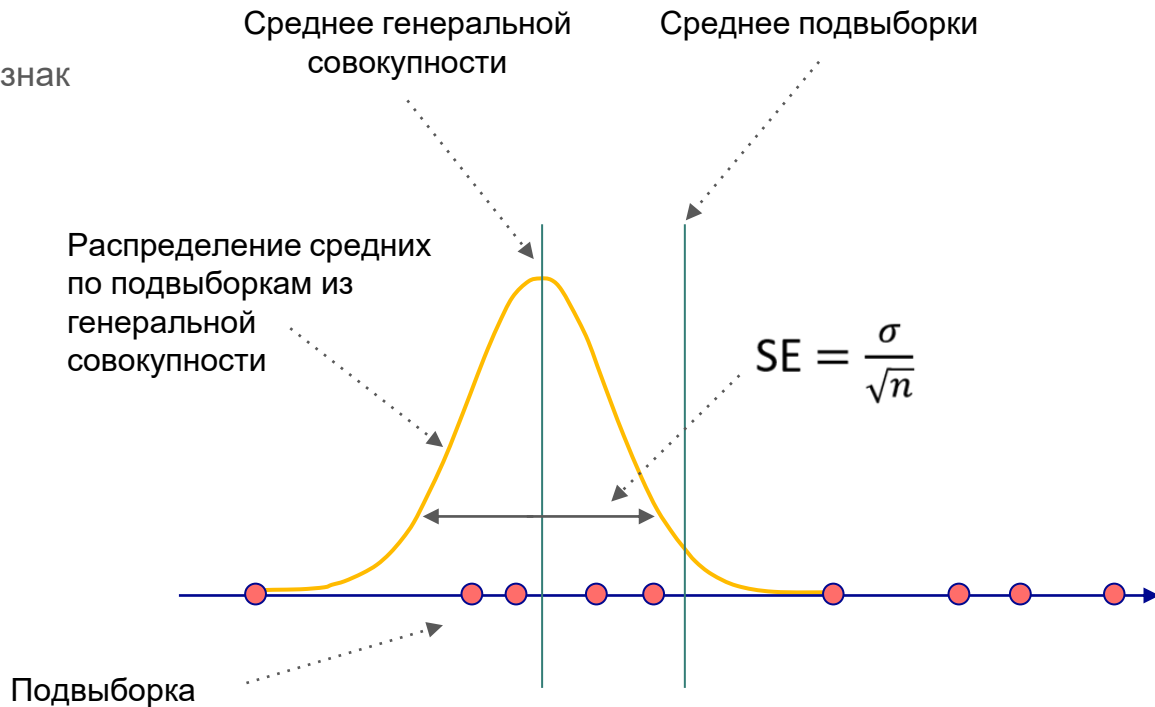
- В генеральной совокупности признак распределен нормально
- $\sigma = 12$

Взяли подвыборку:

- Размер  $n = 9$
- Среднее подвыборки равно 10

Гипотезы:

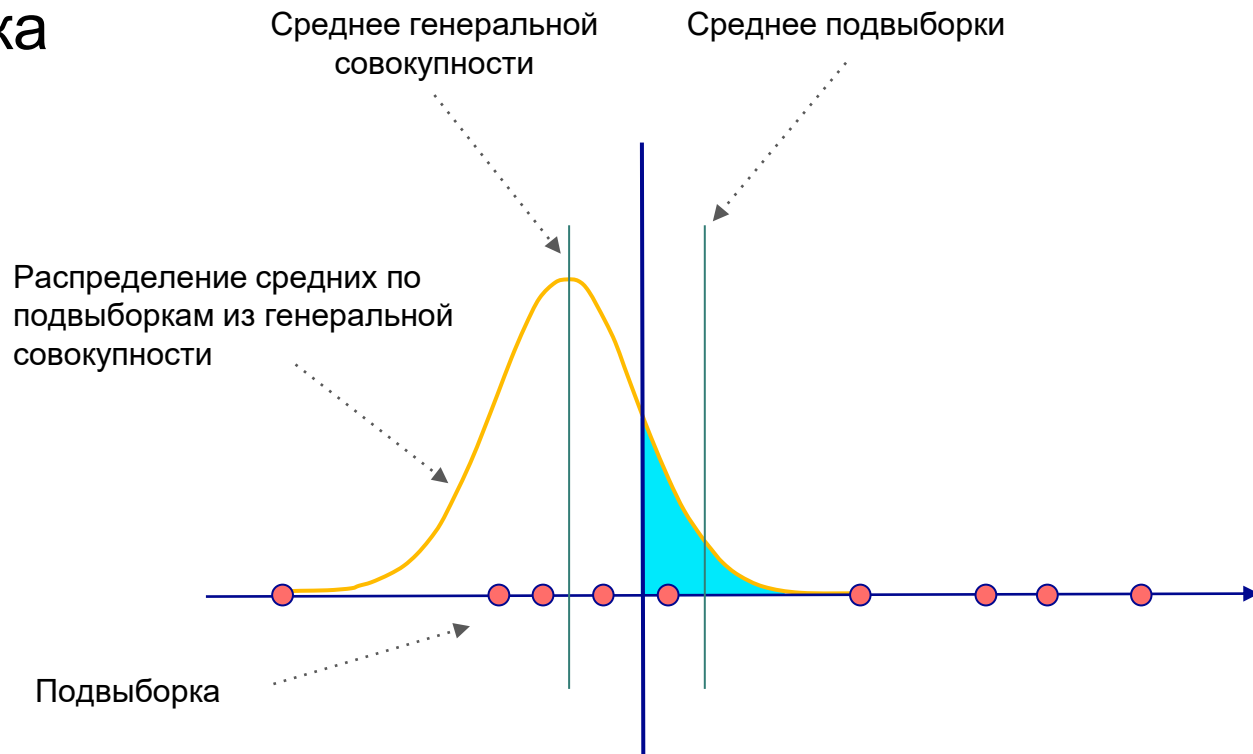
- $H_0$ : среднее генеральной совокупности равно 1
- $H_1$ : среднее генеральной совокупности  $> 1$



# Статистическая значимость. Тестирование гипотез



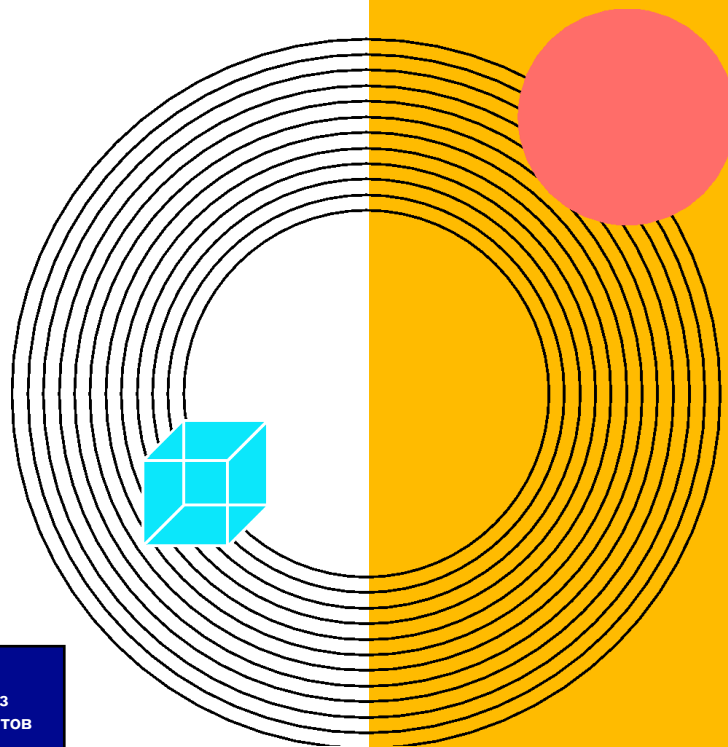
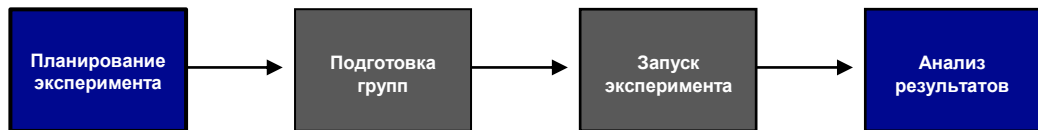
## Практика



Принятие решения: если среднее подвыборки попадает в заштрихованную область, то результат считается статистически значимым и  $H_0$  отвергается в пользу альтернативной  $H_1$



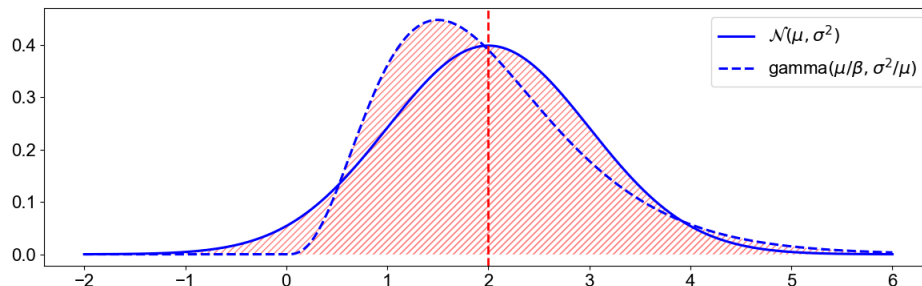
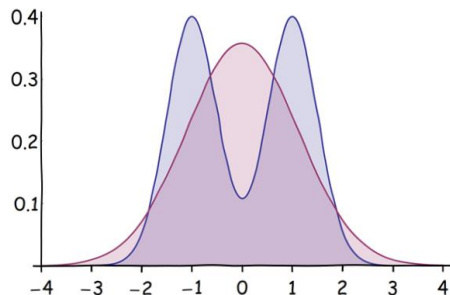
# Статистические тесты



# Статистический тест: сравнение распределений



Описать распределения одним числом, например, средним?



Применять стандартные подходы - непараметрические и параметрические тесты

Исходят из гипотезы о форме распределения (как правило, работают с нормальными распределениями)



# Непараметрические тесты



Метод	Суть метода
△ Манн - Уитни	<ul style="list-style-type: none"><li>□ Единый отранжированный ряд с проставленными рангами</li><li>□ Для каждого ряда рассчитывается функция от рангов</li></ul>
△ Уилкоксона	<ul style="list-style-type: none"><li>□ Парный критерий</li><li>□ Для каждой пары рассчитывается разность между значениями, абсолютные значения которых упорядочиваются</li><li>□ Если сдвиги в ту или иную сторону происходят случайно, то и суммы их рангов окажутся примерно равны</li></ul>
△ Крускала - Уоллиса	<ul style="list-style-type: none"><li>□ Проверка равенства медиан нескольких выборок</li><li>□ Упорядочиваются элементы всех выборок и рассчитывается ранг каждого элемента в полученном вариационном ряду</li></ul>
△ Колмогорова - Смирнова	<ul style="list-style-type: none"><li>□ Происходит сопоставление частот сначала по первому разряду, затем по первому и второму совместно, ...</li><li>□ Если различия существенны, то разница накопленных частот в какой-то момент превысит критическое значение</li></ul>



# Реализация непараметрических тестов



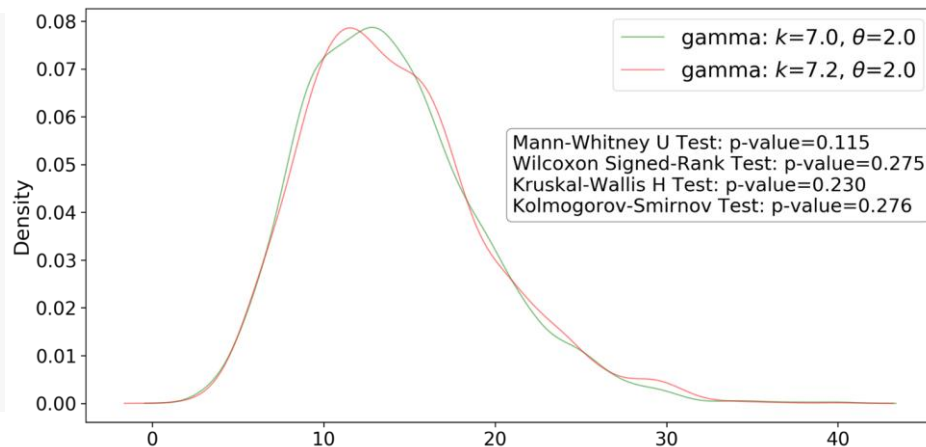
```
from scipy.stats import mannwhitneyu, wilcoxon, kruskal, ks_2samp

k1=7
k2=7.2

theta1 = 2
theta2 = 2

gamma_dist_1 = np.random.gamma(k1, theta1, 3000)
gamma_dist_2 = np.random.gamma(k2, theta2, 3000)

_, mw_p = mannwhitneyu(gamma_dist_1, gamma_dist_2)
_, wc_p = wilcoxon(gamma_dist_1, gamma_dist_2)
_, kw_p = kruskal(gamma_dist_1, gamma_dist_2)
_, ks_p = ks_2samp(gamma_dist_1, gamma_dist_2)
```



? А что, если требуется сравнить распределения по метрике?



# Сравнение распределений по метрике



## Примеры метрик:

- **Статистические метрики**

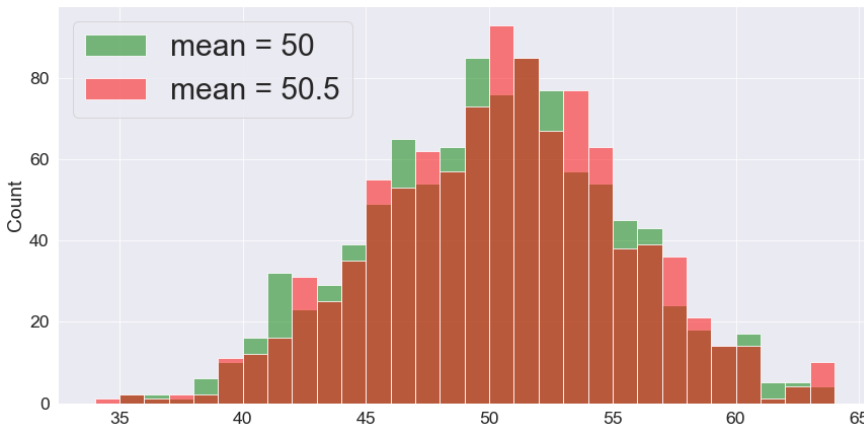
- Среднее
- Медиана
- Мода
- Граничные значения (MIN, MAX)
- Определенный квантиль

- Вычислим значение метрики для каждого распределения и сравним полученные значения

- Как определить статистическую значимость наблюдаемого отличия?

- **Бизнес-метрики**

- Средние траты клиента
- Средний скор склонности клиентов к оттоку (или к какому-либо продукту)



# Параметрические тесты



## Метод

## Суть метода

- Z-test

sigma - SD\* популяции, для конверсии

$$t = \frac{Z}{s} = \frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}}$$

- T-test Student

sigma - SD сэмпла, для денег

- T-test Welch

$$t = \frac{\Delta \bar{X}}{s_{\Delta \bar{X}}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_{\bar{X}_1}^2 + s_{\bar{X}_2}^2}}$$

$$s_{\bar{X}_i} = \frac{s_i}{\sqrt{N_i}}$$

- ANOVA/ANCOVA

- Применяется F\_test для нескольких групп
- Сравнение дисперсий
- Группы независимые

\*SD - Standard Deviation

# Реализация параметрических тестов

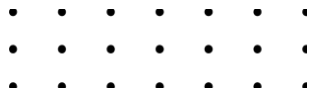
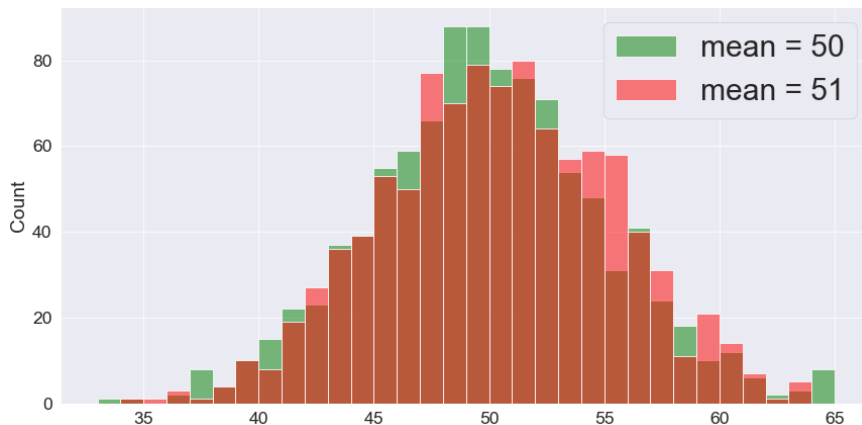


```
from scipy.stats import ttest_ind, f_oneway
from statsmodels.stats.weightstats import ztest

ttest_student_res = ttest_ind(normal_distr1, normal_distr2)
ztest_res = ztest(normal_distr1, normal_distr2)
ttest_welch_res = ttest_ind(normal_distr1, normal_distr2, equal_var=False)
anova_res = f_oneway(normal_distr1, normal_distr2)

print('Student`s t-test - \tstatistic: {}, p-value: {}'.format(*ttest_student_res))
print('Z-test - \t\t\tstatistic: {}, p-value: {}'.format(*ztest_res))
print('Welch`s t-test - \tstatistic: {}, p-value: {}'.format(*ttest_welch_res))
print('ANOVA test - \t\tstatistic: {}, p-value: {}'.format(*anova_res))
```

Student`s t-test -	statistic: -1.78743601953593, p-value: 0.07401856983530722
Z-test -	statistic: -1.78743601953593, p-value: 0.0738670443207255
Welch`s t-test -	statistic: -1.78743601953593, p-value: 0.07401863607106599
ANOVA test -	statistic: 3.194927523934394, p-value: 0.07401856983529545



# Параметрические критерии



## Ограничения

- ▲ Гипотеза о форме распределения
- ▲ Допущение: значения нескольких метрик (например, среднее и дисперсия) полностью описывают распределение
- ▲ Параметрические тесты работают с ограниченным набором метрик

## Недостатки

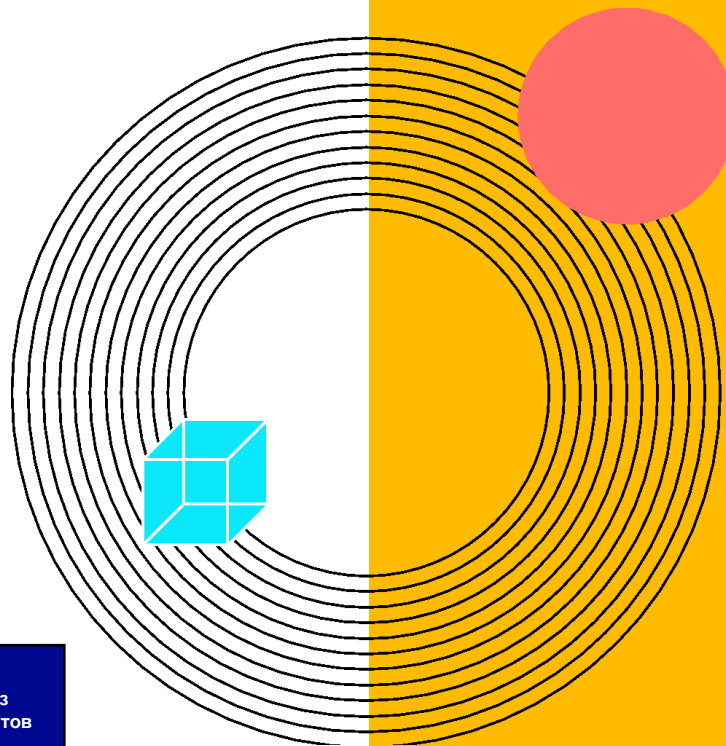
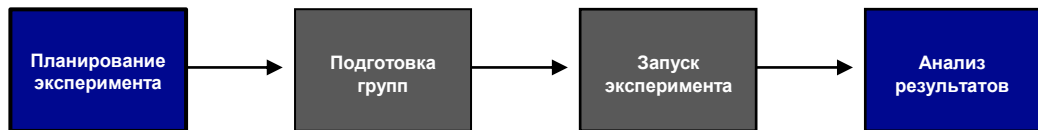
- ▲ Распределение не всегда соответствует исходной гипотезе о форме распределения
- ▲ Значение среднего не всегда бывает характерным свойством распределения
- ▲ Может требоваться сравнение по нестандартной метрике

## Распространенные ошибки

- ✗ С ростом количества наблюдений распределение всегда стремится к нормальному. Закон больших чисел говорит только о том, что с ростом числа наблюдений характеристики выборки стремятся к истинным значениям этих характеристик для общей совокупности
- ✗ Т-тест неприменим для ненормальных распределений



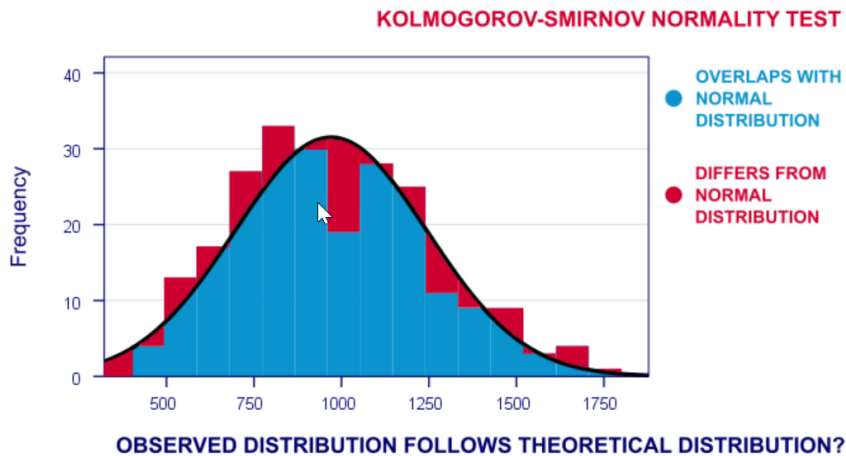
# Нормализация распределения



# Проверка на нормальность

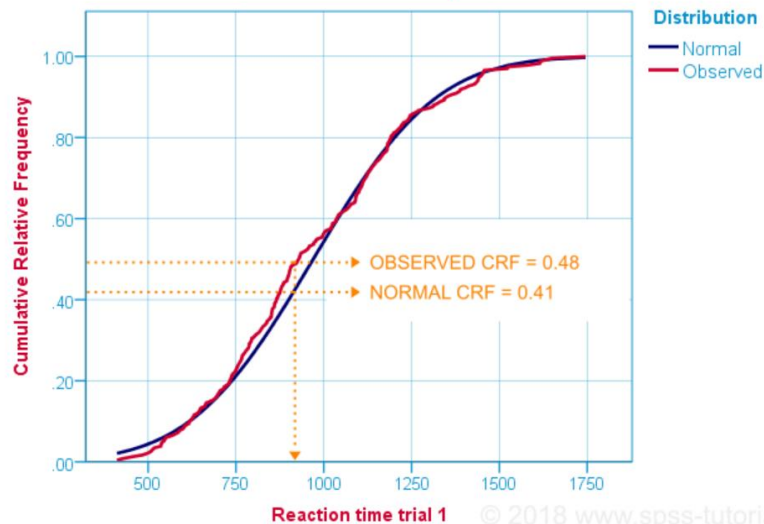


- ▼ Kolmogorov-Smirnov test for Normality
- ▼ Shapiro-Wilk Test
- ▼ Anderson-Darling Normality Test
- ▼ Chi-Square Normality Test



Observed Versus Normal Cumulative Relative Frequencies

All Respondents | N = 233



© 2018 www.spss-tutorials.com



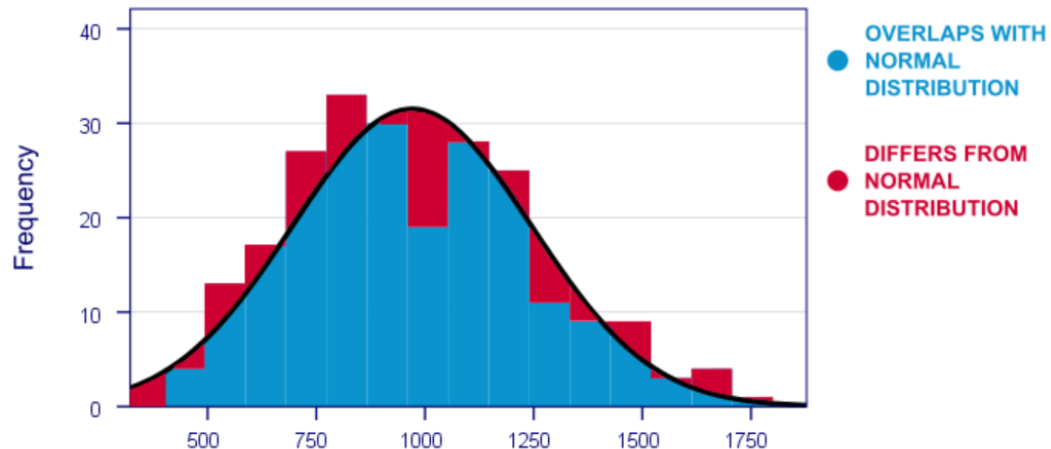
# Проверка на нормальность



- ▼ Kolmogorov-Smirnov test for Normality
- ▼ **Shapiro-Wilk Test**
- ▼ Anderson-Darling Normality Test
- ▼ Chi-Square Normality Test

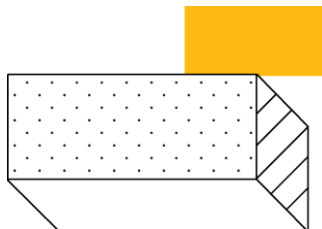
www.spss-tutorials.com

## SHAPIRO-WILK NORMALITY TEST



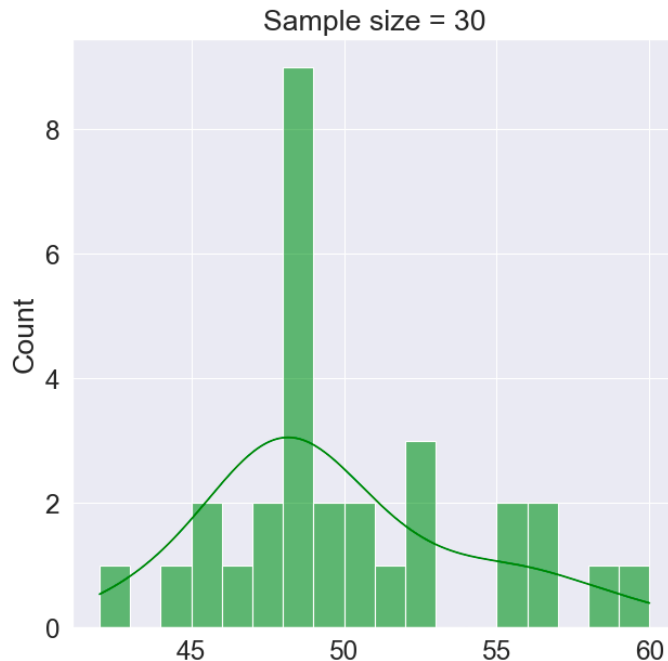
OBSERVED DISTRIBUTION FOLLOWS THEORETICAL DISTRIBUTION?

Расчет % схожести экспериментального и нормального распределений.





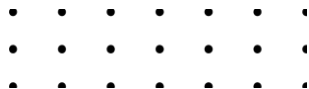
# Реализация проверки на нормальность



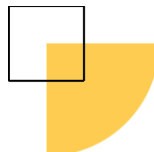
```
from scipy.stats import shapiro, normaltest, anderson, chisquare, jarque_bera, kstest
from statsmodels.stats.diagnostic import lilliefors
```

```
_, sh_p = shapiro(sample)
_, dk_p = normaltest(sample)
_, cs_p = chisquare(sample)
_, li_p = lilliefors(sample)
_, jb_p = jarque_bera(sample)
_, ks_p = kstest(sample, cdf='norm')
an_result = anderson(sample)
```

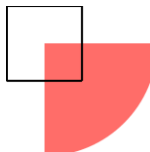
Shapiro-Wilk Test	: p-value=0.072. Probably gaussian
D'Agostino's K-squared Test	: p-value=0.244. Probably gaussian
Chi-Square Normality Test	: p-value=0.999. Probably gaussian
Lilliefors Test for Normality	: p-value=0.005. Probably not Gaussian
Jarque-Bera test for Normality	: p-value=0.342. Probably gaussian
Kolmogorov-Smirnov test for Normality	: p-value=0.0. Probably not Gaussian
Anderson-Darling Normality Test	: stat=0.949. critical value - 0.521 at 15.0 level of significance. Probably not Gaussian



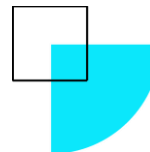
# Нормализация распределения



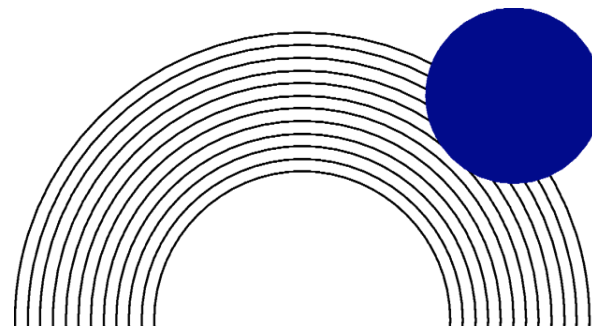
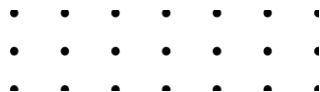
Исключаем  
выбросы



Функциональное  
преобразование



Бакетирование:  
вычисляем значение  
метрики на  
непересекающихся  
бакетах одинакового  
размера



# Параметрические тесты - не панацея. Когда они не работают?



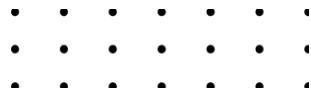
Нестандартная метрика

Нормализовать  
распределение не  
удалось

Распределение не  
соответствует  
стандартной форме






**Тогда:**

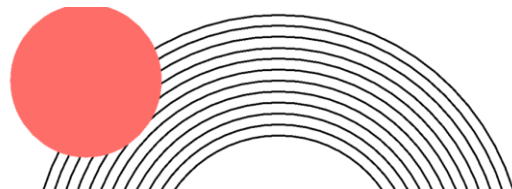
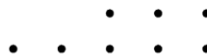
- ▶ Как определить статистическую значимость наблюдаемого отличия?
- ▶ Насколько вычисленное значение соответствует истине?
- ▶ Насколько вычисленное значение метрики характеризует распределение?



# Возможное решение

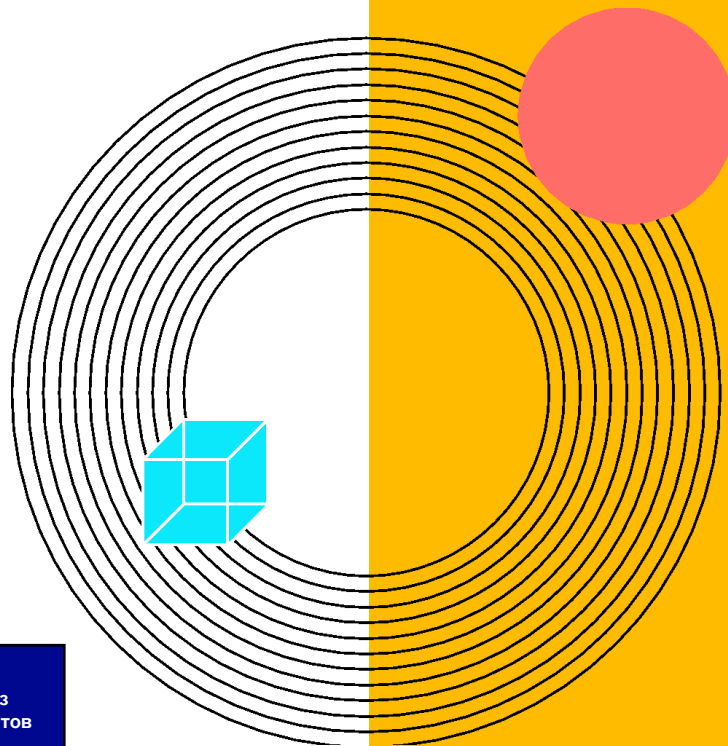
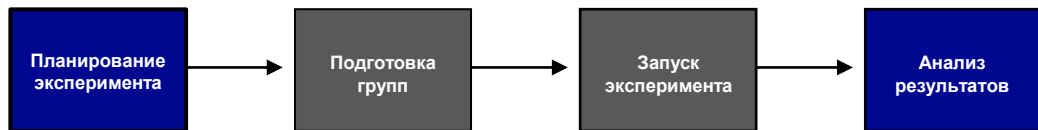


-  Повторить эксперимент множество раз (и каждый раз вычислить значение метрики для полученного распределения)
  -  Чем больше раз мы повторим эксперимент - тем точнее нам удастся охарактеризовать каждое из распределений
  -  Сравнить полученные распределения значений метрики
-  Но ведь повторить эксперимент невозможно!
  -  Для имитации повторного эксперимента используются методы ресемплинга - в частности, bootstrap





# Bootstrap



# Bootstrap



bootstrap sample 1

bootstrap sample 2

bootstrap sample 3

Ресэмплинг: имитация повторного эксперимента (имитация повторной выборки из генеральной совокупности)

Описать каждое из полученных распределений одним числом (например, средним или медианой)

# Выводы по первому занятию



- ▲ Статистическая значимость (p-value) показывает долю значений столь же или менее правдоподобных, чем выбранное значение.
- ▲ К непараметрическим тестам относят следующие:
  - ▶ Манна - Уитни
  - ▶ Уилкоксона
  - ▶ Крускала - Уоллиса
  - ▶ Колмогорова - СмирноваОни позволяют сравнивать непосредственно формы распределений.
- ▲ Сравнить распределения может оказаться полезным не только по стандартной метрике - среднее - но и по нестандартным: например, медиана, мода, граничные значения (MIN, MAX), квантиль.
- ▲ Параметрические тесты работают для сравнения средних:
  - ▶ Z-test
  - ▶ T-test Student
  - ▶ T-test Welch
  - ▶ ANOVA/ANCOVAТакие тесты обладают целым рядом ограничений и недостатков.
- ▲ Если есть необходимость сравнить распределения по любой другой метрике, кроме среднего, то это можно реализовать с помощью bootstrap.

# Литература



- ▶ Fundamentals of Biostatistics ([2015](#));
- ▶ The Practice of Statistics for Business and Economics ([2020](#));
- ▶ All of Statistics A Concise Course in Statistical ([2010](#));
- ▶ “[История одного обмана](#)” или “[Требования к распределению в t-тесте](#)”;
- ▶ [How Not To Run an A/B Test](#);
- ▶ [Как правильно считать деньги, или Несколько слов в пользу теста Стьюдента.](#)







# ВОПРОСЫ

