Геометрия хи-квадрат распределения

Винни-Пух

2018-01-06

Содержание

Пространства и подпространства . . . 1
Проекции . . . 2
Хи-квадрат распределение . . . 3
Связь со стандартным определением . . . 3
Определение в матрицах . . . 4
Выборочная дисперсия — геометрия . . . 6
Ковариационная матрица спроецированного вектора . . . 7
Хи-квадрат тест Пирсона, геометрия . . . 8
Выборочная дисперсия — явно скалярно . . . 9
Выборочная дисперсия — матрицы . . . 10
Сумма квадратов остатков - геометрия . . . 10
Сумма квадратов остатков - матрицы . . . 10
Про t и F похоже отдельно :) . . . 10

Пространства и подпространства

Если говорить совсем просто, то пространство $\mathbb{R}^n$ — это все столбики, состоящие из n действительных чисел. А если вспоминать определение, то линейное пространство — это такой набор векторов, в котором:

1. Разрешено складывать два любых вектора и результат остаётся внутри набора;
2. Разрешено умножать любой вектор на любое число и результат остаётся внитри набора;
3. Сложение векторов и умножение вектора на число согласованы между собой.

Подпространство — это часть набора, которая сама по себе является пространством. Есть два популярных способа описать подпространство внутри $\mathbb{R}^n$:

1. Линейная оболочка некоторого набора векторов.

Например, внутри $\mathbb{R}^3$ есть два вектора $x = (1, 1, 1)$ и $y = (1, 0, 0)$ и подпространство V, образованное ими

$$V = \mathcal{L}in(x, y),$$

то есть это все вектора вида $\alpha x + \beta y$, где α и β — произвольные числа.

2. Ортогональное дополнение к набору векторов.

Например, внутри $\mathbb{R}^3$ есть два вектора $x = (1, 1, 1)$ и $y = (1, 0, 0)$ и подпространство W всех векторов, перпендикулярным им обоим

$$W = \mathcal{L}in^{\perp}(x, y).$$

Напомним, два вектора a и b в пространстве $\mathbb{R}^n$ перпендикулярны, если их скалярное произведение $\langle a, b\rangle$ равно нулю. Поэтому подпространство W можно также описать системой уравнений.

Подпространство W состоит из всех векторов $w = (w_1, w_2, w_3)$, удовлетворяющих системе:

$$\begin{cases} w_1 \cdot 1 + w_2 \cdot 1 + w_3 \cdot 1 = 0 \\ w_1 \cdot 1 + w_2 \cdot 0 + w_3 \cdot 0 = 0 \end{cases}$$

Если вектора a_1, ..., a_k линейно независимы и лежат внутри $\mathbb{R}^n$, то размерности описанных нами подпространств равны

$$\dim \mathcal{L}in(a_1, \ldots, a_k) = k$$

$$\dim \mathcal{L}in^{\perp}(a_1, \ldots, a_k) = n - k$$

Упражнения

1. Лежит ли вектор ... в подпространстве ...
2. Найдите базис в ортогональном дополнении подпространства ...
3. Найдите размерность пространств ... ...

Проекции

Проекцией вектора a на подпространство L называется вектор $\hat{a}$, лежащий в L и ближайший к a.

Есть два популярных способа найти проекцию:

1. Решить задачу минимизации расстояния

$$\min_{\hat{a} \in L} ||a - \hat{a}||$$

2. Потребовать, чтобы разница $a - \hat{a}$ была перпендикулярна любому вектору из L:

$$a - \hat{a} \in L^{\perp}$$

Упражнения:

- Спроецируйте вектор z на подпространство ... Найдите квадрат длины проекции.
- Спроецируйте вектор z на прямую, порождённую вектором b. Найдите косинус угла между a и b.
- Спроецируйте вектор z на ортогональное дополнение к ...
- Спроецируйте вектор z на вектор единичной длины v.
- Спроецируете произвольный вектор z на пространство $\mathcal{L}in^{\perp}(v)$, где v — вектор единичной длины. Найдите квадрат длины проекции.

Хи-квадрат распределение

Определение. Пусть компоненты n-мерного вектора z имеют стандартное нормальное распределение, $z_i \sim \mathcal{N}(0;1)$ и независимы. Рассмотрим произвольное фиксированное k-мерное подпространство L. Абсолютно любое. Обозначим проекцию вектора z на подпространство L буквой $\hat{z}$, а квадрат длины проекции — буквой Q:

$$Q = ||\hat{z}||^2 = \langle \hat{z}, \hat{z} \rangle = \hat{z}'\hat{z}$$

Закон распределения случайной величины Q называется хи-квадрат распределением с k-степенями свободы.

Пример. Вектор $z \in R^3$, компоненты $z_i \sim \mathcal{N}(0;1)$ и независимы. Найдите явную формулу для величины Q, квадрата длины проекции z на плоскость $z_1 + z_2 + z_3 = 0$. Какое распределение имеет Q?

Пример. Вектор $z \in R^3$, компоненты $z_i \sim \mathcal{N}(0;1)$ и независимы. Найдите явную формулу для величины Q, квадрата длины проекции z на прямую, порожденную вектором $a = (1,1,1)$. Какое распределение имеет Q?

Пример. Вектор $z \in R^7$, компоненты $z_i \sim \mathcal{N}(0;1)$ и независимы. Какое распределение имеет величина Q, квадрат длины проекции z на подпространство, задаваемое системой уравнений

$$\begin{cases} z_1 + z_2 + z_3 + z_4 + z_5 + z_6 + z_7 = 0 \\ z_1 + 2z_2 + 3z_3 + 4z_4 + 5z_5 + 6z_6 + 7z_7 = 0 \end{cases}$$

Пример. Вектор $z \in R^4$, компоненты $z_i \sim \mathcal{N}(7;1)$ и независимы. Какое распределение имеет величина Q, квадрат длины проекции z на подпространство, ортогональное прямой, порождаемой вектором $a = (1,1,1,1)$?

Сразу скажем, что этот подход не нов. Например, он обсуждается в статье Cobb [Cob11]. Однако аккуратного изложения его на русском я не знаю :)

Связь со стандартным определением

Если взять практически любой учебник, то там будет дано другое определение χ^2-распределения.

Величина Q имеет χ^2-распределение с k степенями свободы, если она представима в виде

$$Q = z_1^2 + z_2^2 + \ldots + z_k^2,$$

где z_i независимы и стандартны нормальны, $z_i \sim \mathcal{N}(0;1)$.

Сначала заметим, что стандартное определение из учебника — частный случай нашего. Что получится если вектор $z = (z_1, z_2, \ldots, z_n)$, лежащий в $\mathbb{R}^n$, спроецировать на k-мерное подпространство V всех векторов, у которых первые k координат произвольные, а остальные — нули?

Получится вектор $\hat{z} = (z_1, z_2, \ldots, z_k, 0, 0, \ldots, 0)$. И квадрат длины проекции будет равен

$$Q = ||\hat{z}||^2 = z_1^2 + z_2^2 + \ldots + z_k^2.$$

А наше новое определение допускает проецирование на любое k-мерное подпространство :)

Возникает естественный вопрос, а вдруг, если спроецировать на какое-то хитрое подпространство, скажем $\mathcal{L}in(a, b, c) \cap \mathcal{L}in^{\perp}(d, e, f)$, эквивалентность определений нарушится?

Вдруг возможно, что квадрат длины проекции вектора z на подпространство размерности k не будет представляться в виде суммы k независимых стандартных нормальных величин?

Оказывается два определения полностью эквивалентны в силу двух фактов:

1. Закон распределения вектора z не изменится, если вектор z повернуть в любом направлении на произвольный угол;

2. Любое k-мерное подпространство всегда можно повернуть так, чтобы оно совпало с подпространством V всех векторов, у которых первые k координат произвольные, а остальные — нули.

Идеи доказательства:

1. Функция плотности z имеет вид:

$$f(z_1, z_2, \ldots, z_n) = f(z_1) \cdot f(z_2) \cdot \ldots \cdot f(z_n) \propto e^{-z_1^2/2} e^{-z_2^2/2} \ldots e^{-z_1^2/2} = e^{-\frac{1}{2}(z_1^2 + z_2^2 + \ldots + z_n^2)};$$

Мы видим, что значение функции плотности в произвольной точке z зависит только от расстояния от z до нуля, но не от угла.

2. Рассмотрим стандартный базис e_1, e_2, ..., e_n в $\mathbb{R}^n$. Рассмотрим k-мерное подпространство $V \subset \mathbb{R}^n$. Выберем в подпространстве V произвольный ортогональный базис из k векторов: v_1, ..., v_k. Сначала повернём V так, чтобы v_1 совпал с e_1. Затем будем поворачивать так, чтобы v_1 не трогать, а v_2 повернуть до совпадения с e_2. И так далее.

Определение в матрицах

Зафиксируем k линейно-независимых векторов x_1, x_2, ..., x_k. Для удобства занесём их столбцами в матрицу X. То есть x_j — это j-ый столбец матрицы X. Введём два обозначения.

Линейная оболочка всех столбцов матрицы X:

$$\mathrm{col} X = \mathcal{L}in(x_1, x_2, \ldots, x_k)$$

Ортогональное дополнение всех столбцов матрицы X:

$$\mathrm{col}^{\perp} X = \mathcal{L}in^{\perp}(x_1, x_2, \ldots, x_k)$$

Проецирование — это линейное преобразование векторов:

1. Если вектор растянуть в α раз, то проекция растянется в α раз;

2. Проекция суммы двух векторов равна сумме проекций каждого вектора по отдельности.

Поэтому проецирование вектора z на пространство $\mathrm{col} X$ можно записать в виде его умножения на некую матрицу H:

$$\hat{z} = H \cdot z$$

Мы называем матрицу H матрицей-шляпницей (hat-matrix), потому что она навешивает шляпку на z.

Естественно матрица H зависит от того подпространства $\mathrm{col}X$ на которое мы проецируем. Осталось найти эту связь. Заметим, что вектор $z - \hat{z}$ перпендикулярен пространству $\mathrm{col}X$. То есть

$$z - \hat{z} \perp X$$

Столбцы X перпендикулярны вектору $\hat{z}$, только если склалярное произведение $\hat{z}$ с каждым столбцом X равно нулю:

$$X' \cdot (z - \hat{z}) = 0$$

Вектор $\hat{z}$ лежит в подпространстве $\mathrm{col}X$, поэтому он должен выражаться через столбцы матрицы X:

$$\hat{z} = X \cdot \alpha$$

Получаем уравнение на веса α:

$$X'(z - X\alpha) = 0$$

После раскрытия скобок имеем:

$$X'X\alpha = X'z$$

Временно предположим, что матрица $X'X$ обратима:

$$\alpha = (X'X)^{-1}X'z$$

И наконец,

$$Hz = \hat{z} = X\alpha = X(X'X)^{-1}X'z$$

Таким образом, проецирование на линейное подпространство $\mathrm{col}X$ можно задать в виде умножения на матрицу

$$H = X(X'X)^{-1}X'$$

У матрицы-шляпницы H много приятных свойств. Например, необходимое и достаточное условие, чтобы некая матрица H задавала проецирование:

$$\begin{cases} H' = H \\ H^2 = H \end{cases}$$

Геометрическая интерпретация:

1. $H' = H$. Для любых двух векторов x и y скалярное произведение спроецированного x на исходный y равно скалярному произведение исходного x на спроецированный y.

$$\langle Hx, y\rangle = (Hx)'y = x'H'y = x'(H'y) = x'(Hy) = \langle x, Hy\rangle$$

2. $H^2 = H$. Проецирование два раза эквивалентно проецированию один раз.

Другое необходимое и достаточное условие:

$$\begin{cases} H' = H \\ \text{Все собственные числа } H \text{ равны } 0 \text{ или } 1 \end{cases}$$

Если спроецировать нормальный стандартный вектор z на $\mathrm{col} X$, то мы получим вектор $\hat{z} = Hz$. И квадрат длины $\hat{z}$ будет равен

$$||\hat{z}||^2 = (Hz)'Hz = z'H'Hz = z'Hz$$

Поэтому можно дать определение:

Величина Q имеет хи-квадрат распределение с k степенями свободы, если она представима в виде

$$Q = z'Hz,$$

где z — нормальные стандартный вектор, а H — матрица, проецирующая на k-мерное подпространство, то есть $H' = H$, $H^2 = H$, $\mathrm{trace} H = k$.

По сути это определение просто переводит на язык матриц идею проецирования. Не стоит бояться матриц! Весь смысл матриц в том, чтобы записать формально какую-то геометрическую идею!

Проецирование на линейное пространство $\mathrm{col}^{\perp} X$ можно задать в виде умножения на матрицу $M = I - H$. Поэтому квадрат длины проекции стандартного нормального вектора z на подпространство $\mathrm{col}^{\perp} X$ записывается как

$$S = ||Mz||^2 = (Mz)'Mz = z'M'Mz = z'(I - H)z$$

И, конечно, величина S имеет хи-квадрат распределение с $n - k$ степенями свободы.

Выборочная дисперсия — геометрия

Начнём с упражнения. Пусть z — вектор из $\mathbb{R}^n$, а $\mathbb{1}$ — вектор из единиц. Чему равен квадрат длины проекции z на $\mathcal{L}in^{\perp}(\mathbb{1})$? Чему равна проекция вектора $\mathbb{1}$ на $\mathcal{L}in^{\perp}(\mathbb{1})$?

Сначала спроецируем вектор z на $\mathcal{L}in(\mathbb{1})$. Получаем вектор $\bar{z} \cdot \mathbb{1} = (\bar{z}, \bar{z}, \ldots, \bar{z})$. Поэтому проекция z на $\mathcal{L}in^{\perp}(\mathbb{1})$ равна $z - \bar{z} \cdot \mathbb{1} = (z_1 - \bar{z}, z_2 - \bar{z}, \ldots, z_n - \bar{z})$.

Вектор из единиц ортогонален пространству $\mathcal{L}in^{\perp}(\mathbb{1})$, поэтому вектор $\mathbb{1}$ проецируется в нулевой вектор.

Поэтому для стандартного нормального вектора z величина $\sum(z_i - \bar{z})^2$ имеет хи-квадрат распределение с $n - 1$-ой степенью свободы.

А теперь замечаем, что выборочная дисперсия вектора x — это квадрат длины проекции делённый на размерность подпространства!

$$s\mathbb{V}ar(x) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \frac{1}{n-1} \cdot ||x - \bar{x} \cdot \mathbb{1}||^2$$

Осталось добавить предположения:

Пусть x_i независимы и одинаково распределены $\mathcal{N}(\mu, \sigma^2)$. Заметим, что вектор x можно представить в виде

$$x = \mu \cdot 1 + \sigma z,$$

где z — стандартный нормальный вектор.

Проекция x на $\mathcal{L}in^{\perp}(\mathbb{1})$ совпадает с проекцией σz на $\mathcal{L}in^{\perp}(\mathbb{1})$. Вектор $\mathbb{1}$ проецируется в нулевой. А проекция σz в σ раз длиннее, чем проекция z. Поэтому:

$$\sum (x_i - \bar{x})^2 = \sigma^2 \sum_{i=1}^n (z_i - \bar{z})^2$$

Таким образом, мы доказали, что

$$\frac{\sum (x_i - \bar{x})^2}{\sigma^2} = \frac{(n-1)s\mathbb{V}ar(x)}{\sigma^2} \sim \chi^2_{n-1}$$

Ковариационная матрица спроецированного вектора

Тут хорошо бы максимально просто доказать, что если $\hat{z} = Hz$, и z — стандартный нормальный вектор, то $\mathbb{V}ar(\hat{z}) = H$.

Ковариационная матрица вектор y определяется как

$$\mathbb{V}ar(y) = \mathbb{E}[(y-\mu)(y-\mu)'],$$

где $\mu = \mathbb{E}(y)$.

Эквивалентно дисперсию можно определить как

$$\mathbb{V}ar(y) = \mathbb{E}(yy') - \mathbb{E}(y)\mathbb{E}(y)'$$

Посмотрим, чему равна $\mathbb{V}ar(Ay)$:

$$\mathbb{V}ar(Ay) = \mathbb{E}((Ay)(Ay)') - \mathbb{E}(Ay)\mathbb{E}(Ay)' = \mathbb{E}(Ayy'A') - A\mathbb{E}(y)(A\mathbb{E}(y))' = A\mathbb{E}(yy')A' - A\mathbb{E}(y)\mathbb{E}(y)'A' = A\mathbb{V}ar(y)A'$$

В силу этого мы находим ещё одно шикарное свойство матрицы-шляпницы! Пусть z — стандартный нормальные вектор, $z \sim \mathcal{N}(0; I)$. В частности, $\mathbb{V}ar(z) = I$.

Найдём ковариационную матрицу проекции $\hat{z}$:

$$\mathbb{V}ar(\hat{z}) = \mathbb{V}ar(Hz) = H\mathbb{V}ar(z)H' = H \cdot I \cdot H' = HH' = H^2 = H$$

Матрица-шляпница H является ковариационной матрицей спроецированного вектора!

Хи-квадрат тест Пирсона, геометрия

Для начала спроецируем стандартный нормальный вектор z на $\mathcal{L}in^{\perp}(v)$, где v — единичный вектор. При этом мы получим вектор $\hat{z} = H \cdot z$:

$$\hat{z} = (I - vv') \cdot z$$

По нашем определению квадрат длины $\hat{z}$ имеет хи-квадрат распределение со степенями свободы равными размерности подпространства $\mathcal{L}in^{\perp}(v)$. А размерность пространства $\mathcal{L}in^{\perp}(v)$ на единицу меньше размерности исходного пространства.

Ковариационная матрица вектора $\hat{z}$ имеет именно такой же вид:

$$\mathbb{V}ar(\hat{z}) = (I - vv')$$

Запомним эту ковариационную матрицу! И запомним, что она возникает у проекции на ортогональное дополнение к вектору v! А теперь к покемонам!

Каждый отловленный покемон может быть одного из r видов. Виды покемонов встречаются с вероятностью p_1, ..., p_r. Всего мы ловим n покемонов, ν_j — количество покемонов вида j.

Замечаем, что ν_j имеет биномиальное распределение $Bin(n, p_j)$. В частности, $\mathbb{E}(\nu_j) = np_j$ и $\mathbb{V}ar(\nu_j) = np_j(1 - p_j)$. Также можно установить, что

$$\mathbb{C}ov(\nu_j, \nu_i) = -np_ip_j$$

Мы немного необычным образом отнормируем эти ν_j: вычтем математическое ожидание и поделим на корень из математического ожидания!

$$\nu_j^* = \frac{\nu_j - np_j}{\sqrt{np_j}}$$

При этом окажется, что: $\mathbb{E}(\nu_j^*) = 0$, $\mathbb{V}ar(\nu_j^*) = 1 - p_j$, $\mathbb{C}ov(\nu_i, \nu_i) = -\sqrt{p_i}\sqrt{p_j}$.

Заметим, что по центральное предельной теореме $\nu_j^* \to \mathcal{N}(0; 1 - p_j)$.

Присмотримся повнимательнее!

$$\mathbb{V}ar(\nu^*) = \begin{pmatrix} 1-p_1 & -\sqrt{p_1}\sqrt{p_2} & -\sqrt{p_1}\sqrt{p_3} & \ldots \\ -\sqrt{p_2}\sqrt{p_1} & 1-p_2 & -\sqrt{p_2}\sqrt{p_3} & \ldots \\ -\sqrt{p_3}\sqrt{p_1} & -\sqrt{p_3}\sqrt{p_2} & 1-p_3 & \ldots \\ \vdots & \vdots & \vdots & \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & \ldots \\ 0 & 1 & 0 & \ldots \\ 0 & 0 & 1 & \ldots \\ \vdots & \vdots & \vdots & \end{pmatrix} - \begin{pmatrix} \sqrt{p_1}\sqrt{p_1} & \sqrt{p_1}\sqrt{p_2} & \sqrt{p_1}\sqrt{p_3} & \ldots \\ \sqrt{p_2}\sqrt{p_1} & \sqrt{p_2}\sqrt{p_2} & \sqrt{p_2}\sqrt{p_3} & \ldots \\ \sqrt{p_3}\sqrt{p_1} & \sqrt{p_3}\sqrt{p_2} & \sqrt{p_3}\sqrt{p_3} & \ldots \\ \vdots & \vdots & \vdots & \end{pmatrix}$$

Выходит, что ковариационная матрица нового вектора ν^* представима в виде:

$$\mathbb{V}ar(\nu^*) = I - vv',$$

где вектор v состоит из корней вероятностей, $v = (\sqrt{p_1}, \sqrt{p_2}, \ldots, \sqrt{p_r})$. Заметим, что вектор v имеет единичную длину:

$$||v||^2 = v_1^2 + v_2^2 + \ldots + v_r^2 = p_1 + p_2 + \ldots + p_r = 1$$

То есть ковариационная матрица вектора ν^* совпадает с ковариационной матрицой проекции вектора z из $\mathbb{R}^r$ на подпространство $\mathcal{L}in^{\perp}(v)$. Закон распределения многомерного нормального вектора однозначно определяется вектором математических ожиданий и ковариационной матрицей.

Следовательно, сумма $\sum_{j=1}^{r}(\nu_j^*)^2$ распределена при больших n также, как квадрат длины проекции z на $\mathcal{L}in^{\perp}(v)$.

Поэтому

$$\sum_{j=1}^{r}(\nu_j^*)^2 = \sum_{j=1}^{r}\frac{(\nu_j - np_j)^2}{np_j} \to \chi^2_{r-1}$$

Недостатки доказательства:

1. Строго говоря, ЦПТ гарантирует, что каждый ν_j^* в отдельности имеет асимптотически нормальное распределение, а здесь требуется асимптотическая нормальность вектора ν^*, то есть требуется ЦПТ в векторной форме.

2. Деление на корень из математического ожидания выглядит магией, которая потом раскрывается, а хотелось бы раскрыть её по ходу.

Аналогичное доказательство можно найти в курсе Panchenko [Pan05].

Выборочная дисперсия — явно скалярно

Мы помним, что $\sum(z_i - \bar{z})^2$ — это проекция вектора z на подпространство $\mathcal{L}in^{\perp}(\mathbb{1})$. Заметим, что мы легко можем выбрать ортогональный базис в подпространстве $\mathcal{L}in^{\perp}(\mathbb{1})$ явно.

Явный базис на примере $z \in \mathbb{R}^5$:

$$\mathcal{L}in^{\perp}(\mathbb{1}) = \operatorname{col}\begin{pmatrix} 1 & 1 & 1 & 1 \\ -1 & 1 & 1 & 1 \\ 0 & -2 & 1 & 1 \\ 0 & 0 & -3 & 1 \\ 0 & 0 & 0 & -4 \end{pmatrix}$$

Действительно, давайте проверим:

1. Каждый столбец ортогонален вектору $\mathbb{1} = (1, 1, 1, 1, 1)$.

2. Столбцы ортогональны между собой.

3. Столбцов четыре :)

Поэтому столбцы задают базис в подпространстве $\mathcal{L}in^{\perp}(\mathbb{1})$.

… …

Итого:

$$\sum(z_i-\bar{z})^2 = \frac{(z_1-z_2)^2}{1+1^2}+\frac{(z_1+z_2-2z_3)^2}{2+2^2}+\frac{(z_1+z_2+z_3-3z_4)^2}{3+3^2}+\ldots\frac{(z_1+z_2+z_3+\ldots+z_{n-1}-(n-1)z_n)^2}{(n-1)+(n-1)^2} \quad (1$$

В этом разложении явно видна сумма $(n-1)$ слагаемого. Каждое слагаемое является квадратом нормальной стандартной случайной величины и слагаемые независимы.

Выборочная дисперсия — матрицы

Сумма квадратов остатков - геометрия

Для максимальной доступности доказательства мы проведём его для двух регрессоров. Случай k регрессоров ничем с геометрической точки зрения не отличается.

Сумма квадратов остатков - матрицы

Про t и F похоже отдельно :)

Список литературы

[Cob11] George W Cobb. "Teaching statistics: Some important tensions". В: *Chilean Journal of Statistics* 2.1 (2011). Преподавание эконометрики, последовательность изложения, геометрический смысл., с. 31—62.

[Pan05] Dmitry Panchenko. *18.650 Statistics for Applications (Fall 2006)*. Отличный курс по статистике. В более поздних версиях вместо заметок к лекциям появились слайды. Симпатичнее, но некоторые доказательства в них исчезли. 2005. URL: https://ocw.mit.edu/courses/mathematics/18-443-statistics-for-applications-fall-2006/.