Глава 11

Нормальная модель

11.1 Базовая часть

Сегодня мы будем рассматривать выборку (выборки) $X_1, ..., X_n$ из нормального $\mathcal{N}(a, \sigma^2)$ распределения.

11.1.1 Многомерное нормальное распределение и его свойства

Определение 1. Многомерным нормальным вектором называют такой вектор $\vec{X}$, что $\vec{X} \stackrel{d}{=} A\vec{Z} + \vec{b}$, где A — некоторая неслучайная матрица, $\vec{b}$ — неслучайный вектор, $\vec{Z}$ вектор с независимыми $\mathcal{N}(0,1)$ компонентами.

Другим определением может быть определение через характеристическую функцию:

Определение 2. Многомерным нормальным вектором называют такой вектор $\vec{Y}$, что

$$\psi_{\vec{Y}}(\vec{t}) = \exp\left(i\vec{\mu}^T\vec{t} - \frac{1}{2}\vec{t}^T\Sigma\vec{t}\right),$$

где $\vec{\mu}$ — некоторый вектор, Σ — некоторая симметричная неотрицательно определенная матрица.

При этом Σ — ковариационная матрица вектора $\vec{Y}$, $\vec{\mu}$ — вектор его математических ожиданий.

Отметим, что определение, задающее распределение плотностью

$$f_{\vec{Y}}(\vec{x}) = \frac{1}{\sqrt{2\pi}\sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2}(\vec{x} - \vec{\mu})\Sigma^{-1}(\vec{x} - \vec{\mu})^T\right),$$

более ограничительно чем два предыдущих. В этом случае предполагается, что Σ невырождена, хотя два исходных распределения этого не предполагают.

Нам понадобится несколько свойств многомерного нормального распределения:

1. Если $\vec{X}$ — многомерный нормальный, то $\vec{Y}$, являющийся его подвектором, также многомерный нормальный.

   Доказательство. Из определения $\vec{X} = A\vec{Z} + \vec{b}$, где $\vec{Z}$ имеет н.о.р. $\mathcal{N}(0,1)$ компоненты, кроме того $\vec{Y} = B\vec{X}$, где B — прямоугольная матрица, оставляющая из координат $\vec{X}$ только те, которые фигурируют в $\vec{Y}$. Значит $\vec{Y} = BA\vec{Z} + B\vec{b}$, что и требовалось доказать.

   Легко понять, что вектор средних $\vec{Y}$ — это соответствующие выбранным координатам средние, а матрица ковариаций — ковариации. □

2. Если $\vec{X} \sim \mathcal{N}(\vec{\mu}, \Sigma)$, то $\vec{Y} = B\vec{X} \sim \mathcal{N}(B\vec{\mu}, B\Sigma B^t)$.

   Доказательство. Проще всего доказать формулу через х.ф.:

$$\psi_{B\vec{X}}(\vec{t}) = \mathbf{E} \exp\left(i(\vec{t}, B\vec{X})\right) = \mathbf{E} \exp\left(i(B^T\vec{t}, \vec{X})\right) = \psi_{\vec{X}}(B^T\vec{t}).$$

   Подставляя формулу из определения х.ф. $\vec{X}$, получаем требуемое утверждение. □

3. Если $\vec{X}$ имеет $\mathcal{N}(\vec{0}, cE)$ распределение, то $B\vec{X}$ имеет $\mathcal{N}(0, cE)$ распределение для любой ортогональной матрицы B.

   Доказательство. В силу предыдущего пункта

$$B\vec{X} \sim \mathcal{N}(B\vec{0}, cBEB^T) = \mathcal{N}(\vec{0}, cE),$$

   что и требовалось. □

4. Если $(\vec{X}, \vec{Y})$ имеет нормальное распределение, то $\vec{X}$ и $\vec{Y}$ независимы тогда и только тогда, когда $\operatorname{cov}(X_i, Y_j) = 0$ при любых i, j.

   Доказательство. Доказательство напрямую вытекает из вида х.ф. Проведем его для векторов с нулевыми средними, поскольку вычитание среднего не изменяет ни зависимости, ни кореллированности. Независимость векторов означает, что

$$\psi_{X,Y}(\vec{t}, \vec{s}) = \exp\left(-\frac{1}{2}(\vec{t}, \vec{s})\Sigma(\vec{t}, \vec{s})^T\right) = \exp\left(-\frac{1}{2}\vec{t}\Sigma_X\vec{t}^T - \frac{1}{2}\vec{s}\Sigma_Y\vec{s}^T\right),$$

   где Σ_X, Σ_Y — матрицы ковариаций X и Y, Σ — матрица ковариаций X, Y. Матрица Σ содержит блок Σ_X в левом верхнем углу и Σ_Y в правом нижнем, а мы хотим сказать, что остальные ее элементы — нули. Это прямо следует из равенства произведений под знаками экспонент. □

11.1.2 Связанные с нормальным распределения

Определение 3. Пусть $X_i \sim \mathcal{N}(0, 1)$ н.о.р. величины, $Y = X_1^2 + \cdots + X_n^2$. Тогда распределение Y называется распределением хи-квадрат с n степенями свободы (обозначается χ_n^2).

Распределение χ_n^2 представляет собой квадрат длины стандартного нормального вектора. Как мы уже говорили прежде, χ_n^2 распределение совпадает с $Gamma(n/2, 2)$.

Определение 4. Пусть $X \sim \mathcal{N}(0, 1)$, $Y \sim \chi_n^2$ независимы, тогда величина

$$\frac{X}{\sqrt{Y/n}}$$

называется величиной с распределением Стьюдента с n степенями свободы (обозначение t_n).

Распределение Стьюдента напоминает при больших n стандартное нормальное (и как нетрудно убедиться из ЗБЧ и леммы Слуцкого стремится к нему с ростом n), но имеют плотность с заметно более тяжелыми хвостами

$$\frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\Gamma\left(\frac{n}{2}\right)}\left(1 + \frac{y^2}{n}\right)^{-(n+1)/2}. \tag{11.1}$$

При $n = 1$ это плотность распределения Коши.

И еще одно распределение будет нам полезно.

Определение 5. Пусть $X \sim \chi^2_n$, $Y \sim \chi^2_m$ независимы, тогда величина

$$\frac{X/n}{Y/m}$$

называется величиной с распределением Фишера-Снедекора ($F_{n,m}$).

Распределение Фишера-Снедекора имеет плотность

$$f(x) = \frac{n^{n/2}x^{n/2-1}m^{m/2}}{(nx+m)^{(n+m)/2}B(n/2,m/2)}. \tag{11.2}$$

11.1.3 Доверительные интервалы в нормальной модели. Простой случай

Нам понадобится следующее соотношение, которым мы уже пользовались на прошлых занятиях:

$$\sum_{i=1}^{n}(a_i-b)^2 = \sum_{i=1}^{n}(a_i-\overline{a})^2 + n(\overline{a}-b)^2. \tag{11.3}$$

Пример 1. Пусть $X_i \sim \mathcal{N}(a,\sigma^2)$. При известном σ мы легко можем построить доверительный интервал для a, пользуясь тем, что

$$\frac{\sqrt{n}(\overline{X}-a)}{\sigma} \sim \mathcal{N}(0,1),$$

откуда имеем интервал

$$a \in \left(\overline{X} - z_{1-\alpha/2}\sigma n^{-1/2}, \overline{X} + z_{1-\alpha/2}\sigma n^{-1/2}\right).$$

Также не представляет труда построить доверительный интервал для σ при известном a. Действительно, $(X_i-a)/\sigma$ имеет $\mathcal{N}(0,1)$ распределение и величины независимы, откуда

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-a)^2 \sim \chi^2_n.$$

Таким образом, доверительный интервал имеет вид

$$\sigma \in \left(\sum_{i=1}^{n}(X_i-a)^2/y_{1-\alpha/2}, \sum_{i=1}^{n}(X_i-a)^2/y_{1-\alpha/2}\right),$$

где y_β — квантиль χ^2_n.

11.1.4 Лемма Фишера

Что же делать, если параметры a и σ^2 оба неизвестны? В этом случае приходит на ум идея подставить вместо неизвестных параметров ОМП и использовать в первом случае функцию

$$g_1(\overline{X}, S, a) = \frac{\sqrt{n}(\overline{X}-a)}{S},$$

а во втором случае

$$g_2(S,\sigma) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\overline{X})^2.$$

К сожалению, подстановка оценок изменила распределение функций g_1 и g_2, однако достаточно очевидно, что эти распределения не зависят от параметров. Чтобы их найти, мы используем следующую лемму Фишера.

Лемма 1. *Если* $X_i \sim \mathcal{N}(\mu, \sigma^2)$, *то величины*

$$g_3(\overline{X}, \mu, \sigma) = \sqrt{n}\frac{\overline{X} - \mu}{\sigma},\ g_2(S, \sigma) = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{\sigma^2}$$

являются независимыми, причем $g_3 \sim \mathcal{N}(0,1)$, $g_2 \sim \chi^2_{n-1}$.

Доказательство Леммы 1. Будем доказывать для $\mu = 0$, $\sigma = 1$, общее доказательство получается заменой $(X_i - \mu)/\sigma$ за новую переменную.
То, что $T_1 \sim \mathcal{N}(0,1)$ очевидно.

Заметим, что $(X_1 - \overline{X}, X_2 - \overline{X}, ..., X_n - \overline{X}, \overline{X})$ — нормальный вектор, поскольку это линейное преобразование нормального вектора. Но

$$\operatorname{cov}(X_i - \overline{X}, \overline{X}) = n^{-1}\operatorname{cov}(X_i, X_i) - \mathbf{D}\overline{X} = n^{-1}\sigma^2 - n^{-1}\sigma^2 = 0,$$

то есть $\vec{X}$ некоррелирован с $X_i - \overline{X}$, а значит не зависит от

$$\sum_{i=1}^n (X_i - \overline{X})^2.$$

Итак, S^2 не зависит от $\overline{X}$. При этом $nS^2 + n\overline{X}^2 = n\overline{X^2} \sim \chi^2_n$. Таким образом,

$$Y + Z = U,\ Y = nS^2,\ Z = n\overline{X}^2 \sim \chi^2_n,\ U = n\overline{X^2} \sim \chi^2_n,$$

причем Y и Z независимы. Значит

$$\psi_Y(t)\psi_Z(t) = \psi_U(t),\ \psi_Y(t) = \frac{\psi_U(t)}{\psi_Z(t)} = \frac{\psi_{\chi^2_n}(t)}{\psi_{\chi^2_1}(t)} = \psi_{\chi^2_{n-1}}(t).$$

Здесь мы пользуемся тем, что $\psi_Z(t)$ не равно 0 при любых t. Лемма доказана. □

С помощью полученного факта можно построить доверительные интервалы для параметров среднего и дисперсии в нормальной модели.

Пример 2. Построим доверительный интервал для σ^2 при неизвестном μ. Имеем

$$\frac{nS^2}{\sigma^2} = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{\sigma^2} \sim \chi^2_{n-1},$$

откуда

$$\mathbf{P}\left(\frac{nS^2}{\sigma^2} \in (x_{\alpha/2}, x_{1-\alpha/2})\right) = 1 - \alpha,$$

где x_α — квантиль χ^2_{n-1}. Отсюда доверительный интервал имеет вид

$$\sigma^2 \in \left(\frac{nS^2}{x_{1-\alpha/2}}, \frac{nS^2}{x_{\alpha/2}}\right).$$

Пример 3. Построим доверительный интервал для μ при неизвестном σ. Мы знаем, что

$$\sqrt{n}\frac{\overline{X} - \mu}{\sigma} \sim \mathcal{N}(0,1),$$

но это соотношение включает в себя неизвестный нам σ. Чтобы избавиться от него, воспользуемся тем, что

$$\sqrt{n}S/\sigma$$

не зависит от $\overline{X}$ и имеет распределение, соответствующее корню из величины χ^2_{n-1}, откуда

$$\frac{\sqrt{n}\frac{(\overline{X}-\mu)}{\sigma}}{\sqrt{n}S/\sigma} = \frac{\sqrt{n}(\overline{X}-\mu)}{\sqrt{\sum_{i=1}^n (X_i-\overline{X})^2}}$$

имеет то же распределение, что отношение $U/\sqrt{V}$, $U \sim \mathcal{N}(0,1)$, $V \sim \chi^2_{n-1}$. Соответственно,

$$\frac{\sqrt{n}(\overline{X}-\mu)}{\sqrt{\frac{1}{n-1}\sum_{i=1}^n (X_i-\overline{X})^2}}$$

имеет то же распределение то же, что у $U/\sqrt{V/(n-1)}$, то есть распределение Стьюдента t_{n-1}.

Тогда имеем

$$P\left(\frac{\sqrt{n}(\overline{X}-\mu)}{\sqrt{\frac{1}{n-1}\sum_{i=1}^n (X_i-\overline{X})^2}} \in (t_{\alpha/2}, t_{1-\alpha/2})\right) = 1-\alpha,$$

где t — квантиль распределения Стьюдента с $n-1$ степенью свободы. Отсюда имеем доверительный интервал для μ с границами

$$\overline{X} \pm t_{1-\alpha/2} n^{-1/2} \sqrt{\frac{1}{n-1}\sum_{i=1}^n (X_i-\overline{X})^2},$$

где мы воспользовались симметричностью распределения Стьюдента $t_\beta = -t_{1-\beta}$.

Можно смотреть на полученный доверительный интервал как на интервал, полученный на основе замены неизвестного параметра σ на оценку $S_0 = \sqrt{\frac{1}{n-1}\sum_{i=1}^n (X_i-\overline{X})^2}$. Оказывается, что при этом нормальное $\mathcal{N}(0,1)$ заменяется на распределение Стьюдента t_{n-1}.

11.1.5 Критерии в нормальной модели

На основе данных интервалов можно построить критерий для проверки гипотез $H_0 : \mu = \mu_0$ с альтернативой $H_1 : \mu \neq \mu_0$, откуда получим критерий

$$\mu_0 \notin \left(\overline{X} - t_{1-\alpha/2} n^{-1/2} \sqrt{\frac{1}{n-1}\sum_{i=1}^n (X_i-\overline{X})^2}, \overline{X} + t_{1-\alpha/2} n^{-1/2} \sqrt{\frac{1}{n-1}\sum_{i=1}^n (X_i-\overline{X})^2}\right).$$

К тому же критерию можно прийти с точки зрения критерия отношения правдоподобий.

Пример 4. В силу примера с прошлого занятия

$$\lambda(x_1,...,x_n) = \frac{\left(\frac{1}{\sqrt{2\pi\sum_{i=1}^n (x_i-\mu)^2}}\right)^n \exp(-n/2)}{\left(\frac{1}{\sqrt{2\pi}S}\right)^n \exp(-n/2)} = \left(\frac{\sum_{i=1}^n (x_i-\mu_0)^2}{\sum_{i=1}^n (x_i-\overline{x})^2}\right)^{n/2}.$$

Тогда

$$\{\lambda(x_1,...,x_n) > c\} = \left\{\frac{\sum_{i=1}^n (x_i-\overline{x})^2 + n(\overline{x}-\mu_0)^2}{\sum_{i=1}^n (x_i-\overline{x})^2} > \tilde{c}\right\} = \left\{\frac{|\overline{x}-\mu_0|}{\sqrt{\sum_{i=1}^n (x_i-\overline{x})^2}} > \hat{c}\right\}.$$

При этом $\hat{c}$ находится из соотношения

$$P_{\mu_0}\left(\frac{|\overline{x}-\mu_0|}{\sqrt{\sum_{i=1}^n (x_i-\overline{x})^2}} > \hat{c}\right) = P_{\mu_0}\left(\frac{\sqrt{n}|\overline{x}-\mu_0|}{\sqrt{\frac{1}{n-1}\sum_{i=1}^n (x_i-\overline{x})^2}} > \sqrt{n(n-1)}\hat{c}\right) = 2\left(1-F_{t_{n-1}}\left(\sqrt{n(n-1)}\hat{c}\right)\right) = \alpha,$$

Отсюда получаем тот же критерий

$$\frac{\sqrt{n}|\overline{x}-\mu_0|}{\sqrt{\frac{1}{n-1}\sum_{i=1}^n (x_i-\overline{x})^2}} > t_{1-\alpha/2}.$$

Аналогичным образом может быть проверена гипотеза для нескольких выборок.

Пример 5. Проверим гипотезу $H_0 : \theta_1 = \theta_2$ с альтернативой $\theta_1 > \theta_2$, $X_i \sim \mathcal{N}(0,\theta_1^2)$, $Y_i \sim \mathcal{N}(0,\theta_2^2)$. Стоит обратить внимание, что для такого рода задач нельзя использовать асимптотическую версию, поскольку мы имеем дело с нерегулярной моделью — область изменения параметра не является открытым множеством (точки с $\theta_1 = \theta_2$ являются граничными).

Правдоподобие имеет вид

$$L(x_1,..,x_n,y_1,...,y_m;\theta_1,\theta_2) = \left(\frac{1}{\sqrt{2\pi}\theta_1}\right)^n \left(\frac{1}{\sqrt{2\pi}\theta_2}\right)^m \exp\left(-\frac{1}{2\theta_1^2}\sum_{i=1}^n x_i^2 + \frac{1}{2\theta_2^2}\sum_{i=1}^m y_i^2\right).$$

Тогда

$$\frac{\partial \ln L}{\partial \theta_1} = \frac{n}{\theta_1} - \frac{1}{\theta_1^3}\sum_{i=1}^n x_i^2,\ \frac{\partial \ln L}{\partial \theta_2} = \frac{m}{\theta_2} - \frac{1}{\theta_2^3}\sum_{i=1}^m x_i^2.$$

Единственной критической точкой функции L, таким образом, является $\overline{x^2}, \overline{y^2}$. Однако в случае $\overline{x^2} < \overline{y^2}$ данная точка лежит за пределами рассматриваемой области $\theta_1 \geq \theta_2$, а, следовательно, максимум будет достигаться в какой-либо точке на границе, то есть при $\overline{x^2} = \overline{y^2}$. При этом отношение правдоподобий будет равно 1.

При выполнении гипотезы H_0 правдоподобие приобретает вид

$$L(x_1,..,x_n,y_1,...,y_m;\theta_1,\theta_2) = \left(\frac{1}{\sqrt{2\pi}\theta_1}\right)^{m+n} \exp\left(-\frac{1}{2\theta_1^2}\sum_{i=1}^n x_i^2 - \frac{1}{2\theta_1^2}\sum_{i=1}^m y_i^2\right).$$

Аналогичным образом устанавливается, что ОМП в данном случае есть $\sqrt{(m\overline{x^2}+n\overline{y^2})/(m+n)}$.

Таким образом, при $\overline{x^2} < \overline{y^2}$ статистика отношения правдоподобий λ есть 1, при $\overline{x^2} \geq \overline{y^2}$

$$\lambda(x_1,...,x_n,y_1,...,y_m) = \left(\frac{n\overline{x^2}}{(m\overline{x^2}+n\overline{y^2})/(m+n)}\right)^{n/2} \left(\frac{m\overline{x^2}}{(m\overline{x^2}+n\overline{y^2})/(m+n)}\right)^{m/2} = (m+n)^{(m+n)/2} m^{-m/2} n^{-n/2} \left(\frac{\overline{y^2}}{\overline{x^2}}+\frac{n}{m}\right)^{-n/2} \left(\frac{\overline{x^2}}{\overline{y^2}}+\frac{m}{n}\right)^{-m/2}.$$

Таким образом, критическое множество имеет вид

$$\left\{\vec{x},\vec{y} : n\ln\left(\frac{\overline{y^2}}{\overline{x^2}}+\frac{n}{m}\right) + m\ln\left(\frac{\overline{x^2}}{\overline{y^2}}+\frac{m}{n}\right) < c, \overline{x^2} \geq \overline{y^2}\right\}.$$

Функция $g(x) = n\ln(x + n/m) + m\ln(1/x + m/n)$ имеет производную

$$g'(x) = \frac{mn}{mx+n} + \frac{mn}{n+mx} - \frac{m}{x} = m\left(\frac{nx + mnx^2 + mx + mnx^2 - (mx+n)(nx+m)}{(mx+n)(nx+m)x}\right) = \frac{m^2n(x^2-1)}{x(mx+n)(nx+m)} \leq 0$$

при $x \leq 1$. Следовательно, функция g монотонно убывает и наше критическое множество переписывается в виде

$$\frac{\overline{x^2}}{\overline{y^2}} > \tilde{c},$$

где $\tilde{c} > 1$. Указанная дробь имеет распределение Фишера. Тогда наше критическое множество приобретает вид

$$\frac{\overline{x^2}}{\overline{y^2}} > f_{1-\alpha},$$

где $f_{1-\alpha}$ — квантиль распределения Фишера с n, m степенями свободы. При этом если $\overline{x^2} < \overline{y^2}$, то мы автоматически принимаем гипотезу (или если уровень значимости столь велик, что этого мы сделать не можем, то придется строить рандомизированный критерий).

11.2 Факультативная часть

11.2.1 Обобщенное хи-квадрат распределение

Справедлива следующая лемма:

Лемма 2. *Пусть $\vec{X} \sim \mathcal{N}(0, \Sigma)$, то*

$$||\vec{X}||^2 \stackrel{d}{=} \lambda_1 Z_1^2 + \cdots + \lambda_n Z_n^2,$$

где λ_i – собственные числа матрицы Σ.

Доказательство. Заметим, что

$$A\vec{X} \sim \mathcal{N}(0, A\Sigma A^T),$$

откуда выбирая в качестве A ортогональную матрицу, приводящую Σ к диагональному виду, мы получим

$$A\vec{X} \sim \mathcal{N}(0, D),$$

где D – диагональная матрица с λ_i на диагонали. Но раз A – ортогональна, то

$$||\vec{X}||^2 = ||A\vec{X}||^2 = \lambda_1 Z_1^2 + \cdots + \lambda_n Z_n^2.$$

□

11.2.2 Проверка независимости

Выборочный коэффициент корреляции

$$R^2 = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^n (X_i - \overline{X})^2 \sum_{i=1}^n (Y_i - \overline{Y})^2}}$$

используют для оценки коэффициента корреляции $\sigma_{1,2}$ координат X и Y нормального вектора на основе выборки $(X_i, Y_i) \sim \mathcal{N}(\vec{\mu}, \Sigma)$.

Поскольку в нормальном случае некореллированность равносильна независимости, то гипотеза независимости X и Y равносильна гипотезе $\sigma_{1,2} = 0$. Можно показать (см. 12.1.3), что критерий отношения правдоподобий в этом случае имеет вид $\{R^2 > c\}$. Можно переписать его в виде

$$\{|W| > c\}, W = \frac{\sqrt{n-2} \cdot R}{1 - R^2}.$$

причем величина W в случае независимости имеет распределение t_{n-2} (это некоторое утверждение). Таким образом, критерий имеет вид $|W| > t_{1-\alpha/2}$.