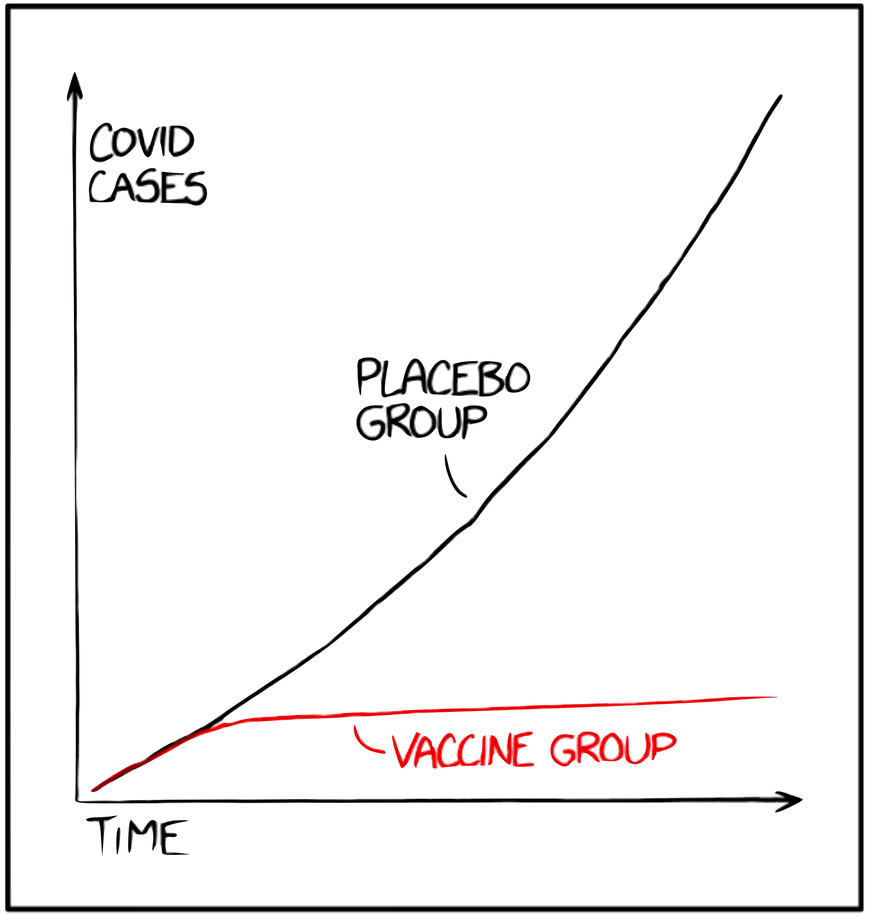
Погружение в statistics



STATISTICS TIP: ALWAYS TRY TO GET DATA THAT'S GOOD ENOUGH THAT YOU DON'T NEED TO DO STATISTICS ON IT

Статистика

- **Статистика** или **Оценка** – измеримая функция от выборки, не зависящая от любых других параметров.

Чаще всего статистики используются для поиска неизвестного параметра распределения θ и имеют вид $T : \mathbb{R}^n \to \Theta$

Функция правдоподобия

$$f_\theta(y) = \begin{cases} \text{плотность } f_\theta(y), & \text{если } \mathcal{P}_\theta \text{ абсолютно непрерывно,} \\ P_\theta\left(X_1 = y\right), & \text{если } \mathcal{P}_\theta \text{ дискретно.} \end{cases}$$

Функция правдоподобия выборки X: $L(X, \theta) = \prod_{i=1}^{n} f_\theta(x_i)$

Функция правдоподобия

$$f_\theta(y) = \begin{cases} \text{плотность } f_\theta(y), & \text{если } \mathcal{P}_\theta \text{ абсолютно непрерывно,} \\ P_\theta\left(X_1 = y\right), & \text{если } \mathcal{P}_\theta \text{ дискретно.} \end{cases}$$

Функция правдоподобия выборки X: $L(X, \theta) = \prod_{i=1}^{n} f_\theta(X_i)$

В дискретном случае принимает вид:

$L(X, \theta) = \mathbb{P}_\theta(X_1 = x_1, \ldots, X_n = x_n)$

Оценка максимального правдоподобия

Оценка максимального правдоподобия $\theta^*(X)$ параметра θ — точка параметрического множества Θ, в которой функция правдоподобия $L(X, \theta)$ при заданной реализации выборки x достигает максимума, т.е.:

$$L(x, \theta^*) = \max_{\theta \in \Theta} L(x, \theta)$$

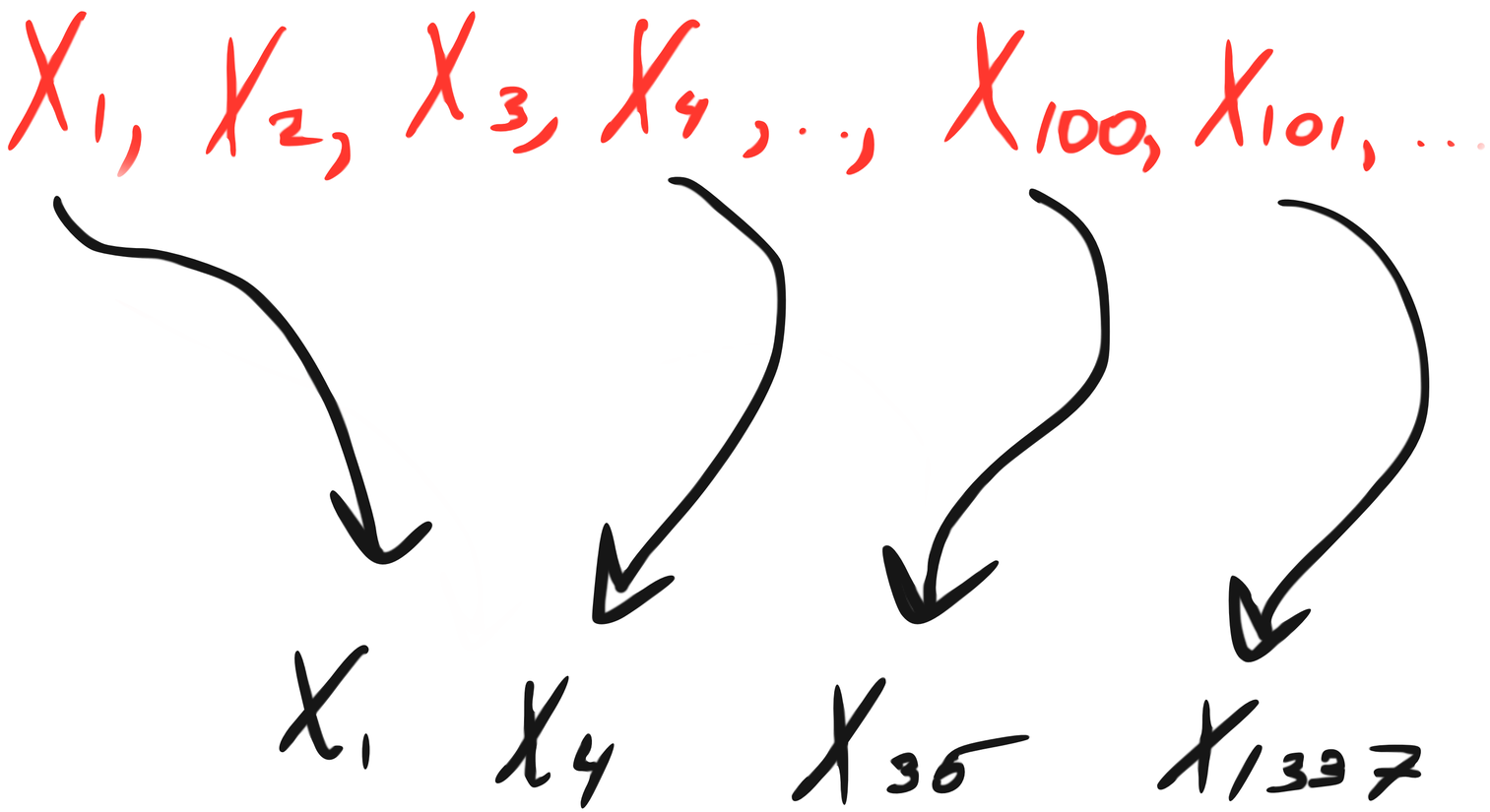
Виды оценок (статистик)

- **Несмещённая оценка** параметра θ – статистика $T(X)$, т.ч.
  $\forall\theta \in \Theta \mathbb{E}_\theta T(x) = \theta$

- **Асимптотическая оценка** параметра θ – статистика $T(X)$, т.ч.
  $\forall\theta \in \Theta : \mathbb{E}_\theta T_n(X) \xrightarrow[n\to+\infty]{} \theta$

- **Состоятельная оценка** параметра θ – статистика $T(X)$, т.ч.
  $\forall\theta \in \Theta : T_n(X) \xrightarrow[n\to+\infty]{p} \theta$

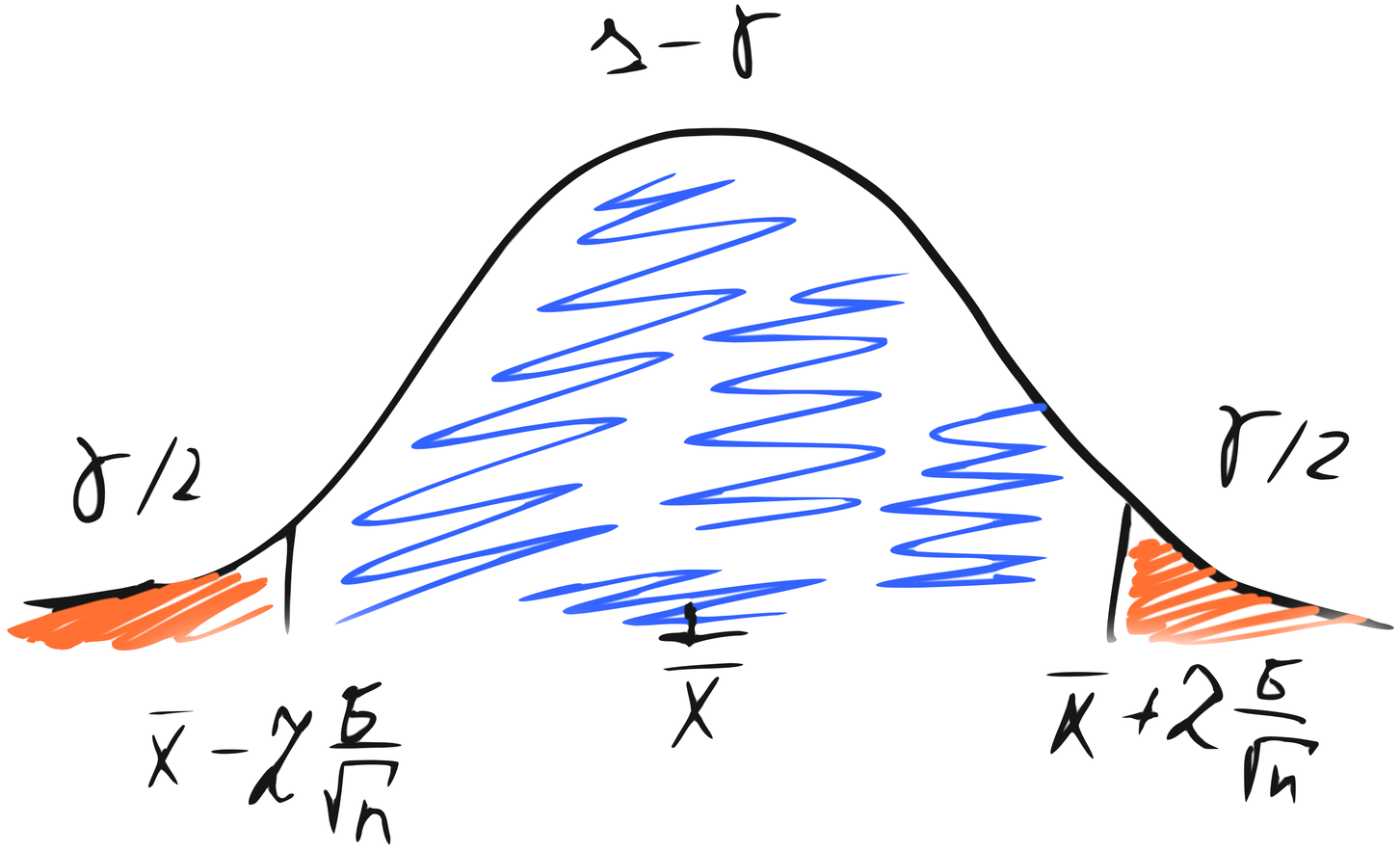
Sampling



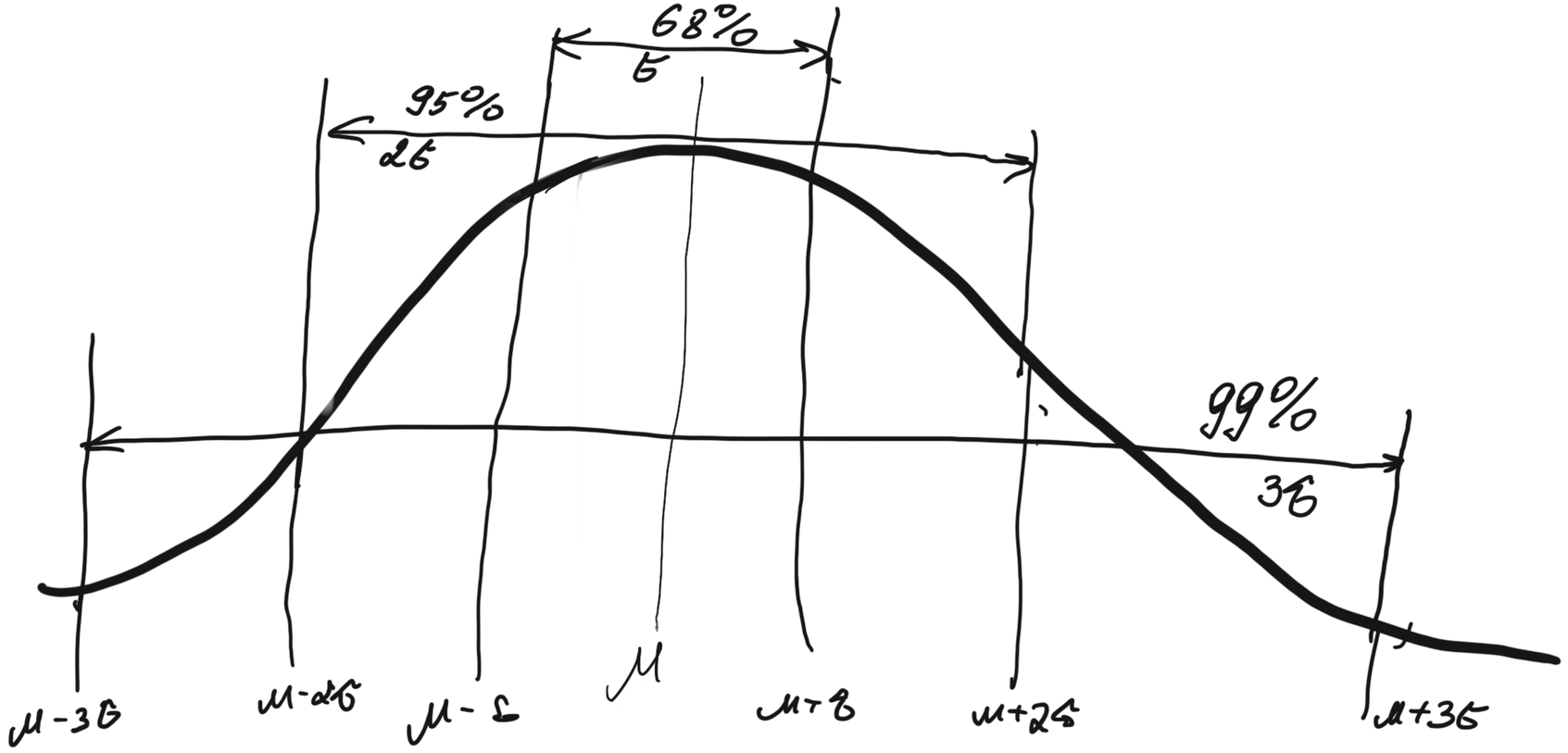
Доверительный интервал

- **Доверительный интервал** для параметра θ с **коэффициентом доверия** $\gamma \in (0;1)$ – интервал $(T_1(X), T_2(X))$, где T_i – статистика, т.ч.:
  - $T_1(X) \leq_{(\text{почти наверное})} T_2(X)$
  - $\mathbb{P}_\theta(T_1(X) \leq \theta \leq T_2(X)) \geq \gamma$

Доверительный интервал



Правило трёх σ



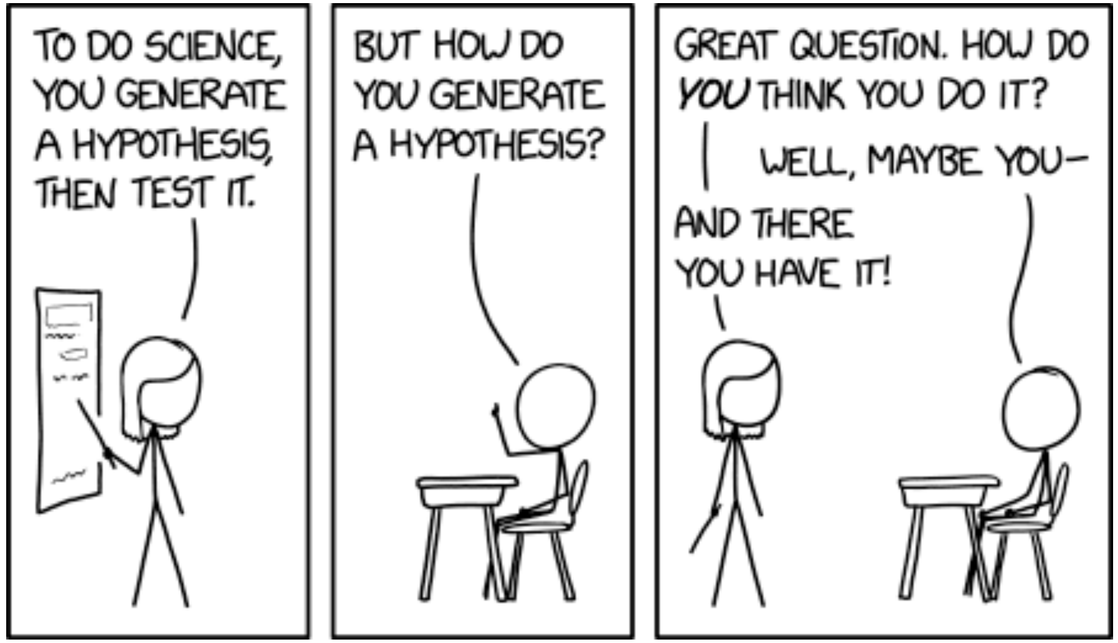
$$\mathbb{P}(\mu - \sigma \leq X \leq \mu + \sigma) \approx 0.6827$$

$$\mathbb{P}(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0.9545$$

$$\mathbb{P}(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 0.9973$$

Проверка статистических гипотез

Статистическая гипотеза H – любое предположение о распределении наблюдаемой случайной величины.



Проверка статистических гипотез

Как правило, рассматривается сразу две взаимоисключающие гипотезы. Одна из них называется основной и обозначается H_0, а другая — альтерна- тивной и обозначается H_1.

Проверка статистических гипотез

На понятном языке

Критерий — правило, согласно которому по выборке делается заключение о верности гипотезы.

На заумном языке

Критерий — статистика $\phi(X)$ со значениями из [0;1], трактуемая как вероятность отвергнуть H_0

Проверка статистических гипотез

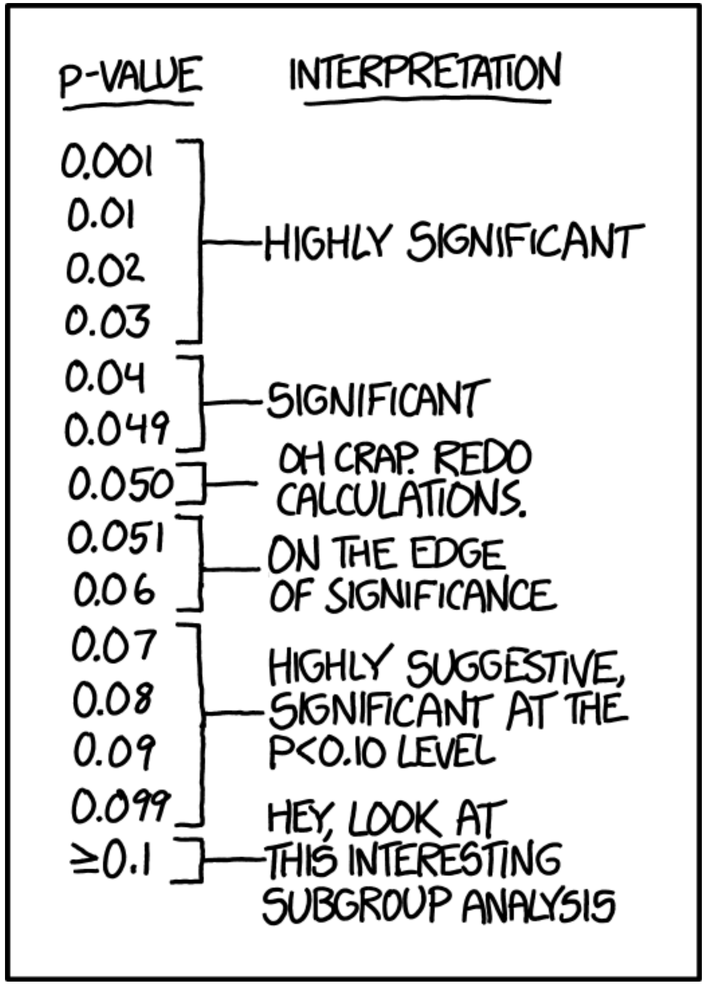
$S \subset \mathbb{R}^n : \mathbb{P}_\theta(X \in S) \leq \alpha, \alpha \in (0;1)$, альфа – критическая область По смыслу критическая область — это множество таких значений выборки, которые маловероятны при условии истинности H0

Истинная гипотеза	Результат принятия решения	
	H_0 принята	H_0 отклонена
H_0	$1-\alpha$	α
H_1	β	$1-\beta$

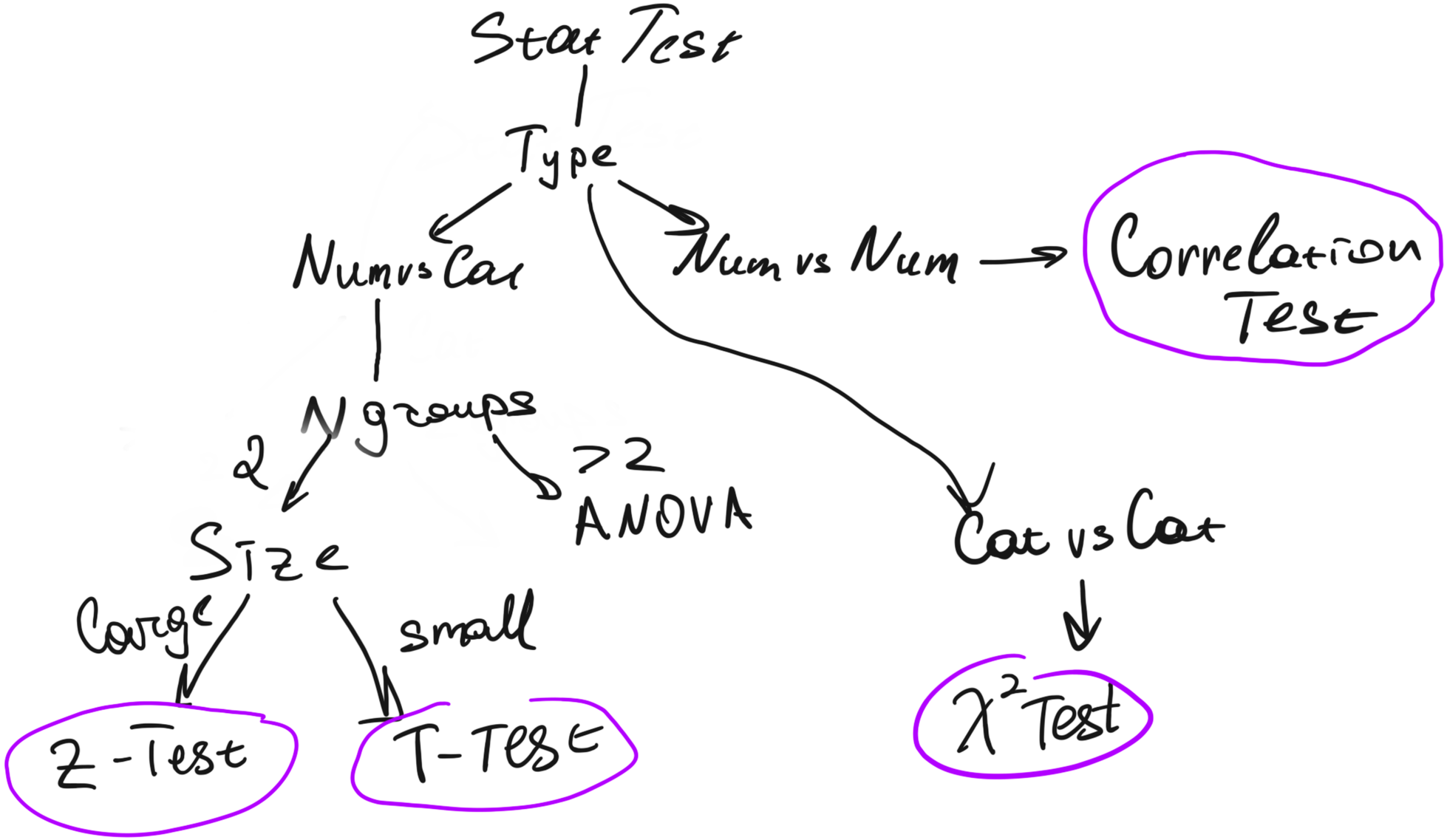
Вероятность отвергнуть верную H_0: $\alpha(S) = P_\theta(X \in S | H_0) = P_\theta(X \in S)$

Проверка статистических гипотез

P-value – расчётная вероятность получить значение статистики критерия равное наблюдаемому или более нетипичное по сравнению с наблюдаемым, если нулевая гипотеза верна.



Основные виды статтестов



T-test

Проверяет, что среднее двух независимых выборок значительно отличается для небольших выборок

Предположения о распределении:

- Наблюдения – НОРСВ
- Наблюдения распределены по стьюденту
- Наблюдения имеют одинаковую дисперсию

H_0: средние выборок равны

H_1: средние выборок неравны

Z-test

Проверяет, что среднее двух независимых выборок значительно отличается для больших выборок. Предполагаем, что работаем с нормальным распределением и известным σ

Предположения о распределении:

- Наблюдения – НОРСВ
- Наблюдения распределены нормально
- Наблюдения имеют одинаковую дисперсию

H_0: средние выборок равны

H_1: средние выборок неравны

ANOVA

Применяется аналогично предыдущим, но для > 2 групп.

Предположения о распределении:

H_0: средние выборок равны

H_1: средние выборок неравны

- Наблюдения – НОРСВ
- Наблюдения распределены нормально
- Наблюдения имеют одинаковую дисперсию
- Наблюдения имеют одинаковый размер

Correlation Test

Будем говорить о тесте корреляции Пирсона

H_0: отсутствует значительная связь

H_1: присутствует связь

Предположения о распределении:

- Наблюдения – НОРСВ
- Наблюдения распределены нормально
- Наблюдения если зависимы, то зависимы линейно

χ^2-test

Проверяет взаимосвязь двух категориальных признаков

H_0: отсутствует значительная связь

H_1: присутствует связь

Предположения о распределении:

- 25 и более размер выборки
- Наблюдения в таблице соответствий независимы