Глава 4

Занятие четвертое. О непараметрическом оценивании, и о том, как бутстрэпить без знания распределения

4.1 Непараметрическое оценивание

4.1.1 Общие слова

Мы привыкли рассматривать именно параметрическую модель $X_i \sim F_\theta$. С другой стороны, начиная исследование, мы часто не имеем никакой предварительной информации о распределении. В таком случае более подходящей является непараметрическая модель $X_i \sim F$, а оценки мы будем строить для $F(x)$, $\mathbf{E}X_1$ или другого показателя, связанного с моделью. При этом мы сохраняем определения состоятельности, несмещенности и асимптотической нормальности, просто не апеллируем в них к параметризации модели.

Пример 1. Так $\overline{X} = (X_1 + ... + X_n)/n$ будет несмещенной состоятельной оценкой $\mathbf{E}X_1$ и в непараметрической модели, а при $\mathbf{E}X_1^2 < \infty$ еще и асимптотически нормальной. Аналогично $S_0^2 = \sum_{i=1}^n (X_i - \overline{X})^2/(n-1)$ будет несмещенной состоятельной оценкой дисперсии, а при $\mathbf{E}X_1^4 < \infty$ асимптотически нормальной.

4.1.2 Как оценивать функции распределения и как с помощью этого строить непараметрические оценки?

Итак, давайте начнем с того, что оценим функцию распределения и плотность нашей выборки X_i. Достаточно хорошей (несмещенной состоятельной и асимптотически нормальной) оценкой функции распределения в конкретной точке x является эмпирическая функция распределения

$$\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{X_i \leq x}.$$

Вопрос 1. Почему она при каждом x обладает свойствами несмещенности, состоятельности и асимптотической нормальности?

Вопрос 2. Какой доверительный интервал для $F(x)$ можно построить при фиксированном x на основе асимптотической нормальности ЭФР в этой точке?

Вопрос 3. Показать, что $\widehat{F}_n$ — ОМП для ф.р. в непараметрической модели (считая, что рассматриваемые распределения являются всеми дискретными распределениями с какими-либо значениями и вероятностями).
В R ЭФР может быть легко подсчитана по выборке с помощью команды ecdf(x). Обратите внимания, что при построении plot от этой функции, нужно будет написать plot(Vectorize(ecdf)).
В Python ЭФР автоматизирована в ECDF в statsmodels.distributions.empirical_distribution. У результата подсчета есть параметры x и y, построить график можно с помощью plot из matplotlib.pyplot.

Теорема Гливенко-Кантелли утверждает, что ЭФР состоятельна как оценка всей функции распределения, даже более того, $\widehat{F}_n \to F$ п.н. по равномерной норме. Теорема Колмогорова дает для непрерывных распределений утверждение

$$\mathbf{P}(\sqrt{n}||\widehat{F}_n - F|| > \varepsilon) \to 1 - K(\varepsilon),$$

где K — функция распределения Колмогорова, а норма равномерная. Более того, в случае непрерывной F при каждом n распределение этой статистики не зависит от n. Существует явная оценка сверху, не требующая непрерывности распределения: неравенство Дворецкого-Кифера-Вольфовица

$$\mathbf{P}(||\widehat{F}_n - F|| > \varepsilon) \leq 2e^{-2n\varepsilon^2},\ \varepsilon > 0.$$

При этом ф.р. Колмогорова близка к правой части неравенства при больших n и не очень больших ε, поэтому методы дают близкие результаты.

Задача 1. Смоделировать 100 выборок. Для каждой построить 95% доверительное множество для ф.р. распределения а) нормальной б) Коши с помощью неравенств Д.-К.-В и теоремы Колмогорова. Построить доверительный интервал 95% для функции распределения в каждой из точек x. Будет ли объединение этих интервалов давать 95% доверительное множество? В какой доле из выборок настоящая ф.р. не попадала в каждое из доверительных множеств?
Решение задачи на Python реализовано в ConfforCDF.py.

Из сходимости $\widehat{F}_n \to F$ при п.в. ω вытекает сходимость при тех же ω $f(\widehat{F}_n) \to f(F)$ для любого непрерывного (по равномерной норме) функционала f. Вполне естественно, таким образом, исследуя функционалы $f(F)$, оценивать их $f(\widehat{F}_n)$.

Пример 2. Для оценки математического ожидания $\mathbf{E}X_1 = \int_{\mathbb{R}} x dF(x)$ и дисперсии $\mathbf{D}X_1 = \int_{\mathbb{R}} x^2 dF(x) - (\int_{\mathbb{R}} x dF(x))^2$ возьмем

$$\widehat{\theta}_1 = \int_{\mathbb{R}} x d\widehat{F}_n(x),\ \widehat{\theta}_2 = \int_{\mathbb{R}} x^2 d\widehat{F}_n(x) - \widehat{\theta}_1^2.$$

Вопрос 4. Будут ли эти функционалы непрерывными по равномерной норме?

Вопрос 5. Почему $\widehat{\theta}_1 = \overline{X}$, $\widehat{\theta}_2 = S^2$?

Задача 2. Построить оценку для асимметрии: $\mathbf{E}(X - \mathbf{E}X)^3(\mathbf{D}X)^{-3/2}$.

4.1.3 Оценки плотности

С той же целью полезно бывает оценить плотность, если мы знаем, что распределение абсолютно-непрерывно.

Простейшей оценкой плотности является гистограмма, с которой вы, вероятно, и так знакомы. Гистограмма предлагает рассмотреть некоторое разбиение области значений нашей величины $(a_i, a_{i+1}]$, $i = 1, 2, ...$ Тогда гистограмму на $(a_i, a_{i+1}]$ положим равной $N_i/N = \#\{x_j \in (a_i, a_{i+1}]\}/N$, то есть частоте попадания в полуинтервал наших наблюдений. В целом, гистограмма неплохая оценка, стремящаяся с ростом числа наблюдений и уменьшением ширины интервалов разбиения к истинному значению плотности, но она кусочно-постоянна.

Хорошим методом получения неперывной оценки для плотности является так называемая *ядерная оценка*. Назовем *ядерной* неотрицательную функцию $K(x)$, т.ч. $\int_{\mathbb{R}} K(x)dx = 1$, $\int_{\mathbb{R}} xK(x)dx = 0$, $\int_{\mathbb{R}} x^2K(x)dx > 0$ (то есть $K(x)$ —- некоторая плотность с нулевым матожиданием и конечной дисперсией). Назовем *ядерной оценкой* плотности

$$\widehat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h_n} K\left(\frac{x - X_i}{h_n}\right),$$

где K — ядерная функция, а h_n — некий параметр, называющийся шириной окна сглаживания.

Если f — непрерывна, $h_n \to 0$, $nh_n \to \infty$, то $\widehat{f}_n(x)$ будет состоятельной для $f(x)$. При этом $\mathbf{E}(\widehat{f}_n(x) - f(x))^2$ есть $O(h_n^4) + O\left(\frac{1}{nh_n}\right) + O\left(\frac{1}{n}\right)$. Таким образом, наилучший порядок h есть $O(n^{-1/5})$. Более точная

формула для h, дающего минимум квадратичного отклонения

$$h^* = \left(\frac{\int K(x)^2 dx}{n \left(\int x^2 K(x) dx \right)^2 \int (f''(x))^2 dx} \right)^{1/5}.$$

Эта формула не вполне удовлетворительна для применения, поскольку использует неизвестное нам $f''(x)$, но дает представить порядок h^*.
Функция K, как мы видим, не влияет на порядок сходимости. Часто берут в роли $K(x)$ равномерную на $[-1, 1]$ плотность (прямоугольное ядро), стандартную нормальную плотность (гауссово ядро) или $3(1-x^2)I_{|x|\leq 1}/4$ (ядро Епанечникова).

В R ядерная оценка плотности встроена, например, в стандартный график qplot с геометрией geom=density, по умолчанию устанавливается гауссовкое ядро, но можно поменять его с помощью параметра kernel.

Непосредственно ядерную оценку для плотности можно рассчитать с помощью функции density. Параметр kernel отвечает за вид ядра, параметр bw отвечает за ширину окна.

Задача 3. Оцените а) стандартную нормальную плотность на выборке размера 100. б) смесь из трех нормальных плотностей $\mathcal{N}(0,1)$, $\mathcal{N}(6,1)$, $\mathcal{N}(-3,1)$ с равными вероятностями. Что происходит при уменьшении ширины окна h?

Существуют и другие подходы к оценке плотности, например, оценить ее коэффициенты разложения в ряд Фурье. Мы не будем их касаться в рамках нашего семинара.

Поговорим о том, как оценить качество полученной оценки для плотности. Хорошим показателем для нас являлась бы величина

$$\int_{\mathbb{R}} (\widehat{f}_n(x) - f(x))^2 dx = \int_{\mathbb{R}} \widehat{f}_n(x)^2 dx - 2\int_{\mathbb{R}} \widehat{f}_n(x) f(x) dx + \int_{\mathbb{R}} f(x)^2 dx,$$

которую мы не можем подсчитать, поскольку не знаем $f(x)$.

Определение 1. *Оценкой кросс-валидацией* для оценки плотности $\widehat{f}_n$ называют

$$\widehat{J}(h) = \int_{\mathbb{R}} \widehat{f}_n(x)^2 dx - \frac{2}{n} \sum_{i=1}^{n} \widehat{f}_{n,i}(X_i),$$

где $\widehat{f}_{n,i}$ — оценка, построенная по выборке с исключенным X_i.

Вопрос 6. Показать, что $\mathbf{E}\widehat{J}(h) = \mathbf{E}\left(\int_{\mathbb{R}} \widehat{f}_n(x)^2 dx - 2\int_{\mathbb{R}} \widehat{f}_n(x) f(x) dx \right)$.

Тем самым, маленькие значения $\widehat{J}(h)$ указывают на то, что $\widehat{f}_n(x)$ близко к $f(x)$.

Для ядерной оценки оценку кросс-валидации можно найти по формуле

$$\widehat{J}(h) = \frac{1}{hn^2} \sum_i \sum_j K^* \left(\frac{X_i - X_j}{h} \right) + \frac{2}{nh} K(0) + O\left(\frac{1}{n^2} \right).$$

$K^* = K * K - 2K$, $*$ — свертка. С помощью оценки кросс-валидации удобно измерять качество приближения. В частности, с ее помощью можно откалибровать ширину h.

Осуществить выбор h с помощью кросс-валидации в R можно, указав bw=bw.ucv(x).

С помощью оценки для плотности можно строить оценки $\int g(x)\widehat{f}_n(x)dx$ для функционалов вида $\int g(x)f(x)dx$. Например, для математического ожидания мы получим оценку

$$\int_{\mathbb{R}} x\widehat{f}_n(x)dx = \overline{X}$$

Вопрос 7. Почему верно последнее тождество?

4.2 Бутстреп или как прикрывать свои недостатки в непараметрическом случае

4.2.1 Исправление смещения с помощью bootstrap

Если мы захотим получать несмещенные оценки, то нам, как и прежде, будет полезна процедура бутстрепинга. Однако, раньше мы брали выборку из распределения с оцениваемым параметром, то теперь мы будем брать выборку из распределения с оцениваемой функцией распределения. Если в качестве оценки для функции распределения используется эмпирическая функция распределения, то это равносильно рассмотрению выборок $X^*_{i,1}, ..., X^*_{i,n}$, $i = 1, ..., M$, взятых из нашей выборки с возвращением. Это может показаться странным — зачем брать с возвращением элементы из нашей выборки? Тем не менее метод действительно вполне осмыслен, если функционал f неперерывен.

Таким образом, если мы рассматриваем статистику $f(\widehat{F}_n)$ в качестве оценки $f(F)$, мы можем брать выборки из распределения $\widehat{F}_n$ и на основе этих выборок изучать распределение $f(\widehat{F}_n)$, ожидая, что оно близко к распределению $f(F)$. В частности, мы можем исследовать смещение $f(\widehat{F}_n)$ по сравнению с $f(F)$, приближая его смещением $f(\tilde{F}_n) - f(\widehat{F}_n)$, $\tilde{F}_n$ – ЭФР выборки из $\widehat{F}_n$. Аналогичным образом мы можем изучать и другие параметры распределения $f(\widehat{F}_n)$, например, дисперсию. Для бутстрэпа в пакете bootstrap в R есть одноименная функция, хотя реализация этого алгоритма совсем проста и не требует специальных функций.

Задача 4. Сгенерировать выборку $R[0, 1]$ размера 50 и оценить дисперсию оценок а) $\overline{X}$, б) MED.

Задача 5. Исправить смещение статистики S^2 как оценки дисперсии с помощью bootstrap по выборке размера 100 из распределения $\mathcal{N}(0, 1)$.

4.2.2 Доверительные интервалы с помощью bootstrap

Предположим, что $\widehat{\theta} = f(\widehat{F}_n)$ — оценка какого-то параметра распределения $f(F)$. Обсудим как построить доверительные интервалы на основе bootstrap.

1. percentile-интервал строится тем же методом, что и прежде — генерируется m выборок из функции распределения $\widehat{F}_n$ (то есть m выборок того же размера с возвращением), ранжируется значение нашей оценки $\widehat{\theta}^*$ на этих выборках и в качестве левой границы доверительного интервала берется $[\alpha m]$-е по возрастанию из значений $\widehat{\theta}^*$, а в качестве правой границы — $[(1 - \alpha)m]$-е. Он также достаточно требователен к данным и оценке.

2. pivotal-интервал так же строится аналогично параметрической версии — генерируется m выборок из функции распределения $\widehat{F}_n$, ранжируется значение нашей оценки $\widehat{\theta}^*$ на этих выборках и в качестве левой границы доверительного интервала берется $[\alpha m]$-е по возрастанию из значений $2\widehat{\theta} - \widehat{\theta}^*$, а в качестве правой границы — $[(1 - \alpha)m]$-е.

3. Можно действовать более хитрым путем. Сперва с помощью bootstrap оценим величиной $\widehat{\sigma}$ стандартное отклонение $\widehat{\theta}$. Теперь начнем новый bootstrap для построения доверительного интервала. Пусть $X^*_{i,1}, ..., X^*_{i,n}$ — одна из выборок с возвращением из нашей $(X_1, ..., X_n)$. Найдем $\widehat{\theta}^*_i = f(\widehat{F}^*_{n,i})$ и оценим еще одним bootstrap'ом ее стандартное отклонение $\widehat{\sigma}^*_i$. Построим

$$Y_i = \frac{\widehat{\theta}^*_i - \widehat{\theta}}{\widehat{\sigma}^*_i}$$

для всех выборок $X^*_{i,\cdot}$, $i \le m$, выберем среди них $[\alpha m]$-е и $[(1 - \alpha)m]$-е по возрастанию, а затем в качестве доверительного интервала возьмем

$$\left(\widehat{\theta} - (\widehat{\theta} - Y_{([(1-\alpha)m])})\widehat{\sigma}, \widehat{\theta} - (\widehat{\theta} - Y_{([\alpha m])})\widehat{\sigma}\right)$$

Рассмотрим теперь величину $(f(\widehat{F}_n) - f(F))/S(f)$. Построим для нее bootstrap-интервал ширины 0.95, откуда получим доверительный интервал для $f(F)$. Этот интервал называется studentized pivotal.

Вопрос 8. Как при больших n ведет себе 60% pivotal доверительный интервал для параметра "супремум возможных значений случайной величины" на основе $\max X_i$, где $X_i \sim R[0,1]$, $i \le n$.?

Задача 6. Испытайте метод pivotal для интервалов для а) среднего $R[0,1]$, б) дисперсии $\exp(1)$.
Рассмотренные интервалы не точные, их уровень доверия близок к $1-\alpha$ с ростом n (хотя и не равен ему). Стьюдентовский интервал имеет более высокую скорость сходимости к $1-\alpha$, равную $O(1/n)$, обычный pivotal интервал имеет скорость $O(1/\sqrt{n})$.

4.3 Функции влияния и Дельта-метод в непараметрическом случае

Этот параграф требуется только если вы сдаете на продвинутом уровне

4.3.1 Функция влияние и производная Гато

Пусть f — функционал от ф.р. Рассмотрим так называемую производную Гато нашего функционала

$$f'_G(F) = \lim_{p\to 0} \frac{f((1-p)F + pG) - f(F)}{p},$$

где $G(x)$ — некоторая ф.р. Это аналог обычной производной по направлению, только для случая, когда аргументом у нас служат функции.

Мы будем рассматривать частный случай $G = \delta_x$, где δ_x — ф.р. константы x, то есть $\delta_x(y) = 1$ при $x \le y$ и 0 при $x > y$. Тогда

$$L_{f,F}(x) = f'_{\delta_x}(F) = \lim_{p\to 0}(f((1-p)F + p\delta_x) - f(F))/p$$

называют функцией влияния.

Давайте поймем физический смысл функции влияния. Функция $(1-p)F(x)+p\delta_x$ соответствует функции распределения следующей величины: с вероятностью p она равна x, а с вероятностью $1-p$ она равна $X \sim F$. Таким образом, функция влияния предлагает добавить в данные из распределения F небольшой процент p данных тождественно равных x, сравнить, насколько при этом изменился функционал и поделить на p, устремив $p \to 0$.

Вопрос 9. Пусть $f(F) = \int_{\mathbb{R}} a(x)dF(x)$. Чему равно $L_{f,F}$?
Нам понадобится более общее понятие дифференцируемости по Адамару. Пусть $\mathcal{D} = \{F - G\}$, где F, G — функции распределения.

Определение 2. Функционал f дифференцируем по Адамару, если для любых $D_n \in \mathcal{D}$, $\varepsilon_n \to 0$ выполнено соотношение

$$\frac{f(F + \varepsilon_n D_n) - f(F)}{\varepsilon_n} - f'_{D_n}(F) \to 0,\ n \to \infty.$$

Это что-то в духе равномерности предела Гато по всем направлениям.

Вопрос 10. Почему функционал $\int_{\mathbb{R}} a(x)dF(x)$ является дифференцируемым по Адамару для любой F, для которой он конечен?

4.3.2 Непараметрический дельта-метод

Как мы уже говорили, функция $\widehat{F}_n(x)$ является асимптотически нормальной оценкой функции $F(x)$ при каждом x. А вдруг нам повезет и функционалы $f(\widehat{F}_n)$ тоже будут асимптотически нормальными

оценками для $f(F)$? В параметрическом случае у нас функции от оценок оказывались асимптотически нормальными в силу дельта-метода при дифференцируемых f. В непараметрическом случае верна аналогичная теорема:

Теорема 1. *Если f дифференцируем по Адамару, то*

$$\sqrt{n}\frac{f(\widehat{F}_n)-f(F)}{\tau}\to Z\sim\mathcal{N}(0,1),$$

где $\tau^2=\int_{\mathbb{R}}L_{f,F}(x)^2dF(x)$. Более того,

$$\sqrt{n}\frac{f(\widehat{F}_n)-f(F)}{\widehat{\tau}}\to Z\sim\mathcal{N}(0,1),$$

где $\widehat{\tau}=\frac{1}{n}\sum_{i=1}^n L^2_{f,\widehat{F}_n}(X_i)$.

Это непараметрический Дельта-метод — мы видим, что при применении функционала f, асимптотическая нормальность сохраняется и можем пересчитать дисперсию.
Дельта-метод позволяет доказывать асимптотическую нормальность самых разнообразных функционалов от ЭФР, что, в свою очередь, позволяет строить на их основе доверительные интервалы и проверять гипотезы.

Задача 7. Получить следующую теорему: если распределение имеет непрерывную плотность и $p(x_{1/2})>0$, где $x_{1/2}$ — корень уравнения $F(x_{1/2})=1/2$, то

$$\sqrt{n}(X_{([(n+1)/2])}-x_{1/2})\xrightarrow{d}\mathcal{N}\left(0,\frac{1}{4p(x_{1/2})^2}\right).$$

Здесь $X_{([(n+1)/2])}$ — это $[(n+1)/2]$-й из X, упорядоченных по возрастанию.

Нетрудно заметить, что среди метод упрощает построение Стьюдентовских интервалов, поскольку позволяет сократить расходы на подсчет дисперсии.

Задача 8. Рассматривая $f(F)=F(1/3)-F(1/4)$, найти для $R[0,1]$ с помощью функции влияния доверительный интервал для $f(F)$ а) обычный б) бутстрэповский стьюдентовский.

4.4 Оценки для ЭФР и размерность Вапника-Червоненкиса

Этот параграф требуется только если вы сдаете на продвинутом уровне

4.4.1 Эмпирическое распределение и его близость к теоретическому

Зачастую нам требуется оценить не функцию распределения, а само распределение $\mathbf{P}(X\in A)$. Естественно делать это с помощью эмпирического распределения

$$\mathbf{P}_X(A):=\mathbf{P}(X\in A)\approx\widehat{\mathbf{P}}_n(A):=\frac{1}{n}\sum_{i=1}^n I_{X_i\in A}.$$

Погрешность приближения ЭФР мы умеем оценивать с помощью неравенства Дворецкого-Кифера-Вольфовица, а можно ли оценить погрешность приближения эмирическим распределением теоретического? При фиксированном A можно показать, что

$$\mathbf{P}(|\widehat{\mathbf{P}}_n(A)-\mathbf{P}_X(A)|>\varepsilon)\le 2e^{-2n\varepsilon^2}.$$

Это неравенство основано на том, что величина $n\widehat{\mathbf{P}}_n(A)$ имеет биномиальное распределение $Binom(n,\mathbf{P}_X(A))$ и является частным случаем так называемого неравенства Хёфдинга.

Мы бы хотели получить оценку сверху по всем A из некоторого множества $\mathcal{A}$, т.е. оценить $\mathbf{P}(\sup_{A\in\mathcal{A}}|\widehat{\mathbf{P}}_n(A)-\mathbf{P}_X(A)|>\varepsilon)$.

Пример 3. При одноточечном множестве $\mathcal{A}$ у нас уже есть оценка Хёфдинга.
При k-точечном множестве мы можем записать

$$\mathbf{P}(\sup_{A\in\mathcal{A}}|\widehat{\mathbf{P}}_n(A)-\mathbf{P}_X(A)|>\varepsilon)\le k\sup_A\mathbf{P}(|\widehat{\mathbf{P}}_n(A)-\mathbf{P}_X(A)|>\varepsilon)\le 2ke^{-2n\varepsilon^2}.$$

При $\mathcal{A}=\mathcal{B}(\mathbb{R})$ (то есть когда мы рассматриваем все A разом) для любого непрерывного распределения $\mathbf{P}_X$ при любом ω найдется множество $A=\{X_1(\omega),...,X_n(\omega)\}$, для которого $\widehat{\mathbf{P}}_n(A)=1$, а $\mathbf{P}_X(A)=0$. Соответственно, при каждом ω найдется такое множество A, что $\widehat{\mathbf{P}}_n(A)-\mathbf{P}_X(A)=1$. Значит наша вероятность будет равна 0 при любом $\varepsilon<1$.

4.4.2 VC-размерность

Тем не менее, для каких-то не очень больших множеств $\mathcal{A}$ мы надеемся получить хорошую оценку сверху. Для этого нам понадобится понятие VC-размерности.

Рассмотрим множество $R=\{x_1,...,x_m\}$. Рассмотрим какую часть его подмножеств мы можем получить, пересекая R с $A\in\mathcal{A}$. Назовем их количество $N_{\mathcal{A}}(R)$.

Пример 4. Если $\mathcal{A}=\mathcal{B}(\mathbb{R}^n)$, то $N_{\mathcal{A}}(R)$ будет равно 2^m, потому что пересечением с борелевскими множествами можно получить любое подмножество R.
Если $\mathcal{A}$ состоит из k элементов, то $N_{\mathcal{A}}$ не больше k по определению. При разных R эта величина будет принимать различные значения от 1 до k
Если $\mathcal{A}=\{(-\infty,x],x\in\mathbb{R}\}$ — множество всех лучей, то $N_{\mathcal{A}}(R)=m+1$, потому что мы можем получить либо пустое множество, либо множество из 1 самого маленького элемента R, либо из двух наименьших и так далее.
Если $\mathcal{A}=\{(-\infty,x],x\in\mathbb{Z}\}$, то $N_{\mathcal{A}}(R)$ будет зависеть от множества R. Скажем, если все элементы R лежат от 0 до 1, то $N_{\mathcal{A}}(R)=2$, поскольку можно будет получить только пустое множество и R, а если они принимают значения 1,2,...,m, то мы сможем получить те же $m+1$ элемент что и прежде.

Величину $s(\mathcal{A},m)=\sup_R N_A(R)$ называют shatter coefficient. Оказывается, что справедливо следующее неравенство (Вапник, Червоненкис, 1971):

$$\mathbf{P}(\sup_{A\in\mathcal{A}}|\widehat{\mathbf{P}}_n(A)-\mathbf{P}_X(A)|>\varepsilon)\le 8s(\mathcal{A},n)e^{-\frac{n\varepsilon^2}{32}}.$$

Пример 5. Если $\mathcal{A}=\mathcal{B}(\mathbb{R})$, то неравенство приобретает вид

$$\mathbf{P}(\sup_{A\in\mathcal{A}}|\widehat{\mathbf{P}}_n(A)-\mathbf{P}_X(A)|>\varepsilon)\le 2^{n+3}e^{-\frac{n\varepsilon^2}{32}}$$

Это малоинтересная оценка, поскольку она больше единицы.

Пример 6. При $\mathcal{A}$, состоящем из k множеств A неравенство приобретает вид

$$\mathbf{P}(\sup_{A\in\mathcal{A}}|\widehat{\mathbf{P}}_n(A)-\mathbf{P}_X(A)|>\varepsilon)\le 8ke^{-\frac{n\varepsilon^2}{32}},$$

Это ухудшенная версия неравенства Хеффдинга.

Пример 7. При $\mathcal{A}$, состоящем из множества $(-\infty,x]$ неравенство приобретает вид

$$\mathbf{P}(\sup|\widehat{F}_n(x)-F_X(x)|>\varepsilon)\le 8(n+1)e^{-\frac{n\varepsilon^2}{32}},$$

Это ухудшенная версия неравенства Дворецкого-Кифера-Вольфовица.

В отличие от неравенства ДКВ неравенство Вапника-Червоненкиса пригодно и для других случайных элементов, а не только случайных величин — случайных векторов, процессов и других.

Давайте получим более простой метод оценки, чем прямой подсчет $s(\mathcal{A}, n)$. Назовем размерностью Вапника-Червоненкиса $\mathcal{A} \subset \mathcal{B}(\mathbb{R})$ а) ∞, если $s(\mathcal{A}, n) = 2^n$ при всех n,
б) $\max\{k : s(\mathcal{A}, k) = 2^k\}$ иначе.

Будем обозначать размерность Вапника-Червоненкиса $VC(\mathcal{A})$. Легко понять, что если $s(\mathcal{A}, k) < 2^k$ при каком-то k, то $s(\mathcal{A}, l) < 2^l$ при всех $l \geq k$, поэтому достаточно найти первый номер k при котором неравенство нарушается.
Иначе говоря, размерность Вапника-Червоненкиса системы $\mathcal{A}$ — это самое большое k при котором для какого-то k-точечного множества можно, пересекая его с элементами $\mathcal{A}$, получить все его подмножества.

Пример 8.
- Для $\mathcal{A} = \mathcal{B}(\mathbb{R}^n)$ мы для любого n-точечного множества B_n можем получить пересечением с $\mathcal{A}$ все подмножества B_n. Значит, в силу а) имеем $VC(\mathcal{A}) = \infty$.
- Для $\mathcal{A}$, состоящего из $(-\infty, x]$ при всех x, $VC(\mathcal{A}) = 1$. Действительно, для любого двухточечного множества $\{a, b\}$, $a < b$, мы не сможем получить точку b пересечением нашего множества с каким либо лучом $(-\infty, x]$.
- Для $\mathcal{A}$, состоящего из всех отрезков $[x, y]$ $VC(\mathcal{A}) = 2$. Действительно, для любого множества R из трех элементов $R = \{a, b, c\}$, $a < b < c$ нельзя получить множество $\{a, c\}$ пересечением $\{a, b, c\}$ с каким-либо отрезком. С другой стороны для двухточечного множества $\{a, b\}$ можно получить такими пересечениями и $\emptyset$, и $\{a\}$, и $\{b\}$ и $\{a, b\}$.
- Пусть $\mathcal{A}$ — множество всех полуплоскостей на плоскости. Тогда $VC(\mathcal{A}) = 3$.
  Докажем это. Если 3-точечное множество R состоит из точек, не лежащих на одной прямой, то можно отделить любую из этих точек от 2 других прямой. Тем самым можно получить в пересечении R с некоторой полуплоскостью любое подмножество из 1 и 2 точек. Получить $\emptyset$ и R также не представляет труда.
  С другой стороны, никакое 4-точечное множество R не имеет $N_{\mathcal{A}}(R) = 2^4$. Убедиться в этом нетрудно в каждой конфигурации:
  Если они образуют выпуклый четырехугольник ABCD, то не существует полуплоскости, пересекающейся с ним по A,C, но не пересекающейся по B,D. Если четырехугольник невыпуклый и A лежит внутри треугольника BCD, то B,C,D не могут лежать в этом пересечении без A. И, наконец, если 3 точки лежат на одной прямой, то нельзя получить пересечение с полуплоскостью две крайних из них, не получив в том же пересечении среднюю. Поэтому $VC(\mathcal{A}) = 3$.

Вопрос 11. А чему равна VC-размерность множества всех прямоугольников с осями параллельными осям координат?

Вопрос 12. Тот же вопрос для параллелипипедов в $\mathbb{R}^d$

С помощью VC-размерности можно оценить $s(\mathcal{A}, n)$:

$$s(\mathcal{A}, n) \leq n^{VC(\mathcal{A})} + 1.$$

При этом подсчет размерности Вапника-Червоненкиса значительно более прост, чем подсчет $s(\mathcal{A}, n)$.

В этот момент у читателя, вероятно, возник вопрос — а зачем все эти сложности? У нас есть неравенство Хеффдинга, позволяющее нам оценить вероятность попадания в любое множество, зачем же нам возиться с супремумами?

Пример 9. Представим, что мы хотим построить 95% доверительный параллелипипед вида $[a_1, b_1] \times [a_2, b_2] \times ... \times [a_d, b_d]$ для X_{n+1} на основе выборки $X_1, ..., X_n \in \mathbb{R}^d$. Допустим, мы выберем какой-то параллелипипед I (зависящий от $X_1, ..., X_n$), такой что в него попали k из наших X_i. Тогда $\widehat{P}_n(I) = \frac{k}{n}$. Оценим максимальное возможное значение $\mathbf{P}(X_{n+1} \notin I(X_1, ..., X_n))$:

$$\mathbf{P}(X_{n+1} \notin I(X_1, ..., X_n)) = \mathbf{E}\left(\mathbf{P}(X_{n+1} \notin I(X_1, ..., X_n) | X_1, ..., X_n)\right)$$

Это математическое ожидание разбивается на две части:
а) $X_1, ..., X_n : \mathbf{P}(X_{n+1} \notin I(X_1, ..., X_n) | X_1, ..., X_n) \leq 1 - k/n + \varepsilon$,

б) $X_1, ..., X_n : \mathbf{P}(X_{n+1} \notin I(X_1, ..., X_n) | X_1, ..., X_n) > 1 - k/n + \varepsilon$.
В случае а) математическое ожидание не превосходит $1 - k/n + \varepsilon$, в случае б) не превосходит

$$\mathbf{P}(\exists \tilde{I} : |\widehat{P}_n(\overline{\tilde{I}}) - \mathbf{P}(X \in \overline{\tilde{I}})| > \varepsilon) = \mathbf{P}(\exists \tilde{I} : |\widehat{P}_n(\tilde{I}) - \mathbf{P}(X \in \tilde{I})| > \varepsilon),$$

где $\tilde{I}$ — некоторый параллелепипед. Неравенство Вапника-Червоненкиса позволяет нам дать нужную оценку для последнего выражения:

$$8(n^v + 1)e^{-\frac{n\varepsilon^2}{32}}$$

для последней вероятности, где v — VC-размерность множества всех параллелепипедов, равная $2d$ в силу вопроса выше. То есть вероятность непопадания в I не больше

$$1 - \frac{k}{n} + \varepsilon + 8(n^v + 1)e^{-\frac{n\varepsilon^2}{32}}$$

и если ε достаточно мало, n достаточно велико, а k/n достаточно близко к 1, то мы можем быть сделана сколь угодно близкой к 1.

Оценки, полученные Вапником и Червоненкисом, были позже значительно улучшены. Так Devroye (1982) показал, что правую часть можно заменить на

$$\mathbf{P}(\sup_{A \in \mathcal{A}} |\widehat{P}_n(A) - \mathbf{P}_X(A)| > \varepsilon) \leq 4e^{4\varepsilon(1+\varepsilon)} s(\mathcal{A}, n) e^{-2n\varepsilon^2}.$$

Наиболее удачная из известных мне оценок принадлежит Lugosi (1995)

$$\mathbf{P}(\sup_{A \in \mathcal{A}} |\widehat{P}_n(A) - \mathbf{P}_X(A)| > \varepsilon) \leq 4e(v+1) \left(\frac{32e^5 n^2 \varepsilon^3}{v^2} \right)^v e^{-2n\varepsilon^2}$$

и справедлива при $n\varepsilon^2 > v/2$, где $v = VC(\mathcal{A})$. Стоит отметить, что есть и асимптотически более удачное неравенство (Talagrand, 1994), в котором, увы, не указана явно константа.

Задача 9. Построить непараметрический доверительный прямоугольник вероятности не менее 95% на основе $X_1, ..., X_n$ а) для выборки из независимых $\mathcal{N}(0, 1)$ б) равномерной на единичном квадрате выборки при $n = 100000$. Насколько часто в него попадает выборка?

Конечно, указанный подход а) работает только при больших выборках (порядка 10^5) б) дает только нижнюю и верхнюю оценки уровня доверия. Зато он позволяет построить доверительное множество любой формы, лишь бы мы могли подсчитать в этом случае VC-размерность. Кроме того, мы получаем оценки для вероятностей попадания во все множества такой формы разом.

4.5 Ответы на вопросы

Ответ 1. Несмещенность вытекает из линейности матожидания

$$\mathbf{E}\frac{1}{n}\sum_{i=1}^{n} I_{X_i \leq x} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{E} I_{X_i \leq x} = \mathbf{P}(X_1 \leq x) = F_X(x).$$

Состоятельность — прямое следствие ЗБЧ, а асимпотическая нормальность — ЦПТ, примененных к $I_{X_i \leq x}$.

Ответ 2. В силу ЦПТ

$$\sqrt{n}(\widehat{F}_n(x) - F(x)) \xrightarrow{d} Z \sim \mathcal{N}(0, F(x)(1 - F(x))).$$

Отсюда

$$\frac{\sqrt{n}(\widehat{F}_n(x) - F(x))}{\sqrt{\widehat{F}_n(x)(1 - \widehat{F}_n(x))}} \overset{Z}{\sim} \mathcal{N}(0, 1).$$

Следовательно, получаем доверительный интервал

$$F(x) \in \left(\widehat{F}_n(x) - \frac{z_{1-\alpha/2}}{\sqrt{n}} \sqrt{\widehat{F}_n(x)(1-\widehat{F}_n(x))}, \widehat{F}_n(x) - \frac{z_{\alpha/2}}{\sqrt{n}} \sqrt{\widehat{F}_n(x)(1-\widehat{F}_n(x))} \right).$$

Ответ 3. Если значения нашей случайной величины $y_1, ..., y_m$ с вероятностями $p_1, ..., p_m$, а выборка $x_1, ..., x_n$, то а) если какой-то из x_i не встречается среди y_i, то правдоподобие 0.
б) иначе правдоподобие принимает вид $L = p_1^{N_1} ... p_m^{N_m}$, где N_1,...,N_m — число x_i, равных $y_1, ..., y_m$ соответственно. Но тогда будем рассматривать параметры $p_1, ..., p_{m-1}$ (и $p_m = 1 - p_1 - ... - p_{m-1}$):

$$\ln L(p_1, ..., p_{m-1}) = \sum_{i=1}^{m-1} N_i \ln p_i + N_m \ln(1 - p_1 - ... - p_{m-1}), \ \frac{\partial}{\partial p_i} \ln L = \frac{N_i}{p_i} - \frac{N_m}{p_m},$$

То, что найденная точка максимум, можно показать напрямую или рассматривая пару p_i, p_j, фиксируя их сумму.

Ответ 4. Нет, из равномерной сходимости F_n к F не следует, что

$$\int_{\mathbb{R}} a(x) dF_n(x) \to \int_{\mathbb{R}} a(x) dF(x),$$

если функция a не является ограниченной. Например, для $a(x) = x$ можно взять $F_n = (1 - 1/n)F(x) + nI_{x \geq n}$. Тогда

$$\int_{\mathbb{R}} x dF_n(x) = \left(1 - \frac{1}{n}\right) \int_{\mathbb{R}} x dF(x) + \frac{1}{n} \cdot n \to \int_{\mathbb{R}} x dF(x) + 1.$$

Ответ 5. Если $x_1, ..., x_n$ (значения элементов выборки) различны, то $\widehat{F}_n(x)$ — ф.р. дискретной случайной величины $\widehat{X}$ со значениями $x_1, ..., x_n$ и вероятностями $1/n$. Тогда

$$\mathbf{E}\widehat{X} = x_1 \frac{1}{n} + x_2 \frac{1}{n} + ... + x_n \frac{1}{n} = \overline{x}, \ \mathbf{D}\widehat{X} = (x_1 - \overline{x})^2 \frac{1}{n} + ... + (x_n - \overline{x})^2 \frac{1}{n} = S^2.$$

Если же какие-то значения совпадают, то слагаемые, соответствующие одинаковых x_i, объединятся.

Ответ 6. Первые слагаемые в обоих выражениях совпадают. Рассмотрим матожидание левой части второго из них

$$\mathbf{E} \frac{2}{n} \sum_{i=1}^{n} \widehat{f}_{n,i}(X_i) = \frac{2}{n} \sum_{i=1}^{n} \mathbf{E} \widehat{f}_{n,i}(X_i).$$

В силу линейности матожидания

$$\mathbf{E}\widehat{f}_{n,i}(X_i) = \frac{1}{n-1} \sum_{j \neq i} \frac{1}{h_n} \mathbf{E} K\left(\frac{X_i - X_j}{h_n}\right) = \frac{1}{h_n} \int_{\mathbb{R}^2} K\left(\frac{x-y}{h_n}\right) f(x) f(y) dy dx.$$

С другой стороны,

$$\mathbf{E} \int_{\mathbb{R}} \widehat{f}_n(x) f(x) dx = \int_{\mathbb{R}} \frac{1}{n} \frac{1}{h_n} \sum_{j=1}^{n} \mathbf{E} K\left(\frac{x - X_j}{h_n}\right) dx = \frac{1}{h_n} \int_{\mathbb{R}} K\left(\frac{x-y}{h_n}\right) f(y) f(x) dy dx.$$

Подставляя полученные выражения в правую и левую части искомого тождества, получаем требуемое

Ответ 7.

$$\int_{\mathbb{R}} x \widehat{f}_n(x) dx = \frac{1}{n} \frac{1}{h_n} \sum_{i=1}^{n} \int_{\mathbb{R}} x K\left(\frac{x - X_i}{h_n}\right) dx = \frac{1}{n} \sum_{i=1}^{n} \int_{\mathbb{R}} (y \cdot h_n + X_i) K(y) dy = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

где в последнем тождестве мы воспользовались тем, что $\int_{\mathbb{R}} yK(y)dy = 0$, $\int_{\mathbb{R}} K(y)dy = 1$.

Ответ 8. Пусть $X_1, ..., X_n \sim R[0,1]$, $A = \max X_i$. Тогда в выборке $X_1^*, ..., X_n^*$ с возвращением из $X_1, ..., X_n$

тот же максимум останется с вероятностью

$$1 - \left(1 - \frac{1}{n}\right)^n \to 1 - e^{-1},\ n \to \infty.$$

Таким образом, bootstrap в 63% случаев даст смещение 0. Значит 60% доверительный интервал для θ на основе максимума может оказаться нулевой длины.

Ответ 9.

$$\int_{\mathbb{R}} a(y)d((1-p)F_X(y) + p\delta_x(y)) = pa(x) + (1-p)\int_{\mathbb{R}} a(y)dF_X(y),$$

поскольку

$$\int_{\mathbb{R}} a(y)d\delta_x(y) = \mathbf{E}a(x) = a(x).$$

Следовательно,

$$L_{f,F}(x) = a(x) - \int_{\mathbb{R}} a(y)dF_X(y).$$

Ответ 10. Поскольку рассматриваемый функционал линеен и $f(F+\varepsilon D) = f(F)+\varepsilon f(D)$, то выражение, фигурирующее в определении дифференцируемости по Адамару, тождественно равно 0.

Ответ 11. Рассмотрим произвольные 5 точек. Выберем среди них одну из самых левых, одну из самых правых, одну из самых верхних, одну из самых нижних. Тогда не найдется прямоугольника, содержащего их, но содержащего пятую точку.
Для 4 точек (1,0), (0,1), (-1,0), (0,-1) можно получить пересечением с прямоугольниками любой набор. Значит $VC(\mathcal{A}) = 4$.

Ответ 12. Ответ 2d получается совершенно аналогично.