МАТЕМАТИЧЕСКАЯ СТАТИСТИКА
Лекция 3

§4 Достаточные статистики

4.1. Предварительные соображения и определение

Достаточная статистика — понятие фундаментальное. Часто возникает следующий вопрос. Имеется большая совокупность $\xi = (\xi_1, \xi_2 \dots \xi_n)$ наблюдений случайного характера, по которой нужно делать какие-либо выводы относительно чего-то неизвестного; обозначим это неизвестное через a. Можно ли сжать информацию, то есть хранить меньший объем данных, не потеряв при этом информацию об a?

Простейший пример: неизвестная вероятность события, n раз испытываем, получаем n наблюдений. Можно ли оставить только число успехов. Или мы при этом потеряем информацию?

Чтобы разобраться, сначала ответим на предварительный вопрос: нужны ли нам наблюдения ξ, если распределение $p_\xi(x)$ для ξ от a не зависит? меняется, а закон распределения остается одним и тем же. Ответ очевиден: наблюдения ξ не нужны, наблюдения ξ можем не хранить.

Следующий вопрос: имеется две совокупности наблюдений ξ и τ. Известно, что распределение $p_\tau(t,a)$ для τ зависит от a;
также известно, что *условное распределение ξ при условии известного значения τ от a не зависит*.
Нужны ли нам в этом случае наблюдения ξ? Ответ очевиден: нет, не нужны, мы можем сжать информацию, выбросив ξ и
оставив только τ, поскольку распределение $p_\tau(t,a)$ зависит от a.

Пусть $\xi \equiv (\xi_1, \xi_2 \dots \xi_n)$ — вектор наблюдений, принимающий значения $x \equiv (x_1, x_2 \dots x_n)$ и распределенный по закону
$P_\xi(x;a)$, где $a \equiv (a_1, a_2 \dots a_k)$ есть k-мерный параметр.

Пусть есть функция $T(x) = [T_1(x), T_2(x) \dots T_r(x)]$ (нам интересны значения $r < n$). Рассмотрим случайную величину $\tau = T(\xi)$.

Обозначим через $P_\tau(t;a)$ распределение τ.

Теперь имеем пару случайных величин ξ и τ.
При дальнейших рассуждениях полагаем случайную величину ξ дискретной (следовательно, τ - тоже).

Запишем закон распределения для пары; по формуле умножения вероятностей имеем:

$$P\{\xi = x; \tau = t = T(x); a\} = P\{\tau = t = T(x); a\} \cdot P\{\xi = x \mid \tau = T(\xi) = t; a\}.$$

Однако ясно, что слева написано распределение ξ,
т.к. из события $\{\xi = x\}$ следует событие $\{\tau = t = T(x)\}$:

$$P\{\xi = x; \tau = t = T(x)\} = P\{\xi = x; a\}.$$

В результате имеем соотношение

$$P\{\xi = x; a\} = P_{\tau}\{t = T(x); a\} \cdot P\{\xi = x \mid \tau = t = T(x); a\}. \quad (1)$$

Т.е. распределение для всей совокупности ξ есть произведение распределения статистики $\tau = T(\xi)$ на условное распределение ξ при условии известного значения τ.

Пусть второй сомножитель $P\{\xi = x \mid \tau = t; a\}$, т. е. условное распределение ξ, при условии известного значения τ, от *a* не зависит.

Это означает, что значение ξ ничего не добавляет к знаниям о параметре *a*, полученным на основании статистики $\tau = T(\xi)$, всю информацию об *a* содержит $\tau = T(\xi)$. Можно отбросить ξ, оставив только $\tau = T(\xi)$. В этом случае $\tau = T(\xi)$ называется достаточной статистикой для *a*.

Если ξ непрерывна, то рассуждения остаются справедливыми, нужно лишь в (1) вероятности заменить на плотности

$$p_{\xi}(x \mid a) = p_{\tau}(t = T(x); a)\, p_{\xi}(x \mid \tau = t = T(x), a). \quad (2)$$

Отсюда для условного распределения имеем

$$p_{\xi}(x \mid \tau = T(x), a) = \frac{p_{\xi}(x \mid a)}{p_{\tau}(t = T(x); a)}. \quad (3)$$

Если условное распределение ξ при известном $\tau = T(x)$ не зависит от a, то ξ можно не хранить, оставить только $\tau = T(x)$.

Определение. *Статистика $\tau \equiv T(\xi)$ называется **достаточной для параметра a**, если условное распределение ξ при условии известного значения $\tau \equiv T(\xi)$ от параметра не зависит, т.е. если*

$$p_{\xi}(x \mid \tau = T(x), a) = const(a).$$

Практический смысл достаточной статистики состоит в том, что любые статистические выводы о неизвестном параметре *a* можно делать без ущерба для качества, основываясь не на всех исходных данных, а только на достаточной статистике.

Верно очевидное утверждение, что

для любого способа $\delta(\xi)$ обработки всей информации ξ существует другой, эквивалентный способ обработки, основанный на достаточной статистике $\tau \equiv T(\xi)$.

Тривиальный эквивалентный способ состоит в следующем:

по исходным наблюдениям ξ вычисляем достаточную статистику $\tau \equiv T(\xi)$, условный закон распределения $p_{\xi}(x \mid \tau = T(x))$, от параметра не зависит;

мы берем генератор случайных чисел и генерируем с.в. ξ' с этим законом. При этом распределения для ξ' и ξ совпадают.

Применим исходную процедуру δ к ξ', $\delta(\xi')$, получаем результат, основанный на статистике $\tau \equiv T(\xi)$, *но он* эквивалентен $\delta(\xi)$.

Пример 1. Пусть для определения параметра λ некоторого однородного пуассоновского потока (например, источника радиоактивного излучения) n-кратно в течение промежутков времени одинаковой продолжительности ΔT измеряется количество поступающих частиц.

На языке математической статистики это означает, что имеется n независимых наблюдений ξ_1, $\xi_2 \ldots \xi_n$ над случайной величиной ξ, распределенной по закону Пуассона с неизвестным параметром $a = \lambda \Delta T$. Возникает вопрос: можно ли не хранить все значения x_1, $x_2 \ldots x_n$, которые приняли случайные величины ξ_1, $\xi_2 \ldots \xi_n$, а хранить только суммарное значение $\sum_{i=1}^{n} x_i$? Другими словами, является ли $\sum_{i=1}^{n} x_i$ достаточной статистикой?

Для ответа на этот вопрос нужно определить условную вероятность (3) получения значений x_1, $x_2 \ldots x_n$ при условии, что их суммарное значение известно и равно $t = \sum_{i=1}^{n} x_i$:

$$P\left\{\xi_1 = x_1, \xi_2 = x_2 \ldots \xi_n = x_n \middle| \sum_{i=1}^{n} \xi_i = t, a\right\} =$$

$$= \frac{P\{\xi_1 = x_1, \xi_2 = x_2 \ldots \xi_n = x_n \mid a\}}{P\left\{\sum_{i=1}^{n} \xi_i = t \middle| a\right\}} = \frac{\prod_{i=1}^{n} e^{-a} \frac{a^{x_i}}{x_i!}}{e^{-na} \frac{(na)^t}{t!}} = . \frac{(\sum x_i)!}{\prod x_i!} \prod_{i=1}^{n} \left(\frac{1}{n}\right)^{x_i} \qquad (4)$$

В приведенной выкладке учтено, что сумма $\sum \xi_i$ независимых пуассоновских случайных величин распределена по закону Пуассона с параметром, равным сумме na параметров.

Поскольку условная вероятность (4) от a не зависит, то $\sum x_i$ является достаточной статистикой, и значения x_1, …, x_n можно не хранить.

Полезно отметить, что распределение (4) является полиномиальным с равными вероятностями $p_i = \frac{1}{n}$, при i = 1, 2…n.

Действительно, на отрезок, состоящий из *n* промежутков длиной ΔT, бросим независимо общее количество $t = \sum_{i=1}^{n} x_i$ равномерно распределенных точек, и определим
вероятность попадания x_1, x_2 … x_n точек на интервалы ΔT_1, $\Delta T_2 \ldots \Delta T_n$ одинаковой длины ΔT.

Вставить напоминание о полиномиальном распр-нии

Эта полиномиальная вероятность равна правой части равенства (4). Мы получили важный результат для теории вероятностей:

Утверждение. Для простейшего потока при известном числе точек на отрезке, положения точек независимы и равномерно распределены.

Это утверждение можно использовать при генерации (при моделировании) простейшего потока событий.

Итак, на числовой оси (удобно, но не обязательно, воспринимать ее как временнУю ось) нужно разбросать точки пуассоновского потока. Это можно сделать, по меньшей мере, двумя путями.

Первый. Известно, что если λ- параметр потока, то расстояние между соседними точками – случайная величина, распределенная по показательному закону с параметром λ, и нужно последовательно генерировать эти случайные величины.

Второй. пусть T – длина отрезка, на котором нужно сгенерировать поток. Общее число ν точек на T - случайная величина, распределенная по закону Пуассона с параметром *a* = λ T. Сгенерируем ν, получим некоторое целое n, а затем раскидаем независимо n точек на отрезке [0,T] с равномерным законом распределения

4.2. Критерий факторизации.

Необходимость вычислять условное распределение ξ (или распределение τ), чтобы потом определить условное распределение делением $p(x;a)$ на $p_\tau(t;a)$, является весьма неудобным способом определения достаточности. Простой способ дает теорема, называемая критерием факторизации.

Теорема. Статистика T(ξ) достаточна для a тогда и только тогда, когда справедлива факторизация (представление):

$$p_\xi(x;a) = g(T(x),a) \cdot h(x).$$

Существенным в этой записи является то, что
множитель, зависящий от параметра, от x зависит только через $T(x)$, (напомним: $p_\xi(x;a)$ понимается как плотность или вероятность).

Докажем утверждение для дискретного случая.

Достаточность.

Пусть $t = T(x)$ (иначе выписанная ниже вероятность равна 0). Рассмотрим условную вероятность:

$$P\{\xi = x \mid T(\xi) = t, a\} = \frac{p_\xi(x \mid a)}{P\{T(\xi) = t\}} = \frac{p_\xi(x;a)}{\sum\limits_{y:\, T(y)=t} p_\xi(y;a)} = \frac{g(t,a)h(x)}{\sum\limits_{y:T(y)=t=T(x)} g(t,a)h(y)} = \frac{h(x)}{\sum\limits_{y:T(y)=t=T(x)} h(y)}.$$

Полученное выражение не зависит от *a*.

$$P\{\xi = x; a\} = P_\tau\{t = T(x); a\} \cdot P\{\xi = x \mid \tau = t = T(x)\}.$$

Необходимость.

Пусть $P\{\xi = x \mid T(\xi) = t = T(x), a\} = k(x)$ не зависит от *a*. Тогда в силу (1)

$$P_\xi(x;a) = p_\tau(t = T(x); a) \cdot k(x),$$

то есть справедлива факторизация (2).

Замечание. Пусть $T(\xi)$ — достаточная статистика.

Любая функция $T_1(\xi)$, по которой можно получить достаточную $T(\xi)$, является тоже достаточной. Другими словами,

если $T(\xi) \equiv F(T_1(\xi))$, то $T_1(\xi)$ достаточна.

Этот факт тривиально следует из критерия факторизации:

$$p_\xi(x;a) = g(T(x),a) \cdot h(x) = g(F(T_1(x)),a) \cdot h(x) = g_1(T_1(x),a) \cdot h(x).$$

В частности, если $T(\cdot)$ и $T_1(\cdot)$ связаны взаимооднозначно, то T_1 — достаточна.

Пример 1. Пусть ξ_1, ξ_2 … ξ_n — *n* независимых наблюдений над случайной величиной, распределенной по закону Пуассона, другими словами, ξ_1, $\xi_2 \ldots \xi_n$ — выборка из совокупности, распределенной по закону Пуассона. Распределение выборки имеет вид:

$$p(x;a) \equiv P\{\xi_1 = x_1, \xi_2 = x_2 \ldots \xi_n = x_n; a\} = \prod_{i=1}^{n} e^{-a} \frac{a^{x_i}}{x_i!} = \frac{e^{-na} a^{\sum x_i}}{\prod\limits_{i=1}^{n} x_i!}.$$

Видим, что множитель, зависящий от параметра *a*, т.е. $e^{-na}a^{\sum x_i}$, от *x* зависит только через $\sum\limits_{i=1}^{n} x_i$. По критерию факторизации $\sum\limits_{i=1}^{n} x_i$ является достаточной статистикой.

Пример 2. Пусть случайное событие *A* с неизвестной вероятностью *q* испытывается независимо *n* раз. Пусть ξ_i - результат i-го испытания — случайная величина (*i* = 1, 2…*n*), распределенная следующим образом:

$$\xi_i = \begin{cases} 1, & c \quad вероятностью \quad q, \\ 0, & c \quad вероятностью \quad 1-q. \end{cases}$$

Пусть ν — число наступлений события *A* в серии из *n* испытаний. Выпишем распределение вероятностей для выборки ξ_1, $\xi_2 \ldots$ ξ_n,

принимая во внимание, что любая из случайных величин ξ_i может принимать лишь два значения ($x_i = 0$ или $x_i = 1$):

$$p(x_1, x_2 \ldots x_n; q) = \prod_{i=1}^{n} P\{\xi_i = x_i; q\} = \prod_{i=1}^{n} q^{x_i}(1-q)^{1-x_i} = q^{\sum x_i}(1-q)^{n-\sum x_i}.$$

Из этого выражения следует, что $\nu \equiv \sum x_i$ является достаточной статистикой.

Пример 3. Пусть ξ_0 — случайная величина, распределенная по равномерному закону на отрезке $[0, a]$, $a > 0$. Плотность распределения (для $x \geq 0$):

$$p_0(x; a) = \begin{cases} 1/a, & 0 \leq x \leq a, \\ 0, & x > a \text{ или } x < 0. \end{cases}$$

Пусть ξ_1, $\xi_2 \ldots \xi_n$ есть n независимых наблюдений над ξ_0. Плотность распределения выборки:

$$p_\xi(x_1, x_2 \ldots x_n \mid a) = \prod_{i=1}^{n} p_0(x_i \mid a) = \begin{cases} (1/a)^n, & \max\limits_i x_i \leq a, \ \min\limits_i x_i \geq 0 \\ 0, & \text{в остальных случаях.} \end{cases}$$

Примечание [MA1]: Поправить в скобках ($x_1, x_2 \ldots x_n$)

Если ввести функцию $u(z)$ — индикатор неотрицательности аргумента z, —определив ее так:

$$u(z) = \begin{cases} 1, & z \geq 0, \\ 0, & z < 0, \end{cases}$$

то плотность распределения выборки запишется в виде:

$$p(x_1, x_2 \ldots x_n; a) = (1/a)^n \cdot u(a - \max_i x_i) \cdot u(\min_i x_i),$$

Примечание [MA2]: Поправить в скобках ($x_1, x_2 \ldots x_n$)

откуда по критерию факторизации следует, что $\max\limits_i x_i$ является достаточной статистикой.

Пример 4. Пусть ξ_1, $\xi_2 \ldots \xi_n$ — выборка из совокупности, распределенной по нормальному закону с параметрами m и σ^2 (роль параметра a в критерии факторизации играет здесь двумерный параметр (m, σ^2)).

Плотность распределения выборки имеет вид:

$$p(x_1, x_2 \ldots x_n; m, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i - m)^2}{2\sigma^2}} = \left(\frac{1}{\sqrt{2\pi}}\right)^n \left(\frac{1}{\sigma}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2} \cdot e^{\frac{m}{\sigma^2}\sum_{i=1}^{n} x_i} \cdot e^{-\frac{nm^2}{2\sigma^2}}.$$

Из этого выражения по критерию факторизации следует, что двумерная статистика $\tau_1 \equiv \left(\sum\limits_{i=1}^{n} x_i, \sum\limits_{i=1}^{n} x_i^2\right)$ является достаточной для $a \equiv (m, \sigma^2)$, так же как и двумерная статистика

$$\tau_2 = \left(\bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i, \ s^2 \equiv \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2\right),$$

связанная с $т_1$ взаимно-однозначно и имеющая своими компонентами несмещенные оценки $\bar{x}$ и s^2 для математического ожидания m и дисперсии σ^2. Поэтому *любые статистические задачи, связанные с нормальным распределением, можно решать, опираясь на оценки* $\bar{x}$ *и* s^2.