Современные проблемы прикладной математики и наукоемкого программного обеспечения

Модуль
«Устойчивые методы оценивания параметров статистических моделей»

Преподаватель

Лисицин Даниил Валерьевич (кафедра ТПИ)

Правила аттестации

50 (всего) = 40 (практика) + 10 (зачет)

40 (практика) = 20 (выполнение) + 20 (защита)

20 (защита) + 10 (зачет) =

20 (зачет + мин. защита автоматом)

+ 10 (доп. вопрос на защиту)

Штраф за перебор релизов:

$$\text{Штраф}(К) = 2^{К-1},\ К>2$$

Основные понятия математической статистики

Математическая статистика.

Генеральная совокупность. Выборка. Выборочные характеристики. Эмпирическое распределение. Параметрическая модель.

Задача оценивания неизвестных параметров. Точечные оценки.

Оценка как случайная величина, распределение оценки, математическое ожидание и дисперсия оценки. Свойства оценки: состоятельность, несмещенность, асимптотическая несмещенность, эффективность, асимптотическая эффективность, асимптотическая нормальность.

Метод максимального правдоподобия.

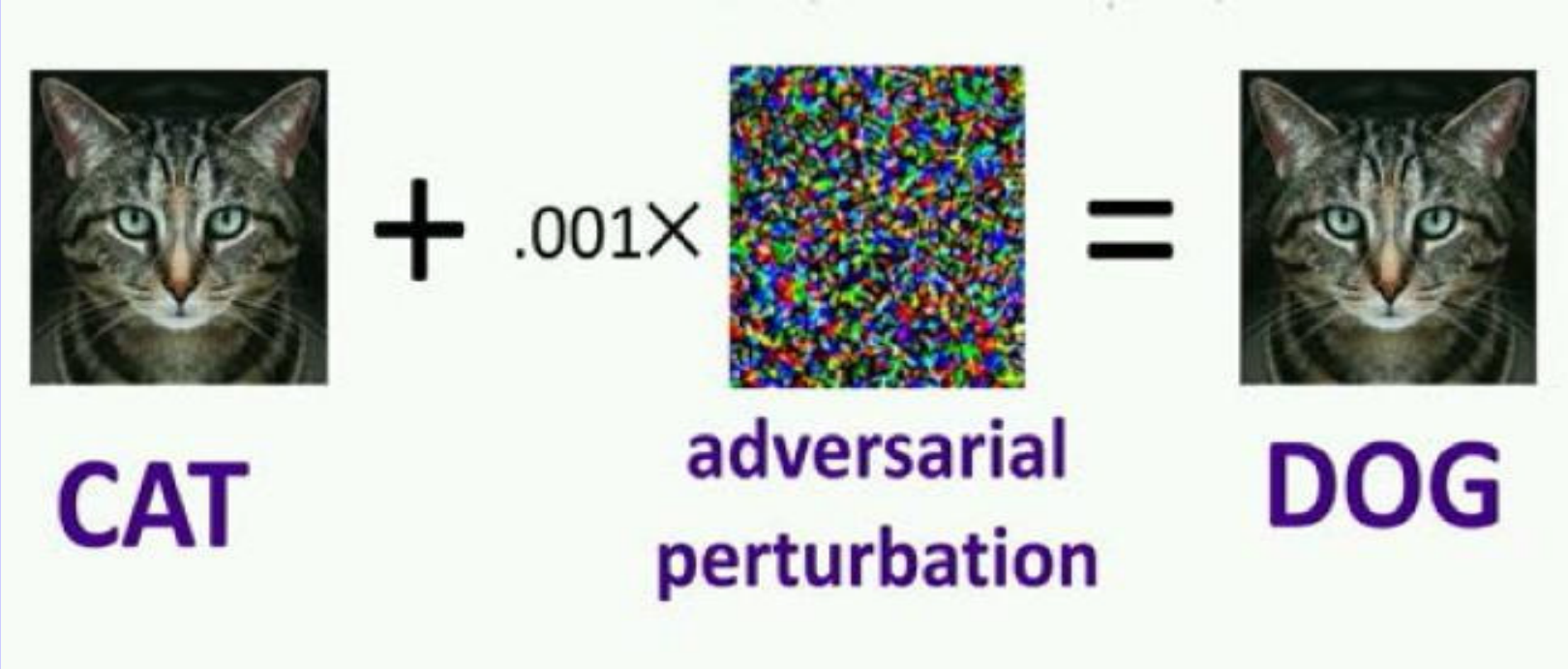
Бочаров П.П., Печинкин А.В. Теория вероятностей. Математическая статистика. – М.: ФИЗМАТЛИТ, 2005. *Часть II. Разделы 1.1–1.3, 2.1, 2.4.*

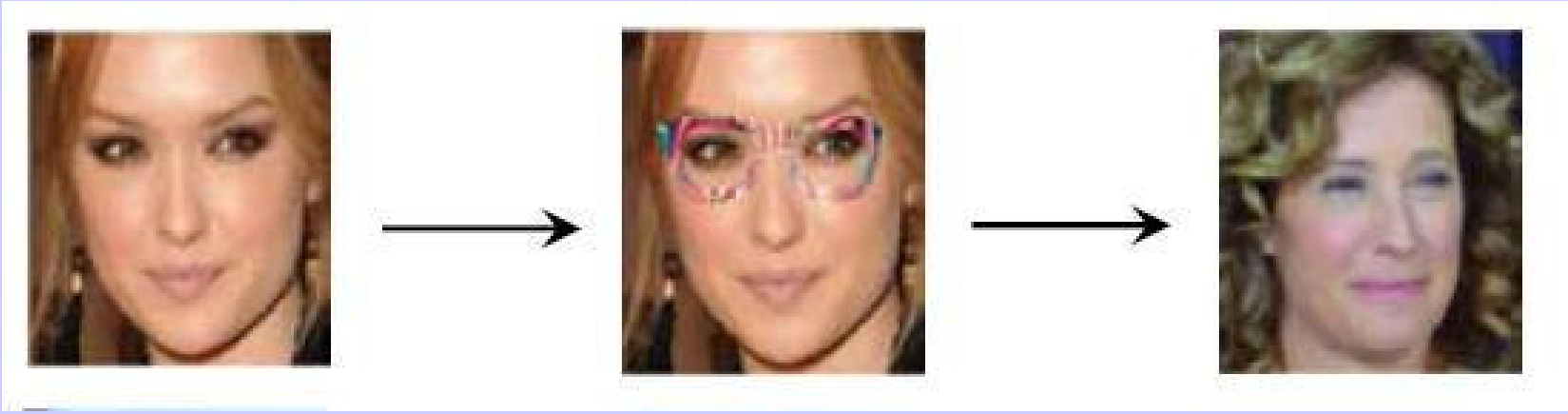
Горяинов В.Б. и др. Математическая статистика. – М.: Изд-во МГТУ им. Н.Э. Баумана, 2002. *Разделы 1, 2.1, 2.3.*

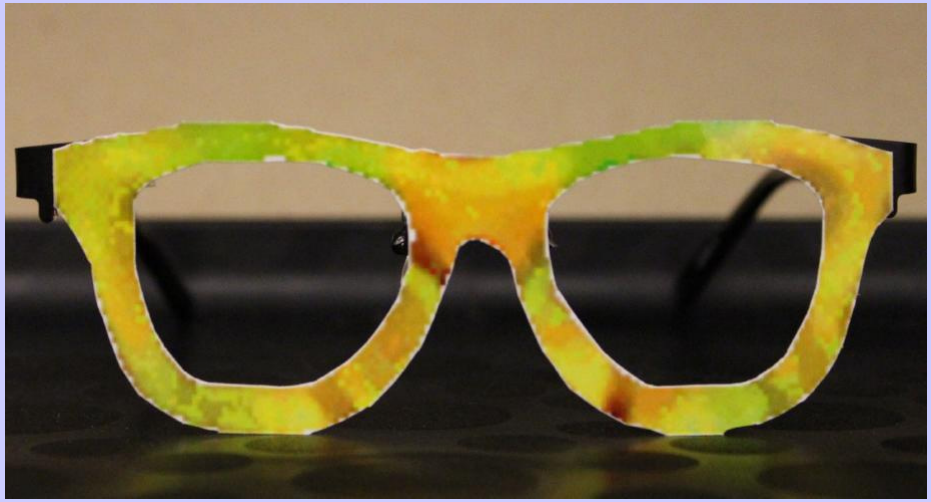
0. Необходимость в устойчивых методах моделирования

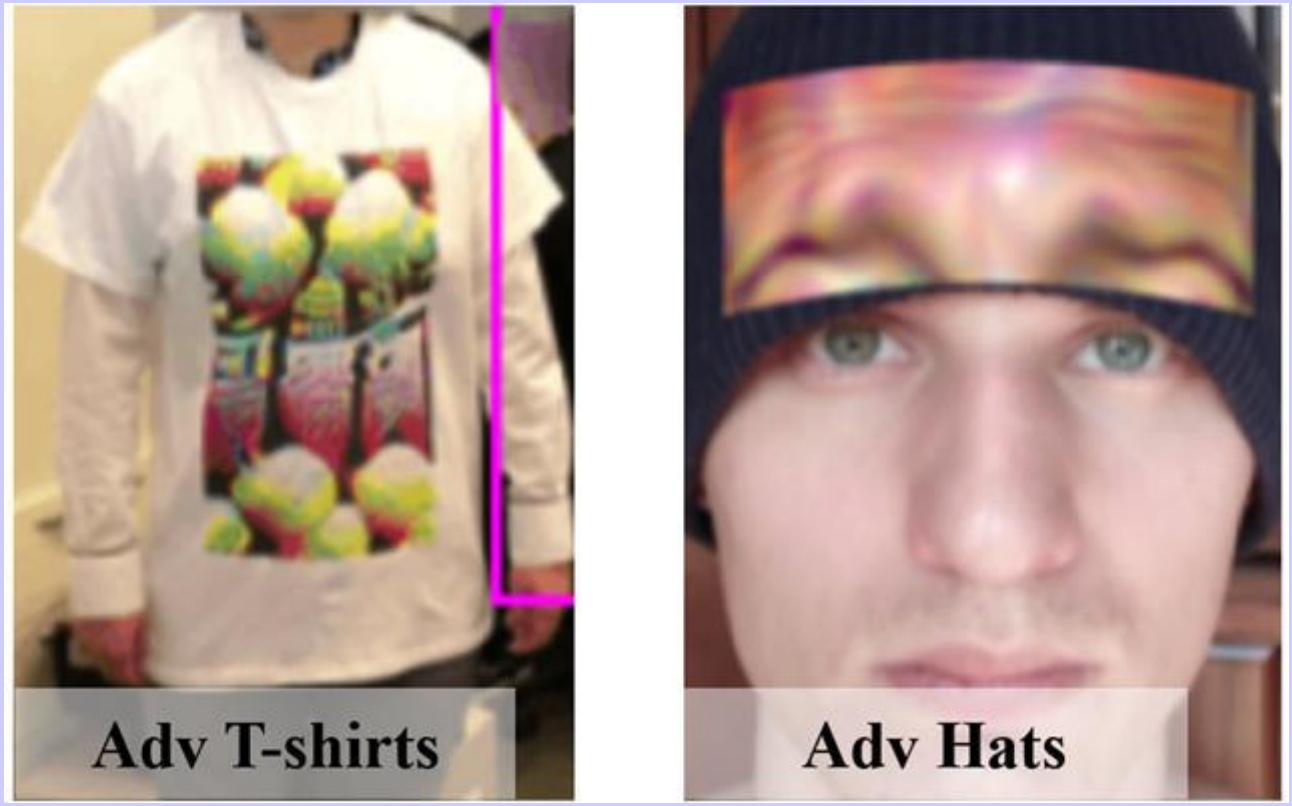
0.1. Искусственный интеллект. Искусственные нейронные сети. Вредоносное машинное обеспечение

Атаки уклонения









Футболка-невидимка, шапка-невидимка

Автоматизированная камера видеонаблюдения пользуется алгоритмами машинного обучения, в режиме реального времени следя за людьми, входящими и покидающими здание. Мимо здания проходит человек в футболке/шапке, однако камера его не обнаруживает: на футболку/шапку нанесен принт, скрывающий человека от камеры.

Атаки черного входа

Классификатор дорожных знаков может ошибочно определить знак «Стоп» в качестве знака ограничения скорости 60 км/ч. На сам знак остановки для этого наклеены специальные стикеры и/или нанесены изображения, маскирующиеся под граффити.

Атаки отравления

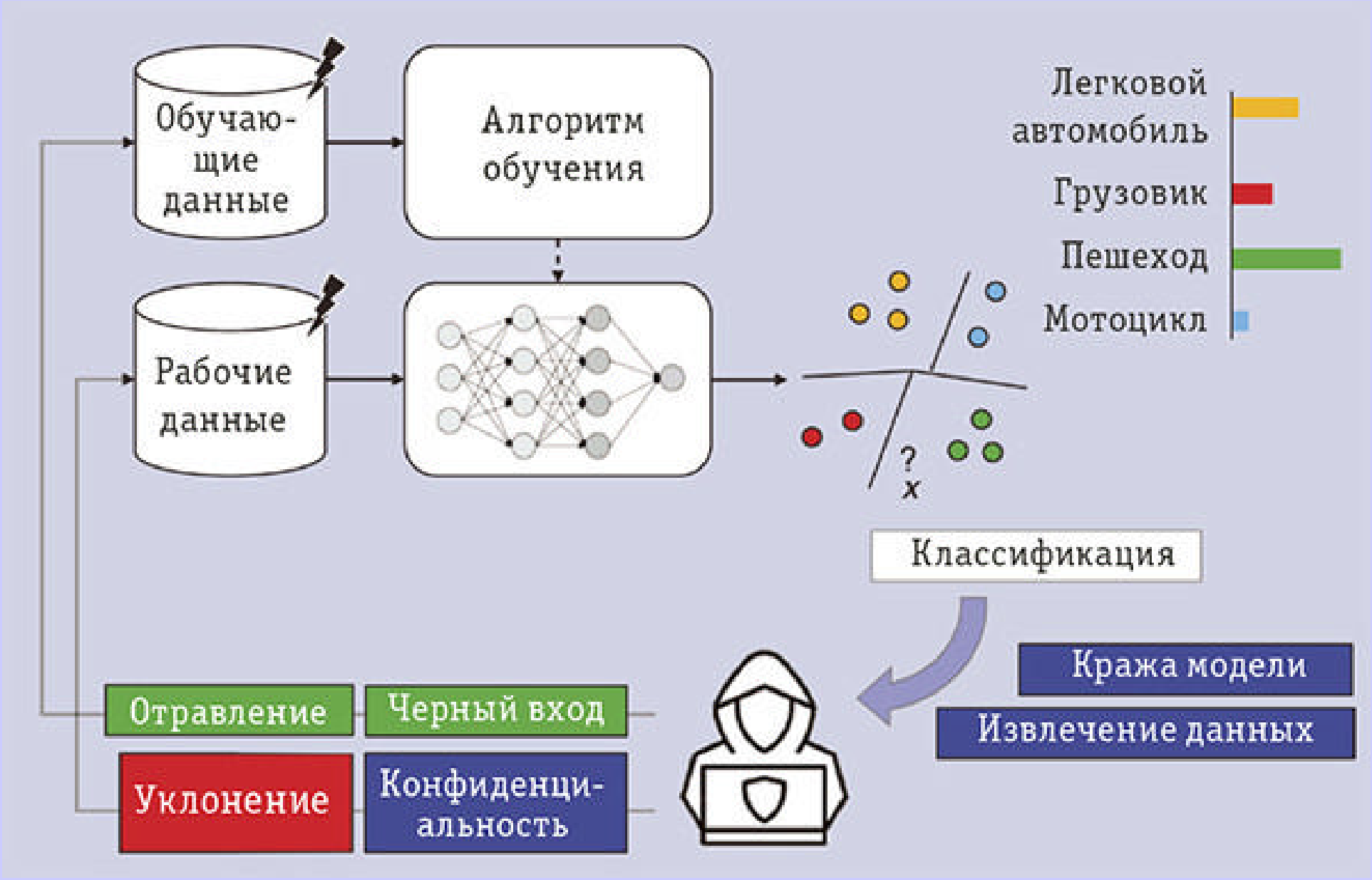
Фильтры спама. Одним из первых реальных объектов атак стали системы фильтрации спама.

Антиспам-фильтры, работающие с помощью машинного обучения, со временем совершенствуются, запоминая реакции пользователя, который отмечает пропущенные сообщения в качестве нежелательных или, наоборот, переносит в ящик входящих сообщения, расцененные фильтром как спам.

Атакующие могут поставить себе на службу процесс обучения спам-фильтра, изменяя содержание нежелательных сообщений, например, путем внесения в них слов, которые обычно присутствуют в легитимных письмах, но отсутствуют в спамерских. В конечном итоге это приводит к тому, что фильтр неверно классифицирует легитимные сообщения с такими словами в качестве потенциально нежелательных.

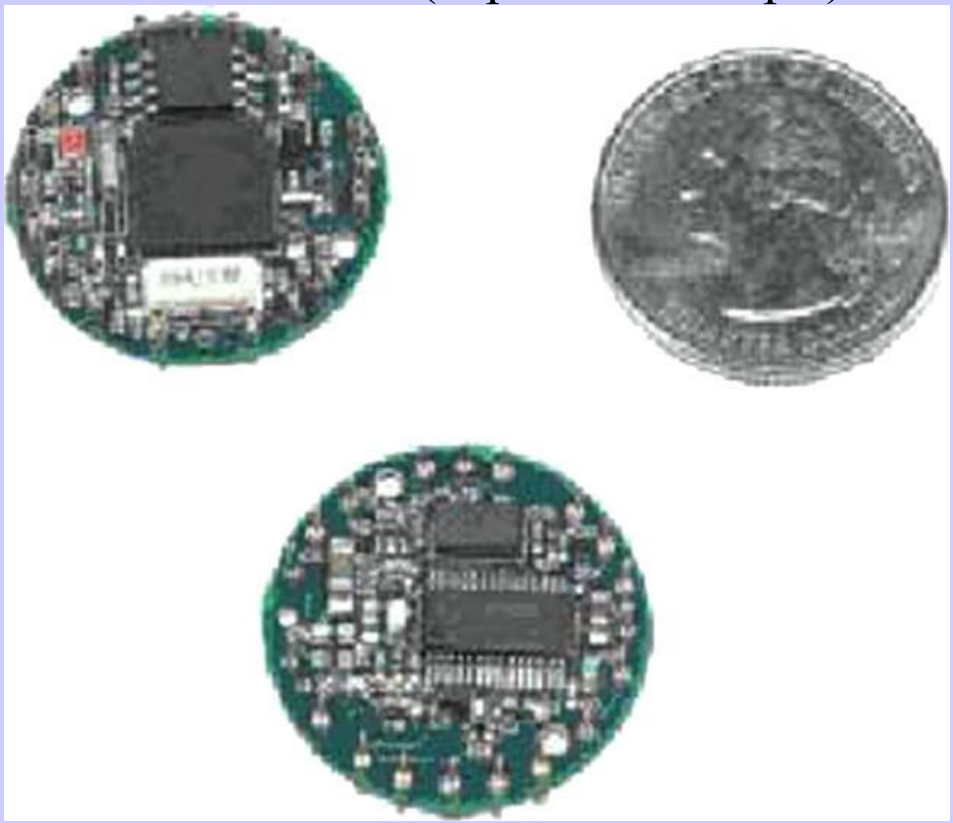
Результативность фильтра существенно снижается, вследствие чего пользователь его отключает.

Отравление текстового бота. В 2016 году для Twitter был разработан чат-бот Tay, рассчитанный на общение с пользователями 18–24 лет. Бот быстро обучился на онлайн-переписках, но начал выдавать непредвиденные результаты. После взаимодействия с пользователями Twitter, которые «отравили» лексикон бота, он начал сыпать оскорблениями. В итоге всего через 16 часов после запуска Tay пришлось отключить.


Обучающие данные
Алгоритм обучения
Рабочие данные
Легковой автомобиль
Грузовик
Пешеход
Мотоцикл
?
x
Классификация
Кража модели
Извлечение данных
Отравление
Черный вход
Уклонение
Конфиденциальность

0.2. Интернет вещей. Умный город. Беспроводные сенсорные сети

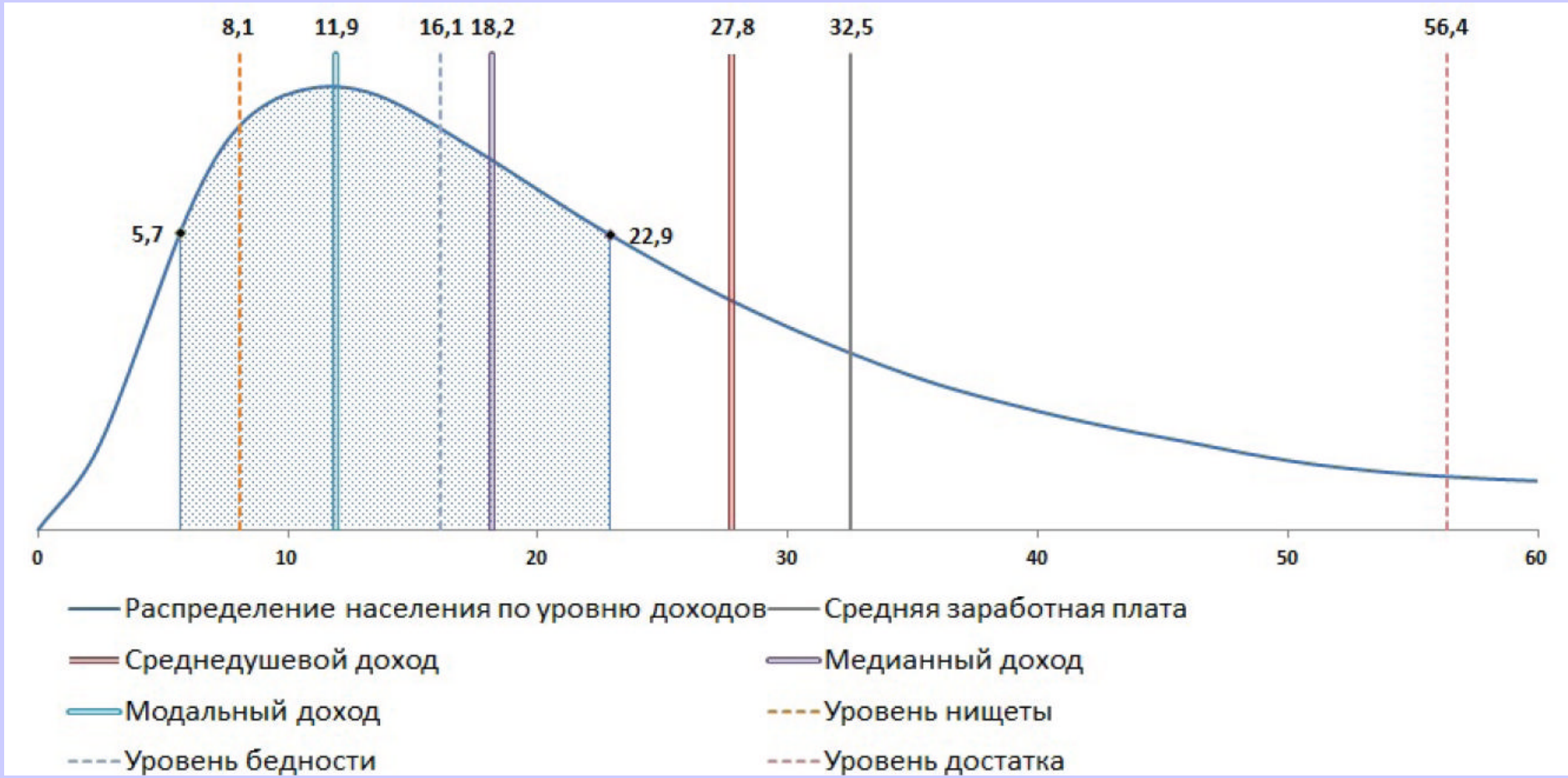
Атаки фальсификации данных. Экологический мониторинг, сельское хозяйство (беспроводной виноградник), промышленное производство. Геолокация, wi-fi-навигация в зданиях (торговые центры).



Узлы беспроводной сенсорной сети в сравнении с монетой

Сенсорный узел для виноградника

0.3 Зарплата – средняя, медианная, модальная



Какую зарплату имеет типичный житель страны, какова «центральная тенденция»?

1. Традиционные подходы к устойчивому оцениванию

1.1. Устойчивость статистических процедур

1.1.1. Введение в проблему устойчивости

Рассмотрим задачу оценивания параметра сдвига распределения вероятностей. Модель наблюдений имеет вид

$$y_i = \theta + e_i,\ i = 1,\ldots,N, \qquad (1.1)$$

где y_i – i-е наблюдение исследуемой случайной величины, θ – параметр сдвига, e_i – i-е значение случайной величины с нулевым параметром сдвига, N – количество наблюдений.

В условиях нормального распределения **среднее арифметическое** является для параметра сдвига асимптотически эффективной оценкой (оценкой по методу максимального правдоподобия – ММП-оценкой).

Отметим характерную особенность нормального распределения: основная масса распределения сосредоточена на конечном интервале $(-3\sigma, +3\sigma)$, где σ – корень квадратный из дисперсии. Вне этого интервала находится лишь 0,27 % распределения. Другими словами, нормальное распределение имеет «легкие хвосты». Таким образом, принимая гипотезу нормальности, мы автоматически предполагаем, что основная масса наблюдений сосредоточена на некотором интервале. Вероятность большого отклонения при этом весьма мала.

В реальной ситуации эта гипотеза – чересчур жесткая: предполагаемая модель редко бывает абсолютно точно специфицированной, в наборе данных могут присутствовать **выбросы** – грубые ошибки. Выбросы могут быть результатом нарушения условия эксперимента, неправильного измерения, появления посторонних данных и т.п. Поэтому необходимо предположить, что отклонения с большой вероятностью могут принимать и большие значения.

«Нормально распределённые данные столь же редки как единороги».

В таких случаях более адекватны распределения с тяжелыми хвостами (т.е. с немалыми вероятностями больших отклонений от центра), а среднее арифметическое не всегда оказывается подходящей оценкой.

Для распределений с тяжелыми хвостами более эффективными, чем среднее арифметическое, будут устойчивые оценки, которые не меняют резко своих значений при возникновении больших отклонений. Если вероятность больших отклонений мала, такие оценки будут менее эффективны, зато если наблюдения содержат выбросы, то эти оценки будут малочувствительны к ним, а потому более удовлетворительными.

Пример 1.1. Пусть имеется следующий набор данных [19]:

0.96 1.01 0.97 1.02 1.04 1.00 10.52.

Последнее число, очевидно, является выбросом.

Среднее арифметическое элементов выборки равно 2.36 и, таким образом, сильно отличается от истинного значения, равного 1.

Более устойчивой оценкой будет **усеченное среднее**, которое вычисляется следующим образом: отбрасывается некоторое количество минимальных и максимальных наблюдений в выборке. На основе оставшихся наблюдений находится среднее арифметическое.

В нашем случае усеченное среднее (при отбрасывании одного минимального и одного максимального наблюдений) равно 1.008.

Рассмотрим другую устойчивую оценку – **выборочную медиану**. Медиана нашей последовательности равна 1.01.

Для обоснования применения методов моделирования используется **принцип непрерывности или устойчивости**, когда малая ошибка в математической модели не должна приводить к существенной ошибке в окончательных выводах.

Оказывается, что многие распространенные статистические процедуры не обладают свойством устойчивости.

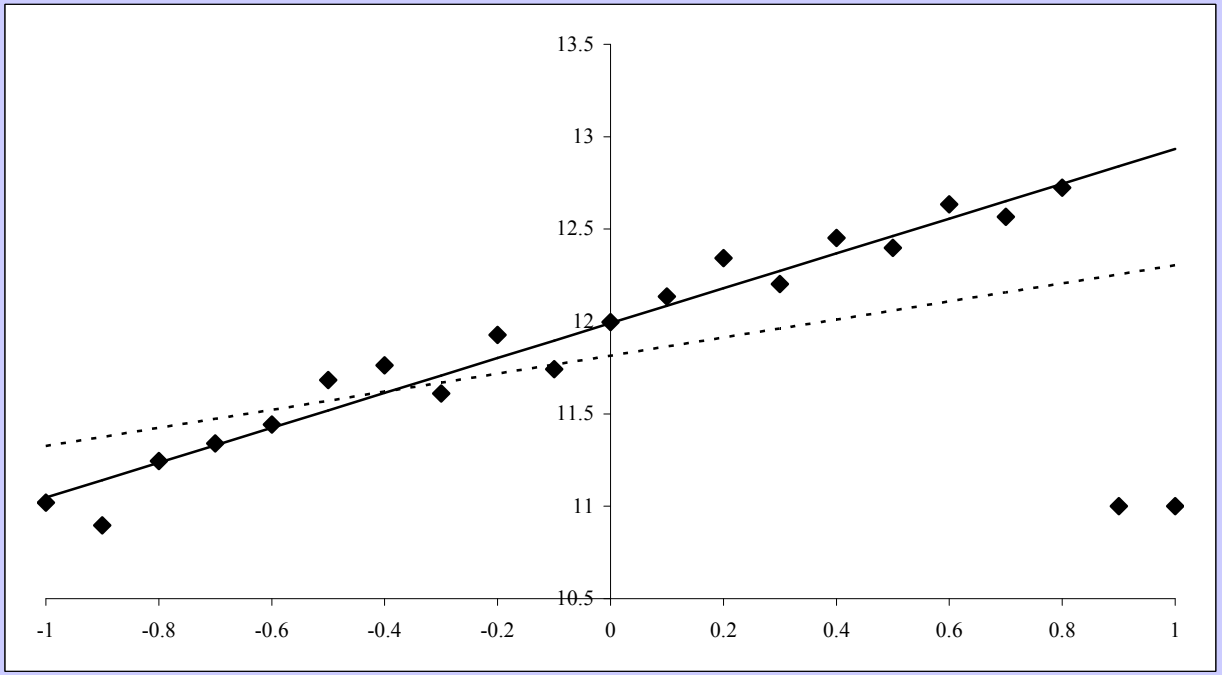
Известно, что за редким исключением не обладают устойчивостью процедуры, основанные на методе максимального правдоподобия. Будучи оптимальными для точной модели, при отклонении от нее свойства резко ухудшаются.

Пример. На графике приведены наблюдения (отмечены ромбиками), полученные по линейной регрессионной модели

$$y = 12 + x + e,$$

где e – ошибка, имеющая нормальное распределение, однако наблюдения в точках $x = 0,9$ и $x = 1$ заменены выбросами – значениями 11. Пунктирная линия соответствует оценке метода наименьших квадратов (МНК-оценке), МНК-оценки параметров равны 11.82 и 0.49. Сплошная линия соответствует МНК-оценке, построенной по данным без выбросов, значения оценок 11.99 и 0.94.

Как видно из графика, МНК-оценка чувствительна к наличию выбросов – линия МНК-оценки отклика отклоняется в сторону наблюдений-выбросов. Линия устойчивой оценки отклика наблюдения-выбросы игнорирует.



1.1.2. Основные подходы к устойчивому оцениванию

Точные параметрические модели – это абстракция, их выбор для описания конкретных данных часто до определенной степени произволен, поэтому истинное распределение в большей или меньшей степени отличается от модельного. Среди других важных типов отклонений назовем зависимость данных. Однако именно отклонение от модельного распределения считается наиболее важным и является наиболее изученным. Его мы и будем в дальнейшем рассматривать.

Фактически приходим к следующей ситуации.

Имеется выборка независимых одинаково распределенных случайных величин

$$y_i,\ i = 1,..., N,$$

из неизвестного распределения G с функцией $G(y)$. Выборке соответствует эмпирическая функция распределения

$$G_N(y) = \frac{1}{N}\sum_{i=1}^{N}\Delta(y_i - y),$$

где $\Delta(y)$ – функция единичного скачка.

Поскольку G нам неизвестно, мы пользуемся некоторым параметрическим семейством $\{F_\theta, \theta \in \Theta\}$, подгоняя параметр θ под данные.

Например, мы можем предположить, что наблюдения имеют нормальное распределение с неизвестным (требующим оценивания по данным) математическим ожиданием и известной дисперсией.

Вообще говоря, параметрическая «идеальная» модель также может зависеть от N: F_θ есть $F_{N,\theta}$ и при $N \to \infty$ мы, получая все больше информации, приближаем идеальную модель к истинной $F_{N,\theta} \to G$.

Наша цель состоит в том, чтобы найти такую оценку неизвестного параметра θ, которая была бы в некотором смысле «хорошей» для распределения G_N.

Поскольку точные параметрические модели не всегда оказываются устойчивыми, при построении устойчивых процедур необходимо использовать модели других типов.

Можно совсем отказаться от точных параметрических моделей. И это делается в **непараметрическом подходе**.

Он требует наложить на распределение лишь некоторые весьма слабые ограничения, в условиях которых возможен анализ получаемых решений (выяснение таких свойств, как нормальность, состоятельность и т. д.). Примеры ограничений – непрерывность распределения, независимость, одинаковая распределенность наблюдений.

Фактически в непараметрических процедурах используется ограниченное непараметризированное множество распределений. Или, говоря иначе, используется бесконечномерный параметр. Однако непараметрические процедуры не обязательно приводят к устойчивым решениям.

Если параметрический подход приводит к слишком неустойчивым решениям, то непараметрический – к слишком неэффективным.

В **теории робастности** сочетаются параметрический и непараметрический подходы. Рассматривается устойчивость в **непараметрической** окрестности **параметрической** модели. Тут идея обеспечения устойчивости становится основной.

В теории робастности имеется несколько различных подходов, и каждый конкретный результат о робастности конкретной процедуры необходимо сопровождать указанием, в каком смысле следует робастность понимать.

В частности, надо зафиксировать следующие моменты.

1. Указать идеальную модель F_θ: $F(y,\theta)$. Например, $N(\mu,1)$ – нормальное распределение с оцениваемым математическим ожиданием μ и единичной дисперсией, т. е. $\theta = \mu$.

2. Описать учитываемые отклонения от модели в виде некоторой окрестности $\tilde{F}$ идеального распределения F_θ, так что ей принадлежит неизвестное истинное распределение: $G \in \tilde{F}$. Например, часто используется окрестность в виде ε-засоренного распределения

$$\tilde{F} = \tilde{F}_\varepsilon = \left\{ G_\theta : G(y,\theta) = (1-\varepsilon)F(y,\theta) + \varepsilon H(y) \right\},$$

где $H(y)$ – произвольное (неизвестное) распределение; ε – уровень (интенсивность) засорения (для этой окрестности используются также названия «модель грубой ошибки» или «модель большой ошибки»).

3. Указать критерий качества процедуры и сформулировать требования к его поведению в заданной окрестности. Например, асимптотическая дисперсия оценки должна быть некоторым образом ограничена.

Рис. Пространство всех вероятностных распределений (обычно бесконечномерное).

В непараметрической статистике допускаются «все» возможные распределения вероятностей и неопределенность относительно них ограничивается лишь одним или несколькими измерениями. В классической параметрической статистике допускается лишь (очень «тонкое») малоразмерное подмножество всех распределений вероятностей – параметрическая модель, обеспечивающая, тем не менее, необходимую для эффективного и сжатого представления данных степень избыточности. В робастной статистике допускается полная (а именно, полноразмерная) окрестность параметрической модели, которая представляется гораздо более реалистичной, и, кроме того, если отвлечься от некоторой её «размытости», обладает теми же преимуществами, что и строгая параметрическая модель.

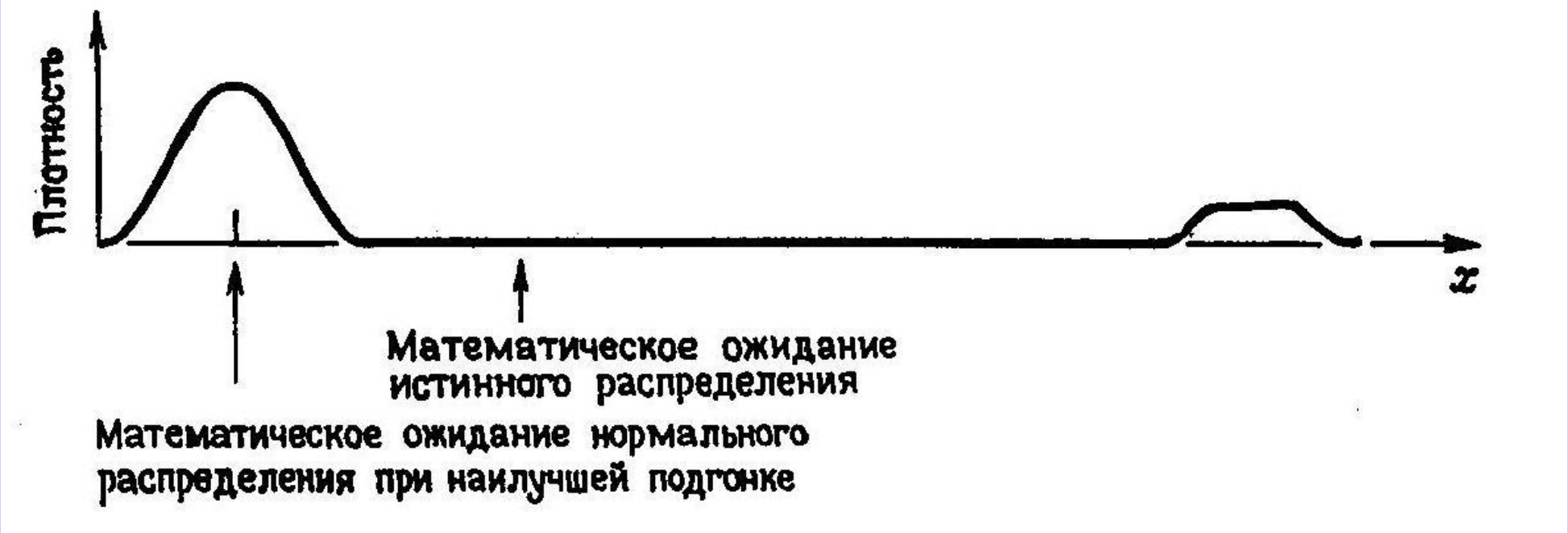


Рис. Различие целей непараметрической и робастной статистик

1.2. *M*-оценки

В дальнейшем будем рассматривать простой и удобный вид оценок параметров – *M*-оценки – для модели одномерной непрерывной случайной величины, имеющей модельное распределение $F(y,\theta)$ с плотностью $f(y,\theta)$ и неизвестным скалярным параметром θ, причем область значений случайной величины – интервал Y – не зависит от параметра θ.

M-оценка $\hat{\theta}$ параметра θ находится как решение задачи минимизации

$$\sum_{i=1}^{N}\rho\left(y_i,\hat{\theta}\right)=\min, \tag{1.2}$$

где ρ – функция потерь.

Альтернативное определение *M*-оценки требует решения неявного уравнения

$$\sum_{i=1}^{N}\psi\left(y_i,\hat{\theta}\right)=0, \tag{1.3}$$

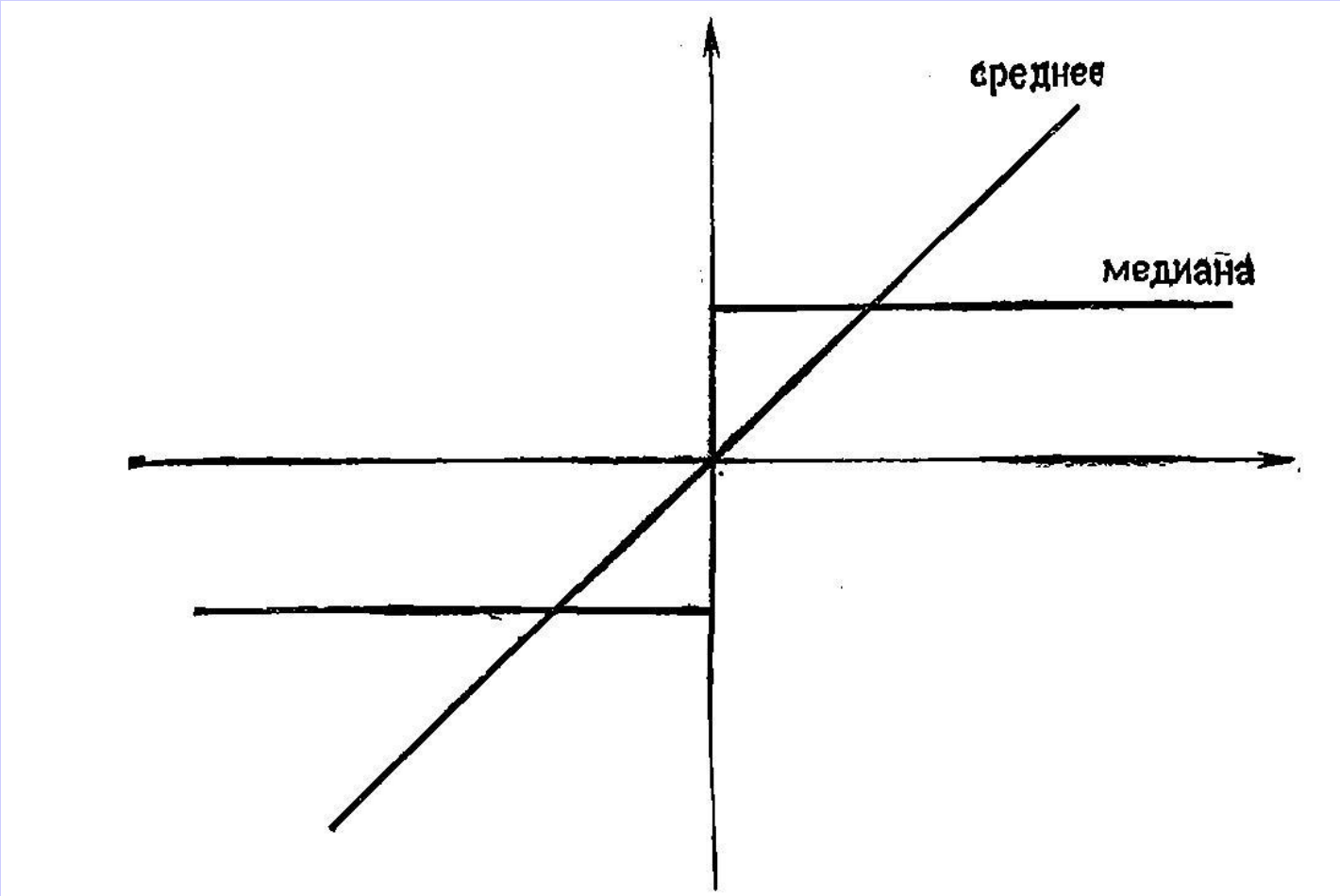
где ψ – оценочная функция.

Примеры. Среднее арифметическое является M-оценкой с функциями

$$\rho(y,\theta)=(y-\theta)^2,\ \psi(y,\theta)=c(y-\theta).$$

Выборочная медиана является M-оценкой с функциями

$$\rho(y,\theta)=\left|y-\theta\right|,\ \psi(y,\theta)=c\,\mathrm{sign}(y-\theta).$$



Обозначив $\psi(y,\theta)=\rho'_\theta(y,\theta)$, переходим от (1.2) к (1.3). Однако при невыпуклой функции потерь может иметься несколько локальных минимумов, соответствующих нескольким решениям уравнения (1.3). По этой причине два указанных определения *M*-оценки не эквивалентны.

Отметим, что одна и та же *M*-оценка может быть задана эквивалентными оценочными функциями, отличающимися сомножителями, не зависящими от данных, т. е. оценочная функция $c(\theta)\psi(y,\theta)$, $c(\theta)\neq 0$, задает ту же оценку, что $\psi(y,\theta)$.

При некоторых условиях регулярности *M*-оценка будет:

– асимптотически несмещенной, т. е. $\mathbf{E}\hat{\theta}\to\theta$ при $N\to\infty$, где $\mathbf{E}$ – символ математического ожидания;

– состоятельной;

– асимптотически нормальной.

Условие асимптотической несмещенности оценки имеет вид

$$\mathbf{E}\psi(y,\theta)=0$$

или

$$\int_Y \psi(y,\theta)f(y,\theta)dy=0. \qquad (1.6)$$

Величина $\sqrt{N}(\hat{\theta}-\theta)$ асимптотически нормальна с нулевым математическим ожиданием и дисперсией

$$V(\psi,F)=\frac{\mathbf{E}\psi^2(y,\theta)}{d^2}, \qquad (1.8)$$

где $d=d(\theta)=\frac{\partial}{\partial t}\mathbf{E}\,\psi(y,t)\big|_{t=\theta}=\mathbf{E}\psi'_\theta(y,\theta)$ (для выполнения последнего равенства требуются дополнительные условия).

Величина $V(\psi,F)$ называется асимптотической дисперсией. Функционал $V(\psi)=V(\psi,F)$ достигает минимума на оценочной функции

$$\psi_0(y,\theta)=c\frac{\partial}{\partial\theta}\ln f(y,\theta),$$

где $c=c(\theta)\neq 0$, соответствующей **ММП-оценке**. При этом функция потерь имеет вид

$$\rho_0(y,\theta)=-\ln f(y,\theta).$$

Таким образом, выбирая оценочную функцию ψ, отличную от оценочной функции ММП-оценки, мы увеличиваем асимптотическую дисперсию в идеальной модели, но можем при должном выборе ψ приобрести свойство устойчивости.

1.4. Минимаксный подход

Перейдем к рассмотрению подходов к робастному оцениванию. И начнем с минимаксного подхода, который хронологически был предложен первым, его автор – П. Хьюбер.

Минимаксный подход основан на использовании наилучшей оценки в наихудшей точке окрестности. Критерием качества при этом является либо асимптотическое смещение (точнее, его модуль), либо асимптотическая дисперсия.

В первом случае имеем задачу

$$\inf_{\psi} \sup_{G \in \tilde{F}} |b(\psi, G)|, \tag{1.19}$$

где $b(\psi, G)$ – асимптотическое смещение, во втором –

$$\inf_{\psi} \sup_{G \in \tilde{F}} V(\psi, G), \tag{1.20}$$

причем в обоих случаях функция ψ удовлетворяет условию (1.6).

1.4.1. Минимаксное смещение

Рассмотрим первую задачу. В качестве окрестности выберем ε-засоренное распределение, точнее реальное распределение будет иметь вид

$$G(y,\theta) = (1-\varepsilon)F(y,\theta) + \varepsilon H(y,\theta). \qquad (1.21)$$

В этом случае приближенное значение асимптотического смещения имеет вид

$$b(\psi, G) = b(\psi, H) \approx \varepsilon \frac{\int_Y \psi(y,\theta) dH(y,\theta)}{-d}. \qquad (1.23)$$

Возвращаясь к минимаксной задаче (1.19), получаем

$$\inf_{\psi} \sup_{H} |b(\psi, H)|,$$

решение задачи имеет вид

$$\psi(y,\theta) = \operatorname{sign}\left[\psi_0(y,\theta) + \beta(\theta)\right].$$

Оценку с такой оценочной функцией будем называть *медианной*. Функция $\beta(\theta)$ определяется из условия асимптотической несмещенности (1.6):

$$\int_Y \operatorname{sign}\left[\psi_0(y,\theta) + \beta(\theta)\right] f(y,\theta) dy = 0.$$

Пример 1.5. Задана нормальная модель $F_\theta : N(\mu, \sigma^2)$ с известной дисперсией σ^2 и оцениваемым математическим ожиданием μ. Плотность распределения имеет вид

$$f(y, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{(y-\mu)^2}{2\sigma^2} \right\}. \qquad (1.28)$$

В нашем случае $f(y, \mu) = f(y, \mu, \sigma)$ и

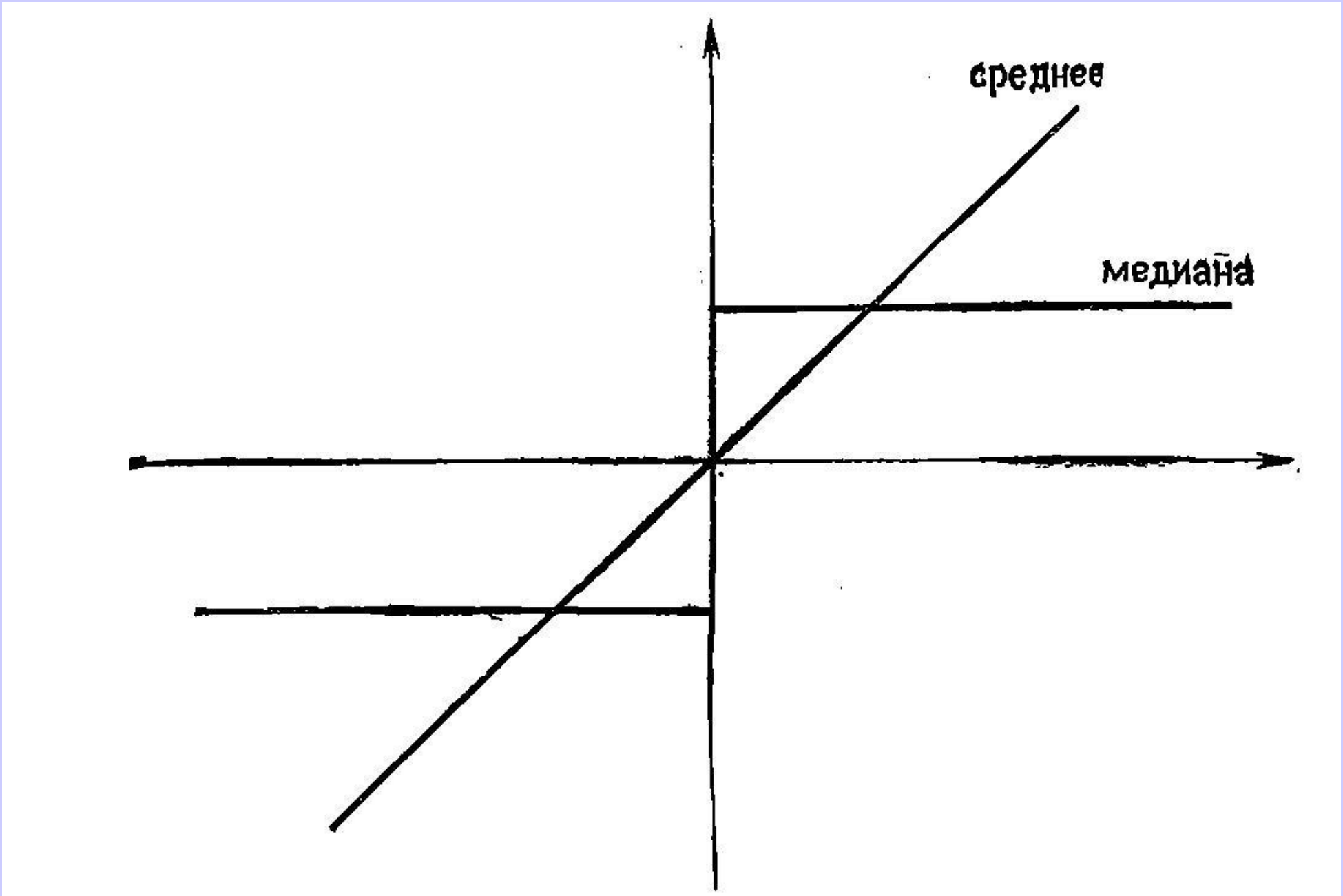
$$f'_\mu(y, \mu) = f(y, \mu) \frac{y-\mu}{\sigma^2},$$

$$\psi(y, \mu) = \operatorname{sign}\left[\frac{f'_\mu(y, \mu)}{f(y, \mu)} + \beta(\mu) \right] = \operatorname{sign}\left[\frac{y-\mu}{\sigma^2} + \beta(\mu) \right].$$

Поскольку $f(y, \mu) = f(y - \mu)$ – четная функция (относительно μ), а $\psi(y, \mu) = \psi(y - \mu)$ при $\beta(\mu) = 0$ – нечетная, при $\beta(\mu) = 0$ условие (1.6) выполняется. Следовательно, $\beta(\mu) = 0$. Окончательно, отбрасывая несущественный знаменатель, имеем

$$\psi(y, \mu) = \operatorname{sign}[y - \mu].$$

В результате $\hat{\mu} = \underset{i}{med}\, y_i$ – выборочная медиана.


среднее
медиана

Пример 1.7. Распределение Коши с оцениваемым параметром сдвига μ и известным параметром масштаба λ. Функция плотности имеет вид

$$f(y,\mu,\lambda)=\frac{1}{\pi\lambda}\frac{1}{1+(y-\mu)^2/\lambda^2}=\frac{1}{\pi}\frac{\lambda}{\lambda^2+(y-\mu)^2}.$$

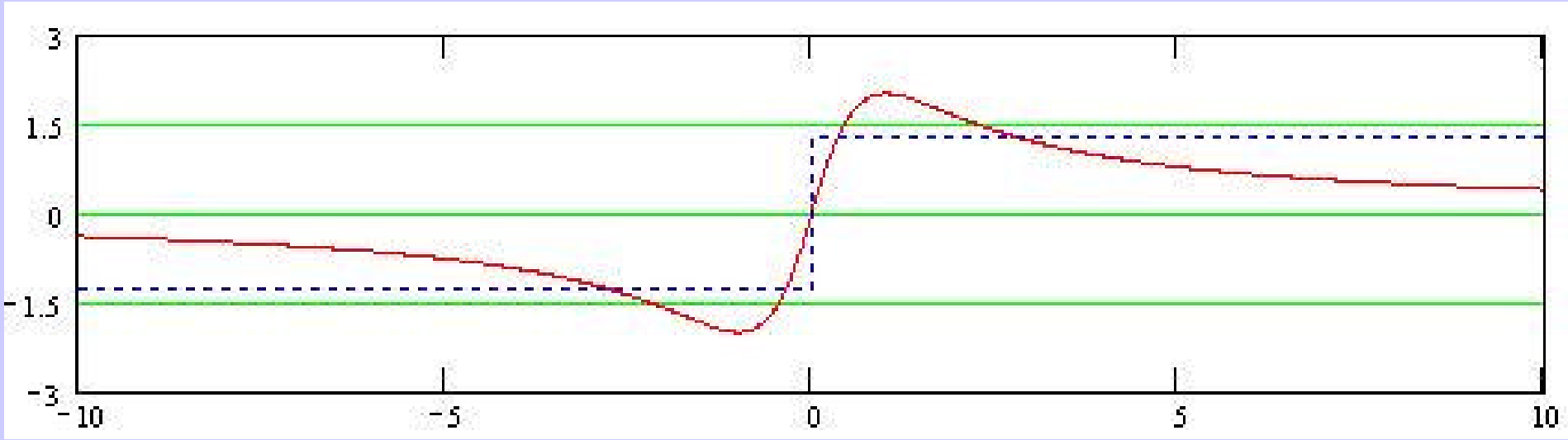
В этом случае $f(y,\mu)=f(y,\mu,\lambda)$. Рассмотрим случай $\lambda=1$.

Распределение Коши является распределением Стьюдента с одной степенью свободы. Распределение симметрично относительно μ, но случайная величина не имеет математического ожидания. Оценка μ в виде выборочного среднего является несостоятельной, ММП приводит к состоятельной и устойчивой оценке с оценочной функцией $\psi(y,\mu)=\dfrac{y-\mu}{1+(y-\mu)^2}$.

Найдем оценочную функцию для оценки с минимаксным смещением

$$\psi(y,\mu)=\operatorname{sign}\left[\frac{f'_\mu(y,\mu)}{f(y,\mu)}+\beta(\mu)\right]=\operatorname{sign}\left[\frac{2(y-\mu)}{1+(y-\mu)^2}\right]=\operatorname{sign}\left[y-\mu\right].$$

Таким образом, $\hat{\mu}$ – выборочная медиана.


3
1.5
0
-1.5
-3
-10
-5
0
5
10