1 Введение

1.1 Основная задача статистики

Рассмотрим несколько задач:

1. Представим себе, что мы подбрасываем монету с неизвестной вероятностью успеха (орла) θ n раз. Видя результаты, мы хотим сделать выводы о том, какая была вероятность успеха.
2. Командир орудийного расчета делает серию пристрелочных залпов, а затем выбирает под каким углом стрелять, чтобы поразить цель.
3. Представим себе, что мы наблюдаем за погодой и хотим построить прогноз среднего значения температуры на следующий день, найдя диапазон значений в котором она может изменяться.
4. Геолог делает несколько замеров в поисках нефти в некоторой области и исходя из этого пытается представить, где расположена нефть. Он хочет локализовать район, в котором в дальнейшем проводить исследования, скорее всего содержащий
5. Игрок проводит серию игр за покерным столом и исходя из результатов этих пробных партий, решает стоит ли ему играть с этими игроками (положителен ли средний выигрыш в партии).
6. Медики на основе анализов пациента решают болен ли он коронавирусом.

Все это – прикладные статистические задачи, которые можно (исходя из некоторых предположений) перевести в математические модели:

- Есть набор н.о.р. величин $X_i \sim Bern(\theta)$, $\theta \in (0,1)$. Исходя из принятых ими при бросании (то есть при фиксировании ω) значений $X_1(\omega), \dots, X_n(\omega)$, необходимо сделать вывод о том значении θ, которое использовано в их распределении.
- Если набор фиксированных величин a_i – настроек пушки при i-м выстреле и набор независимых, но разнораспределенных (зависящих от a_i) случайных величин X_i – точек падения снарядов. Требуется на основе $X_1(\omega), \dots, X_n(\omega)$ оценить значение a для которого X будет близок (например, в среднем) к заданной точке x.
- Есть набор зависимых величин X_i, требуется на основе их значений $X_1(\omega), \dots, X_n(\omega)$ предсказать некий параметр их распределения, построив интервал, в который эта величина скорее всего попадет.
- Есть набор независимых величин X_i – замеров в точках (u_i, v_i) с распределением, зависящем от x_i, y_i и параметра – области пространства. Требуется на основе их значений $X_1(\omega), \dots, X_n(\omega)$ оценить область, определяющую распределение этих величин.
- Есть набор н.о.р. величин X_i. Требуется на основе их значений $X_1(\omega), \dots, X_n(\omega)$ сделать вывод, положительно ли математическое ожидание X_i.
- Есть некоторый вектор зависимых величин X_i (результатов анализов). Наши величины получены из некоторого распределения F, которое может принадлежать некоторому множеству $\mathcal{F}_1$, отвечающему здоровым людям или $\mathcal{F}_0$, отвечающему больным. Требуется на основе их значений $X_1(\omega), \dots, X_n(\omega)$ сделать вывод о принадлежности их распределения к одному из множеств.

Во всех этих моделях у нас есть некоторый набор величин X_i (вообще говоря, зависимых), чье распределение неизвестно. Мы хотим сделать какой-то вывод об этом распределении исходя из принятых на нашем ω значений величин $X_1(\omega), \dots, X_n(\omega)$. Давайте построим общую схему такого рода.

1.2 Статистическое пространство

В теории вероятностей вы изучали поведение случайных величин, зная их распределения. Основная задача статистики, в некотором смысле, обратная — сделать вывод о распределении случайных величин, наблюдая за их значениями.

Начнем с наиболее общей постановки.

Определение 1. Статистическим пространством называют множество $\mathcal{X}$, сигма-алгебру $\mathcal{F}$ на нем и семейство вероятностных мер $\mathcal{P}$ на $\mathcal{F}$.

Основной объект с которым мы будем работать – выборка, случайный элемент со значениями в $\mathcal{X}$, имеющий распределение $P \in \mathcal{P}$. А наблюдать мы будем реализацию – конкретный элемент $\mathcal{X}$, значение, принятое выборкой.

Наиболее частый случай, с которым мы будем работать: $\mathcal{X} = \mathbb{R}^{nk}$, $\mathcal{F} = \mathcal{B}(\mathcal{X})$, $\mathcal{P} = \{P_\theta \times ... \times P_\theta, \theta \in \Theta\}$, где P_θ – некоторые вероятностные меры на $\mathbb{R}^k$, $\Theta \subset \mathbb{R}^m$.

Таким образом, выборка будет представлять собой набор независимых величин, чье распределение известно с точностью до векторного параметра, который нас интересует. Реализация при этом будет числовым вектором – значениями наших величин. Наша задача при каждом θ сделать некоторые выводы о θ на основе $X_1, \dots, X_n \sim P_\theta$.

Пример 1. Например, в модели с подбрасыванием монеты мы имеем $\mathcal{X} = \{0,1\}^n$, $\mathcal{F} = 2^{\mathcal{X}}$,

$$\mathbf{P}_\theta(x_1) = \theta^{x_1}(1-\theta)^{1-x_1}.$$

Последняя формула изящно отражает запись

$$\mathbf{P}_\theta(x_1) = \begin{cases} \theta, & x_1 = 1, \\ 1-\theta, & x_1 = 0. \end{cases}$$

Соответственно, выборка – набор n н.о.р. $Bern(\theta)$ величин, реализация – конкретный принятый ими набор значений $\{0,1\}$. Параметр θ, использованный для генерирования нашей выборки, нам неизвестен, мы бы хотели сделать вывод о нем исходя из реализации.

В описанном случае распределение параметрическое – оно известно с точностью до скалярного или векторного параметра θ. Такие модели называют *параметрическими*. В такую категорию попадает модель 1) бросания монеты, туда же попадет модель 5) покерного стола, если мы умеем задавать распределение выигрыша исходя из характеристик остальных игроков.

В модели 4) залежей нефти параметром является не число или вектор, а целое множество. Если модель 5) покерного стола не включает описание распределения выигрыша, то опять же распределение будет иметь довольно общий вид, а наша задача будет состоять в оценке математического ожидания этого неизвестного распределения. Такие модели называют *непараметрическими*.

Кроме того, в некоторых моделях (модель 2) орудийного расчета, модель 3) погоды и модель 6) анализов) данные зависимы или по-разному распределены. В этом случае выборка уже не будет представлять н.о.р. величины. При этом модели могут быть параметрическими (например, промах орудийного расчета может быть двумерным нормальным с некоторыми неизвестными параметрами, зависящими от x_i).

Заметьте, что мы не говорим о вероятностном пространстве, на котором заданы X_i. Все дело в том, что мы будем существовать в рамках модели, где нас интересуют только функции от X_i. Распределение такой функции полностью определяется распределением X_i и совершенно неважно как устроено их вероятностное пространство. Можно считать, что вероятностное пространство и есть $(\mathcal{X}, \mathcal{F}, P)$.

Обратите внимание на важный момент. В теории вероятностей мы привыкли к использованию меры $\mathbf{P}$, которая задана в нашем вероятностном пространстве. Статистическое пространство не связано с отдельной мерой, поэтому вероятности событий мы будем записывать в виде $\mathbf{P}_\theta(A)$. Это уже не одна вероятностная мера, а набор вероятностных мер, параметризованный неизвестным нам θ.

1.3 Параметрическая и непараметрическая статистика

В нашем курсе мы будем преимущественно работать в параметрическом случае и формально в курсе лекций даже не вводится непараметрическая модель. Откуда же берется параметрическая модель — как мы можем заранее предполагать, что наблюдения будут из заданного семейства?

Рассмотрим несколько случаев, когда это естественно:

1. Мера $\mathbf{P}$ дискретная, причем число принимаем величиной X значений невелико. В этом случае мы имеем естественную параметризацию — дискретное распределение со значениями $x_1, ..., x_k$ параметризуется $k-1$ параметром p_1,...,p_{k-1}. Скажем, в случае монетки у нас потенциально есть лишь один неизвестный параметр — вероятность орла.

2. Параметризация может возникнуть из физических особенностей модели. Например, наблюдая за движение мелкой частицы в жидкости, мы из физических соображений понимаем, что это соответствует броуновскому движению и потому изменения ее координат имеют нормальное распределение.

3. Параметризация может возникнуть из предыдущих опытов в схожей области. Скажем, мы запускаем на рынок новое лекарство, а для предыдущих лекарств видели, что распределение хорошо приближается каким-то известным распределением. Тогда естественно предположить, что и здесь мы увидим ту же картину. Параметры распределения при этом будут отвечать уже особенностям реализации данного конкретного лекарства и мы должны будем их оценить исходя из наблюдений за продажами.

4. Параметризация может возникнуть из вероятностных соображений, например, из центральной предельной теоремы или закона Пуассона. Скажем, накладывая на сигнал большое количество мелких однородных шумов, мы в силу ЦПТ получим распределение, близкое к нормальному. Скажем, в модели 2) орудийного расчета мы можем предполагать, что отклонение снарядов вызвано множеством мелких случайных независимых факторов, что приведет к параметризации $Y_i \sim \mathcal{N}(f(x_i) + \theta_1, \theta_2^2)$, где $f(x_i)$ – точка падения снаряда без учета случайности, θ_1 – систематический фактор (например, снос снаряда ветром), а θ_2 – дисперсия случайных воздействий.

Конечно, любая параметризация — это некоторое допущение, огрубление реальности, зато она позволяет точнее работать с данным конкретным случаем. Исследователь сам решает, что для него важнее – огрубить задачу параметризацией, но выиграть на более точном решении задач с ее использованием или же использовать общую непараметрическую модель.

1.4 Три задачи статистики

Наиболее распространены три статистические задачи.

- В моделях 1) бросания монеты или модели 2) орудийного расчета нашей целью было получения числа, приближающего интересующий нас параметр. В таком случае мы должны постараться построить *точечную оценку* — на основе реализиации $x_1, ..., x_n$ нашего контрольного эксперимента мы пытаемся получить какое-то число $\widehat{\theta}(x_1, ..., x_n)$, которое "близко" к нашему θ в каком-то смысле. При конкретной реализации нам может и не повезти, но вот на выборке мы ожидаем получить случайную величину $\widehat{\theta}(X_1, ..., X_n)$, которая в каком-то смысле часто близка к θ.

  Как именно можно измерять качество оценки, мы обсудим позднее.

  Для удобства работы, здесь и далее мы будем предполагать, что рассматриваемые нами функции $\widehat{\theta}()$ от выборки являются измеримыми функциями из $(\mathcal{X}, \mathcal{S})$ в некоторое пространство $(Y, \mathcal{C})$, где $\mathcal{C}$ — заданная на Y сигма-алгебра.

  Определение 2. Измеримые функции от выборки называют *статистиками*. Если мы хотим подчеркнуть, что такая функция использует для оценки параметра, то мы называем ее *оценкой*.

  Обычно Y у нас будет $\mathbb{R}$ или $\mathbb{R}^m$, то есть статистики будут либо числом, либо вектором.

- В модели 3) синоптиков и в модели 4) нефтяных залежей мы хотим получить соответственно интервал, который содержит искомый параметр или область, которая скорее всего содержат искомое множество залежей нефти. Здесь уже не очень удобно использовать точечную оценку, а удобнее задать некоторый интервал (область), в котором наш параметр достаточно вероятно находится. Иначе говоря, в случае одномерного параметра θ (как в модели 3)) мы задаем некоторую вероятностью α, а получаем две статистики $\widehat{\theta}_1(X_1, ..., X_n)$, $\widehat{\theta}_2(X_1, ..., X_n)$, таких, что

$\mathbf{P}_\theta(\theta \in (\widehat{\theta}_1(X_1, ..., X_n), \widehat{\theta}_2(X_1, ..., X_n))) = 1 - \alpha$. Такой интервал $(\widehat{\theta}_1, \widehat{\theta}_2)$ называется *доверительным интервалом* для параметра θ. В более общем случае многомерного параметра (или даже еще более общего случая модели 4), где параметром является множество) мы ищем случайное множество $D(X_1, ..., X_n)$, которое содержит наш параметр (в данном случае область залежей) с вероятностью $1 - \alpha$.

Увеличивая α, мы будем получать достаточно маленькое множество или узкий диапазон средних температур, что позволяет нефтяникам или метеорологам сузить интервал поиска параметра, однако, при этом с большой долей вероятности соответствующеее множество или интервал не будут включать требуемый параметр. А вот при маленьком α мы наверняка будем иметь интервал или множество, содержащие наш параметр, однако, излишне большой.

Заметим, что при этом, вообще говоря, нам требуется, чтобы приведенное выше равенство из определения доверительного интервала выполнялось лишь при том параметре θ, который на самом деле реализован. То есть вероятность попадания именно в нашем мире попадания соответствующего θ в наш интервал должна быть $1 - \alpha$, а от остальных миров мы ничего не требуем. Увы, поскольку мы не знаем, какое именно значение принимает θ, поэтому удобнее требовать выполнение этого равенства при всех $\theta \in \Theta$.

Стоит также обратить внимание на то, что в рамках нашей модели у доверительного интервала случайные границы, и в рассматриваемой вероятности мы рассматриваем те $\omega \in \Omega$, при которых при фиксированном θ интервал $\widehat{\theta}_1$, $\widehat{\theta}_2$ накроет θ. По сути мы говорим о том, что с такой-то вероятностью интервал накроет θ, а не наоборот θ попадет в интервал.

- Третий вид задач называется проверкой гипотез. Для врача в модели 6) или игрока в покер в модели 5) важно понять выполнено ли интересующее их утверждение (болен ли человек или квалифицированы игроки). Им совершенно бесполезна точечная оценка, доверительный интервал также не слишком полезен, он может содержать одновременно и больных, и здоровых, и положительный средний выигрыш, и отрицательный. Для нас важно понять лежит ли неизвестный параметр в множестве Θ_0 или неизвестное распределение в множестве Θ_0. Таким образом, мы приходим к третьей задаче — проверке гипотез.

  Иначе говоря, мы обладаем некоторой гипотезой $H_0 : \theta \in \Theta_0$ и альтернативой $H_1 : \theta \in \Theta_1$ (в непараметрическом случае $F \in \mathcal{F}_0$ и $F \in \mathcal{F}_1$ соответственно). Мы бы хотели построить *решающее правило* d, то есть функцию из $\mathbb{R}^n$ в $\{0, 1\}$, которая реализации $(x_1, ..., x_n)$ будет сопоставлять номер гипотезы, которую мы должны принять. При этом у нас имеется два вида ошибок: гипотеза верна, а мы ее отвергли — ошибка первого рода или гипотеза неверна, а мы ее приняли — ошибка второго рода. Классический подход предлагает ограничить шансы ошибки первого рода, то есть вероятность ошибки первого рода (отвержения гипотезы H_0, хотя она верна) должна быть не больше некоторого заданного параметра α. Иначе говоря,

$$\mathbf{P}_\theta(d(X_1, \ldots, X_n) = 1) \leq \alpha, \quad \theta \in \Theta_0.$$

  Вообще говоря, мы могли бы всегда принимать гипотезу, тогда ошибка первого рода была бы равна нулю. Но при этом значительно возросла бы ошибка второго рода, то есть вероятность принятия H_0, хотя она неверна. Наша задача среди критериев ошибающихся в первом роде редко (с вероятность не больше α) ошибаться во втором роде по возможности реже.

1.5 Резюме

Итак, в рамках курса мы будем работать с выборкой – набором случайных величин $(X_1, \ldots, X_n)$, имеющих некоторое неизвестное распределение. В непараметрической модели класс распределений достаточно общий, а в параметрической – заданным с точность до неизвестного векторного параметра.

Чаще всего, это будет набор н.о.р. величин с ф.р. F_θ, где F – заданные распределения (например, нормальные), а θ – их неизвестные параметры.

Наша задача — оценить параметр θ (построить функцию от $X_1, \ldots, X_n$, приближающую θ), доверительно оценить параметр (построить множество, зависящее от $X_1, \ldots, X_n$, с заданной вероятностью

накрывающее θ) или проверить гипотезу (построить критическую функцию $d(x_1, \ldots, x_n)$, принимающую значение 1, если следует отвергнуть гипотезу и 0 иначе).

1.6 Итоговая классификации рассмотренных моделей

Для спортивного интереса предложим возможные описания каждой из шести моделей. Я буду использовать одну из возможных интерпретаций, поскольку модели заданы достаточно общо.

Модель	$\mathcal{X}$	$\mathcal{F}$	$\mathcal{P}$	Искомый параметр	Тип задачи
1) Бросание монеты	$\{0,1\}^n$	$2^{\mathcal{X}}$	P_1	Параметр Бернулли	Точечная оценка
2) Орудийный расчет	$\mathbb{R}^{2n}$	$\mathcal{B}(\mathcal{X})$	P_2	$a : \theta_1 + f(a) = x$	Точечная оценка
3) Прогноз погоды	$\mathbb{R}^n$	$\mathcal{B}(\mathcal{X})$	P_3	параметр среднего	Доверительное оценивание
4) Поиск нефти	$\mathbb{R}^{3n}$	$\mathcal{B}(\mathcal{X})$	P_4	область $D \in \mathbb{R}^2$	Доверительное оценивание
5) Игра в покер	$(\mathbb{R}^+)^n$	$\mathcal{B}(\mathcal{X})$	P_5	среднее распределения	Проверка гипотезы
6) Анализы	$\mathbb{R}^n$	$\mathcal{B}(\mathcal{X})$	P_6	принадлежность к больным	Проверка гипотез

Здесь

- P_1 – множество распределений n н.о.р. бернуллиевских величин.
- P_2 – множество многомерных нормальных распределений с вектором средних $(f(a_1)+\theta_1, \ldots, f(a_n)+\theta_1)$ и блочной матрицей ковариации из одинаковых блоков 2×2 на диагонали, отвечающих за матрицу ковариаций координат точек падения каждого из снарядов. Оценивается при этом такая точка a, что $f(a) + \theta_1$ будет равным x. Функция f при этом может быть известной (тогда модель параметрическая) или неизвестной (тогда модель непараметрическая).
- P_3 – множество распределений зависимых n-мерных векторов. Конкретная модель определяется из метеорологических соображений. Например, модель $X_i = aX_{i-1} + bX_{i-2} + \varepsilon_i$, где a, b – некоторые коэффициенты, а ε_i – н.о.р. случайные флюктуации, предполагает, что завтрашняя температура определяется двумя последними днями и зависит от них линейно.
- P_4 – множество распределений векторов $(X_1, \ldots, X_n)$ с независимыми компонентами, чье распределение зависит от (u_i, v_i). Например, $X_i \sim F(x)$, если $(u_i, v_i) \in D$ и $X_i \sim G(x)$ при $(u_i, v_i) \notin D$, то есть замеры на залежах нефти имеют одно распределение, а вне – другое. Распределения F и G могут быть параметрически заданы исходя из геологических соображений.
- P_5 – множество распределений n н.о.р. величин.
- P_6 – множество распределений n-мерных векторов, разделенное на два класса: здоровых и больных.

В случаях 5 и 6 мы взяли непараметрические распределения, хотя в конкретных задачах (в особенности, если у нас есть данные прошлых опытов игры или анализов) мы можем иметь параметризацию, уточняющую как именно распределены наши выигрыши или анализы.