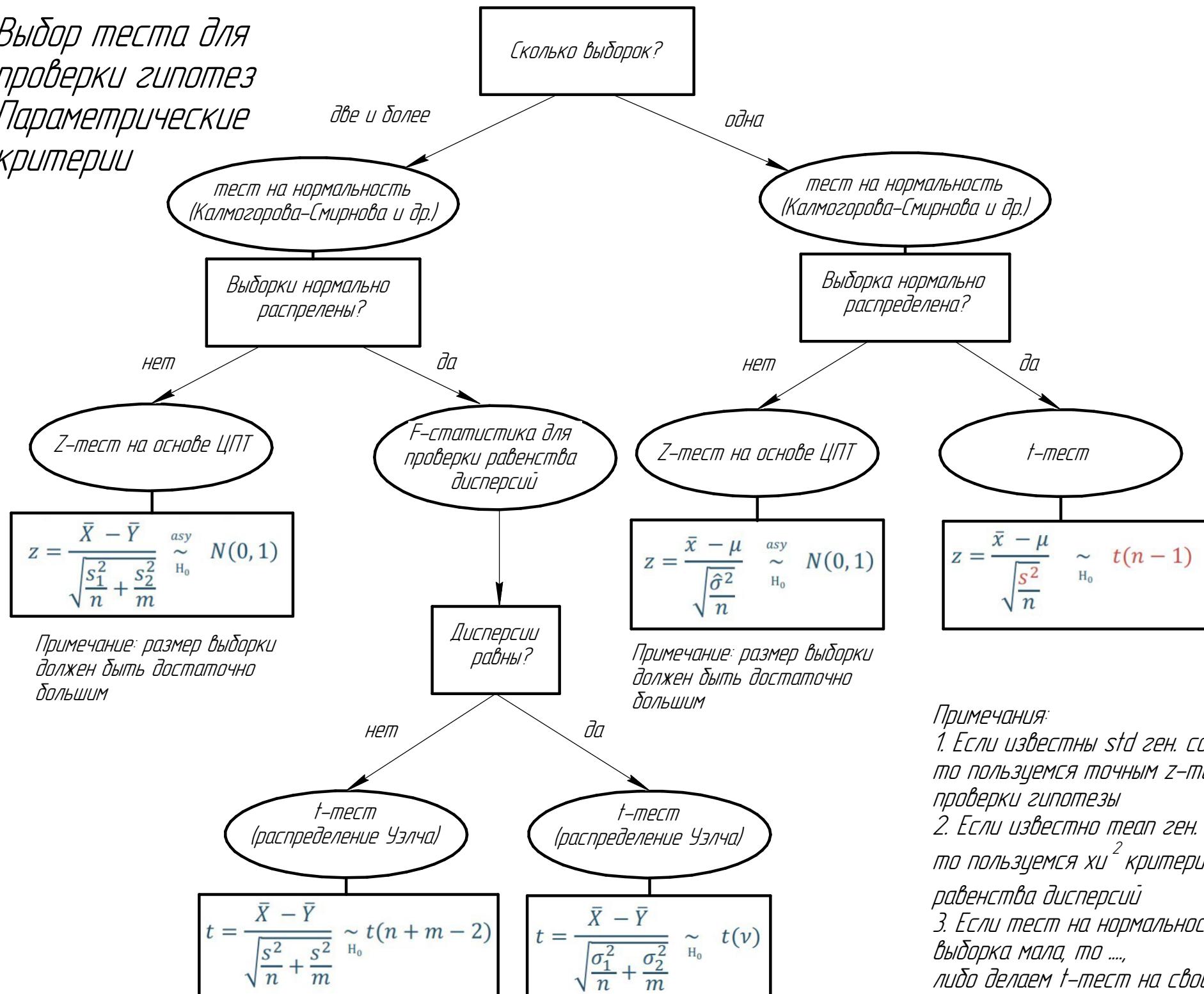


Выбор теста для проверки гипотез

Параметрические критерии



База для статистики

Выучить, чтобы от зубов отскакивало:
i.i.d. – independent, identically distributed



С ростом размера выборки среднее арифметическое сходится по вероятности к мат ожиданию ген. совокупности более формализованно:

$$\bar{X} = \frac{X_1 + \dots + X_n}{n} \xrightarrow{p} \mathbb{E}(X)$$

Примечание: сходимость по вероятности означает

$$\forall \varepsilon > 0 \quad \mathbb{P}(|X_n - X| < \varepsilon) \rightarrow 1 \text{ при } n \rightarrow \infty$$

то есть

$$\lim_{n \rightarrow \infty} \mathbb{P}(|X_n - X| < \varepsilon) = 1$$



Распределение суммы случайных величин при увеличении их количества имеет распределение, близкое к нормальному более формализованно:

$$\frac{X_1 + \dots + X_n}{n} \xrightarrow{d} N\left(\mathbb{E}(X), \frac{\text{Var}(X)}{n}\right)$$

Примечание: сходимость по распределению означает

$$\lim_{n \rightarrow \infty} F_{X_n}(x) = F_X(x)$$

то есть при увеличении п эмпирическая случайная величина стремится стать теоретической

Оценки

Ген. совокупность имеет мат ожидание и среднее квадратичное отклонение, выборка же не может всегда точно отражать ген. совокупность и наша задача состоит в оценке глобальных параметров, имея лишь ограниченную выборку.
Оценка – лишь функция, в которую мы вставляем наши случайные величины и получаем некий искомый параметр

Чего мы хотим добиться?

Как оценить оценку?
Как понять, что наша выборка хорошо отражает ген. совокупность



Метод моментов

1. Берем формулы для теоретических моментов,
2. Вставляем в них нашу выборку → получаем оценки,
3. Говорим, что оценки соответствуют моментам ген. совокупности

Метод макс правдоподобия

Ищем, при какой оценке вероятность получить такую же выборку максимальна
(Формула страшная, приводить ее здесь нет смысла)

Несмешенность

Это просто когда мат ожидание оцениваемого параметра равно оцениваемому параметру

$$\mathbb{E}(\hat{\theta}) = \theta$$

Смещение: $bias(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$

Если к определению добавить "при увеличении размера выборки", то получим асимптотическую несмешенность

Состоятельность

Это когда оценка сходится по вероятности к оцениваемому параметру

$$\hat{\theta} \xrightarrow{p} \theta$$

То есть это когда при увеличении выборки наша оценка выдает нам параметр, все более напоминающий глобальный

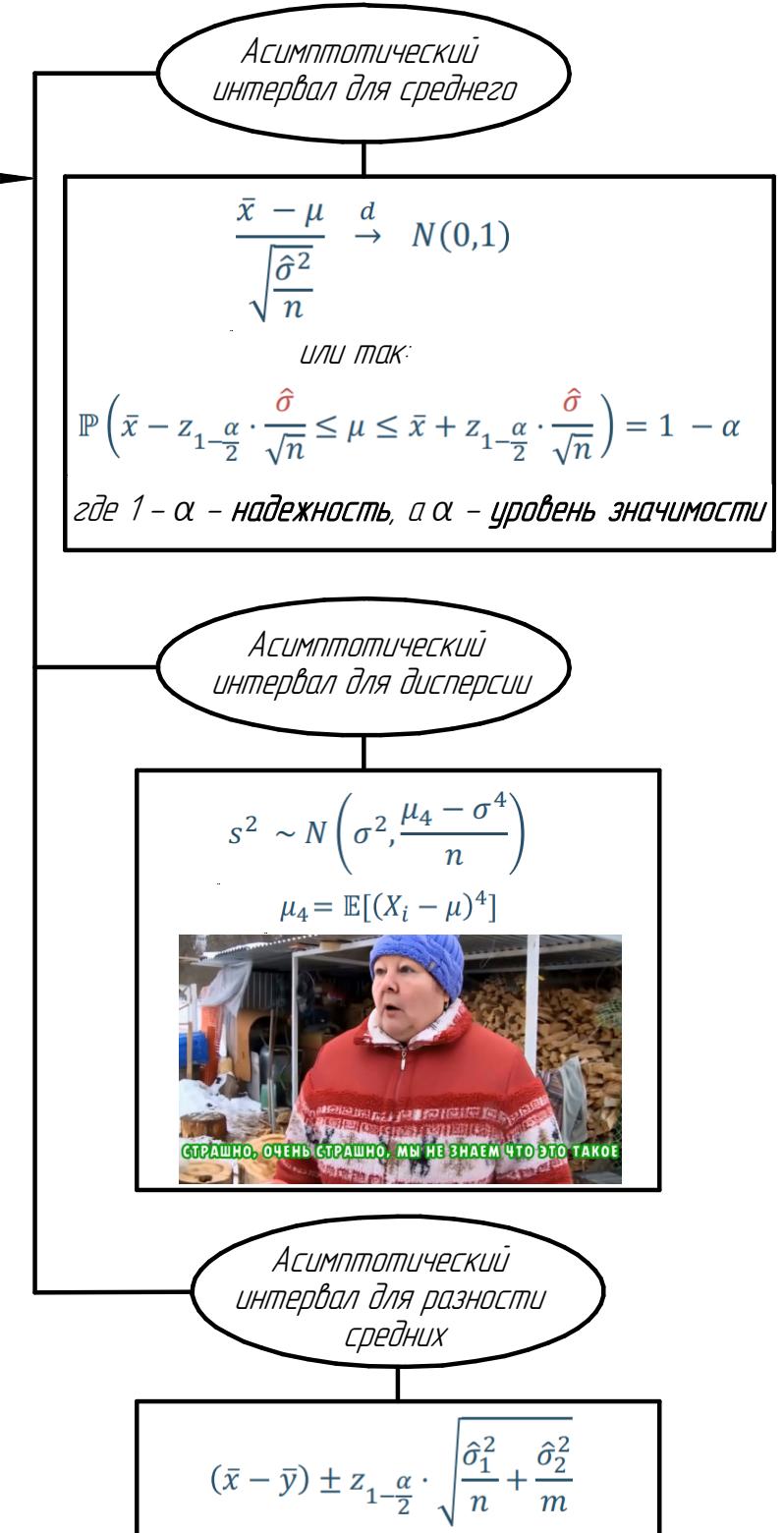
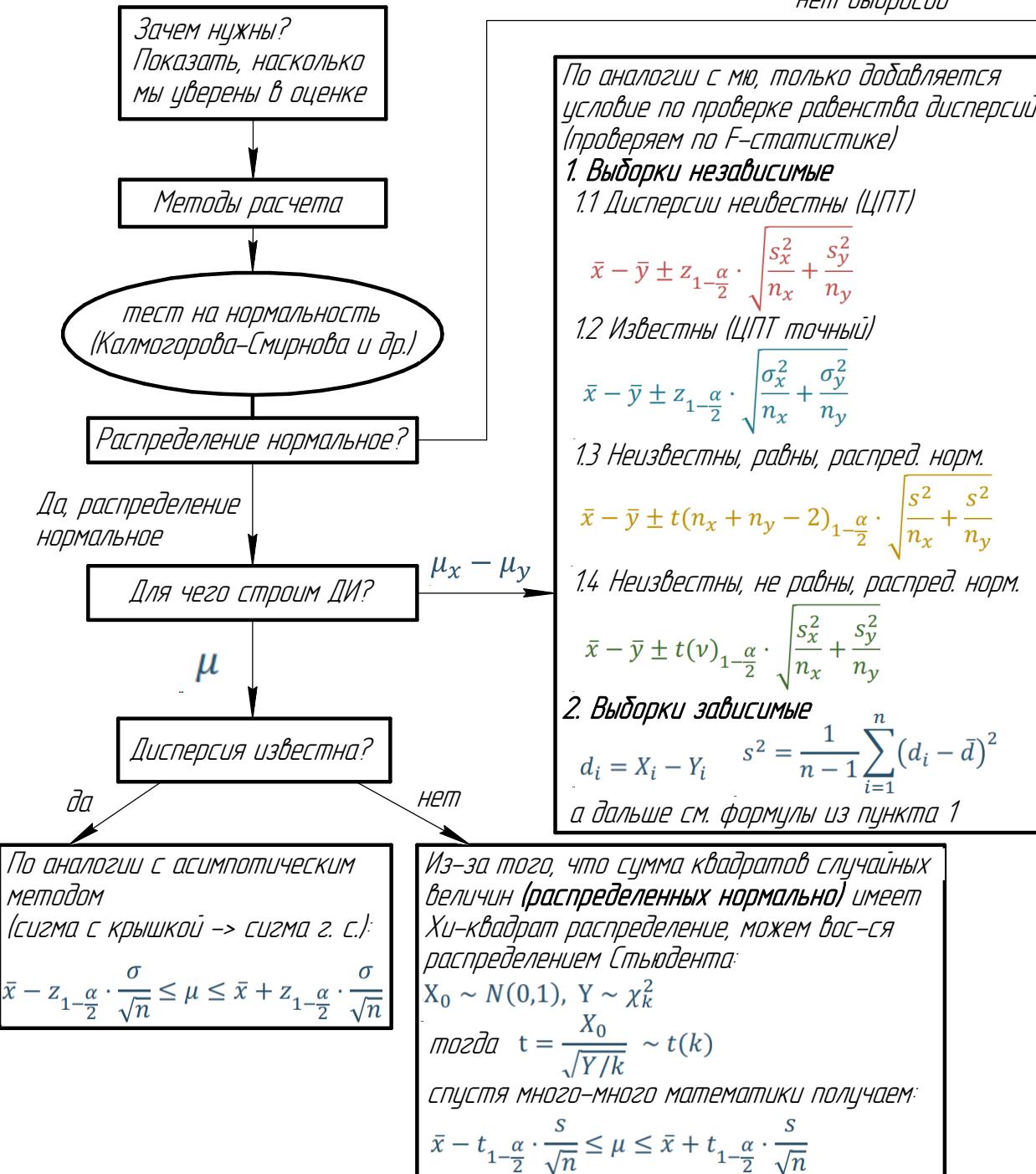
Эффективность

Фиксируем bias оценки и выбираем ту, у которой наименьшая дисперсия
→ эта оценка наиболее эффективная

В этом нам помогает среднеквадрат. ошибка

$$\begin{aligned} MSE &= \mathbb{E}(\hat{\theta} - \theta)^2 = \\ &= \text{Var}(\hat{\theta}) + \text{bias}^2(\hat{\theta}) \end{aligned}$$

Доверительные интервалы



Доверительные интервалы (дополнение)



Важное замечание для понимания:

При построении какого-либо доверительного интервала мы стремимся подобрать формулу таким образом, чтобы в ней фигурировало какое-либо распределение (нормальное, t -расп., хи квадрат и т.д.). Это необходимо для того, чтобы знать вероятность полученного интервала (то есть мы получаем интервал при заранее обозначенном уровне значимости α)

Примечание: выборки независимые
То самое распределение Фишера:

$$\frac{s_m^2 \cdot \sigma_m^2}{s_n^2 \cdot \sigma_n^2} \sim F_{n-1, m-1}$$

В итоге приходим к такому интервалу:

$$\frac{s_m^2}{s_n^2} \cdot F_{n-1, m-1} \left(\frac{\alpha}{2} \right) \leq \frac{\sigma_m^2}{\sigma_n^2} \leq \frac{s_m^2}{s_n^2} \cdot F_{n-1, m-1} \left(1 - \frac{\alpha}{2} \right)$$

Вот это теорема Фишера:

$$\frac{(n-1) \cdot s^2}{\sigma^2} \sim \chi_{n-1}^2$$

$$P\left(\frac{(n-1) \cdot s^2}{\chi_{n-1}^2(1 - \frac{\alpha}{2})} \leq \sigma^2 \leq \frac{(n-1) \cdot s^2}{\chi_{n-1}^2(\frac{\alpha}{2})}\right) = 1 - \alpha$$

Предисловие к проверке гипотез Суть проверки

Если есть утверждение (гипотеза), которая математически formalизована в некоторую задачу, то мы можем его проверить на выборке реальных данных, используя соответствующие стат. методы. Нулевая гипотеза H_0 – базовое проверяемое утверждение, например, из двух выборок выяснить, имеет ли эффект лекарство. За нулевую гипотезу можем принять, что эффекта нет.

Альтернативная гипотеза H_1 будет утверждать обратное

Ошибки при проверке

Изначально мы утверждаем уровень значимости α , который обозначает вероятность отвергнуть H_0 при ее верности. При этом бетта обозначает противоположную вероятность не отвергнуть H_0 , хотя она не верна, т.е.: $\alpha = \mathbb{P}(H_0 \text{ отвергнута} | H_0 \text{ верна})$

$\beta = \mathbb{P}(H_0 \text{ не отвергнута} | H_0 \text{ не верна})$

Примечания:

Альфа и бетта не равнозначны

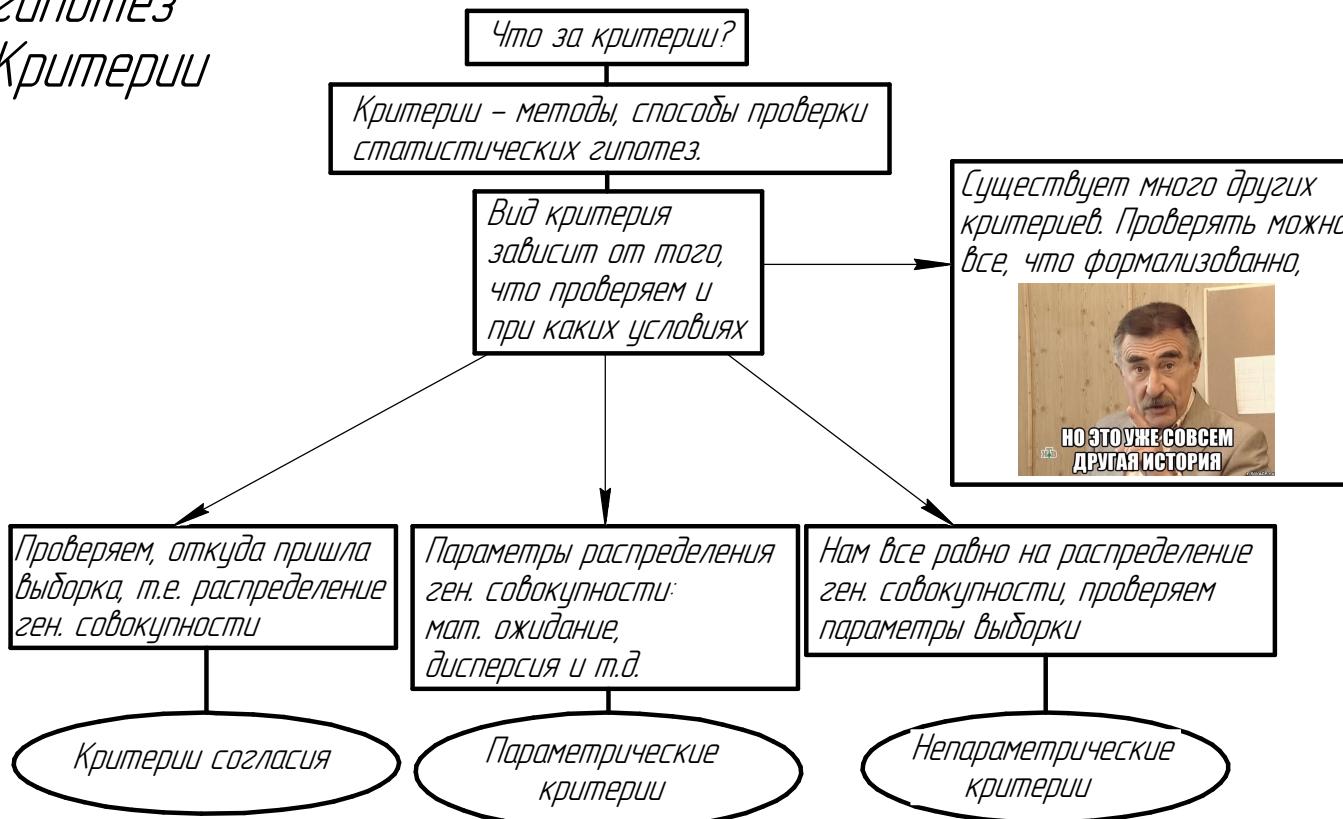
$1 - \beta$ называют мощностью критерия

Визуализация ошибок

$y = 1$	$y = 0$	
$\hat{y} = 1$	TP	FP
$\hat{y} = 0$	FN	TN
		ошибка 2 рода
		ошибка 1 рода

Предисловие к проверке гипотез

Критерии



Общий алгоритм проверки гипотез

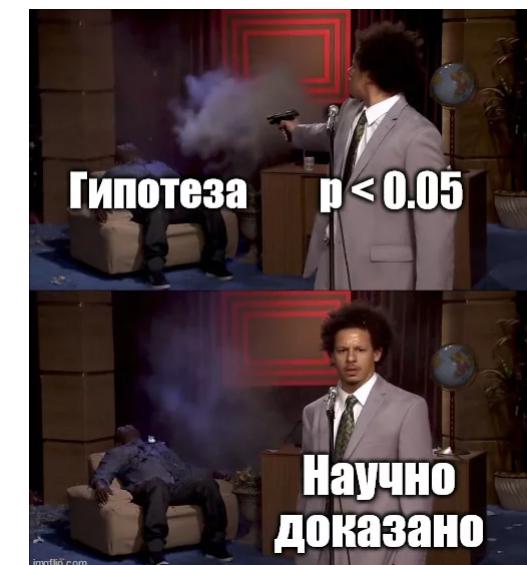
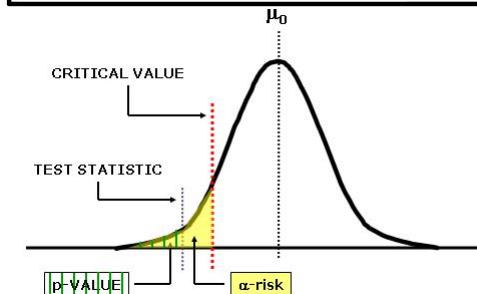
Формулирование гипотез
Например, нулевая гипотеза H_0 утверждает отрицание эффекта от лекарства.
Альтернативная гипотеза H_1 утверждает, что эффект есть (причем можно выбрать, отрицательный это эффект или положительный)

Прим.: на данном этапе мы также понимаем, тест из какого критерия нам использовать

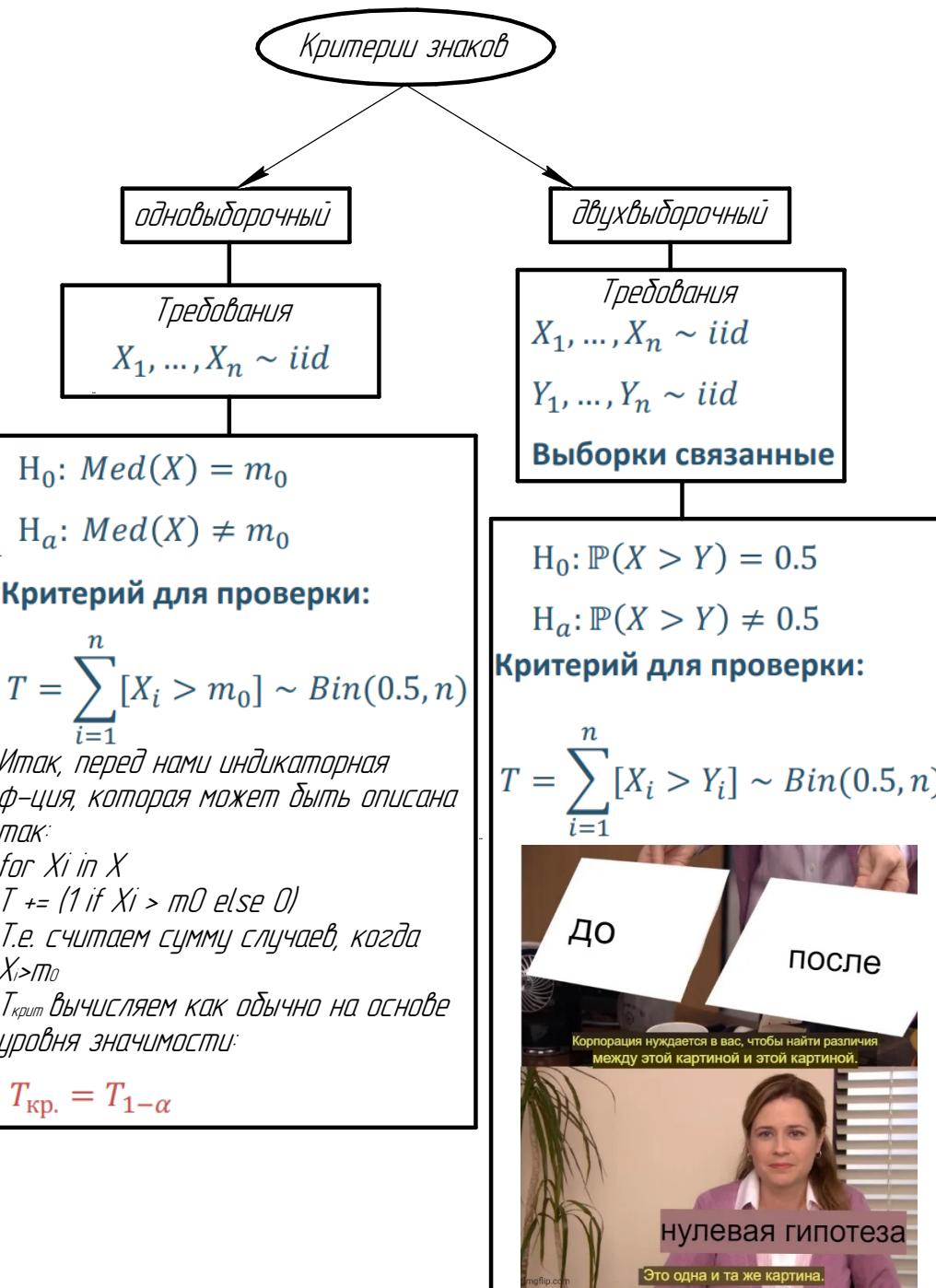
Выбор теста и уровня значимости
Следуя алгоритму, выбираем тест внутри принятого критерия
Параметр α обычно принимается за 5%

Прим.: следует выбирать самый мощный тест

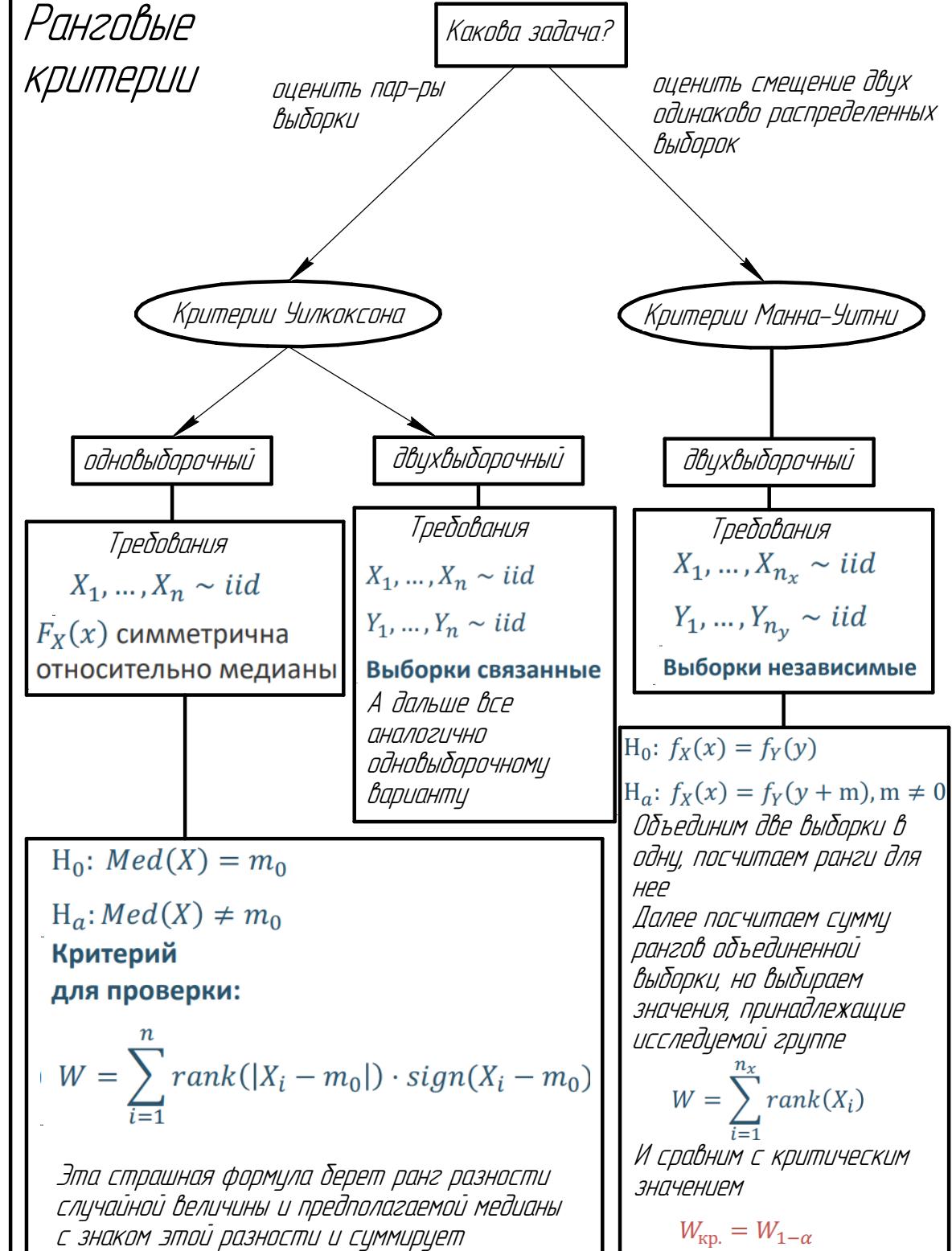
p -value
Проведя тест, получаем p -value, исходя из которого отвергаем или не отвергаем нулевую гипотезу H_0 . Если p -value меньше α , то отвергаем, иначе принимаем



Непараметрические критерии



Ранговые критерии



Критерии согласия

Какова задача?

Задача: понять, из какого распределения пришла выборка, т.е. сравнить распределение выборки с теоретическим распределением

Задача: понять, пришли ли выборки из одного и того же распределения

Задача: проверить нормальность (критериями помимо тех, что слева)

для непрерывных рас-ний
для непрерывных рас-ний
для дискретных рас-ний

Критерий Колмогорова

Критерий Крамера-Мизеса

Критерий Пирсона

Критерий Колмогорова-Смирнова

Находит наибольшее расстояние между функциями распределения. Далее это расстояние сравнивается с критическими значениями, зависящими от уровня значимости.

Формализованная задача:

$$X_1, \dots, X_n \sim iid F_X(x)$$

$$H_0: F_X(x) = F_0(x)$$

$$H_a: F_X(x) \neq F_0(x)$$

Вводится так называемая функция штрафа, которая фиксирует разницу. Чем больше разница, тем больше штраф.
Формализация задачи та же

$$\int_{-\infty}^{\infty} 1 \cdot [F_0(x) - \hat{F}_n(x)]^2 \cdot f_0(x) dx$$

$$H_0: F_X(x) = F_0(x)$$

$$H_a: F_X(x) \neq F_0(x)$$

X	z_1	z_2	...	z_s
$\mathbb{P}(X = z)$	$p_1(\theta)$	$p_2(\theta)$...	$p_s(\theta)$
#($X_i = z$)	v_1	v_2	...	v_s

(1) - возможные значения (интервалы)

(2) - теоретические вероятности

(3) - эмпирические частоты

$$\sum_{j=1}^s \frac{(v_j - n \cdot p_j(\hat{\theta}))^2}{n \cdot p_j(\hat{\theta})} \xrightarrow[H_0]{asy} \chi^2_{s-k-1}$$

Нужно сравнить все эмпир. частоты с теоретическими. Аналогично сравниваем эмпирику с эмпирикой

$$H_0: F_X(x) = F_Y(x)$$

$$H_a: F_X(x) \neq F_Y(x)$$

Снова ищем supremum, но уже между эмпирическими ф-ями

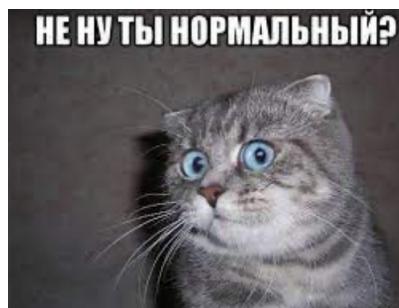
$$D_n = \sup_x |\hat{F}_X(x) - \hat{F}_Y(x)|$$

$$\sqrt{\frac{n_x \cdot n_y}{n_x + n_y}} \cdot D_n \quad \text{при } n_x, n_y \rightarrow \infty$$

имеет распределение Колмогорова, на котором проверяется однородность выборок

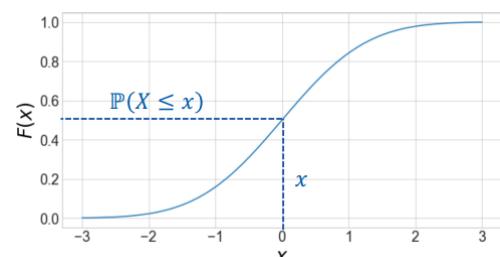
1. Критерий Лиллиегорса
2. Критерий Харке-Бера
3. Критерий Шапиро-Чилка

Самый мощный критерий - Шапиро-Чилка, сперва проверяем им.
Но этот критерий часто ошибается из-за выбросов, поэтому есть смысл знать другие критерии.



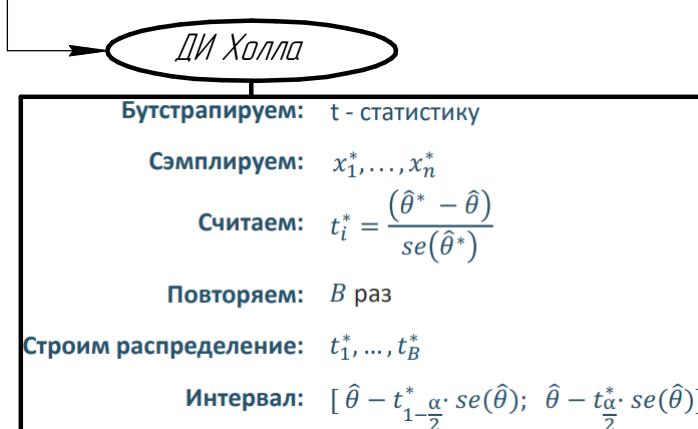
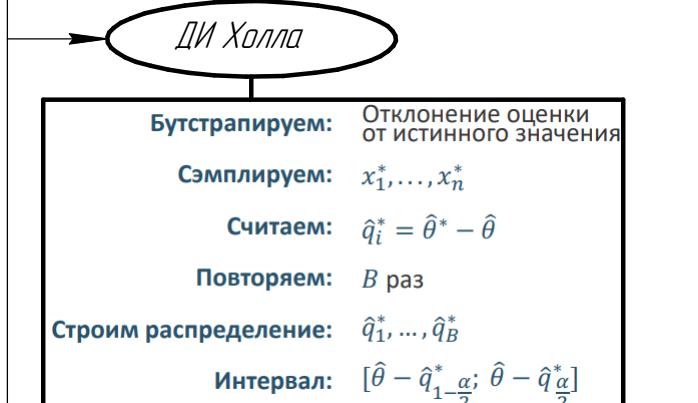
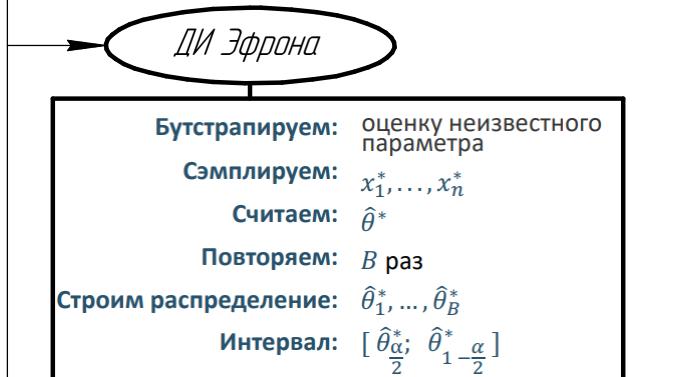
Примечание: Из теор. вер.: Функция распределения - ф-ция, которая определяет вероятность события $X \leq x$, она же интеграл плотности распределения от -беск. до x , т.е.:

$$F(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^x f(t) dt$$



Бутстррап

Бутстррап – это метод, при котором мы принимаем за "ген. совокупность" нашу выборку. Из этой "ген. совокупности" берем B "псевдоизборок" объема n и оцениваем интересующие параметры, строим ДИ
Выборка должна быть репрезентативной



Множественная проверка гипотез

Family-Wise Error Rate (FWER) – это вероятность совершить хотя бы одну ошибку первого рода
 $FWER = \mathbb{P}(V > 0)$

V – кол-во ошибок первого рода при проверке p гипотез.

	верных H_0i	неверных H_0i
не отвергнутых H_0i	<i>U</i>	<i>T</i>
отвергнутых H_0i	<i>V</i>	<i>S</i>

False Discovery Rate (FDR) – ожидаемая доля ложных отклонений или ожидаемая доля ошибок 1 рода

$$FDR = \mathbb{E}\left(\frac{V}{V + S}\right)$$



Проблема: накопленная вероятность ошибки первого рода увеличивается с каждой проверкой гипотезы, чем больше проверок, тем меньше вероятность не ошибиться при каждой проверке.

Суммарная вероятность ошибиться: $1 - (1 - \alpha)^n$

Решение: нужно делать корректировку уровня значимости для проверки каждой гипотезы



*Схожий, более мощный метод Бенджамина–Хохберга: сортируем $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(k)}$ берем $\alpha_{(1)} = \frac{\alpha}{k}, \alpha_{(2)} = \frac{2\alpha}{k}, \dots, \alpha_{(i)} = \frac{i\alpha}{k}, \dots, \alpha_{(k)} = \alpha$
Таким образом контролируем $FDR \leq FWER$



Линейная регрессия

Модель парной регрессии – зависимость вида

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$$

y_i – значения зависимой переменной

x_i – значения независимой переменной (регрессора)

ε_i – случайные ошибки

Алгебра: модель множественной регрессии

$$y_i = \beta_1 + \beta_2 * x_i^{(2)} + \beta_3 * x_i^{(3)} + \dots + \beta_k * x_i^{(k)} + \varepsilon_i$$

y_i – зависимая (объясняемая) переменная

$x_i^{(m)}$ – объясняющие переменные (регрессоры)

ε_i – случайные ошибки

k – число коэффициентов в модели

n – число наблюдений

Линейная регрессия – способ найти линейную зависимость некоторой переменной u от одного и более регрессоров x .

Разумеется, получить истинную зависимость не получится, поэтому

оценим коэффициенты β_i

Для того, чтобы оценки $\hat{\beta}_i$ были максимально близки к истинным значениям, должны выполняться предпосылки классической модели парной (множественной) регрессии (КЛМПР или КЛММР)

Примечание: МНК – метод наименьших квадратов. Суть метода в минимизации фактических значений y_i от \hat{y}_i

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \rightarrow \min.$$

$$\frac{\partial SSE}{\partial a} = -2 \sum_{i=1}^n (y_i - a - bx_i) = 0$$

$$\frac{\partial SSE}{\partial b} = -2 \sum_{i=1}^n (x_i)(y_i - a - bx_i) = 0$$

Решаем систему, получаем коэф-ты

Качество подгонки модели

Как оценить, какая модель работает лучше?

Стандартная ошибка регрессии

$$SEE = \sqrt{S^2} = \sqrt{\frac{1}{n-k} * \sum_{i=1}^n e_i^2}$$

При помощи SEE можно сравнивать между собой модели с одинаковой зависимой переменной, но с разным набором регрессоров

Коэффициент детерминации R^2

$$TSS = \sum_{i=1}^n (y_i - \bar{y})^2$$
 общая сумма квадратов
(total sum of squares)

$$ESS = \sum_{i=1}^n e_i^2$$
 сумма квадратов остатков
(error sum of squares)

$$RSS = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$
 объясненная сумма квадратов
(regression sum of squares)

Коэффициент детерминации показывает долю дисперсии зависимой переменной, полученной с помощью модели

$$R^2 = \frac{\text{Var}(\hat{y})}{\text{Var}(y)} = \frac{RSS}{TSS} = 1 - \frac{ESS}{TSS}$$

Чем лучше модель, тем ближе R^2 к единице

Используется в тестах незначимости уравнения

Примечание: лучше использовать при сравнении моделей с одинаковым количеством регрессоров

Скорректированный (нормированный) R^2_{adj}

$$R^2_{adj} = R^2 - \frac{k-1}{n-k} * (1 - R^2)$$

Если сравниваем модели с разным количеством регрессоров – используем скорректированный коэф

Используется в тестах сравнения "длинной" и "короткой" регрессии

КЛМПР/КЛММР

1. Модель линейна по параметрам и правильно специфицирована (т.е. нет еще регрессоров, которые влияют на u и одновременно коррелируют с x).

2. Регрессоры – детерминированные, т.е. случайные величины

3. Мат. ожидание случайных ошибок равно нулю: $E(\varepsilon_i) = 0$

4. Случайные ошибки имеют постоянную дисперсию:

$$V(\varepsilon_i) = \sigma^2 = \text{const}$$

5. Случайные ошибки, соответствующие разным наблюдениям не зависят друг от друга: $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ при $i \neq j$

6*. Случайные ошибки имеют нормальное распределение

Теорема Гаусса – Маркова

Если выполняются предпосылки 1–5, то полученные оценки коэф-ов (по МНК) являются

(а) несмещенными
(б) эффективными, т.е. имеет наименьшую дисперсию в классе всех линейных по u несмещенных оценок

Временные ряды

Временной ряд – это последовательность измерений переменной во времени. График погоды в приложении, график цены акции на бирже и т.п. являются примерами временного ряда. Чуть более формально: ВР – последовательность **зависимых** случайных величин во времени с различными распределениями.

Для достижения катарсиса при изучении ВР следует сразу определиться с обозначениями

$$\Delta Y_t = Y_t - Y_{t-1}$$

где Y_t – значение переменной Y в период t

Y_{t-1} – значение переменной Y в период $t-1$, или 1 лаг относительно момента t

ΔY_t – первая разность переменной

$\frac{\Delta Y_t}{Y_{t-1}}$ – темп прироста

$L^k Y_t = Y_{t-k}$ лаговый оператор (получаем k -ый лаг)

Невероятно но факт: за t мы принимаем текущий момент времени. При исследовании ВР мы пользуемся лагами, просто потому что значения переменной Y_{t+1} мы не знаем (не знаем будущего) и решаем задачу ретроспективно.

Белый шум

Это нестационарный временной ряд из независимых случайных величин.

Мат. ожидание белого шума = 0

Дисперсия БШ или WN конечна

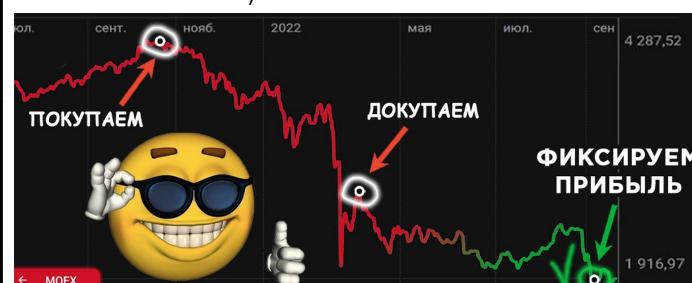
Стационарен ли белый шум?

Да, ведь его статистики не зависят от времени :)

Зачем он нужен?

Часто между предсказанными значениями и наблюдаемыми данными существуют небольшие отклонения – остатки. Если предположить, что остаточные ошибки независимы и гомоскедастичны, то наиболее естественным образом это отражается добавлением белого шума

Зачем нам изучать ВР? Конечно же, чтобы попытаться их предсказать! Но как это сделать?



Процесс случайного среднего MA

Этот процесс описывает ВР, где каждый новый элемент зависит от случайных возмущений в прошлом.

Например, сегодняшняя цена акции может зависеть не только от текущих новостей, но также частично от неожиданных изменений вчерашнего дня и позавчерашнего. А вот и формула MA(q):

$$y_t = \delta + \varepsilon_t + \alpha_1 \varepsilon_{t-1} + \dots + \alpha_q \varepsilon_{t-q}$$

Процесс ARMA(p,q)

Возьмем дальнейшность AR и точность MA и получим ARMA(p,q):

$$y_t = \delta + \theta_1 y_{t-1} + \dots + \theta_p y_{t-p} + \varepsilon_t + \alpha_1 \varepsilon_{t-1} + \dots + \alpha_q \varepsilon_{t-q}$$

Модель авторегрессии первого порядка AR(1)

Звучит страшно, а выглядит вполне несложно:

$$y_t = \delta + \theta y_{t-1} + \varepsilon_t$$

y_t – AR(1)

δ и θ – самые обычные коэффициенты

ε_t – белый шум

Аналогично

Модель авторегрессии порядка p AR(p)

$$y_t = \delta + \theta_1 y_{t-1} + \dots + \theta_p y_{t-p} + \varepsilon_t$$

С лаговым оператором будет выглядеть так:

$$y_t = \delta + \theta_1 L y_t + \dots + \theta_p L^p y_t + \varepsilon_t$$

Проверяем стационарность (цикличность ВР)

Примечание: стационарность – свойство ВР, при котором его статистики (мат. ожидание, дисперсия) не зависят от времени. Перепишем ур-ние AR(p):

$$(1 - \theta_1 L - \dots - \theta_p L^p) y_t = \varepsilon_t$$

Характер. ур-ние: $1 - \theta_1 z - \dots - \theta_p z^p = 0$

Теорема: Процесс AR(p) является стационарным тогда и только тогда, когда все корни хар. ур-ния по модулю больше 1, т.е. $|z_j| > 1$

Примечание: Почему корни должны по модулю быть больше 1? Потому что иначе коф. Θ будут > 0 и не будут "гасить" потенциальные шоки ВР \rightarrow ВР не будет стационарным

Процесс ARIMA(p,k,q)

На практике не всегда ВР стационарен сразу. А вот если взять k -ую разность для Y , то когда-нибудь мы таки придем к стационарному ВР. Плохо интерпритируем! Пример ARIMA(1,1,0):

$y_t = 0,9 y_{t-1} + 0,1 y_{t-2} + \varepsilon_t$ – не стационар. Берем первую разность:

$$\Delta y_t = -0,1 \Delta y_{t-1} + \varepsilon_t \quad \text{– стационарный!}$$

Прогнозирование

Итак, зная, y_t , y_{t-1} ... найдем мат. ожидание y_{t+1} : $E(y_{t+1}|y_t) = E(\Theta y_t + \varepsilon_{t+1}|y_t) = \Theta y_t + E(\varepsilon_{t+1}|y_t) = = \Theta y_t$

Т.е. зная предыдущие значения y_t мы спрогнозировали будущий момент y_{t+1} ! Вот он, момент катарсиса!