MS LEC 04

isagila ablearthy

Собрано 07.08.2024 в 15:41

Содержание

1. Лекции

1.1. Лекция 24.02.08.

Выборки

Def 1.1.1. Под генеральной совокупностью понимают все результаты данной серии экспериментов (или экспериментальных значений случайной величины).

Def 1.1.2. Под выборочной совокупностью понимают имеющиеся у нас данные (выборка из генеральной совокупности, возможно неполная).

Замечание 1.1.3. Выборочная совокупность не всегда отражает реальное поведение случайной величины. В качестве иллюстрации можно рассматривать известный пример с бронированием самолетов (классическая ошибка выжившего).

Def 1.1.4. Репрезентативной выборкой называется выборка, имеющая то же самое распределение, что и у генеральной совокупности.

Замечание 1.1.5. В дальнейшем в курсе предполагаем, что все выборки репрезентативные.

Def 1.1.6 (Первое)**.** Выборкой объема n называется набор экспериментальных данных $(x_1, \dots, x_n)$.

Def 1.1.7 (Второе)**.** Выборкой объема n называется набор $X_1, \dots, X_n$ независимых одинаково распределенных случайных величин.

Замечание 1.1.8. Таким образом $\mathbb{E}(X_i) = \mathbb{E}(X_1)$ и $\mathbb{D}(X_i) = \mathbb{D}(X_1)$, поэтому обычно числовые характеристики i-того заменяются числовыми характеристиками первого экземпляра.

Выборочные характеристики

Выборку можно рассматривать как дискретную случайную величину.

X	x_1	$\dots$	x_n
p^*	$\frac{1}{n}$	$\dots$	$\frac{1}{n}$

Def 1.1.9. Средним выборочным $\overline{x}$ называется число

$$\overline{x} = \frac{x_1 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Это будет оценкой неизвестного математического ожидания.

Def 1.1.10. Выборочной дисперсией $\mathbb{D}^*$ называется число

$$\mathbb{D}^* = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2$$

Это будет оценкой неизвестной дисперсии.

Def 1.1.11. Выборочным среднеквадратическим отклонение называется число

$$\sigma^* = \sqrt{\mathbb{D}^*}$$

Это будет оценкой неизвестного среднеквадратического отклонения.

Def 1.1.12. Выборочной функцией распределения $F^*(y)$ называется функция

$$F^*(y) = \frac{\text{число данных } x_i \in (-\infty; y)}{n}$$

Теорема 1.1.13. (Гливенко—Кантелли) Пусть имеется выборка $\vec{X} = (x_1, \dots, x_n)$ объема n. Обозначим $F^*(y)$ — эмпирическую функцию распределения, а $F(y)$ — теоретическую функцию распределения. Тогда

$$\sup_{y \in \mathbb{R}} |F^*(y) - F(y)| \xrightarrow{\mathsf{P}} 0 \qquad \text{при } n \to \infty$$

Начальная обработка статданных

I. Ранжирование выборки

Упорядочиваем данные по возрастанию, в результате получаем вариационный ряд вида $X_{(1)}, \dots, X_{(n)}$.

Def 1.1.14. Разность $X_{(n)} - X_{(1)}$ называется размахом выборки.

Def 1.1.15. Элемент $X_{(i)}$ в полученном ряде называется i-той порядковой статистикой.

Замечание 1.1.16. Если мы при этом объединяем повторяющиеся результаты (с учетом числа этих результатов), то получаем частотный вариационный ряд.

II. Разбиение на интервалы

Если много неповторяющихся данных, то разбиваем их на интервалы и составляем интервальный вариационный ряд.

Замечание 1.1.17. Есть два подхода к разбиению на интервалы.

1. Берем интервалы одинаковой длины. Это удобнее для построения гистограммы и выдвижения гипотезы и типе распределения.
2. Берем равнонаполненные интервалы. Это удобнее при проверке гипотез о типе распределения.

Замечание 1.1.18. Обычно (но далеко не всегда) количество интервалов вычисляется по формуле Стерджеса.

$$K \approx 1 + \log_2 n$$

В результате получаем K интервалов $[a_{i-1}; a_i)$ и считаем их частоты v_i — число данных, попавших в i-ый интервал. Величина $\frac{v_i}{n}$ называется относительной частотой (она является оценкой теоретической вероятности попадания случайной величины в данный интервал). Если заменяем интервалы на их середины $c_i = \frac{a_{i-1}+a_i}{2}$, то опять получаем огрубленный вариационный ряд, по которому можно дать точные оценки числовым характеристикам распределения. Для этого можно использовать формулы

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} c_i v_i \qquad \mathbb{D}^* = \frac{1}{n}\sum_{i=1}^{n} (c_i - \overline{x})^2 v_i$$

Геометрическая интерпретация данных

Обычно удобнее визуализировать данные в виде гистограммы. На плоскости строим набор прямоугольников для каждого интервала. Основание это $[a_{i-1}; a_i)$ длины $l = a_i - a_{i-1}$, а высоту берем пропорционально частоте, причем таким образом, чтобы общая площадь фигуры равнялась единице, т.е. высота будет равна $\frac{v_i}{nl}$. Гистограмма является приближением плотности распределения, и по ее виду можно выдвинуть гипотезы о типе распределения.

Теорема 1.1.19. Если число интервалов $K(n) \to \infty$ и $\frac{K(n)}{n} \to 0$ при $n \to \infty$, то гистограмма по вероятности поточечно сходится к теоретической плотности.

1.2. Лекция 24.02.15.

Точечные оценки

Пусть имеется выборка $\vec{X} = (X_1, \ldots, X_n)$ объема n — набор независимых экземпляров случайной величины X.

Def 1.2.1. Статистикой называется измеримая функция $\Theta^* = \Theta^*(X_1, \ldots, X_n)$.

Пусть требуется найти приближенную оценку неизвестного параметра Θ по выборке $(X_1, \ldots, X_n)$. Оценка считается при помощи некоторой статистики $\Theta^* = \Theta^*(X_1, \ldots, X_n)$.

Свойства статистических оценок

Def 1.2.2. Статистика $\Theta^* = \Theta^*(X_1, \ldots, X_n)$ неизвестного параметра Θ называется состоятельной, если $\Theta^* \xrightarrow{\mathsf{P}} \Theta$ при $n \to \infty$.

Def 1.2.3. Статистика $\Theta^* = \Theta^*(X_1, \ldots, X_n)$ неизвестного параметра Θ называется несмещенной, если $\mathbb{E}(\Theta^*) = \Theta$.

Замечание 1.2.4. Оценка называется асимптотически несмещенной, если $\mathbb{E}(\Theta^*) \to \Theta$ при $n \to \infty$.

Def 1.2.5. Оценка Θ_1^* не хуже оценки Θ_2^*, если

$$\mathbb{E}\left((\Theta_1^* - \Theta)^2\right) \leqslant \mathbb{E}\left((\Theta_2^* - \Theta)^2\right)$$

Если Θ_1^* и Θ_2^* — несмещенные оценки, то это равносильно тому, что $\mathbb{D}(\Theta_1^*) \leqslant \mathbb{D}(\Theta_2^*)$.

Def 1.2.6. Оценка Θ^* называется эффективной, если она не хуже всех остальных оценок.

Замечание 1.2.7. В классе всех возможных оценок не существует эффективной оценки.

Теорема 1.2.8. В классе несмещенных оценок существует эффективная оценка, причем единственная.

Точечные оценки моментов

Def 1.2.9. Выборочным средним $\overline{x}$ называется величина

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Def 1.2.10. Выборочной дисперсией $\mathbb{D}^*$ называется величина

$$\mathbb{D}^* = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2$$

Def 1.2.11. Исправленной выборочной дисперсией S^2 называется величина

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \overline{x})^2 = \frac{n}{n-1}\mathbb{D}^*$$

Def 1.2.12. Выборочным среднеквадратичным отклонением называется величина

$$\sigma^* = \sqrt{\mathbb{D}^*}$$

Def 1.2.13. Исправленным выборочным среднеквадратичным отклонением называется величина

$$S = \sqrt{S^2}$$

Def 1.2.14. Выборочным k-ым моментов $\overline{x^k}$ называется величина

$$\overline{x^k} = \frac{1}{n}\sum_{i=1}^{n} x_i^k$$

Def 1.2.15. Выборочной модой Mo^* называется варианта с наибольшей частотой.

$$Mo^* = x_i \,\Big|\, n_i = \max(n_1, \ldots, n_k)$$

Def 1.2.16. Выборочной медианой Me^* называется значение варианты x_i в середине ряда.

$$Me^* = \begin{cases} x_{(k)}, & \text{если } n = 2k-1 \\ \frac{1}{2}\left(x_{(k)} + x_{(k+1)}\right), & \text{если } n = 2k \end{cases}$$

Теорема 1.2.17. Выборочное среднее $\overline{x}$ является несмещенной состоятельной оценкой для математического ожидания.

$$\mathbb{E}(\overline{x}) = \mathbb{E}(X) = a \qquad \overline{x} \xrightarrow{\mathsf{P}} \mathbb{E}(X) = a$$

□ Покажем несмещенность.

$$\mathbb{E}(\overline{x}) = \mathbb{E}\left(\frac{x_1 + \ldots + x_n}{n}\right) = \frac{1}{n} \cdot n \cdot \mathbb{E}(x_1) = \mathbb{E}(x_1)$$

Покажем состоятельность.

$$\overline{x} = \frac{x_1 + \ldots + x_n}{n} \xrightarrow{\mathsf{P}} \mathbb{E}(X) \text{ при } n \to \infty$$

Это закон больших чисел. ■

Теорема 1.2.18. Выборочный k-ый момент $\overline{x^k}$ является несмещенной состоятельной оценкой для теоретического k-ого момента.

$$\mathbb{E}\left(\overline{x^k}\right) = \mathbb{E}\left(X^k\right) = m_k \qquad \mathbb{E}\left(\overline{x^k}\right) \xrightarrow{\mathsf{P}} \mathbb{E}\left(X^k\right) = m_k \text{ при } n \to \infty$$

□ Это следует из предыдущей теоремы, если в качестве случайной величины взять x^k. ■

Теорема 1.2.19. Выборочные дисперсии $\mathbb{D}^*$ и S^2 являются состоятельным оценками для дисперсии. При этом $\mathbb{D}^*$ — смещенная оценка (есть систематическая вниз), а S^2 — несмещенная оценка.

□ Заметим, что

$$\mathbb{D}^* = \frac{1}{n}\sum (x_i - \overline{x})^2 = \overline{x^2} - (\overline{x})^2$$

$$\mathbb{D}(\overline{x}) = \mathbb{E}\left((\overline{x})^2\right) - (\mathbb{E}(\overline{x}))^2 \Longrightarrow \mathbb{E}\left((\overline{x})^2\right) = (\mathbb{E}(\overline{x}))^2 + \mathbb{D}(\overline{x})$$

Таким образом

$$\mathbb{E}(\mathbb{D}^*) = \mathbb{E}\left(\overline{x^2}\right) - \mathbb{E}\left((\overline{x})^2\right) = \mathbb{E}\left(x^2\right) - \left((\mathbb{E}(x))^2 + \mathbb{D}(\overline{x})\right) = \left(\mathbb{E}\left(x^2\right) - (\mathbb{E}(x))^2\right) - \mathbb{D}(\overline{x}) = \mathbb{D}(x) - \mathbb{D}(\overline{x})$$

Видно, что будет смещение на величину $\mathbb{D}(\overline{x})$. Преобразуем полученное выражение в более удобную форму.

$$\mathbb{D}(x) - \mathbb{D}(\overline{x}) = \mathbb{D}(x) - \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{D}(x_i) = \mathbb{D}(x) - \frac{1}{n}\cdot\mathbb{D}(x) = \frac{n-1}{n}\cdot\mathbb{D}(x)$$

Значит

$$\mathbb{E}\left(S^2\right) = \mathbb{E}\left(\frac{n}{n-1}\mathbb{D}^*\right) = \frac{n}{n-1}\cdot\frac{n-1}{n}\cdot\mathbb{D}(x) = \mathbb{D}(x)$$

Т.е. S^2 это несмещенная оценка. Далее покажем состоятельность оценок.

$$\mathbb{D}^* = \overline{x^2} - (\overline{x})^2 \xrightarrow{\mathsf{P}} \mathbb{E}\left(x^2\right) - (\mathbb{E}(x))^2 = \mathbb{D}(x)$$

$$S^2 = \underbrace{\tfrac{n}{n-1}}_{\to 1}\mathbb{D}^* \xrightarrow{\mathsf{P}} \mathbb{D}(x)$$

■

Замечание 1.2.20. Т.к. $\frac{n}{n-1} \to 1$ при $n \to \infty$, то $\mathbb{E}(\mathbb{D}^*) \xrightarrow{\mathsf{P}} \mathbb{D}(x)$ при $n \to \infty$. Значит выборочная дисперсия является асимптотически несмещенной оценкой, поэтому на практике при $n \geqslant 100$ можно считать обычную выборочную дисперсию, а при $n < 100$ следует ее заменить на исправленную выборочную дисперсию.

Метод моментов (Пирсона)

Зная выборочные моменты можно дать оценки остальным параметрам распределения. Пусть имеется выборка $\vec{X} = (X_1, \ldots, X_n)$ неизвестного распределения, но при этом знаем, что данное распределение определенного типа, задаваемого k параметрами $\Theta = (\Theta_1, \ldots, \Theta_k)$. Зная параметры можем вычислить теоретические k-ые моменты. Например, если распределение непрерывное, то

$$m_i = \int_{\mathbb{R}} x^k f(x, \Theta_1, \ldots, \Theta_k)\,\mathrm{d}x = h_i(\Theta_1, \ldots, \Theta_k)$$

Вычислим выборочные моменты и подставим в эти формулы, получим систему уравнений.

$$\begin{cases} \overline{x} = & h_1(\Theta_1, \ldots, \Theta_k) \\ \overline{x^2} = & h_2(\Theta_1, \ldots, \Theta_k) \\ & \vdots \\ \overline{x^k} = & h_k(\Theta_1, \ldots, \Theta_k) \end{cases}$$

Решив эту систему, находим оценки $\Theta_1^*, \ldots, \Theta_k^*$ неизвестных параметров.

Замечание 1.2.21. При этом обычно получаем состоятельные оценки, но смещенные.

Пример 1.2.22. Пусть $X \in U(a; b)$. При обработке статданных получили оценки первого и второго момента $\overline{x} = 2.25$ и $\overline{x^2} = 6.75$. Необходимо дать оценки неизвестных параметров a и b.
Решение: Плотность будет иметь вид

$$f(x, a, b) = \begin{cases} 0, & x < a \\ \frac{1}{b-a}, & a \leqslant x \leqslant b \\ 0, & x > b \end{cases}$$

Тогда

$$\mathbb{E}(x) = \frac{a+b}{2} \qquad \mathbb{E}\left(x^2\right) = \frac{a^2+ab+b^2}{3}$$

Составим систему

$$\begin{cases} 2.25 = \frac{a^*+b^*}{2} \\ 6.75 = \frac{(a^*)^2+a^*b^*+(b^*)^2}{3} \end{cases} \Longleftrightarrow \begin{cases} a^* + b^* = 4.5 \\ (a^*)^2 + a^*b^* + (b^*)^2 = 20.25 \end{cases} \Longleftrightarrow \begin{cases} a^* + b^* = 4.5 \\ a^*b^* = 0 \end{cases} \overset{\text{a < b}}{\Longleftrightarrow} \begin{cases} a^* = 0 \\ b^* = 4.5 \end{cases}$$

1.3. Лекция 24.02.22.

Пусть имеется выборка распределения известного нам типа, но с неизвестными параметрами. Требуется найти эти параметры. На прошлой лекции был рассмотрен метод моментов. Его недостаток заключается в том, что он не гарантирует эффективность оценок.

Метод максимального правдоподобия (Фишера)

Пусть имеется выборка $\vec{X} = (X_1, \dots, X_n)$ распределения известного типа, задаваемого параметрами $\Theta = (\Theta_1, \dots, \Theta_k)$. Идея метода заключается в подборе параметров таким образом, чтобы вероятность получения данной выборки при случайном эксперименте была наибольшей. Например, если распределение дискретное, то эта вероятность будет равна

$$\mathsf{P}_\Theta(x_1, \dots, x_n) = \mathsf{P}_\Theta(x = x_1) \cdot \ldots \cdot \mathsf{P}_\Theta(x = x_n)$$

Def 1.3.1. Функцией правдоподобия $L\left(\vec{X}, \Theta\right)$ называется функция

$$L\left(\vec{X}, \Theta\right) = \prod_{i=1}^{n} \mathsf{P}\left(x = x_i\right) \qquad \text{при дискретном распределении}$$

$$L\left(\vec{X}, \Theta\right) = \prod_{i=1}^{n} f_\Theta(x_i) \qquad \text{при непрерывном распределении}$$

Def 1.3.2. Логарифмической Функцией правдоподобия называется функция $\ln L\left(\vec{X}, \Theta\right)$.

Замечание 1.3.3. Т.к. $y = \ln x$ возрастающая функция, то их точки экстремума совпадают, но искать их проще во втором случае.

Def 1.3.4. Оценкой максимального правдоподобия $\widehat{\Theta}$ называется значение Θ, при котором функция правдоподобия достигает наибольшего значения (при фиксированных значениях $x_1, \dots, x_n$).

Пример 1.3.5. Пусть $\vec{X} = (X_1, \dots, X_n)$ из распределения Пуассона с неизвестным параметром $\lambda > 0$. Требуется найти оценку максимального правдоподобия параметра λ.
Решение: Составим функцию правдоподобия.

$$L\left(\vec{X}, \alpha\right) = \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} = \frac{\lambda^{n\overline{x}}}{\prod_{i=1}^{n} x_i!} e^{-n\lambda}$$

Перейдем к логарифмической функции правдоподобия.

$$\ln L\left(\vec{X}, \lambda\right) = n\overline{x} \ln \lambda - \ln \prod_{i=1}^{n} x_i! - n\lambda$$

Вычислим производную по λ и приравняем ее к нулю, чтобы найти экстремум.

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\ln L\left(\vec{X}, \lambda\right)\right) = \frac{n\overline{x}}{\lambda} - n = 0 \Longrightarrow \lambda = \overline{x}$$

Покажем, что это действительно точка максимума.

$$\frac{\mathrm{d}^2}{\mathrm{d}\lambda^2}\left(\ln L\left(\vec{X}, \lambda\right)\right) = -\frac{n\overline{x}}{\lambda^2} < 0$$

Итого $\lambda = \overline{x}$ это оценка максимального правдоподобия.

Пример 1.3.6. Пусть $\vec{X} = (X_1, \dots, X_n)$ — выборка из $N\left(a; \sigma^2\right)$, где $a \in \mathbb{R}$ и $\sigma > 0$. Требуется найти оценку максимального правдоподобия.
Решение: Составим функцию правдоподобия.

$$L\left(\vec{X}, a, \sigma\right) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x_i - a)^2}{2\sigma^2}\right) = \frac{1}{\sigma^n \left(2\pi\right)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - a)^2\right)$$

Перейдем к логарифмической функции правдоподобия.

$$\ln L\left(\vec{X}, a, \sigma\right) = -n \ln \sigma - \frac{n}{2} \ln(2\pi) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - a)^2$$

Найдем частные производные по a и σ.

$$\frac{\partial}{\partial a}\ln L\left(\vec{X},a,\sigma\right)=-\frac{1}{2\sigma^2}\sum_{i=1}^{n}2(x_i-a)\cdot(-1)=\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-a)=\frac{n}{\sigma^2}\left(\overline{x}-a\right)$$

$$\frac{\partial}{\partial \sigma}\ln L\left(\vec{X},a,\sigma\right)=-\frac{n}{\sigma}-\frac{1}{2}\cdot(-2)\sigma^{-3}\cdot\sum_{i=1}^{n}\left(x_i-a\right)^2=\frac{1}{\sigma^3}\sum_{i=1}^{n}\left(x_i-a\right)^2-\frac{n}{\sigma}$$

Приравняем их к нулю. Из производной по a получаем $\widehat{a}=\overline{x}$. Поработаем с производной по σ.

$$\frac{1}{\sigma^3}\sum_{i=1}^{n}\left(x_i-a\right)^2-\frac{n}{\sigma}=0\Longrightarrow\widehat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}\left(x_i-a\right)^2\Longrightarrow\widehat{\sigma}^2=\mathbb{D}^*$$

То, что это действительно точка максимума можно показать с помощью вторых дифференциалов (самостоятельно).

Пример 1.3.7. Пусть $\vec{X}=(X_1,\ldots,X_n)$ из $U\left(0;\Theta\right)$, где $\Theta>0$. Найти оценки Θ методом моментов и методом максимального правдоподобия. Сравнить результаты.
Решение: По методу моментов

$$\mathbb{E}\left(X\right)=\frac{0+\Theta}{2}\Longrightarrow\overline{x}=\frac{\Theta^*}{2}\Longrightarrow\Theta^*=2\overline{x}$$

Далее рассмотрим метод максимального правдоподобия. Обозначим $X_{(n)}$ n-ую порядковую статистику (т.е. максимальный элемент выборки), тогда

$$L\left(\vec{X},\Theta\right)=\prod_{i=1}^{n}f_{\Theta}(x_i)=\begin{cases}0, & \Theta<X_{(n)}\\ \frac{1}{\Theta^n}, & \Theta\geqslant X_{(n)}\end{cases}$$

Значит $\widehat{\Theta}=X_{(n)}$ это оценка максимального правдоподобия.
Видим, что полученные оценки отличаются. Сравним их. Сначала рассмотрим оценку по методу моментов. Отметим, что она будет несмещенной, т.к.

$$\Theta^*=2\overline{x}\qquad\mathbb{E}\left(\Theta^*\right)=2\mathbb{E}\left(\overline{x}\right)=2\mathbb{E}\left(x\right)=\Theta$$

Таким образом

$$\mathbb{E}\left(\left(\Theta^*-\Theta\right)^2\right)=\mathbb{D}\left(\Theta^*\right)=4\mathbb{D}\left(\overline{x}\right)=4\frac{\mathbb{D}\left(x\right)}{n}=\frac{4}{n}\cdot\frac{\Theta^2}{12}=\frac{\Theta^2}{3n}$$

Далее рассмотрим оценку по методу максимального правдоподобия. Вначале рассмотрим случайную величину $X_{(n)}=\max\left(X_1,\ldots,X_n\right)$. Найдем ее функцию распределения.

$$F_{X_{(n)}}=\mathsf{P}\left(X_{(n)}<x\right)=\mathsf{P}\left(X_1<x,\ldots,X_n<x\right)=\mathsf{P}\left(X_1<x\right)\cdot\ldots\cdot\mathsf{P}\left(X_n<x\right)=\left(F(x)\right)^n=\begin{cases}0, & x<0\\ \frac{x^n}{\Theta^n}, & 0\leqslant x\leqslant\Theta\\ 1, & x>\Theta\end{cases}$$

Далее найдем плотность.

$$f_{X_{(n)}}(x)=F'_{X_{(n)}}(x)=\begin{cases}0, & x<0\\ \frac{nx^{n-1}}{\Theta^n}, & 0\leqslant x\leqslant\Theta\\ 0, & x>\Theta\end{cases}$$

Вычислим первый момент.

$$\mathbb{E}\left(X_{(n)}\right)=\int_0^{\Theta}x\cdot\frac{nx^{n-1}}{\Theta^n}\mathrm{d}x=\frac{1}{\Theta^n}\int_0^{\Theta}nx^n\mathrm{d}x=\frac{1}{\Theta^n}\left.\frac{nx^{n+1}}{n+1}\right|_0^{\Theta}=\frac{1}{\Theta^n}\cdot\frac{n\Theta^{n+1}}{n+1}=\frac{n}{n+1}\Theta$$

Значит эта оценка будет смещенной вниз. Подправим ее следующим образом

$$\widetilde{\Theta}=\frac{n+1}{n}\widehat{\Theta}=\frac{n+1}{n}X_{(n)}$$

Такая оценка будет уже несмещенной. Вычислим ее второй момент.

$$\mathbb{E}\left(\widetilde{\Theta}^2\right)=\frac{(n+1)^2}{n^2}\int_0^{\Theta}x^2\frac{nx^{n-1}}{\Theta^n}\mathrm{d}x=\frac{(n+1)^2}{n}\cdot\frac{1}{\Theta^n}\int_0^{\Theta}x^{n+1}\mathrm{d}x=\frac{(n+1)^2}{n\Theta^n}\cdot\left.\frac{x^{n+2}}{n+2}\right|_0^{\Theta}=\frac{(n+1)^2\Theta^2}{n(n+2)}$$

Теперь вычислим дисперсию этой несмещенной оценки.

$$\mathbb{E}\left(\left(\widetilde{\Theta}-\Theta\right)^2\right)=\mathbb{D}\left(\widetilde{\Theta}\right)=\mathbb{E}\left(\widetilde{\Theta}^2\right)-\left(\mathbb{E}\left(\widetilde{\Theta}\right)\right)^2=\frac{(n+1)^2\Theta^2}{n(n+2)}-\Theta^2=\frac{n^2+2n+1-n^2-2n}{n(n+2)}\cdot\Theta^2=\frac{\Theta^2}{n(n+2)}$$

Итак, теперь можно сравнить эти две оценки.

$$\mathbb{D}(\Theta^*) = \frac{\Theta^2}{3n} \qquad \mathbb{D}\left(\widetilde{\Theta}\right) = \frac{\Theta^2}{n(n+2)}$$

Хорошо видно, что вторая оценка сходится быстрее, т.е. эффективность оценки по методу максимального правдоподобия заметно выше. Оценка по методу моментов сходится со скоростью $\sim \frac{1}{\sqrt{n}}$, а по методу максимального правдоподобия со скоростью $\sim \frac{1}{n}$.

Замечание 1.3.8. Т.к. $\mathbb{E}(X) = \frac{\Theta}{2}$, то из примера выше следует, что оценка $\overline{x}$ не является эффективной оценкой математического ожидания. В данном случае оценка $\widetilde{\mathbb{E}(X)} = \frac{n+1}{2n} X_{(n)}$ будет лучше. Можно показать, что эта оценка является эффективной.

Замечание 1.3.9. В общем случае равномерного распределения несмещенной эффективной оценкой математического ожидания будет $\frac{1}{2}\left(X_{(1)} + X_{(n)}\right)$. А несмещенной эффективной оценкой длины интервала будет $\frac{n+1}{n-1}\left(X_{(n)} - X_{(1)}\right)$.

Замечание 1.3.10. Таким образом мы видим, что при равномерном распределении нельзя откидывать крайние значения выборки.

Замечание 1.3.11. При методе максимального правдоподобия как правило получем состоятельные и эффективные оценки.

Неравенство Рао—Крамера

Пусть известно, что случайная величина $X \in \mathcal{F}_\Theta$, где $\mathcal{F}_\Theta$ это семейство распределений с параметром Θ.

Def 1.3.12. Носителем семейства распределений $\mathcal{F}_\Theta$ называется множество $C \in \mathbb{R}$ такое, что $\forall \Theta \,\Big|\, \mathsf{P}(X \in C) = 1$.

Замечание 1.3.13. Далее для того, чтобы не формулировать теорему дважды, под плотностью будем иметь в виду

$$f_\Theta(x) = \begin{cases} \text{плотность } f_\Theta(x), & \text{если распределение абсолютно непрерывное} \\ \mathsf{P}_\Theta(X = x), & \text{если распределение дискретное} \end{cases}$$

Def 1.3.14. Информацией Фишера $I(\Theta)$ семейства распределений $\mathcal{F}_\Theta$ называется величина

$$I(\Theta) = \mathbb{E}\left(\left(\frac{\partial}{\partial \Theta} \ln f_\Theta(x)\right)^2\right)$$

при условии, что она существует.

Def 1.3.15. Семейство распределений $\mathcal{F}_\Theta$ называется регулярным, если

1. $\exists C$ — носитель $\mathcal{F}_\Theta$ такой, что $\forall x \in C \,\Big|$ функция $\ln f_\Theta(x)$ непрерывно дифференцируема по Θ.
2. Информация Фишера $I(\Theta)$ существует и непрерывна по Θ.

Теорема 1.3.16. (Неравенство Рао—Крамера) Пусть $(X_1, \ldots, X_n)$ это выборка из регулярного семейства $\mathcal{F}_\Theta$, $\Theta^* = \Theta^*(X_1, \ldots, X_n)$ — несмещенная оценка параметра Θ, дисперсия $\mathbb{D}(\Theta^*)$ ограничена на любом компакте (компакт $\approx$ ограниченное замкнутое множество). Тогда

$$\mathbb{D}(\Theta^*) \geqslant \frac{1}{nI(\Theta)}$$

Замечание 1.3.17. Если $\mathbb{D}(\Theta^*) = \frac{1}{nI(\Theta)}$, то оценка Θ^* будет эффективной.

Пример 1.3.18. Пусть $\vec{X} = (X_1, \ldots, X_n)$ — выборка из $N\left(a; \sigma^2\right)$, где $a \in \mathbb{R}$ и $\sigma > 0$. Проверим эффективность оценки $a^* = \overline{x}$.
Решение: В качестве носителя выберем $C = (-\infty; \infty)$. Далее вычислим $\ln f(x)$.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-a)^2}{2\sigma^2}\right) \Longrightarrow \ln f(x) = -\ln\sigma - \ln\sqrt{2\pi} - \frac{(x-a)^2}{2\sigma^2}$$

Теперь вычислим частную производную по a.

$$\frac{\partial}{\partial a} \ln f(x) = -\frac{1}{2\sigma^2} \cdot 2(x-a) \cdot (-1) = \frac{x-a}{\sigma^2}$$

Видим, что эта производная непрерывна по a. Значит первое условие из определения 1.3.15 выполнено. Теперь вычислим информацию Фишера.

$$I(\Theta) = \mathbb{E}\left(\left(\frac{\partial}{\partial a} \ln f(x)\right)^2\right) = \mathbb{E}\left(\left(\frac{X-a}{\sigma^2}\right)^2\right) = \frac{\mathbb{E}\left((X - \mathbb{E}(X))^2\right)}{\sigma^4} = \frac{\mathbb{D}(X)}{\sigma^4} = \frac{1}{\sigma^2}$$

Видим, что информация Фишера существует и непрерывна по a. Значит выполнено второе условие из определения 1.3.15. Таком образом семейство нормальных распределений $N\left(a;\sigma^2\right)$ регулярно относительно параметра a. Аналогично можно показать, что оно будет регулярно относительно параметра σ. Вычислим дисперсию оценки.

$$\mathbb{D}\left(a^*\right) = \mathbb{D}\left(\overline{x}\right) = \frac{\mathbb{D}\left(X\right)}{n} = \frac{\sigma^2}{n}$$

В правой части неравенства 1.3.16 имеем

$$\frac{1}{nI(\Theta)} = \frac{\sigma^2}{n}$$

Видим, что дисперсия оценки получилась минимально возможной. Значит $\overline{x}$ это эффективная оценка параметра a. Аналогично можно показать, что S^2 — несмещенная эффективная оценка σ^2.

Замечание 1.3.19. В практической статистике часто используют линейные оценки $\Theta^* = \sum c_i x_i$. Лучшая оценка в таком классе называется наилучшей линейной несмещенной оценкой или BLUE-оценкой. В примере с нормальным распределение оценка $a^* = \overline{x}$ будет BLUE-оценкой, а в примере с равномерным распределением оценка $\Theta^* = 2\overline{x}$ будет BLUE-оценкой.

1.4. Лекция 24.02.29.

Основные распределения математической статистики

Def 1.4.1. Случайная величина имеет нормальное распределение $N\left(a;\sigma^2\right)$ с параметрами $a \in \mathbb{R}$ и $\sigma^2(\sigma > 0)$, если его плотность имеет вид

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-a)^2}{2\sigma^2}\right) \qquad x \in \mathbb{R}$$

Def 1.4.2. Распределение $N\left(0;1\right)$ с параметрами $a = 0, \sigma = 1$ называется стандартным нормальным распределением. Его плотность имеет вид

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

Свойства нормального распределения

1. Смысл параметров

$$\mathbb{E}\left(X\right) = a \qquad \mathbb{D}\left(X\right) = \sigma^2$$

2. Линейность

$$\xi \in N\left(a;\sigma^2\right) \Longrightarrow b\xi + c \in N\left(ab + c; b^2\sigma^2\right)$$

3. Стандартизация

$$\xi \in N\left(a;\sigma^2\right) \Longrightarrow \frac{\xi - a}{\sigma} \in N\left(0;1\right)$$

4. Устойчивость относительно суммирования

$$\left.\begin{array}{l} \xi_1 \in N\left(a_1;\sigma_1^2\right) \\ \xi_2 \in N\left(a_2;\sigma_2^2\right) \end{array}\right\} \Longrightarrow \xi_1 + \xi_2 \in N\left(a_1 + a_2; \sigma_1^2 + \sigma_2^2\right)$$

I. Распределение «хи—квадрат»

Def 1.4.3. Распределением «хи—квадрат» H_k с k степенями свободы называется распределение суммы k квадратов независимых стандартных нормальных величин.

$$\chi_k^2 = X_1^2 + \ldots + X_k^2 \qquad X_i \in N\left(0;1\right) \text{ независимые}$$

Свойства распределения «хи—квадрат»

Lm 1.4.4.

$$\mathbb{E}\left(\chi_k^2\right) = k$$

□

$$\mathbb{D}(X_1) = \mathbb{E}\left(X_1^2\right) - \left(\mathbb{E}(X_1)\right)^2 = \mathbb{E}\left(X_1^2\right) = 1$$
$$\mathbb{E}\left(\chi_k^2\right) = \mathbb{E}\left(X_1^2 + \ldots + X_k^2\right) = k\mathbb{E}\left(X_1^2\right) = k$$

■

Замечание 1.4.5. Распределение «хи-квадрат» устойчиво относительно суммирования.

$$\left.\begin{array}{l} X_1 \in \chi_n^2 \\ X_2 \in \chi_m^2 \\ X_1, X_2 \text{ независимы} \end{array}\right\} \Longrightarrow X_1 + X_2 \in \chi_{n+m}^2$$

II. Распределение Стьюдента

Def 1.4.6. Пусть случайные величины $X_0, \ldots, X_k$ независимы и имеют стандартное нормальное распределение. Распределением Стьюдента T_k с k степенями свободы называется распределение случайной величины

$$t_k = \frac{X_0}{\sqrt{\frac{1}{k}\left(X_1^2 + \ldots + X_k^2\right)}} = \frac{X_0}{\sqrt{\frac{1}{k}\chi_k^2}}$$

Свойства распределения Стьюдента

Замечание 1.4.7.

$$\mathbb{E}(t_k) = 0$$

Lm 1.4.8.

$$t_k \rightrightarrows N(0;1)$$

□

$$\frac{\chi_k^2}{k} \xrightarrow{\mathsf{P}} \mathbb{E}\left(X_1^2\right) = 1 \Longrightarrow t_k = \frac{X_0}{\sqrt{\frac{1}{k}\chi_k^2}} \xrightarrow{\mathsf{P}} X_0$$

■

III. Распределение Фишера—Снедекора (F—распределение)

Def 1.4.9. Распределение Фишера—Снедекора $F_{m,n}$ с m и n степенями свободы называется случайной величины

$$f_{m,n} = \frac{n\chi_m^2}{m\chi_n^2}$$

Замечание 1.4.10.

$$\mathbb{E}(f_{m,n}) = \frac{n}{n-2}$$

Преобразование нормальных выборок

Пусть $\vec{X} = (X_1, \ldots, X_n)^T$ — выборка из $N(0;1)$, т.е. $X_i \in N(0;1)$ — независимы. A это невырожденная матрица порядка n. Рассмотрим $\vec{Y} = A\vec{X}$, где $Y_i = a_{i,1}X_1 + \ldots + a_{i,n}X_n$. По свойствам нормального распределения эти компоненты будут нормальными случайными величинами, но в общем случае зависимыми. Нас в основном интересует случай, когда A — ортогональная матрица.

Многомерное нормальное распределение

Def 1.4.11. Пусть случайный вектор $\vec{\xi} = (\xi_1, \ldots, \xi_n)$ имеет вектор средних $\mathbb{E}\left(\vec{\xi}\right) = \vec{a} = \left(\mathbb{E}(\xi_1), \ldots, \mathbb{E}(\xi_n)\right)^T$, а K — симметричная положительно определенная матрица. Вектор $\vec{\xi}$ имеет многомерное нормальное распределение в $\mathbb{R}^n$ с параметрами $\vec{a}$ и K, если его плотность имеет вид

$$f_{\vec{a},K}\left(\vec{x}\right) = \frac{1}{\left(\sqrt{2\pi}\right)^n \sqrt{\det K}} \exp\left(-\frac{1}{2}\left(\vec{x} - \vec{a}\right)^T K^{-1}\left(\vec{x} - \vec{a}\right)\right)$$

Многомерное нормальное распределение обозначается как $\vec{\xi} \in N\left(\vec{a}; K\right)$.

Замечание 1.4.12. Матрица $K = \mathbb{D}\left(\vec{\xi}\right) = (\operatorname{cov}(\xi_i, \xi_j))$ — матрица ковариаций.

Lm 1.4.13. Если $K = E$ и $\vec{a} = \vec{0}$, то имеем вектор из независимых стандартных нормальных случайных величин.

□ Запишем плотность с данными параметрами

$$f_\xi\left(\vec{x}\right) = \frac{1}{\left(\sqrt{2\pi}\right)^n}\exp\left(-\frac{1}{2}\left(\vec{x}\right)^T\vec{x}\right) = \frac{1}{\left(\sqrt{2\pi}\right)^n}\exp\left(-\frac{1}{2}\left(x_1^2+\ldots+x_n^2\right)\right) = f_{\xi_1}(x_1)\cdot\ldots\cdot f_{\xi_n}(x_n)$$

где $f_{\xi_i}(x_i)$ — плотность стандартного нормального распределения. Заметим, что случайные величины ξ_i независимы, т.к. плотность совместного распределения равна произведению частных плотностей. ■

Замечание 1.4.14. Пусть $\vec{X}$ состоит из независимых нормальных стандартных случайных величин, B — невырожденная матрица порядка n. Тогда $\vec{Y} = B\vec{X} + \vec{a}$ имеет многомерное нормальное распределение с параметрами $\vec{a}$ и $K = BB^T$. Это эквивалентное определение многомерных нормальных распределений — все они получаются таким образом.

Замечание 1.4.15. Случайный вектор имеет многомерное нормальное распределение, если всего его компоненты это нормальные случайные величины и нет функциональной зависимости одной компоненты от остальных.

Lm 1.4.16. Пусть случайный вектор $\vec{X}$ состоит из независимых стандартных нормальных случайных величин, а C это ортогональная матрица. Тогда $\vec{Y} = C\vec{X}$ также состоит из независимых стандартных нормальных случайных величин.

□ Т.к. C — ортогональная матрица, то $C^{-1} = C^T$, поэтому $K = CC^{-1} = E$. По лемме 1.4.13 получаем, что $\vec{Y}$ состоит из независимых стандартных нормальных случайных величин. ■

Lm 1.4.17. Пусть случайный вектор $\vec{\xi}$ имеет многомерное нормальное распределение с параметрами $\vec{a}$ и K. Тогда $\vec{\eta} = B^{-1}\left(\vec{\xi} - \vec{a}\right)$, где $B = \sqrt{K}$, состоит из независимых стандартных нормальных случайных величин.

Lm 1.4.18. Пусть случайный вектор $\vec{\xi}$ имеет многомерное нормальное распределение с параметрами $\vec{a}$ и K. Координаты вектора $\vec{\xi}$ независимы тогда и только тогда, когда они не коррелированны, т.е. матрица ковариаций K диагональная.

Замечание 1.4.19. Если плотность совместного распределения нормальных случайных величин ξ и η ненулевая, то они независимы тогда и только тогда, когда их коэффициент корреляции равен нулю.

Теорема 1.4.20. (Многомерная центральная предельная теорема) Среднее арифметическое независимых одинаково распределенных случайных векторов слабо сходится к многомерному нормальному распределению.

Лемма Фишера

Lm 1.4.21 (Фишера)**.** Пусть $\vec{X}$ состоит из независимых нормальных стандартных величин, а $\vec{Y} = C\vec{X}$, где C — ортогональная матрица. Тогда случайная величина

$$T\left(\vec{X}\right) = \sum_{i=1}^{n} X_i^2 - \sum_{i=1}^{k} Y_i^2 \qquad \forall 1 \leqslant k \leqslant n-1$$

не зависит от случайных величин $Y_1, \ldots, Y_k$ и имеет распределение H_{n-k}.

□ Т.к. C ортогональная матрица, то $\left\|\vec{X}\right\| = \left\|\vec{Y}\right\|$, значит

$$\sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} Y_i^2 \Longrightarrow T\left(\vec{X}\right) = \sum_{i=1}^{n} Y_i^2 - \sum_{i=1}^{k} Y_i^2 = \sum_{i=k+1}^{n} Y_i^2$$

Таким образом $T\left(\vec{X}\right) \in H_{n-k}$, т.к. по 1.4.16 случайные величины Y_i независимы и имеют стандартное нормальное распределение. ■

Основная теорема

Замечание 1.4.22. Название неофициальное и актуально только в рамках курса, т.к. мы будем часто ссылаться на эту теорему.

Пусть имеется выборка $\vec{X} = (X_1, \ldots, X_n)$ из $N\left(a;\sigma^2\right)$. Обозначим $\overline{x}$ — выборочное среднее, а S^2 — исправленную выборочную дисперсию. Тогда имеют место следующие утверждения.

Lm 1.4.23.

$$\sqrt{n}\cdot\frac{\overline{x}-a}{\sigma} \in N(0;1)$$

□

$$X \in N\left(a;\sigma^2\right) \Longrightarrow \sum_{i=1}^{n} X_i \in N\left(na;n\sigma^2\right) \Longrightarrow \overline{x} \in N\left(a;\frac{\sigma^2}{n}\right) \Longrightarrow \overline{x}-a \in N\left(0;\frac{\sigma^2}{n}\right) \Longrightarrow \sqrt{n}\cdot\frac{\overline{x}-a}{\sigma} \in N(0;1)$$

■

Lm 1.4.24.

$$\sum_{i=1}^{n}\left(\frac{x_i-a}{\sigma}\right)^2 \in H_n$$

□

$$\forall i \left| \frac{x_i-a}{\sigma} \in N(0;1) \Longrightarrow \sum_{i=1}^{n}\left(\frac{x_i-a}{\sigma}\right)^2 \in H_n \text{ по определению}\right.$$

■

Lm 1.4.25.

$$\sum_{i=1}^{n}\left(\frac{x_i-\overline{x}}{\sigma}\right)^2 = \frac{(n-1)S^2}{\sigma^2} \in H_{n-1}$$

□

$$\sum_{i=1}^{n}\left(\frac{x_i-\overline{x}}{\sigma}\right)^2 = \sum_{i=1}^{n}\left(\frac{x_i-a}{\sigma}-\frac{\overline{x}-a}{\sigma}\right)^2 = \sum_{i=1}^{n}(z_i-\overline{z})^2$$

где

$$z_i = \frac{x_i-a}{\sigma} \in N(0;1)$$

$$\overline{z} = \frac{z_1+\ldots+z_n}{n} = \frac{1}{n}\sum_{i=1}^{n}\frac{x_i-a}{\sigma} = \frac{\sum x_i - na}{n\sigma} = \frac{n\overline{x}-na}{n\sigma} = \frac{\overline{x}-a}{\sigma}$$

Применим 1.4.21. Сначала рассмотрим

$$T\left(\vec{z}\right) = \sum_{i=1}^{n}(z_i-\overline{z})^2 = n\mathbb{D}^* = \sum_{i=1}^{n}z_i^2 - n(\overline{z})^2 = \sum_{i=1}^{n}z_i^2 - \left(\sqrt{n}\overline{z}\right)^2$$

Рассмотрим второе слагаемое.

$$\sqrt{n}\cdot\overline{z} = \frac{z_1+\ldots+z_n}{\sqrt{n}} = \frac{1}{\sqrt{n}}z_1+\ldots+\frac{1}{\sqrt{n}}z_n$$

Строка $\left(\frac{1}{\sqrt{n}},\ldots,\frac{1}{\sqrt{n}}\right)$ имеет единичную длину, поэтому, как известно из курса линейной алгебры, мы можем ее дополнить до ортогональной матрицы C. Тогда $\sqrt{n}\cdot\overline{z} = Y_1$ — первая компонента вектора $\vec{Y} = C\vec{Z}$. Отсюда по 1.4.21 имеем

$$T\left(\vec{z}\right) = \sum_{i=1}^{n}(z_i-\overline{z})^2 \in H_{n-1}$$

Причем он не зависит от $Y_1 = \sqrt{n}\cdot\overline{z}$. ■

Lm 1.4.26.

$$\sqrt{n}\cdot\frac{\overline{x}-a}{S} \in T_{n-1}$$

□

$$\sqrt{n}\cdot\frac{\overline{x}-a}{S} = \sqrt{n}\cdot\frac{\overline{x}-a}{\sigma}\cdot\frac{1}{\sqrt{\frac{(n-1)S^2}{\sigma^2}\cdot\frac{1}{n-1}}} = \frac{X_0}{\sqrt{\frac{\chi^2_{n-1}}{n-1}}}$$

где $X_0 \in N(0;1)$ по 1.4.23, а $\frac{(n-1)S^2}{\sigma^2} \in H_{n-1}$ по 1.4.25. Итого по определению получаем распределение Стьюдента с $n-1$ степенями свободы. Стоит отметить, что в определении распределения Стьюдента числитель и знаменатель должны быть независимы. У нас с этим проблем нет, т.к. $\overline{x}$ и S^2 независимы согласно 1.4.27. ■

Lm 1.4.27.

$$\overline{x}, S^2 \text{ независимые случайные величины}$$

□ В 1.4.25 показали, что

$$\sum_{i=1}^{n}\left(\frac{x_i-\overline{x}}{\sigma}\right)^2 = \frac{(n-1)S^2}{\sigma^2} \text{ не зависит от } \sqrt{n}\cdot\overline{x}$$

Умножение на константу не влияет на независимость, поэтому получаем, что S^2 не зависит от $\overline{x}$. ■

1.5. Лекция 24.03.07.

Квантили распределений

Предполагаем, что распределение абсолютно непрерывно и $F(X)$ — его функция распределения.

Def 1.5.1 (A)**.** Число t_γ называется квантилем распределения уровня γ, если $F(t_\gamma) = \gamma$. Таким образом $t_\gamma = F^{-1}(\gamma)$.

Замечание 1.5.2. Медиана это квантиль уровня $\frac{1}{2}$.

TODO: image 4:30
В ряде источников под квантилем понимается несколько иное.

Def 1.5.3 (B)**.** Число t_α называется квантилем уровня значимости α, если $\mathsf{P}(x > t_\alpha) = \alpha$ или $F(t_\alpha) = 1 - \alpha$. Таким образом t_α соответствует вероятности попадания величины в правостороннюю область.

Замечание 1.5.4. Будем использовать оба определения, но во втором случае добавлять «уровня значимости» и использовать символ α, а не γ.

Квантили основных распределений в Excel

1. `НОРМ.СТ.ОБР` — обратная функция к функции стандартного нормального распределения.

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{z^2}{2}\right) \mathrm{d}z$$

   Тогда `НОРМ.СТ.ОБР` $(x + 0.5)$ это обратная функция к функции Лапласа $\Phi(x)$.

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{0}^{x} \exp\left(-\frac{z^2}{2}\right) \mathrm{d}z$$

2. (a) `СТЬЮДЕНТ.ОБР` (γ, n) — функция, обратная к функции распределения Стьюдента с n степенями свободы.

$$t_n = \frac{X_0}{\sqrt{\frac{1}{n}\chi_n^2}}$$

   Заметим, что это распределение симметричное относительно нулю, поэтому можно использовать другой способ.

   (b) `СТЬЮДЕНТ.ОБР.2Х` — так называемое двустороннее обратное преобразование Стьюдента, которое возвращает t_α такое, что $\mathsf{P}(|X| > t_\alpha) = \alpha$.
   Таким образом $\mathsf{P}(|X| < t_\alpha) = 1 - \alpha$, значит $t_\gamma =$ `СТЬЮДЕНТ.ОБР.2Х` $(1 - \gamma, n)$.

3. (a) `ХИ2.ОБР` — возвращает квантиль t_γ для распределения H_n с n степенями свободы.
   (b) `ХИ2.ОБР.ПХ` — возвращает квантиль t_α. Заметим, что `ХИ2.ОБР` $(\gamma, n) =$ `ХИ2.ОБР.ПХ` $(1 - \gamma, n)$.

4. (a) `F.ОБР` — возвращает t_γ для F-распределения.
   (b) `F.ОБР.ПХ` — возвращает t_α. Заметим, что `F.ОБР` $(\gamma, n, m) =$ `F.ОБР.ПХ` $(1 - \gamma, n, m)$.

Интервальные оценки

Def 1.5.5. Интервал $\left(\Theta_\gamma^-; \Theta_\gamma^+\right)$ называется доверительным интервалом для параметра Θ уровня надежности γ если

$$\Theta \in \left(\Theta_\gamma^-; \Theta_\gamma^+\right) \text{ с вероятностью } \gamma \qquad \text{или} \qquad \mathsf{P}\left(\Theta_\gamma^- < \Theta < \Theta_\gamma^+\right) = \gamma$$

Замечание 1.5.6. В случае дискретного распределения более точным будет определение

$$\mathsf{P}\left(\Theta_\gamma^- < \Theta < \Theta_\gamma^+\right) \geqslant \gamma$$

Замечание 1.5.7. Неверно говорить, что Θ лежит в данном интервале с вероятностью γ, т.к. параметр Θ не является случайной величиной. Правильнее говорить, что доверительный интервал накрывает параметр Θ с вероятностью γ.

Замечание 1.5.8. Вероятность $\alpha = 1 - \gamma$ называется уровнем значимости доверительного интервала.

Замечание 1.5.9. Обычно пытаемся получить доверительный интервал симметричный относительно несмещенной точной оценки Θ^*, но не всегда это возможно.

Замечание 1.5.10. Стандартные уровни надежности 0.9, 0.95, 0.99 и 0.999. Наиболее распространенный 0.95.

Lm 1.5.11. Если случайная величина ξ имеет симметричное распределение относительно точки $x = 0$, то $\mathsf{P}(|\xi| < t) = 2F(t) - 1$.

□ Под симметричным распределением подразумевается распределение с симметричным графиком плотности. Тогда

$$\mathsf{P}\left(|\xi| < t\right) = 2\mathsf{P}\left(0 < \xi < t\right) = 2\left(F(t) - F(0)\right) = 2F(t) - 1$$

т.к. $F(0) = 0.5$ в силу симметрии. ■

Доверительные интервалы для параметров нормального распределения

Пусть $\vec{X} = (X_1, \ldots, X_n)$ — выборка объема n из $N\left(a; \sigma^2\right)$. Необходимо построить доверительные интервалы. Возможны 4 ситуации.

I. Доверительный интервал для параметра a при известном значении параметра σ^2

По 1.4.23 имеем

$$\sqrt{n} \cdot \frac{\overline{x} - a}{\sigma} \in N\left(0; 1\right)$$

Будем искать интервал в виде

$$\mathsf{P}\left(-t_\gamma < \sqrt{n} \cdot \frac{\overline{x} - a}{\sigma} < t_\gamma\right) = \mathsf{P}\left(\left|\sqrt{n} \cdot \frac{\overline{x} - a}{\sigma}\right| < t_\gamma\right) = 2\Phi\left(t_\gamma\right) = \gamma \Longrightarrow \Phi\left(t_\gamma\right) = \frac{\gamma}{2}$$

Таким образом t_γ — обратное значение функции Лапласа или квантиль уровня $\frac{1+\gamma}{2}$ стандартного нормального распределения. Остается решить изначальное неравенство относительно параметра a, получаем

$$\begin{gathered} -t_\gamma < \sqrt{n} \cdot \frac{\overline{x} - a}{\sigma} < t_\gamma \\ -t_\gamma \frac{\sigma}{\sqrt{n}} < \overline{x} - a < t_\gamma \frac{\sigma}{\sqrt{n}} \\ \overline{x} - t_\gamma \frac{\sigma}{\sqrt{n}} < a < \overline{x} + t_\gamma \frac{\sigma}{\sqrt{n}} \end{gathered}$$

Итак, получили доверительный интервал для параметра a надежности γ вида $\left(\overline{x} - t_\gamma \frac{\sigma}{\sqrt{n}}; \overline{x} + t_\gamma \frac{\sigma}{\sqrt{n}}\right)$, где t_γ находится из условия $\Phi\left(t_\gamma\right) = \frac{\gamma}{2}$.

II. Доверительный интервал для параметра a при неизвестном значении параметра σ^2

По 1.4.26 имеем

$$\sqrt{n} \cdot \frac{\overline{x} - a}{S} \in T_{n-1}$$

Проведем аналогичные рассуждения, что и в прошлом пункте.

$$\mathsf{P}\left(-t_\gamma < \sqrt{n} \cdot \frac{\overline{x} - a}{S} < t_\gamma\right) = \mathsf{P}\left(\left|\sqrt{n} \cdot \frac{\overline{x} - a}{S}\right| < t_\gamma\right) = 2F_{T_{n-1}}\left(t_\gamma\right) - 1 = \gamma \Longrightarrow F_{T_{n-1}}\left(t_\gamma\right) = \frac{1+\gamma}{2}$$

Таким образом t_γ это квантиль распределения Стьюдента уровня $\frac{1+\gamma}{2}$. Решаем неравенство и получаем

$$\overline{x} - t_\gamma \frac{S}{\sqrt{n}} < a < \overline{x} + t_\gamma \frac{S}{\sqrt{n}}$$

Замечание 1.5.12. Заметим, что в обоих случаях получили симметричные интервалы относительно точечной оценки $\overline{x}$.

Замечание 1.5.13. В Excel квантиль t_γ удобнее находить как $t_\gamma =$ `СТЬЮДЕНТ.ОБ.2Х` $(1 - \gamma, n - 1)$.

III. Доверительный интервал для параметра σ^2 при неизвестном значении параметра a

По 1.4.25 имеем

$$\frac{(n-1)S^2}{\sigma^2} \in H_{n-1}$$

Пусть χ_1^2 и χ_2^2 это квантили распределения H_{n-1} уровней $\frac{1-\gamma}{2}$ и $\frac{1+\gamma}{2}$. Тогда

$$\mathsf{P}\left(\chi_1^2 < \frac{(n-1)S^2}{\sigma^2} < \chi_2^2\right) = F_{H_{n-1}}\left(\chi_2^2\right) - F_{H_{n-1}}\left(\chi_1^2\right) = \frac{1+\gamma}{2} - \frac{1-\gamma}{2} = \gamma$$

Решим неравенство относительно σ^2.

$$\chi_1^2 < \frac{(n-1)S^2}{\sigma^2} < \chi_2^2$$
$$\frac{1}{\chi_2^2} < \frac{\sigma^2}{(n-1)S^2} < \frac{1}{\chi_1^2}$$
$$\frac{(n-1)S^2}{\chi_2^2} < \sigma^2 < \frac{(n-1)S^2}{\chi_1^2}$$

Получили доверительный интервал $\left(\frac{(n-1)S^2}{\chi_2^2}; \frac{(n-1)S^2}{\chi_1^2}\right)$, где χ_1^2 и χ_2^2 это квантили распределения H_{n-1} уровней $\frac{1-\gamma}{2}$ и $\frac{1+\gamma}{2}$.

Замечание 1.5.14. Получили доверительный интервал для среднего квадратического отклонения $\left(\frac{S\sqrt{n-1}}{\chi_2}; \frac{S\sqrt{n-1}}{\chi_1}\right)$.

IV. Доверительный интервал для параметра σ^2 при известном значении параметра a

По 1.4.24 имеем

$$\sum_{i=1}^{n}\left(\frac{x_i - a}{\sigma}\right)^2 = \frac{n\left(\sigma^2\right)^*}{\sigma^2} \in H_n$$

где $\left(\sigma^2\right)^* = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - a\right)^2$. Пусть χ_1^2 и χ_2^2 это квантили распределения H_n уровней $\frac{1-\gamma}{2}$ и $\frac{1+\gamma}{2}$. Тогда

$$\mathsf{P}\left(\chi_1^2 < \frac{n\left(\sigma^2\right)^*}{\sigma^2} < \chi_2^2\right) = F_{H_n}\left(\chi_2^2\right) - F_{H_n}\left(\chi_1^2\right) = \frac{1+\gamma}{2} - \frac{1-\gamma}{2} = \gamma$$

Аналогично третьему пункту решим неравенство и получим

$$\frac{n\left(\sigma^2\right)^*}{\chi_2^2} < \sigma^2 < \frac{n\left(\sigma^2\right)^*}{\chi_1^2}$$

Итак, доверительный для σ^2 имеет вид $\left(\frac{n\left(\sigma^2\right)^*}{\chi_2^2}; \frac{n\left(\sigma^2\right)^*}{\chi_1^2}\right)$, где χ_1^2 и χ_2^2 это квантили распределения H_n уровней $\frac{1-\gamma}{2}$ и $\frac{1+\gamma}{2}$.

Замечание 1.5.15. Получили доверительный интервал для σ вида $\left(\frac{\sqrt{n\left(\sigma^2\right)^*}}{\chi_2}; \frac{\sqrt{n\left(\sigma^2\right)^*}}{\chi_1}\right)$.

Замечание 1.5.16. Доверительные интервалы полученные в третьем и четвертом пунктах не являются симметричными относительно S^2, но есть формулы для симметричных доверительных интервалов.

Пример 1.5.17. Пусть $X \in N\left(a; \sigma^2\right)$ причем известно, что $\sigma = 3$. В результате обработки выборки объема $n = 36$ получили $\overline{x} = 4.1$. Найти доверительный интервал для параметра a надежности $\gamma = 0.95$.
Решение: Найдем t_γ.

$$2\Phi\left(t_\gamma\right) = 0.95 \Longrightarrow \Phi\left(t_\gamma\right) = 0.475 \Longrightarrow t_\gamma = 1.96$$

Применим формулу из первого пункта.

$$\begin{aligned} \overline{x} - t_\gamma\frac{S}{\sqrt{n}} < a < & \quad \overline{x} + t_\gamma\frac{S}{\sqrt{n}} \\ 4.1 - 1.96\cdot\frac{3}{\sqrt{36}} < a < & \quad 4.1 + 1.96\cdot\frac{3}{\sqrt{36}} \\ 3.12 < a < & \quad 5.08 \end{aligned}$$

Пример 1.5.18. Пусть $X \in N\left(a; \sigma^2\right)$. В результате обработки выборки объема $n = 25$ получили $\overline{x} = 42.32$ и $S = 6.4$. Найти доверительный интервал для параметра a при уровне надежности $\gamma = 0.95$.
Решение: Найдем t_γ. Получим `СТЬЮДЕНТ.ОБР.2Х` $(1 - 0.95, 25 - 1) = 2.064$. Подставим в формулу из второго пункта.

$$\begin{aligned} \overline{x} - t_\gamma\frac{S}{\sqrt{n}} < a < & \quad \overline{x} + t_\gamma\frac{S}{\sqrt{n}} \\ 42.32 - 2.064\cdot\frac{6.4}{\sqrt{25}} < a < & \quad 42.32 + 2.064\cdot\frac{6.4}{\sqrt{25}} \\ 39.678 < a < & \quad 44.962 \end{aligned}$$

1.6. Лекция 24.03.14.

Проверка статистических гипотез

Def 1.6.1. Гипотезой H называется предположение о распределении случайной величины.

Def 1.6.2. Гипотеза называется простой, если она однозначно определяет распределение, т.е. $H\colon \mathcal{F} \in \mathcal{F}_1$, где $\mathcal{F}_1$ — распределение известного типа с известными параметрами. Все остальные гипотезы называются сложными. Они состоят из конечного или бесконечного числа простых гипотез.

Будем изучать лишь самую простую схему из двух гипотез, где H_0 — это основная (нулевая) гипотеза, а H_1 это альтернативная (конкурирующая) гипотеза, состоящая в том, что основная гипотеза неверна. Стоит отметить, что иногда рассматриваются более сложные схемы, состоящие из набора нескольких гипотез.
Основная гипотеза H_0 принимается или отклоняется при помощи статистики критерия K, где $K(X_1, \dots, X_n) \to \mathbb{R} = \overline{S} \cup S$. Принимаем гипотезу H_0, если статистика критерия попала в область $\overline{S}$, в противном случае принимаем гипотезу H_1.

Def 1.6.3. Область S называется критической областью критерия.

Def 1.6.4. Точки $t_{\text{кр}}$ на границе двух областей называются критическими.

Def 1.6.5. Ошибка первого рода состоит в том, что нулевая гипотеза H_0 отклоняется, хотя она верна.

Def 1.6.6. Ошибка второго рода состоит в том, что отклоняется альтернативная гипотеза H_1, хотя она верна.

Def 1.6.7. Вероятность α ошибки первого рода называется уровнем значимости критерия.

Def 1.6.8. Вероятность ошибки второго рода обозначается β. Мощностью критерия называется вероятность $1 - \beta$, т.е. вероятность не допущения ошибки второго рода.

Замечание 1.6.9. Заметим, что вероятности α и β не связаны каким-либо соотношением.

Замечание 1.6.10. Ясно, что критерий будет тем лучше, чем меньше вероятности ошибок α и β. При увеличении объема выборки обе эти вероятности уменьшаются, однако при фиксированном объеме выборки попытки уменьшить одну вероятность ведут к увеличению другой.

Способы сравнения критериев

Пусть имеется два критерия K_1 и K_2 с соответствующими вероятностями ошибок α_1, β_1 и α_2, β_2. Объем выборки n фиксированный.

I. Минимаксный подход

Критерий K_1 не хуже, чем K_2, если $\max(\alpha_1, \beta_1) < \max(\alpha_2, \beta_2)$. Из всех критериев выбираем тот, который не хуже остальных.

II. Байесовский подход

Пусть известны потери h_1 и h_2 от ошибок первого и второго рода. Тогда средние ожидаемые потери составят $u = \alpha h_1 + \beta h_2$. Из всех критериев выбираем тот, у которого средние ожидаемые потери наименьшие.

III. Выбор наиболее мощного критерия

Обозначим $K_\varepsilon = \left\{ K_i \,\middle|\, \alpha \leqslant \varepsilon \right\}$, т.е. класс критериев данной выборки.

Def 1.6.11. Критерий $K \in K_\varepsilon$ называется наиболее мощным критерием уровня ε, если $\beta \leqslant \beta_i \,\middle|\, \forall K_i \in K_\varepsilon$.

Построение критериев согласия

Def 1.6.12. Говорят, что критерий K является критерием асимптотического уровня ε, если $\alpha \to \varepsilon$ при $n \to \infty$.

Def 1.6.13. Критерий K для проверки гипотезы H_0 против альтернативной гипотезы H_1 называется состоятельным, если $\beta \to 0$ при $n \to \infty$.

Def 1.6.14. Критерием согласия уровня ε называется состоятельный критерий асимптотического уровня ε.

В качестве критерия согласия берется статистика $K(X_1, \dots, X_n)$ со свойствами

1. Если H_0 верна, то $K(X_1, \dots, X_n) \rightrightarrows z$, где z — случайная величина с известным распределением.
2. Если H_0 неверна, то $K(X_1, \dots, X_n) \xrightarrow{\mathsf{P}} \infty$ при $n \to \infty$.

Для заданного уровня значимости ε находим квантиль $t_{\text{кр}}$ такой, что $\mathsf{P}(|z| \geqslant t_{\text{кр}}) = \varepsilon$. В результате получаем критерий с уровнем значимости α. Итого

$$\begin{cases} H_0, & \text{если } |K| < t_{\text{кр}} \\ H_1, & \text{если } |K| \geqslant t_{\text{кр}} \end{cases}$$

Замечание 1.6.15. Таким образом в качестве статистики берется функция отклонения эмпирического распределения от теоретического. Если нулевая гипотеза верна, то эта функция сходится к собственному распределению, если нет, то неограниченно возрастает.

Lm 1.6.16. Построенный критерий обладает свойствами

1. Это критерий асимптотического уровня ε.
2. Этот критерий состоятельный.

□ Пусть гипотеза H_0 верна. Тогда по первому свойству выбранной статистики имеем $\forall x \,\Big|\, F_K(x) \to F_z(x)$ при $n \to \infty$. Значит

$$\alpha = \mathsf{P}\left(|K| \geqslant t_{\text{кр}} \,\middle|\, H_0\right) = 1 - \left(F_K\left(t_{\text{кр}}\right) - F_K\left(-t_{\text{кр}}\right)\right) \xrightarrow{n\to\infty} 1 - \left(F_z\left(t_{\text{кр}}\right) - F_z\left(-t_{\text{кр}}\right)\right) = \mathsf{P}\left(|z| \geqslant t_{\text{кр}}\right) = \varepsilon$$

Далее покажем состоятельность критерия. По второму свойству выбранной статистики получаем, что если H_1 верна, то $|K| \xrightarrow{\mathsf{P}} \infty$, т.е. $\forall c \in \mathbb{R} \,\Big|\, \mathsf{P}(|K| \geqslant c) \to 1$, значит $\mathsf{P}(|K| < c) \to 0$. ■

Гипотеза о среднем нормальной совокупности с известной дисперсией

Пусть имеется выборка $\vec{X} = (X_1, \ldots, X_n)$ объема n случайной величины $X \in N\left(a; \sigma^2\right)$, где σ^2 это известное значение. Проверяется гипотеза H_0, которая состоит в том, что $a = a_0$. Альтернативная гипотеза H_1 состоит в том, что $a \neq a_0$.
В качестве статистики критерия возьмем функцию $K = \sqrt{n} \cdot \dfrac{\overline{x} - a_0}{\sigma}$. Проверим, что она имеет требуемые свойства.

1. Если H_0 верна, то

$$K = \sqrt{n} \cdot \frac{\overline{x} - a_0}{\sigma} = \sqrt{n} \cdot \frac{\overline{x} - a}{\sigma} \in N(0; 1)$$

по 1.4.23.

2. Если H_0 неверна, т.е. $a \neq a_0$, то

$$|K| = \left|\sqrt{n} \cdot \frac{\overline{x} - a_0}{\sigma}\right| = \left|\sqrt{n} \cdot \frac{\overline{x} - a}{\sigma} + \sqrt{n} \cdot \frac{a - a_0}{\sigma}\right|$$

Первое слагаемое по 1.4.23 имеет стандартное нормальное распределение, поэтому оно будет ограничено по вероятности. Второе слагаемое будет стремиться к бесконечности, т.к. $\sqrt{n} \to \infty$ при $n \to \infty$, а остальная часть этого слагаемого это какая ненулевая константа в силу того, что $a \neq a_0$. Итого $|K| \xrightarrow{\mathsf{P}} \infty$ при $n \to \infty$.

Итак, получили следующий критерий: для уровня значимости α выберем $t_{\text{кр}}$ такую, что

$$\alpha = \mathsf{P}(|K| \geqslant t_{\text{кр}}) = 1 - 2\Phi\left(t_{\text{кр}}\right) \Longrightarrow \Phi\left(t_{\text{кр}}\right) = \frac{1 - \alpha}{2}$$

Значит $t_{\text{кр}}$ это обратное значение функции Лапласа в точке $\frac{1-\alpha}{2}$ или квантиль уровня $1 - \frac{\alpha}{2}$ стандартного нормального распределения. Получаем критерий согласия

$$\begin{cases} H_0, & |K| < t_{\text{кр}} \\ H_1, & |K| \geqslant t_{\text{кр}} \end{cases}$$

Гипотеза о среднем нормальной совокупности при неизвестной дисперсии

Пусть имеется выборка $\vec{X} = (X_1, \ldots, X_n)$ объема n случайной величины $X \in N\left(a; \sigma^2\right)$. Проверяется гипотеза H_0, которая состоит в том, что $a = a_0$. Альтернативная гипотеза H_1 состоит в том, что $a \neq a_0$.
В качестве статистики критерия возьмем функцию $K = \sqrt{n} \cdot \dfrac{\overline{x} - a_0}{S}$. Проверим, что она имеет требуемые свойства.

1. Если H_0 верна, то

$$K = \sqrt{n} \cdot \frac{\overline{x} - a_0}{S} = \sqrt{n} \cdot \frac{\overline{x} - a}{S} \in T_{n-1}$$

по 1.4.26.

2. Если H_0 неверна, т.е. $a \neq a_0$, то

$$|K| = \left|\sqrt{n} \cdot \frac{\overline{x} - a_0}{S}\right| = \left|\sqrt{n} \cdot \frac{\overline{x} - a}{S} + \sqrt{n} \cdot \frac{a - a_0}{S}\right|$$

Аналогично предыдущей гипотезе получаем, что $|K| \xrightarrow{\mathsf{P}} \infty$ при $n \to \infty$.

Итак, получили следующий критерий: для уровня значимости α выберем $t_{\text{кр}}$ такую, что $\mathsf{P}(|t_{n-1}| \geqslant t_{\text{кр}}) = \alpha$. Таким образом $t_{\text{кр}}$ это квантиль уровня $1 - \frac{\alpha}{2}$ распределения T_{n-1} или $t_{\text{кр}}$ это квантиль уровня значимости α двустороннего распределения $|T_{n-1}|$. Получаем критерий согласия

$$\begin{cases} H_0, & |K| < t_{\text{кр}} \\ H_1, & |K| \geqslant t_{\text{кр}} \end{cases}$$

Доверительные интервалы как критерии гипотез по параметрам распределения

Пусть имеется выборка $\vec{X} = (X_1, \ldots, X_n)$ объема n из $X \in \mathcal{F}_\Theta$, где $\mathcal{F}_\Theta$ это распределение известного типа, с неизвестным параметром Θ. Проверяется гипотеза H_0 о том, что $\Theta = \Theta_0$, против альтернативной гипотезы $\Theta \neq \Theta_0$. Допустим, что для параметра Θ построен доверительный интервал $\left(\Theta_\gamma^-; \Theta_\gamma^+\right)$ надежности γ.

Lm 1.6.17. В описанных условиях получаем критерий

$$\begin{cases} H_0, & \Theta_0 \in \left(\Theta_\gamma^-; \Theta_\gamma^+\right) \\ H_1, & \Theta_0 \notin \left(\Theta_\gamma^-; \Theta_\gamma^+\right) \end{cases}$$

уровня $\alpha = 1 - \gamma$.

□

$$\alpha = \mathsf{P}\left(\Theta_0 \notin \left(\Theta_\gamma^-; \Theta_\gamma^+\right) \,\middle|\, X \in \mathcal{F}_{\Theta_0}\right) = 1 - \mathsf{P}\left(\Theta_0 \in \left(\Theta_\gamma^-; \Theta_\gamma^+\right) \,\middle|\, X \in \mathcal{F}_{\Theta_0}\right) = 1 - \gamma$$

■

Замечание 1.6.18. Доказать, что данный критерий будет состоятельным, в общем случае нельзя.

Пример 1.6.19. По выборке объема $n = 36$ из нормальной совокупности с известным средним квадратическим отклонением $\sigma = 1.44$ найдено выборочное среднее $\overline{x} = 21.6$. Проверить гипотезу H_0 о том, что $a = 21$, против альтернативной гипотезы H_1 о том, что $a \neq 21$. Уровень значимости принять $\alpha = 0.05$.

Решение: Составляем статистику

$$K = \sqrt{n} \cdot \frac{\overline{x} - a_0}{\sigma} = \sqrt{36} \cdot \frac{21.6 - 21}{1.44} = 2.5$$

Теперь ищем критическую точку.

$$\Phi(t_{\text{кр}}) = \frac{1 - \alpha}{2} = 0.475 \Longrightarrow t_{\text{кр}} = 1.96$$

Т.к. $|K| > t_{\text{кр}}$, то основная гипотеза H_0 отклоняется, и принимается альтернативная гипотеза H_1.

Некоторые дополнения с практики

Теорема 1.6.20. Если $f_\xi(Me) \neq 0$, то $Me^* \xrightarrow{\mathsf{P}} Me$, причем со скоростью $\frac{1}{\sqrt{n}}$. Таким образом выборочная медиана является состоятельной асимптотически нормальной оценкой теоретической медианы.

В случае симметричных распределений выборочную медиану удобно использовать, если распределение имеет «жирные хвосты». В случае «жирных» хвостов возможны так называемые «выбросы», которые заметно повлияют на выборочное среднее, и поэтому его не желательно использовать в качестве оценки центра симметрии. В случае «жирных хвостов» или «выбросов» можно использовать медиану или иные методики.

Недостатки медианы:

1. Хорошо работает только в случае симметричных распределений.
2. В случае нормального распределения медиана сходится приблизительно на 20% медленнее, чем среднее выборочное, и до 40% медленнее в некоторых других случаях.

Подумаем над устранением этих недостатков. Пусть распределение симметрично. Рассмотрим два метода.

I. Метод усеченного среднего

Метод заключается в том, что мы отбрасываем из выборки сверху и снизу по 5—10 значений или по 5% значений.

Пусть дана выборка $\vec{X} = (X_1, \ldots X_n)$ объема n. Допустим, что мы отбросили по k значений сверху и снизу и получили выборку объема $n - 2k$ вида $\left(X_{(k+1)}, \ldots, X_{(n-k)}\right)$. Вычислим среднее выборочное этой выборки $\frac{\sum_{i=k+1}^{n-k} X_i}{n - 2k}$. Это компромисс между средним выборочным и медианой, т.к.

1. При $k = 0$ получаем $\overline{x}$.
2. При $n - 2k = 1$ получаем Me.

II. Метод средних Уолша

Пусть дана выборка $\vec{X} = (X_1, \dots X_n)$ объема n. Для всех пар считаем среднее арифметическое $y_k = \frac{1}{2}(x_i + x_j)$, где $1 \leqslant k \leqslant \frac{1}{2}n(n+1)$. Далее находим медиану полученной новой выборки. Преимущества данного метода:
1. Эффективность по сравнению со среднем выборочным упадет не более, чем на 12%.
2. Сглаживает выбросы.

1.7. Лекция 24.03.21.

Критерии для проверки гипотез о распределении

I. Критерий χ^2 (для проверки параметрической гипотезы)

Пусть дана выборка $\vec{X} = (X_1, \dots, X_n)$ неизвестного распределения $\mathcal{F}$. Проверяется основная (сложная) гипотеза H_0 о том, что $\mathcal{F} \in \mathcal{F}_\Theta$, где $\mathcal{F}_\Theta$ это распределение известного типа, определяемое набором из m неизвестных параметров $\Theta = (\Theta_1, \dots, \Theta_m)$. В качестве альтернативной гипотезы H_1 о том, что $\mathcal{F} \notin \mathcal{F}_\Theta$.
Пусть $\widehat{\Theta} = \left(\widehat{\Theta}_1, \dots, \widehat{\Theta}_m\right)$ — оценка неизвестных параметров $\Theta = (\Theta_1, \dots, \Theta_m)$ методом максимального правдоподобия.
Разобьем выборку на k интервалов $A_i = [a_i; a_{i+1})$, обозначим n_i соответствующие частоты этих интервалов. Пусть $p_i = F_{\widehat{\Theta}}(a_{i+1}) - F_{\widehat{\Theta}}(a_i)$ это теоретические вероятности попадания случайной величины с распределением $\mathcal{F}_{\widehat{\Theta}}$ в данные интервалы. Тогда $n' = np_i$ это теоретические частоты в эти интервалы.
В качестве статистики критерия берется функция

$$K = \sum_{i=1}^{k} \frac{(n_i - n_i')^2}{n_i'} = \sum_{i=1}^{k} \frac{n_i^2}{n_i'} - n$$

Теорема 1.7.1. (Фишера) Если гипотеза $H_0 \colon \mathcal{F} \in \mathcal{F}_\Theta$ верна, то

$$K = \sum_{i=1}^{k} \frac{(n_i - n_i')^2}{n_i'} \rightrightarrows \chi^2_{k-m-1}$$

где k это число интервалов, m — число параметров распределения.

Получили следующий критерий: для заданного уровня значимости α находим $t_{\text{кр}}$ такое, что $\mathsf{P}\left(\chi^2_{k-m-1} \geqslant t_{\text{кр}}\right) = \alpha$. Тогда критерий согласия имеет вид

$$\begin{cases} H_0, & K < t_{\text{кр}} \\ H_1, & K \geqslant t_{\text{кр}} \end{cases}$$

Замечание 1.7.2. Для вычисления квантиля можно использовать формулу Excel `ХИ2.ОБР.ПХ` $(\alpha, k - m - 1)$.

Замечание 1.7.3. Частота каждого интервала должна быть не менее пяти. Если это не так, то сливаем соседние интервалы. Число интервалов желательно брать достаточно большим (чтобы $k - m - 1$ было больше), но при этом стоит помнить об ограничениях на частоту каждого интервала.

Замечание 1.7.4. При этом критерии выборку лучше разбивать на «равнонаполненные» интервалы, т.е. на интервалы содержащие примерно одинаковое число элементов выборки.

Замечание 1.7.5. Желательно, чтобы объем выборки был не менее 50. В противном случае метод работает плохо.

A_i	$[5.2; 7.4)$	$[7.4; 9.6)$	$[9.6; 11.8)$	$[11.8; 14.0)$	$[14.0; 16.2)$	$[16.2; 18.4)$	$[18.4; 20.6)$	$[20.6; 22.8]$	$\sum$
n_i	12	17	14	13	18	14	13	19	120

Таблица 1.7.6: Интервальный ряд для примера 1.7.7

Пример 1.7.7. Имеется выборка в виде вариационного ряда $\vec{X} = (5.2, \dots, 22.8)$ объема $n = 120$. При разбиении ее на $k = 8$ интервалов одинаковой длины получили интервальный ряд 1.7.6. Проверить гипотезу о равномерности этого распределения при уровне значимости $\alpha = 0.05$.
Решение: Итак, проверяем гипотезу $H_0 \colon \mathcal{F} \in U(a; b)$ против альтернативной $H_1 \colon \mathcal{F} \notin U(a; b)$. По методу наибольшего правдоподобия получаем (смещенные) оценки $a^* = 5.2$ и $b^* = 22.8$. Теоретические вероятности случайной величины с распределением $U(5.2; 22.8)$ равны $p_i = \frac{1}{k} = \frac{1}{8}$. Тогда теоретические частоты $n_i' = np_i = 15$. Далее вычисляем статистику

$$\chi^2_{\text{набл}} = \sum_{i=1}^{k} \frac{(n_i - n'_i)^2}{n'_i} = 3.2$$

При уровне значимости $\alpha = 0.05$ и числе степеней свободы $k - m - 1 = 5$ находим $t_{\text{кр}} =$ `ХИ2.ОБР.ПХ` $(0.05, 5) = 11.07$. Таким образом, т.к. $\chi_{\text{набл}} < t_{\text{кр}}$, то принимаем основную гипотезу о равномерном распределении.

II. Критерий χ^2

Постановка задачи отличается от предыдущего пункта тем, что проверяется простая основная гипотеза H_0 о том, что $\mathcal{F} = \mathcal{F}_1$, где $\mathcal{F}_1$ это распределение известного типа с известными параметрами. В качестве статистики берем ту же функцию

$$K = \sum_{i=1}^{k} \frac{(n_i - n'_i)^2}{n'_i}$$

Теорема 1.7.8. (Пирсона) Если $H_0 \colon \mathcal{F} = \mathcal{F}_1$ верна, то

$$K = \sum_{i=1}^{k} \frac{(n_i - n'_i)^2}{n'_i} \rightrightarrows \chi^2_{k-1}$$

В остальном критерий аналогичен предыдущему.

III. Критерий Колмогорова

Пусть имеется выборка $\vec{X} = (X_1, \ldots, X_n)$ объема n распределения $\mathcal{F}$. Проверяется нулевая гипотеза $H_0 \colon \mathcal{F} = \mathcal{F}_1$ против альтернативной $H_1 \colon \mathcal{F} \neq \mathcal{F}_1$, где $\mathcal{F}_1$ это абсолютно непрерывное распределение известного типа с известными параметрами. Обозначим $F(x)$ — функцию распределения $\mathcal{F}_1$, а $F^*(x)$ — выборочную функцию распределения.

Теорема 1.7.9. (Колмогорова) Если гипотеза $H_0 \colon \mathcal{F} = \mathcal{F}_1$ верна, то

$$K = \sqrt{n} \cdot \sup_x |F(x) - F^*(x)| \rightrightarrows \mathcal{K}$$

где $\mathcal{K}$ это распределение Колмогорова с функцией распределения

$$F_{\mathcal{K}}(x) = \sum_{j=-\infty}^{\infty} (-1)^j \exp\left(-2j^2x^2\right)$$

В результате получаем критерий Колмогорова: пусть $t_{\text{кр}}$ это квантиль распределения Колмогорова уровня значимости α, тогда

$$\begin{cases} H_0, & K < t_{\text{кр}} \\ H_1, & K \geqslant t_{\text{кр}} \end{cases}$$

Критерии для проверки однородности

IV. Критерий Колмогорова—Смирнова

Пусть имеются две независимые выборки $\vec{X} = (X_1, \ldots, X_n)$ и $\vec{Y} = (Y_1, \ldots, Y_m)$ неизвестных непрерывных распределений $\mathcal{F}$ и $\mathcal{J}$. Проверяется основная гипотеза H_0 о том, что $\mathcal{F} = \mathcal{J}$ против альтернативной H_1 о том, что $\mathcal{F} \neq \mathcal{J}$. В качестве статистики критерия берется функция

$$K = \sqrt{\frac{nm}{n+m}} \cdot \sup_x |\mathcal{F}^*(x) - \mathcal{J}^*(x)|$$

где $\mathcal{F}^*(x)$ и $\mathcal{J}^*(x)$ это соответствующие выборочные функции распределения.

Теорема 1.7.10. (Колмогорова—Смирнова) Если гипотеза $H_0 \colon \mathcal{F} = \mathcal{J}$ верна, то $K \rightrightarrows \mathcal{K}$.

В результате получаем критерий Колмогорова—Смирнова: пусть $t_{\text{кр}}$ это квантиль распределения Колмогорова уровня значимости α, тогда

$$\begin{cases} H_0, & K < t_{\text{кр}} \\ H_1, & K \geqslant t_{\text{кр}} \end{cases}$$

Замечание 1.7.11. Этот критерий не часто используют. Основные критерии это критерии Фишера и Стьюдента (они описаны ниже). Сначала применяется критерий Фишера, и если он не отвергает основную гипотезу, применяется критерий Стьюдента.

V. Критерий Фишера

Пусть имеются две независимые выборки $\vec{X} = (X_1, \ldots, X_n)$ и $\vec{Y} = (Y_1, \ldots, Y_m)$ нормальных распределений $X \in N\left(a_1; \sigma_1^2\right)$ и $Y \in N\left(a_2; \sigma_2^2\right)$. Проверяется основная гипотеза H_0 о том, то $\sigma_1 = \sigma_2$ против альтернативной H_1 о том, то $\sigma_1 \neq \sigma_2$. В качестве статистики критерия берется функция

$$K = \frac{S_x^2}{S_y^2}$$

где S_x^2 и S_y^2 — соответствующие исправленные выборочные дисперсии, причем $S_x^2 \geqslant S_y^2$.

Теорема 1.7.12. Если гипотеза $H_0\colon \sigma_1 = \sigma_2$ верна, то $K \in F_{n-1,m-1}$.

□ Если H_0 верна, то $\sigma_1 = \sigma_2$, поэтому

$$K = \frac{S_x^2}{S_y^2} = \frac{S_x^2}{\sigma_1^2} \cdot \frac{\sigma_2^2}{S_y^2}$$

По 1.4.25 получаем, что

$$\frac{(n-1)S^2}{\sigma^2} \in H_{n-1} \Longrightarrow K = \frac{\chi_{n-1}^2}{n-1} \cdot \frac{m-1}{\chi_{m-1}^2} \in F_{n-1,m-1} \text{ по определению}$$

■

Получили критерий: пусть $t_{\text{кр}}$ это квантиль распределения $F_{n-1,m-1}$ уровня значимости α, тогда

$$\begin{cases} H_0, & K < t_{\text{кр}} \\ H_1, & K \geqslant t_{\text{кр}} \end{cases}$$

Замечание 1.7.13. Если гипотеза неверна, то статистика K стремится не к бесконечности, а к $\frac{\sigma_1^2}{\sigma_2^2}$.

VI. Критерий Стьюдента

Пусть имеются две независимые выборки $\vec{X} = (X_1, \ldots, X_n)$ и $\vec{Y} = (Y_1, \ldots, Y_m)$ нормальных распределений $X \in N\left(a_1; \sigma^2\right)$ и $Y \in N\left(a_2; \sigma^2\right)$ с одинаковой дисперсией σ^2. Проверяется основная гипотеза H_0 о том, то $a_1 = a_2$ против альтернативной H_1 о том, то $a_1 \neq a_2$. В качестве статистики критерия берется функция

$$K = \sqrt{\frac{nm}{n+m}} \cdot \frac{\overline{x} - \overline{y}}{\sqrt{\frac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2}}}$$

Теорема 1.7.14.

$$\sqrt{\frac{nm}{n+m}} \cdot \frac{(\overline{x} - a_1) - (\overline{y} - a_2)}{\sqrt{\frac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2}}} \in T_{n+m-2}$$

Из этой теоремы следует, что если гипотеза H_0 верна, то $K \in T_{n+m-2}$. А если она не верна, то $|K| \to \infty$. Получаем критерий: пусть $t_{\text{кр}}$ это квантиль распределения $|T_{n+m-2}|$ уровня значимости α, тогда

$$\begin{cases} H_0, & |K| < t_{\text{кр}} \\ H_1, & |K| \geqslant t_{\text{кр}} \end{cases}$$

Критерий вероятности появления события

Пусть $\mathsf{P}(A) = p$ — неизвестная теоретическая вероятность события A. При достаточно большом числе n независимых испытаний события A появилось m раз. Проверяется основная гипотеза H_0 о том, что $\mathsf{P}(A) = p_0$ против альтернативной H_1 о том, что $\mathsf{P}(A) \neq p_0$. Обозначим $p^* = \frac{m}{n}$ точечную оценку $\mathsf{P}(A)$. В качестве статистики критерия берем величину

$$K = \frac{m - np_0}{\sqrt{np_0q_0}} \qquad (q_0 = 1 - p_0)$$

Если H_0 верна, то $K = \frac{m-np}{\sqrt{npq}} \rightrightarrows N(0;1)$ согласно теореме Муавра—Лапласа. Получаем критерий: пусть $t_{\text{кр}}$ это точка такая, что $\Phi(t_{\text{кр}}) = \frac{1-\alpha}{2}$, где α это уровень значимости. Тогда

$$\begin{cases} H_0, & |K| < t_{\text{кр}} \\ H_1, & |K| \geqslant t_{\text{кр}} \end{cases}$$

Пример 1.7.15. При посеве гороха вероятность рецессивного признака $p = 0.25$, а доминантного — $q = 0.75$. Из $n = 4000$ посеянных семян 970 оказались с рецессивным признаком, а 3030 — с доминантным. Проверяется основная гипотеза H_0 о том, что $p = 0.25$, против альтернативной H_1 о том, что $p \neq 0.25$.
Решение: Вычисляем статистику

$$K = \frac{m - np_0}{\sqrt{np_0q_0}} = -1.095$$

Далее найдем $t_{\text{кр}}$.

$$\Phi\left(t_{\text{кр}}\right) = \frac{1 - 0.05}{2} = 0.475 \Longrightarrow t_{\text{кр}} = 1.96$$

Т.к. $|K| < t_{\text{кр}}$, то гипотеза H_0 принимается.

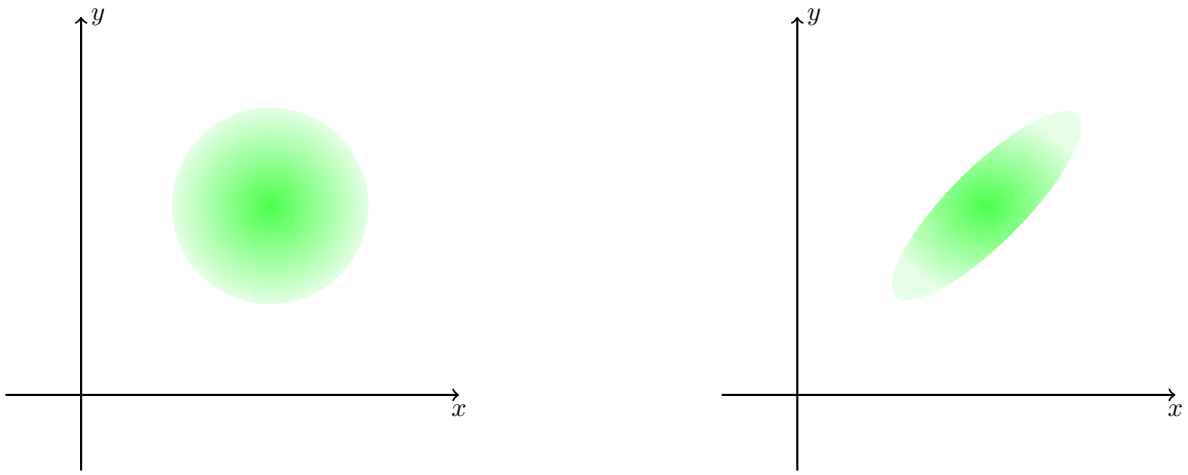
1.8. Лекция 24.03.28.

Статистическая зависимость

Def 1.8.1. Зависимость называется статистической, если изменение одной случайной величины вызывает изменение распределения другой. Если при этом изменяется среднее значение другой случайной величины, то такая статистическая зависимость называется корреляционной. Если при увеличении одной случайной величины, среднее другой также увеличивается, то имеет место прямая корреляция, в противном случае — обратная корреляция.

Корреляционное облако

Пусть в ходе n экспериментов появились значения случайных величин X и Y. Нанеся эти точки на координатную плоскость (рис. 1.8.2) получим корреляционное облако. По его виду можно сделать некоторые предположения о наличии и типе корреляции.



(a) Нет зависимости (b) Прямая корреляция

Рис. 1.8.2: Корреляционное облако

Корреляционная таблица

Пусть даны экспериментальные данные $(x_1, y_1), \ldots, (x_n, y_n)$. Их удобно представить в виде корреляционной таблицы, где по горизонтали будут расположены x_i, по вертикали — y_i, а в клетках будут находиться числа $n_{i,j}$, показывающие сколько раз встретилась точка (x_i, y_j).

Y / X	10	20	30	40	n_x	$\overline{y}_x$
2	7	3	0	0	10	13
4	3	10	10	2	25	24.4
6	0	2	10	3	15	30.67
n_y	10	15	20	5	$\sum = 50$	

Таблица 1.8.3: Корреляционная таблица

Пример 1.8.4. Пусть $n = 50$. Данные представлены в виде корреляционной таблицы 1.8.3. Столбец $\overline{y}_x$ не входит в корреляционную таблицу, но обычно записывается рядом. Это условное среднее, т.е. выборочное среднее одной случайной величины при условии, что другая величина зафиксирована. Его можно вычислить по формуле

$$\overline{y}_x = \frac{1}{n_x}\sum_i n_{x,y} y_i \qquad \overline{x}_y = \frac{1}{n_y}\sum_i n_{x,y} n_i$$

Т.к. с ростом x растут условные средние $\overline{y}_x$, то имеет место прямая корреляция.

Замечание 1.8.5. Если основные частоты сгруппированы относительно главной диагонали, то имеет место прямая корреляция, а если относительно побочной, то обратная. Если частоты распределены примерно равномерно, то корреляции практически нет.

Замечание 1.8.6. При большом объеме данных непрерывных случайных величин X и Y данные удобно собрать в интервальную корреляционную таблицу, где по вертикали отмечаем интервалы $[a_i; a_{i+1})$ случайной величины X, а по вертикали — интервалы $[b_j; b_{j+1})$ случайной величины Y. В клетках указываем количество точек $v_{i,j}$, попавших в соответствующую прямоугольную область $[a_i; a_{i+1}) \times [b_j; b_{j+1})$.

Критерий χ^2 для проверки независимости

Пусть выборка $(x_1, y_1), \ldots, (x_n, y_n)$ представлена в виде интервальной корреляционной таблицы. Случайная величина X при этом разбита на k интервалов, а случайная величина Y на m интервалов. Обозначим $v_{i\cdot}$ число значений X в интервале $[a_i; a_{i+1})$, где $1 \leqslant i \leqslant k$, $v_{\cdot j}$ число значений Y в интервале $[b_j; b_{j+1})$, где $1 \leqslant j \leqslant m$, а $v_{i,j}$ это число точек в соответствующем прямоугольнике $[a_i; a_{i+1}) \times [b_j; b_{j+1})$. Данные собрали в корреляционную таблицу 1.8.7.

X \ Y	$[b_0; b_1)$	$[b_1; b_2)$	$\ldots$	$[b_{m-1}; b_m]$	$v_{i\cdot} = \sum_{j=1}^{m} v_{i,j}$
$[a_0; a_1)$	$v_{1,1}$	$v_{1,2}$	$\ldots$	$v_{1,m}$	$v_{1\cdot}$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
$[a_{k-1}; a_k]$	$v_{k,1}$	$v_{k,2}$	$\ldots$	$v_{k,m}$	$v_{k\cdot}$
$v_{\cdot j} = \sum_{i=1}^{k} v_{i,j}$	$v_{\cdot 1}$	$v_{\cdot 2}$	$\ldots$	$v_{\cdot m}$	n

Таблица 1.8.7: Корреляционная таблица

Проверяется основная гипотеза H_0 о том, что X и Y — независимы против альтернативной H_1, а том, что они зависимы. Если гипотеза H_0 верна, то теоретическая вероятность попадания случайной величины $\langle x, y\rangle$ в любой прямоугольник равна произведению теоретических вероятностей попадания этих случайных величин в соответствующие интервалы. Таким образом

$$p_{i,j} = \mathsf{P}\left(X \in [a_{i-1}; a_i) \text{ и } Y \in [b_{j-1}; b_j)]\right) = \mathsf{P}\left(X \in [a_{i-1}; a_i)\right)\mathsf{P}\left(Y \in [b_{j-1}; b_j)]\right) = p_i q_j$$

Тогда по Закону Больших Чисел

$$\frac{v_{i\cdot}}{n} \xrightarrow{\mathsf{P}} p_i \qquad \frac{v_{\cdot j}}{n} \xrightarrow{\mathsf{P}} q_j \qquad \frac{v_{i,j}}{n} \xrightarrow{\mathsf{P}} p_{i,j}$$

поэтому основанием для отклонения нулевой гипотезы должна служить заметная разница между величинами $\frac{v_{i,j}}{n}$ и $\frac{v_{i\cdot}}{n} \cdot \frac{v_{\cdot j}}{n}$. В качестве статистики критерия берется функция

$$K = n\sum_{i,j} \frac{\left(v_{i,j} - \frac{1}{n} v_{i\cdot} v_{\cdot j}\right)^2}{v_{i\cdot} v_{\cdot j}}$$

Теорема 1.8.8. Если основная гипотеза H_0 верна, то $K \rightrightarrows \chi^2_{(k-1)(m-1)}$.

Получили критерий: пусть $t_{\text{кр}}$ это квантиль распределения $H_{(n-1)(m-1)}$ уровня значимости α, тогда

$$\begin{cases} H_0, & K < t_{\text{кр}} \\ H_1, & K \geqslant t_{\text{кр}} \end{cases}$$

Замечание 1.8.9. Частота каждой клетки должна быть не менее пяти.

Однофакторный дисперсионный анализ

Предположим, что на случайную величину X (результат) может влиять фактор Z, причем Z не обязательно случайная величина. Требуется определить, оказывает ли фактор Z на среднее значение X. Пусть при различных k уровнях фактора Z получены k независимых выборок случайной величины X. Обозначим их $X^{(1)} = \left(X_1^{(1)}, \ldots, X_{n_1}^{(1)}\right), \ldots, X^{(k)} = \left(X_1^{(k)}, \ldots, X_{n_k}^{(k)}\right)$. В общем случае их распределение отличается, поэтому с формальной точки зрения это выборки различных случайных величин.

Общая межгрупповая и внутригрупповая дисперсия

Для каждой выборки вычислим ее выборочное среднее и дисперсию.

$$\overline{x}^{(j)} = \frac{1}{n_j}\sum_{i=1}^{n_j} x_i^{(j)} \qquad \mathbb{D}^{(j)} = \frac{1}{n_j}\sum_{i=1}^{n_j}\left(x_i^{(j)} - \overline{x}^{(j)}\right)^2$$

Объединив все выборки в одну общую получаем выборку объема $n = n_1 + \ldots + n_k$. Для нее вычислим общее выборочное среднее и общую выборочную дисперсию

$$\overline{x} = \frac{1}{n}\sum_{i,j} x_i^{(j)} = \frac{1}{n}\sum_{j=1}^{k}\overline{x}^{(j)} n_j \qquad \mathbb{D}_{\text{О}} = \frac{1}{n}\sum_{i,j}\left(x_i^{(j)} - \overline{x}\right)^2$$

Def 1.8.10. Внутригрупповой (остаточной) дисперсией называется среднее групповых дисперсий.

$$\mathbb{D}_{\text{в}} = \frac{1}{n}\sum_{j=1}^{k} n_j \mathbb{D}^{(j)}$$

Def 1.8.11. Межгрупповой (факторной) дисперсией (или дисперсией выборочных средних) называется величина

$$\mathbb{D}_{\text{м}} = \frac{1}{n}\sum_{j=1}^{k}\left(\overline{x}^{(j)} - \overline{x}\right)^2 n_j$$

Теорема 1.8.12. (О разложении дисперсии)

$$\mathbb{D}_{\text{О}} = \mathbb{D}_{\text{м}} + \mathbb{D}_{\text{в}}$$

□ Это можно доказать чисто алгебраически исходя из определений. ■

Замечание 1.8.13. Ясно, что чем большее влияние оказывает Z на результат X, тем больше получаются отдельные выборочные средние, а значит растет доля межгрупповой дисперсии в данной сумме. Таким образом величина корреляции характеризуется отношением $\dfrac{\mathbb{D}_{\text{м}}}{\mathbb{D}_{\text{О}}}$. Если зависимость функциональная, то эта дробь будет равна единице.

Проверка гипотезы о влиянии фактора

Предполагаем, что случайная величина X имеет нормальное распределение и фактор Z может влиять только на среднее значение, но не на дисперсию и тип распределения. Из-за этого можно считать, что данные независимые k-выборки при различных уровнях фактора Z также имеют нормальное распределение с одинаковой дисперсией, т.е. $X^{(j)} \in N\left(a_j; \sigma^2\right)$.
Проверяется основная гипотеза H_0 о том, что $a_1 = \ldots = a_j$, т.е. Z не влияет на среднее X, против альтернативной H_1 о том, что фактор Z влияет на среднее X. По 1.4.25 для каждой из k выборок имеем

$$\sum_{i=1}^{n}\left(\frac{x - \overline{x}}{\sigma}\right)^2 = \frac{n\mathbb{D}^*}{\sigma^2} \in H_{n_j - 1} \qquad 1 \leqslant j \leqslant k$$

Т.к. распределение «хи-квадрат» устойчиво относительно суммирования, то

$$\sum_{j=1}^{k}\frac{n_j\mathbb{D}^{(j)}}{\sigma^2} = \frac{n\mathbb{D}_{\text{в}}}{\sigma^2} \in H_{n-k} \text{ т.к. } \underbrace{(n_1 - 1) + \ldots + (n_k - 1)}_{k \text{ раз}} = n - k$$

Пусть основная гипотеза H_0 верна, тогда все данные можно считать одной выборкой объема n и опять по 1.4.25 имеем

$$\frac{n\mathbb{D}_{\text{О}}}{\sigma^2} \in H_{n-1}$$

Согласно 1.8.12 получаем

$$\mathbb{D}_{\text{О}} = \mathbb{D}_{\text{м}} + \mathbb{D}_{\text{в}}$$
$$\frac{n\mathbb{D}_{\text{О}}}{\sigma^2} = \frac{n\mathbb{D}_{\text{м}}}{\sigma^2} + \frac{n\mathbb{D}_{\text{в}}}{\sigma^2}$$

Т.к. левая часть имеет распределение H_{n-1}, а второе слагаемое в правой части имеет распределение H_{n-k}, то первое слагаемое в правой части имеет распределение H_{k-1}. Итак, при первой основной гипотезе H_0 получили, что

$$\frac{n\mathbb{D}_{\text{м}}}{\sigma^2} \in H_{k-1} \qquad \frac{n\mathbb{D}_{\text{в}}}{\sigma^2} \in H_{n-k}$$

Значит

$$\frac{n\mathbb{D}_{\text{м}}}{\sigma^2(k-1)} \cdot \frac{\sigma^2(n-k)}{n\mathbb{D}_{\text{в}}} = \frac{n-k}{k-1} \cdot \frac{\mathbb{D}_{\text{м}}}{\mathbb{D}_{\text{в}}} \in F_{k-1,n-k}$$

Таким образом в качестве статистики критерия берем функцию

$$K = \frac{n-k}{k-1} \cdot \frac{\mathbb{D}_{\text{м}}}{\mathbb{D}_{\text{в}}}$$

В результате получили критерий: пусть $t_{\text{кр}}$ это квантиль распределения $F_{k-1,n-k}$ уровня значимости α, тогда

$$\begin{cases} H_0, & K < t_{\text{кр}} \\ H_1, & K \geqslant t_{\text{кр}} \end{cases}$$

1.9. Лекция 24.04.04.

Исследование статистической зависимости

Пусть случайная величина X зависит от величины Z (не обязательно случайной).

Def 1.9.1. Регрессией X на Z называется функция

$$f(z) = \mathbb{E}\left(X \,\middle|\, Z = z\right)$$

Она показывает зависимость среднего значения X от значения Z. Уравнение $x = f(z)$ называется уравнением регрессии, а ее график — линей регрессии.

Пусть при n экспериментах при значениях $z_1, \ldots, z_n$ величины Z наблюдались соответствующие значения $x_1, \ldots, x_n$ случайной величины X. Обозначим через $\varepsilon_i = x_i - f(z_i)$ разницу между экспериментальными и теоретическими значениями X. Тогда $x_i = f(z_i) + \varepsilon_i$, где ε_i можно считать ошибкой наблюдения, случая и влияния неучтенный факторов.

Замечание 1.9.2. Обычно можно считать, что ε_i — независимые одинаковые нормальные случайные величины с нулевым первым моментом, т.к.

$$a = \mathbb{E}(\varepsilon_i) = \mathbb{E}(X_i) - \mathbb{E}\left(X \,\middle|\, Z = z_i\right) = \mathbb{E}\left(X \,\middle|\, Z = z_i\right) - \mathbb{E}\left(X \,\middle|\, Z = z_i\right) = 0$$

Замечание 1.9.3. Второй параметр σ^2 не всегда одинаковый, в некоторых ситуациях (временные ряды) ошибки ε_i могут быть зависимы.

Задача состоит в том, чтобы по данным значениям $(z_1, x_1), \ldots, (z_n, x_n)$ как можно точнее оценить функцию регрессии $f(z)$. При этом предполагаем (часто из теории), что функция $f(z)$ определенного типа, но параметры которой не известны. В противном случае лучший решением была бы любая кривая, проходящая через данные точки.

Метод наименьших квадратов

Этот метод состоит в выборе параметром функции $f(z)$ таким образом, чтобы сумма квадратов ошибок была наименьшей.

Def 1.9.4. Пусть $\Theta = (\Theta_1, \ldots, \Theta_k)$ — набор неизвестных параметров функции $f(z)$. Оценка $\widehat{\Theta}$, при которой достигается минимум $\sum \varepsilon_i^2$, называется оценкой метода наименьших квадратов.

Линейная парная регрессия

Пусть имеется линейная регрессия $f(z) = a + bz$, тогда $X_i = a + bz_i + \varepsilon_i$. Найдем оценки неизвестных параметров a и b методом наименьших квадратов.

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (x_i - a - bz_i)^2 \to \min$$

Найдем частные производные.

$$\frac{\partial}{\partial a} \sum_{i=1}^{n} \varepsilon_i^2 = 2 \sum_{i=1}^{n} (x_i - a - bz_i) \cdot (-1) = -2 \left(\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} a - b \sum_{i=1}^{n} z_i \right) = -2 \left(n\overline{x} - na - bn\overline{z} \right)$$

$$\frac{\partial}{\partial b} \sum_{i=1}^{n} \varepsilon_i^2 = 2 \sum_{i=1}^{n} (x_i - a - bz_i) \cdot (-z_i) = -2 \left(\sum_{i=1}^{n} x_i z_i - \sum_{i=1}^{n} a z_i - b \sum_{i=1}^{n} z_i^2 \right) = -2 \left(n\overline{xz} - an\overline{z} - bn\overline{z^2} \right)$$

Приравняем их к нулю.

$$\begin{cases} -2\left(n\overline{x} - na - bn\overline{z}\right) = 0 \\ -2\left(n\overline{xz} - an\overline{z} - bn\overline{z^2}\right) = 0 \end{cases} \iff \begin{cases} a + b\overline{z} = \overline{x} \\ a\overline{z} + b\overline{z^2} = \overline{xz} \end{cases}$$

Получаем нормальную систему уравнений. Решим ее

$$\begin{cases} a = \overline{x} - b\overline{z} \\ \left(\overline{x} - b\overline{z}\right)\overline{z} + b\overline{z^2} = \overline{xz} \end{cases} \iff \begin{cases} a = \overline{x} - b\overline{z} \\ b\left(\overline{z^2} - \left(\overline{z}\right)^2\right) = \overline{xz} - \overline{x}\ \overline{z} \end{cases} \iff \begin{cases} \widehat{a} = \overline{x} - b\overline{z} \\ \widehat{b} = \frac{\overline{xz} - \overline{x}\ \overline{z}}{\widehat{\sigma}_z^2} \end{cases}$$

Получили оценки параметров a и b методов наименьших квадратов. Обозначим $\overline{x}_z = \mathbb{E}\left(X \,\middle|\, Z = z\right)$ — условное среднее. Запишем уравнение линейной регрессии в более удобном виде

$$\begin{aligned} \overline{x}_z &= a + bz \\ \overline{x}_z &= \overline{x} - b\overline{z} + bz \\ \overline{x}_z - \overline{x} &= b\left(z - \overline{z}\right) \\ \overline{x}_z - \overline{x} &= \frac{\overline{xz} - \overline{x}\ \overline{z}}{\widehat{\sigma}_z^2} \cdot \left(z - \overline{z}\right) \\ \overline{x}_z - \overline{x} &= \frac{\overline{xz} - \overline{x}\ \overline{z}}{\widehat{\sigma}_x \widehat{\sigma}_z} \cdot \frac{\widehat{\sigma}_x}{\widehat{\sigma}_z} \cdot \left(z - \overline{z}\right) \\ \overline{x}_z - \overline{x} &= \rho^* \cdot \frac{\widehat{\sigma}_x}{\widehat{\sigma}_z} \cdot \left(z - \overline{z}\right) \\ \frac{\overline{x}_z - \overline{x}}{\widehat{\sigma}_x} &= \rho^* \cdot \frac{z - \overline{z}}{\widehat{\sigma}_z} \end{aligned}$$

где

$$\rho^* = \frac{\overline{xz} - \overline{x}\ \overline{z}}{\widehat{\sigma}_x \widehat{\sigma}_z}$$

это выборочный коэффициент линейной корреляции. Итого получили выборочное уравнение линейной регрессии

$$\frac{\overline{x}_z - \overline{x}}{\widehat{\sigma}_x} = \rho^* \cdot \frac{z - \overline{z}}{\widehat{\sigma}_z}$$

Замечание 1.9.5. При $n \to \infty$ имеем $\overline{x} \to \mathbb{E}\left(X\right)$, $\overline{z} \to \mathbb{E}\left(Z\right)$, $\overline{xz} \to \mathbb{E}\left(xz\right)$ и следовательно выборочное уравнение регрессии стремится к

$$\frac{\mathbb{E}\left(X \,\middle|\, Z = z\right) - \mathbb{E}\left(X\right)}{\sigma_x} = \rho \frac{z - \mathbb{E}\left(Z\right)}{\sigma_z}$$

Это теоретическое уравнение линейной регрессии, где

$$\rho = \frac{\mathbb{E}\left(xz\right) - \mathbb{E}\left(x\right)\mathbb{E}\left(z\right)}{\sigma_x \sigma_z}$$

это теоретический коэффициент линейной корреляции.

Прямая строится таким образом, чтобы сумма квадратов длин синих отрезков была наименьшей.

Выборочный коэффициент линейной корреляции

Def 1.9.7. Выборочным коэффициентом линейной корреляции называется величина

$$\rho^* = \frac{\overline{xz} - \overline{x}\ \overline{z}}{\widehat{\sigma}_x \widehat{\sigma}_z}$$

Т.к. при замене в этой формуле средних на математические ожидания и выборочных средних квадратических отклонений на теоретические получаем теоретический коэффициент линейной корреляции ρ_{xz}, то данная величина является его точной оценкой. Таким образом выборочный коэффициент линейной корреляции характеризует силу линейной связи между случайными величинами. Его знак показывает, является ли она прямой или обратной.

Для оценки коэффициента линейной корреляции обычно используется шкала Чеддока (1.9.8).

Проверка гипотезы о значимости выборочного коэффициента корреляции

Пусть двумерная случайная величина $\langle z, x \rangle$ распределена нормально. По выборке объема n найден выборочный коэффициент линейной корреляции ρ^*. Теоретический коэффициент линейной корреляции обозначим ρ. Проверяется основная гипотеза H_0 о том, что $\rho = 0$, т.е. коэффициент ρ^* статистически не значим, против альтернативной гипотезы H_1 о том, что $\rho \neq 0$, т.е. коэффициент ρ^* статистически значим.

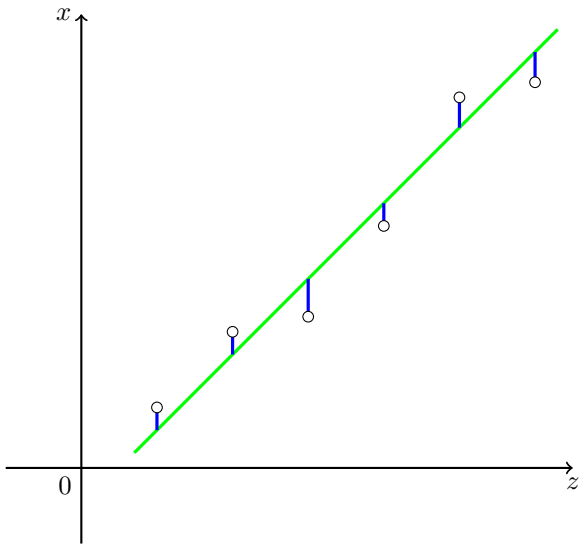


Рис. 1.9.6: Геометрический смысл прямой линейной регрессии

ρ^*	Характеристика силы связи
0.1—0.3	Слабая
0.3—0.5	Умеренная
0.5—0.7	Заметная
0.7—0.9	Сильная
> 0.9	Очень сильная

Таблица 1.9.8: Шкала Чеддока

Теорема 1.9.9. Если гипотеза H_0 верна, то

$$K = \frac{\rho^* \sqrt{n-2}}{\sqrt{1-(\rho^*)^2}} \in T_{n-2}$$

На основе этой теоремы получаем критерий: пусть $t_{\text{кр}}$ это квантиль распределения $|T_{n-2}|$ уровня значимости α, тогда

$$\begin{cases} H_0, & |K| < t_{\text{кр}} \\ H_1, & |K| \geqslant t_{\text{кр}} \end{cases}$$

Исторический смысл понятия регрессия

Исследовалась зависимость роста детей от роста родителей (Гальтон, $\approx$ 1886 год). На основе собранных данных были получены следующие формулы

$$\mathbb{E}\left(P_s \,\middle|\, z_f = z_1; z_m = z_2\right) = 0.27 z_1 + 0.2 z_2 + const$$

$$\mathbb{E}\left(P_d \,\middle|\, z_f = z_1; z_m = z_2\right) = \frac{1}{1.08} P_s$$

Казалось бы, что чем выше родители, тем выше должны быть их дети (прямая корреляция). В целом так и есть, но есть исключение: дети самых высоких родителей как правило были среднего роста, т.е. все сходилось к среднему значению. Отсюда и название «эффект регрессии» и «уравнение регрессии».

Выборочное корреляционное отношение

Пусть имеется k выборок случайной величины X при значениях $z_1, \ldots, z_k$ фактора Z.

Def 1.9.10. Выборочным корреляционным отношением называется величина

$$\eta_{xz} = \sqrt{\frac{\mathbb{D}_{\text{м}}}{\mathbb{D}_{\text{O}}}}$$

Lm 1.9.11.

$$0 \leqslant \eta_{xz} \leqslant 1$$

□ По 1.8.12

$$\left.\begin{array}{l}\mathbb{D}_{\text{О}} = \mathbb{D}_{\text{м}} + \mathbb{D}_{\text{в}} \\ \mathbb{D}_{\text{О}} \geqslant 0 \\ \mathbb{D}_{\text{в}} \geqslant 0 \\ \mathbb{D}_{\text{м}} \geqslant 0\end{array}\right\} \Longrightarrow 0 \leqslant \mathbb{D}_{\text{м}} \leqslant \mathbb{D}_{\text{О}} \Longrightarrow 0 \leqslant \frac{\mathbb{D}_{\text{м}}}{\mathbb{D}_{\text{О}}} \leqslant 1$$

■

Lm 1.9.12. Если $\eta_{xz} = 1$, то имеется функциональная зависимость $x = f(z)$.

□ Если $\eta_{xz} = 1$, то $\mathbb{D}_{\text{м}} = \mathbb{D}_{\text{О}}$ и $\mathbb{D}_{\text{в}} = 0$. Т.е. при определенном значении z случайная величина X всегда принимает одно и то же значение, что является функциональной зависимостью. ■

Lm 1.9.13. Если $\eta_{xz} = 0$, то корреляция отсутствует.

□ Если $\eta_{xz} = 0$, то $\mathbb{D}_{\text{м}} = 0$, т.е. при разных уровнях z получили одно и то же выборочное среднее, что и означает отсутствие корреляции. ■

Lm 1.9.14. Имеет место неравенство $\eta \geqslant |\rho^*|$. Причем $\eta = |\rho^*|$ тогда и только тогда, когда имеет место точная корреляционная зависимость (все точки экспериментальных данных (z_i, x_i) лежат на одной прямой).

1.10. Лекция 24.04.11.

Ковариация и ее свойства

Def 1.10.1. Ковариацией случайных величин X и Y называется

$$\operatorname{cov}(X,Y) = \mathbb{E}\left((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\right) = \mathbb{E}(XY) - \mathbb{E}(X)\,\mathbb{E}(Y)$$

Она является индикатором наличия и направления связи.

Пусть получены экспериментальные данные $(x_1, y_1), \ldots, (x_n, y_n)$ случайных величин X и Y.

Def 1.10.2. Выборочной ковариацией называется величина

$$\operatorname{cov}^*(X,Y) = \frac{1}{n}\sum_{i,j}(x_i - \overline{x})(y_j - \overline{y}) = \frac{1}{n}\sum_{i,j} x_i y_j - \overline{x}\ \overline{y} = \overline{xy} - \overline{x}\ \overline{y}$$

Замечание 1.10.3. Ясно, что выборочная ковариация является точной оценкой теоретической, но это смещенная оценка. Несмещенной оценкой будет $\frac{n}{n-1}\operatorname{cov}^*(X,Y)$.

Свойства ковариации

1. $\operatorname{cov}(X,Y) = \operatorname{cov}(Y,X)$
2. $\operatorname{cov}(X,a) = 0$, где $a = const$
3. $\operatorname{cov}(X,bY) = b\operatorname{cov}(X,Y)$
4. $\operatorname{cov}(X+Y,Z) = \operatorname{cov}(X,Z) + \operatorname{cov}(Y,Z)$
5. $\operatorname{cov}(X,X) = \mathbb{D}(X)$
6. $\mathbb{D}(X+Y) = \mathbb{D}(X) + \mathbb{D}(Y) + 2\operatorname{cov}(X,Y)$

Замечание 1.10.4. Эти же свойства имеет и выборочная ковариация.

Анализ модели парной линейной регрессии

Пусть при n экспериментах получены значения $(z_1, x_1), \ldots, (z_n, x_n)$ величин X (случайная) и Z (фактор, не обязательно случайная). Пусть $X = \alpha + \beta Z + \varepsilon$ это теоретическая модель линейной регрессии. Случайный член ε отражает влияние неучтенных факторов, возможную нелинейность модели, ошибок измерения и просто случая. Цель заключается в том, чтобы дать оценки неизвестным параметрам α, β и ε.
Пусть при обработке данных методом наименьших квадратов нашли выборочное уравнение линейной регрессии $\widehat{X} = a + b\widehat{Z}$. Тогда $X_i = a + bZ_i + \varepsilon_i$, где $\varepsilon_i = X_i - a - bZ_i$ — наблюдаемые ошибки. a и b это точные оценки неизвестных параметров α и β.

Свойства ошибок ε_i

Lm 1.10.5.

$$\overline{\varepsilon_i} = 0$$

□

$$a = \overline{x} - b\overline{z} \Longrightarrow a + b\overline{z} = \overline{x}$$
$$\overline{\varepsilon_i} = \overline{X_i - a - bZ_i} = \overline{X_i} - \overline{a + bz_i} = \overline{x} - \overline{x} = 0$$

■

Lm 1.10.6.

$$\operatorname{cov}(\widehat{x}, \varepsilon) = 0$$

□

$$b = \frac{\overline{xz} - \overline{x}\ \overline{z}}{\mathbb{D}(Z)} = \frac{\operatorname{cov}(X, Z)}{\mathbb{D}(Z)} \Longrightarrow b\mathbb{D}(Z) - \operatorname{cov}(X, Z) = 0$$

Значит

$$\operatorname{cov}(\widehat{x}, \varepsilon) = \operatorname{cov}(a + bZ, X - a - bZ) = \operatorname{cov}(bZ, X - bZ) = b\operatorname{cov}(Z, X) - b^2\operatorname{cov}(Z, Z) = b(\operatorname{cov}(Z, X) - b\mathbb{D}(Z)) = 0$$

■

Анализ дисперсии результатов

Дисперсия наблюдаемых значений X_i:

$$\mathbb{D}(X) = \frac{1}{n}\sum_i (x_i - \overline{x})^2$$

Дисперсия расчетных значений $\widehat{X}_i = a + bZ_i$:

$$\mathbb{D}\left(\widehat{X}\right) = \frac{1}{n}\sum_i (\widehat{x}_i - \overline{x})^2$$

Дисперсия экспериментальных ошибок:

$$\mathbb{D}(\varepsilon) = \frac{1}{n}\sum_i (\varepsilon_i - \overline{\varepsilon})^2 = \frac{1}{n}\sum_i \varepsilon_i^2$$

Знаем, что $X_i = a + bZ_i + \varepsilon_i$, т.е. $X_i = \widehat{X}_i + \varepsilon_i$. Применим дисперсию к обеим частям, и получим

$$\mathbb{D}(X) = \mathbb{D}\left(\widehat{X}\right) + \mathbb{D}(\varepsilon) + 2\operatorname{cov}\left(\widehat{X}, \varepsilon\right) = \mathbb{D}\left(\widehat{X}\right) + \mathbb{D}(\varepsilon)$$

Получили разложение дисперсии. Ясно, что качестве модели будет тем лучше, чем меньше дисперсия ошибок (в этом и суть метода наименьших квадратов). Значит модель будет тем лучше, чем больше доля первого слагаемого.

Def 1.10.7. Коэффициентом детерминации называется величина

$$R^2 = \frac{\mathbb{D}\left(\widehat{X}\right)}{\mathbb{D}(X)} = 1 - \frac{\mathbb{D}(\varepsilon)}{\mathbb{D}(X)}$$

Смысл коэффициента детерминации заключается в том, что это доля дисперсии объясненной при помощи данной модели. Величину $\dfrac{\mathbb{D}(\varepsilon)}{\mathbb{D}(X)}$ можно трактовать как долю необъясненной дисперсии.

Замечание 1.10.8. Ясно, что $0 \leqslant R^2 \leqslant 1$, причем чем больше R^2, тем выше качество модели. Если $R^2 = 1$, то $\mathbb{D}(\varepsilon) = 0$ и, т.к. $\overline{\varepsilon} = 0$, то $\forall i \,\Big|\, \varepsilon_i = 0$. Таким образом получили идеальную модель, все точки данных легли на прямую регрессии. Если $R^2 = 0$, то $\mathbb{D}\left(\widehat{X}\right) = 0$ и $\widehat{X} = \overline{X}$, т.е. построили ничего не объясняющую модель.

Проверка гипотезы о значимости уравнения регрессии

Проверяется основная гипотеза H_0 о том, что $R^2_{\text{т}} = 0$ (т.е. R^2 статистически не значим), против альтернативной H_1 о том, что $R^2_{\text{т}} \neq 0$ (т.е. R^2 статистически значим).

Теорема 1.10.9. Если нулевая гипотеза верна, то

$$K = \frac{R^2(n-2)}{1 - R^2} \in F_{1,n-2}$$

На основе этой теоремы получаем критерий: пусть $t_{\text{кр}}$ это квантиль распределения $F_{1,n-2}$ уровня значимости α, тогда

$$\begin{cases} H_0, & |K| < t_{\text{кр}} \\ H_1, & |K| \geqslant t_{\text{кр}} \end{cases}$$

Замечание 1.10.10. Т.к. $R^2 = 0 \Longleftrightarrow b = 0$, то это одновременно равносильно проверке гипотезы H_0 о том, что $\beta = 0$.

Связь между коэффициентом детерминации и коэффициентом линейной корреляции

Lm 1.10.11.

$$\sqrt{R^2} = \rho_{\widehat{X},X}$$

где $\rho_{\widehat{X},X}$ это коэффициент линейной корреляции между расчетными и наблюдаемыми значениями.

□

$$\operatorname{cov}\left(\widehat{X}, X\right) = \operatorname{cov}\left(\widehat{X}, \widehat{X} + \varepsilon\right) = \operatorname{cov}\left(\widehat{X}, \widehat{X}\right) + \underbrace{\operatorname{cov}\left(\widehat{X}, \varepsilon\right)}_{=0} = \mathbb{D}\left(\widehat{X}\right)$$

Таким образом

$$\rho_{\widehat{X},X} = \frac{\operatorname{cov}\left(\widehat{X}, X\right)}{\sqrt{\mathbb{D}\left(\widehat{X}\right)\mathbb{D}\left(X\right)}} = \frac{\mathbb{D}\left(\widehat{X}\right)}{\sqrt{\mathbb{D}\left(\widehat{X}\right)\mathbb{D}\left(X\right)}} = \sqrt{\frac{\mathbb{D}\left(\widehat{X}\right)}{\mathbb{D}\left(X\right)}} = \sqrt{R^2}$$

■

Lm 1.10.12.

$$\rho_{\widehat{X},X} = \rho_{X,Z}$$

□

$$\operatorname{cov}\left(\widehat{X}, X\right) = \operatorname{cov}\left(a + bZ, X\right) = b\operatorname{cov}\left(Z, X\right)$$
$$\mathbb{D}\left(\widehat{X}\right) = \mathbb{D}\left(a + bZ\right) = b^2\mathbb{D}\left(Z\right)$$

Значит

$$\rho_{\widehat{X},X} = \frac{\operatorname{cov}\left(\widehat{X}, X\right)}{\sqrt{\mathbb{D}\left(\widehat{X}\right)\mathbb{D}\left(X\right)}} = \frac{b\operatorname{cov}\left(Z, X\right)}{\sqrt{b^2\mathbb{D}\left(Z\right)\mathbb{D}\left(X\right)}} = \frac{\operatorname{cov}\left(Z, X\right)}{\sqrt{\mathbb{D}\left(Z\right)\mathbb{D}\left(X\right)}} = \rho_{X,Z}$$

■

Замечание 1.10.13. В случае парной линейной регрессии коэффициент детерминации совпадает с квадратом коэффициента корреляции, т.е. $R^2 = \rho^2_{Z,X}$

Замечание 1.10.14. В случае парной линейной регрессии совпадают результаты проверок гипотез $H_0\colon R^2_{\text{t}} = 0$, $H_0\colon \rho_{Z,X} = 0$ и $H_0\colon \beta = 0$.

Теорема Гаусса—Маркова

Пусть $X_i = \alpha + \beta Z_i + \varepsilon_i$ это теоретическая модель линейной регрессии, а $\widehat{X} = a + bZ$ это выборочная уравнение линейной регрессии, полученное методом наименьших квадратов. Хотим узнать, насколько хороши оценки a и b неизвестных параметров α и β.

Теорема 1.10.15. (Гаусса—Маркова) Пусть выполнены следующие условия:

1. Случайные члены ε_i независимы и имеют одинаковое нормальное распределение $N\left(0; \sigma^2\right)$.
2. Случайные величины Z_i и ε_i независимы.

Тогда оценка (a, b) является наилучшей линейной несмещенной оценкой неизвестных параметров α и β, т.е.

1. Состоятельность: $a \xrightarrow{\mathsf{P}} \alpha$ и $b \xrightarrow{\mathsf{P}} \beta$ при $n \to \infty$.
2. Несмещенность: $\mathbb{E}\left(a\right) = \alpha$ и $\mathbb{E}\left(b\right) = \beta$.
3. Эффективность: a и b имеют наименьшую дисперсию в классе линейных оценок, равную

$$\mathbb{D}\left(a\right) = \frac{\overline{z^2}\sigma^2}{n\mathbb{D}\left(Z\right)} \qquad \mathbb{D}\left(b\right) = \frac{\sigma^2}{n\mathbb{D}\left(Z\right)}$$

Замечание 1.10.16. Если ε_i зависимы или имеют разные дисперсии, то оценки по методу наименьших квадратов становятся неэффективными.

Замечание 1.10.17. Если случайные величины Z_i и ε_i зависимы, то оценки становятся смещенными и могут быть даже несостоятельными.

Стандартные ошибки коэффициентов регрессии

Видим, что дисперсии $\mathbb{D}(a)$ и $\mathbb{D}(b)$ зависят от дисперсии σ^2 случайного члена. По данным выборки экспериментальных ошибок ε_i получаем оценку дисперсии ошибок

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_i \varepsilon_i^2$$

Однако это оценка смещенная, т.к.

$$\mathbb{E}\left(\frac{1}{n}\sum_i \varepsilon_i^2\right) = \frac{n-2}{n}\sigma^2$$

поэтому несмещенной оценкой дисперсии σ^2 будет величина

$$S^2 = \frac{1}{n-2}\sum_i \varepsilon_i^2$$

Def 1.10.18. Величина S это стандартная ошибка регрессии. Она характеризует разброс результата вокруг линии регрессии.

Из этой формулы и 1.10.15 получем оценки дисперсий $\mathbb{D}(a)$ и $\mathbb{D}(b)$:

$$S_a^2 = \frac{\overline{z^2}S^2}{n\mathbb{D}(z)} \qquad S_b^2 = \frac{S^2}{n\mathbb{D}(z)}$$

Def 1.10.19. Коэффициенты S_a и S_b называются стандартными ошибками коэффициентов регрессии.

Прогнозирование в регрессионных моделях

Пусть $X_i = \alpha + \beta Z_i + \varepsilon_i$ это теоретическая модель, а $\widehat{X} = a + bZ$ это модель метода наименьших квадратов. С помощью данной модели надо дать прогноз значения X_p при данном значении фактор Z_p. Тогда $X_p = \alpha + \beta Z_p + \varepsilon_i$ это реальное значение, а $\widehat{X}_p = a + bZ_p$ — его точечная оценка (точечный прогноз). Обозначим $\Delta_p = \widehat{X}_p - X_p$ ошибку прогноза.

Lm 1.10.20.

$$\mathbb{E}(\Delta_p) = 0$$

□ **TODO:** самостоятельно ■

Lm 1.10.21.

$$\mathbb{D}(\Delta_p) = \left(1 + \frac{1}{n} + \frac{(Z_p - \overline{z})^2}{n\mathbb{D}(Z)}\right)\cdot\sigma^2$$

где $\sigma^2 = \mathbb{D}(\varepsilon)$ это дисперсия случайного члена.

Замечание 1.10.22. Если σ^2 заменить на S^2, то получаем стандартную ошибку прогноза

$$S_{\Delta p} = S\sqrt{1 + \frac{1}{n} + \frac{(Z_p - \overline{z})^2}{n\mathbb{D}(Z)}}$$

Замечание 1.10.23. Проанализируем полученное выражение для $\mathbb{D}(\Delta p)$.

1. Точность прогноза ограничена значением σ^2, т.е. $\mathbb{D}(\Delta p) \geqslant \sigma^2$.
2. $\mathbb{D}(\Delta p) \to \sigma^2$ при $n \to \infty$, т.е. чем больше объем выборки, тем более качественная модель.
3. Чем дальше Z_p от $\overline{Z}$, тем хуже качество прогноза (рис. 1.10.24). Наилучшая точность достигается при $Z_p = \overline{Z}$, тогда $\mathbb{D}(\Delta p) = \left(1 + \frac{1}{n}\right)\sigma^2$.

Доверительные интервалы прогноза и коэффициентов уравнения линейной регрессии

Пусть t_γ это квантиль двустороннего распределения Стьюдента с $n-2$ степенями свободы уровня γ. Тогда доверительные интервалы надежности γ для параметров α и β имеют вид

$$(a - t_\gamma S_a; a + t_\gamma S_a) \qquad (b - t_\gamma S_b; b + t_\gamma S_b)$$

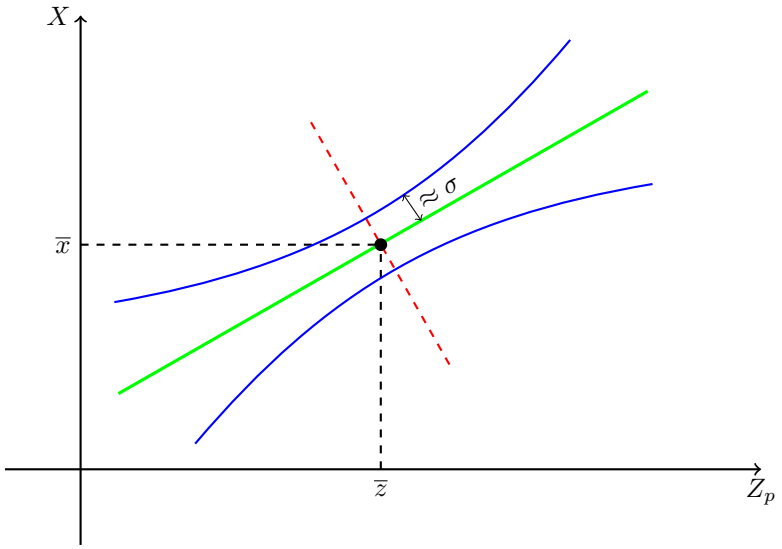


Рис. 1.10.24: Точность прогноза

Доверительный интервал прогноза надежности γ имеет вид

$$\left(\widehat{X}_p - t_\gamma S_{\Delta p}; \widehat{X}_p + t_\gamma S_{\Delta p}\right)$$

1.11. Лекция 24.04.18.

Математическое ожидание и дисперсия случайного вектора

Пусть имеется случайный вектор $\vec{X} = (X_1, \ldots, X_n)^T$, где случайные величины X_i это компоненты случайного вектора.

Def 1.11.1. Математическим ожиданием случайного вектора $\vec{X}$ называется вектор с координатами из математических ожиданий компонент.

$$\mathbb{E}\left(\vec{X}\right) = \begin{pmatrix} \mathbb{E}(X_1) \\ \vdots \\ \mathbb{E}(X_n) \end{pmatrix}$$

Def 1.11.2. Дисперсией (или матрицей ковариаций) случайного вектора $\vec{X}$ называется матрица

$$\mathbb{D}\left(\vec{X}\right) = \mathbb{E}\left(\vec{X} - \mathbb{E}(X)\right)\left(\vec{X} - \mathbb{E}(X)\right)^T$$

Или, более просто, это матрица, состоящая из элементов $d_{i,j} = \operatorname{cov}(X_i, X_j)$.

Замечание 1.11.3. На главной диагонали матрицы ковариаций стоят дисперсии компонент.

Свойства

1. $\mathbb{E}\left(A\vec{X}\right) = A\mathbb{E}\left(\vec{X}\right)$, где A это матрица размера $n \times n$.
2. $\mathbb{E}\left(\vec{X} + \vec{B}\right) = \mathbb{E}\left(\vec{X}\right) + \vec{B}$
3. $\mathbb{D}\left(A\vec{X}\right) = A\mathbb{D}\left(\vec{X}\right)A^T$
4. $\mathbb{D}\left(\vec{X} + \vec{B}\right) = \mathbb{D}\left(\vec{X}\right)$

Общая модель линейной регрессии

Пусть результат X зависит от факторов $Z_1, \ldots, Z_k$. Рассматривается теоретическая модель линейной регрессии.

$$\mathbb{E}\left(\vec{X} \,\middle|\, \vec{Z}\right) = f\left(\vec{Z}\right) = \beta_1 Z_1 + \ldots + \beta_k Z_k + \varepsilon$$

где $\vec{Z} = (z_1, \ldots, z_k)^T$. Обозначим $\vec{\beta} = (\beta_1, \ldots, \beta_k)^T$ — вектор неизвестных параметров регрессии, ε — случайный член (ошибка), отражающая нелинейность модели, влияние неучтенных факторов и т.д.

Пусть проведено $n \geqslant k$ экспериментов. Обозначим $\vec{Z}^{(i)} = \left(z_1^{(i)}, \ldots, z_k^{(i)}\right)$ набор значений факторов при i-ом эксперименте. Пусть $\vec{X} = (X_1, \ldots, X_n)$ это набор результатов при n экспериментах. Согласно модели

$$\begin{cases} X_1 = \beta_1 z_1^{(1)} + \ldots + \beta_k z_k^{(1)} + \varepsilon_1 \\ X_2 = \beta_1 z_1^{(2)} + \ldots + \beta_k z_k^{(2)} + \varepsilon_2 \\ \vdots \\ X_n = \beta_1 z_1^{(n)} + \ldots + \beta_k z_k^{(n)} + \varepsilon_n \end{cases}$$

где ε_i это теоретическая ошибка при i-ом эксперименте (она неизвестна). В матричной форме

$$\vec{X} = Z^T \vec{\beta} + \vec{\varepsilon}$$

где $\vec{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)$ это столбец ошибок, а матрица Z называется матрицей плана и имеет вид

$$Z_{k \times n} = \begin{pmatrix} z_1^{(1)} & z_1^{(2)} & \ldots & z_1^{(n)} \\ z_2^{(1)} & z_2^{(2)} & \ldots & z_2^{(n)} \\ \vdots & \vdots & \ddots & \vdots \\ z_k^{(1)} & z_k^{(2)} & \ldots & z_k^{(n)} \end{pmatrix}$$

Требуется по данной матрице плана Z и вектору результатов $\vec{X}$ найти оценки $\vec{B} = (b_1, \ldots, b_k)$ для параметров регрессии $\vec{\beta} = (\beta_1, \ldots, \beta_k)$ и параметров ошибок ε_i.

Замечание 1.11.4. В данной модели мы не теряем свободный член a, т.к. можно считать, что $z_1 \equiv 1$ и ей соответствует строка $(1, \ldots, 1)$ в матрице плана Z.

Метод наименьших квадратов и нормальные уравнения

Будем считать, что выполнено условие о том, что $\operatorname{rank} Z = k$, т.е. все строки матрицы плана независимы. Обозначим $A = ZZ^T$ — это будет квадратная матрица размера $k \times k$. Заметим, что

1. Матрица A симметричная, т.е. $A^T = A$.
2. Матрица A будет положительно определенной.
3. Существенная вещественная матрица $\sqrt{A}$ такая, что $\left(\sqrt{A}\right)^2 = A$.

Экспериментальная модель имеет вид

$$X_i = b_1 Z_1 + \ldots b_k Z_k + \widehat{\varepsilon}_i$$

где $\widehat{\varepsilon}_i = X_i - (b_1 Z_1 + \ldots + b_k Z_k)$.
При методе наименьших квадратов находим оценку $\vec{B} = (b_1, \ldots, b_k)^T$, которая минимизирует функцию

$$L\left(\vec{B}\right) = \sum_{i=1}^{n} \widehat{\varepsilon}_i^2 = \|\widehat{\varepsilon}\|^2 = \left\|\vec{X} - Z^T \vec{B}\right\|^2$$

Заметим, что $\left\|\vec{X} - Z^T \vec{B}\right\|^2$ это квадрат расстояния от точки $\vec{X} \in \mathbb{R}^n$ до точки $Z^T \vec{B} \in \mathbb{R}^n$. Причем $Z^T \vec{B}$ это произвольная точка подпространства $Z^T \vec{t}$, где $\vec{t} \in \mathbb{R}^k$. Заметим, что $\dim Z^T \vec{t} = k$, т.к. все строки матрицы плана независимы. Таким образом, искомое минимальное расстояние это расстояние от точки $\vec{X} \in \mathbb{R}^n$ до подпространства $Z^T \vec{t}$, $\vec{t} \in \mathbb{R}^k$. Это длина перпендикуляра, т.е. нужное расстояние получаем при условии, что вектор $\vec{X} - Z^T \vec{B}$ будет ортогонален всем векторам данного подпространства. Значит $\forall \vec{t} \in \mathbb{R}^k$ справедливо

$$\left(Z^T \vec{t}; \vec{X} - Z^T \vec{B}\right) = \left(Z^T \vec{t}\right)^T \cdot \left(\vec{X} - Z^T \vec{B}\right) = \vec{t}^T Z \left(\vec{X} - Z^T \vec{B}\right) = \vec{t}^T \left(Z\vec{X} - ZZ^T \vec{B}\right) = \vec{t}^T \left(Z\vec{X} - A\vec{B}\right) = 0$$

Т.к. всем векторам данного пространства может быть ортогонален только нулевой вектор, то получаем, что

$$Z\vec{X} - A\vec{B} = \vec{0} \Longleftrightarrow A\vec{B} = Z\vec{X} \Longleftrightarrow \vec{B} = A^{-1} Z\vec{X}$$

Это нормальное уравнение или система нормальных уравнений с неизвестными $b_1, \ldots, b_k$. Получили оценки по методу наименьших квадратов.

Свойства оценок по методу наименьших квадратов

Далее предполагаем, что выполнено условие $\operatorname{rank} Z = k$ и ошибки ε_i — независимые нормальные случайные величины с распределением $N\left(0; \sigma^2\right)$. Таким образом $\mathbb{D}\left(\vec{\varepsilon}\right) = \sigma^2 E_n$.

Lm 1.11.5.

$$\vec{B} - \vec{\beta} = A^{-1} Z \vec{\varepsilon}$$

□

$$\begin{aligned}
\vec{B} - \vec{\beta} &= A^{-1} Z \vec{X} - \vec{\beta} \\
&= A^{-1} Z \left(Z^T \vec{\beta} + \vec{\varepsilon} \right) - \vec{\beta} \\
&= A^{-1} Z Z^T \vec{\beta} + A^{-1} Z \vec{\varepsilon} - \vec{\beta} \\
&= A^{-1} A \vec{\beta} + A^{-1} Z \vec{\varepsilon} - \vec{\beta} \\
&= \vec{\beta} + A^{-1} Z \vec{\varepsilon} - \vec{\beta} \\
&= A^{-1} Z \vec{\varepsilon}
\end{aligned}$$

■

Lm 1.11.6. Оценка $\vec{B}$ это несмещенная оценка для параметра $\vec{\beta}$.

□

$$\mathbb{E}\left(\vec{B} - \vec{\beta}\right) = \mathbb{E}\left(A^{-1} Z \vec{\varepsilon}\right) = A^{-1} Z \mathbb{E}\left(\vec{\varepsilon}\right) = A^{-1} Z \vec{0} = \vec{0} \Longrightarrow \mathbb{E}\left(\vec{B}\right) = \vec{\beta}$$

■

Lm 1.11.7.

$$\mathbb{D}\left(\vec{B}\right) = \sigma^2 A^{-1}$$

□

$$\begin{aligned}
\mathbb{D}\left(\vec{B}\right) &= \mathbb{D}\left(\vec{B} - \vec{\beta}\right) \\
&= \mathbb{D}\left(A^{-1} Z \vec{\varepsilon}\right) \\
&= A^{-1} Z \mathbb{D}\left(\vec{\varepsilon}\right) \left(A^{-1} Z\right)^T \\
&= A^{-1} Z \sigma^2 E_n Z^T A^{-1} \\
&= \sigma^2 A^{-1} Z Z^T A^{-1} \\
&= \sigma^2 A^{-1} A A^{-1} \\
&= \sigma^2 A^{-1}
\end{aligned}$$

■

Замечание 1.11.8. Таким образом $\mathbb{D}(b_i) = \sigma^2 \left(A^{-1}\right)_{i,i}$.

Введем обозначение $\widehat{\sigma}^2$, которое определим как

$$\widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \widehat{\varepsilon}_i^2 = \frac{1}{n} \left\| \vec{X} - Z^T \vec{B} \right\| = \frac{1}{n} L\left(\vec{B}\right)$$

Заметим, что $\widehat{\sigma}^2$ это оценка неизвестной дисперсии ошибки σ^2.

Теорема 1.11.9. Пусть выполнено условие $\operatorname{rank} Z = k$ и ошибки ε_i — независимые нормальные случайные величины с распределением $N\left(0; \sigma^2\right)$. Тогда

1. Вектор $\frac{1}{\sigma} \sqrt{A} \left(\vec{B} - \vec{\beta}\right)$ состоит из независимых случайных величин со стандартным нормальным распределением.
2. $$\frac{n\widehat{\sigma}^2}{\sigma^2} \in H_{n-k} \text{ и не зависит от } \vec{B}$$
3. $$S^2 = \frac{n\widehat{\sigma}^2}{(n-k)\sigma^2} = \frac{1}{n-k} \sum_{i=1}^{n} \widehat{\varepsilon}_i^2 \qquad \text{Это несмещенная оценка для } \sigma^2$$

□ **I** пункт

$$\frac{1}{\sigma} \sqrt{A} \left(\vec{B} - \vec{\beta}\right) = \frac{1}{\sigma} \sqrt{A} A^{-1} Z \vec{\varepsilon} = \frac{1}{\sigma} \sqrt{A^{-1}} Z \vec{\varepsilon}$$

Получили линейное преобразование нормального вектора $\vec{\varepsilon}$, и поэтому оно также является нормальным вектором.

$$\mathbb{E}\left(\frac{1}{\sigma}\sqrt{A}\left(\vec{B}-\vec{\beta}\right)\right)=\frac{1}{\sigma}\sqrt{A}\mathbb{E}\left(\vec{B}-\vec{\beta}\right)=\frac{1}{\sigma}\sqrt{A}\,\vec{0}=\vec{0}$$

Таким образом первый параметр компонент равен нулю.

$$\mathbb{D}\left(\frac{1}{\sigma}\sqrt{A}\left(\vec{B}-\vec{\beta}\right)\right)=\frac{1}{\sigma^2}\sqrt{A}\mathbb{D}\left(\vec{B}-\vec{\beta}\right)\sqrt{A^T}=\frac{1}{\sigma^2}\sqrt{A}\sigma^2A^{-1}\sqrt{A}=E_k$$

Значит все координаты не коррелированны и имеют дисперсию равную единице.

II пункт

По построению метода наименьших квадратов имеем

$$\forall\vec{t}\in\mathbb{R}^k\,\Big|\,\vec{X}-Z^T\vec{B}\perp Z^T\vec{t}$$

В частости это верно для $\vec{t}=\vec{B}-\vec{\beta}$. По обобщенной теореме Пифагора

$$\begin{gathered}\left\|\vec{X}-Z^T\vec{B}\right\|^2+\left\|Z^T\left(\vec{B}-\vec{\beta}\right)\right\|^2=\left\|\vec{X}-Z^T\vec{B}+Z^T\left(\vec{B}-\vec{\beta}\right)\right\|^2=\left\|\vec{X}-Z^T\vec{\beta}\right\|^2\\ \left\|\vec{X}-Z^T\vec{B}\right\|^2=\left\|\vec{X}-Z^T\vec{\beta}\right\|^2-\left\|Z^T\left(\vec{B}-\vec{\beta}\right)\right\|^2=\left\|\vec{\varepsilon}\right\|^2-\left\|Z^T\left(\vec{B}-\vec{\beta}\right)\right\|^2\end{gathered}\tag{1}$$

Далее работает со вторым слагаемым.

$$\begin{aligned}\left\|Z^T\left(\vec{B}-\vec{\beta}\right)\right\|^2&=\left(\vec{B}-\vec{\beta}\right)^TZZ^T\left(\vec{B}-\vec{\beta}\right)\\&=\left(\vec{B}-\vec{\beta}\right)^TA\left(\vec{B}-\vec{\beta}\right)\\&=\left(\vec{B}-\vec{\beta}\right)^T\sqrt{A^T}\sqrt{A}\left(\vec{B}-\vec{\beta}\right)\\&=\left\|\sqrt{A}\left(\vec{B}-\vec{\beta}\right)\right\|^2\\&=\left\|\sqrt{A}A^{-1}Z\vec{\varepsilon}\right\|^2\\&=\left\|\sqrt{A^{-1}}Z\vec{\varepsilon}\right\|^2\end{aligned}\tag{2}$$

Заметим, что строки матрицы $\sqrt{A^{-1}}Z$ ортогональны, т.к.

$$\left(\sqrt{A^{-1}}Z\right)\left(\sqrt{A^{-1}}Z\right)^T=\sqrt{A^{-1}}ZZ^T\left(\sqrt{A^{-1}}\right)^T=\sqrt{A^{-1}}A\sqrt{A^{-1}}=E$$

$\sqrt{A^{-1}}Z$ это прямоугольная матрица размера $k\times n$. Из курса линейной алгебры известно, что ее можно дополнить до ортогональной матрицы C размера $n\times n$. Тогда первые k координат n-мерного вектора $\vec{Y}=\frac{1}{\sigma}C\vec{\varepsilon}$ совпадают с координатами вектора $\vec{Y}=\frac{1}{\sigma}\sqrt{A^{-1}}Z\vec{\varepsilon}$.
В результате из (1) и (2) получаем

$$\begin{aligned}\frac{n\widehat{\sigma}^2}{\sigma^2}&=\frac{1}{\sigma^2}\sum_{i=1}^n\widehat{\varepsilon}_i^2\\&=\frac{1}{\sigma^2}\left\|\vec{X}-Z^T\vec{B}\right\|^2\\&\overset{(1)}{=\!=}\frac{1}{\sigma^2}\left\|\vec{\varepsilon}\right\|^2-\frac{1}{\sigma^2}\left\|Z^T\left(\vec{B}-\vec{\beta}\right)\right\|^2\\&=\left\|\frac{\vec{\varepsilon}}{\sigma}\right\|^2-\left\|\frac{Z^T\left(\vec{B}-\vec{\beta}\right)}{\sigma}\right\|^2\\&\overset{(2)}{=\!=}\left\|\frac{\vec{\varepsilon}}{\sigma}\right\|^2-\left\|\frac{\sqrt{A^{-1}}Z\vec{\varepsilon}}{\sigma}\right\|^2\\&=\sum_{i=1}^n\left(\frac{\varepsilon_i}{\sigma}\right)^2-Y_1^2-Y_2^2-\ldots-Y_k^2\end{aligned}\tag{3}$$

Вектор $\frac{\vec{\varepsilon}}{\sigma}$ имеет n-мерное стандартное нормальное распределение, а $Y_1,\ldots,Y_k$ это первые k координат ортогонального преобразования данного вектора, поэтому (3) согласно 1.4.21 имеет распределение H_{n-k} и не зависит от вычитаемых координат вектора $\frac{\sqrt{A^{-1}}Z\vec{\varepsilon}}{\sigma}$, а значит и от вектора

$$\vec{B} = A^{-1}Z\vec{\varepsilon} + \vec{\beta} = \sigma\sqrt{A^{-1}}\left(\frac{1}{\sigma}\sqrt{A^{-1}}Z\vec{\varepsilon}\right)$$

которые являются их линейными комбинациями.

III пункт

Т.к. $\mathbb{E}\left(\chi^2_{n-k}\right) = n-k$, то

$$\mathbb{E}\left(\widehat{\sigma}^2\right) = \frac{\sigma^2}{n}\mathbb{E}\left(\frac{n\widehat{\sigma}^2}{\sigma^2}\right) = \frac{\sigma^2}{n}\cdot(n-k) = \frac{n-k}{n}\cdot\sigma^2$$

Это не равно σ^2, т.е. оценка будет смещенной. Значит оценка

$$S^2 = \frac{n}{n-k}\widehat{\sigma}^2 = \frac{1}{n-k}\sum_{i=1}^{n}\widehat{\varepsilon}_i^2$$

будет несмещенной. ■

1.12. Лекция 24.04.25.

Построение и анализ уравнения множественной линейной регрессии

Пусть выявлена зависимость результата X от факторов $Z_1, \ldots, Z_m$. Проведено $n \geqslant m$ экспериментов и получены экспериментальные данные результата $\vec{X} = (X_1, \ldots, X_n)$ при соответствующих значениях факторов $\vec{Z}^{(k)} = \left(Z_1^{(k)}, \ldots, Z_n^{(k)}\right)$. Предполагаем, что зависимость X от всех факторов линейная. Требуется по этим данным построить модель наилучшим образом объясняющую и предсказывающую поведение X.

Мультиколлинеарность

Def 1.12.1. Мультиколлинеарность это наличие заметной корреляции между всеми или некоторыми факторами.

Неприятные последствия мультиколлинеарности:

1. Оценки параметров становятся ненадежными, имеют большие стандартные ошибки и малую значимость, причем даже в том случае, когда модель в целом имеет высокую значимость.
2. Небольшое изменение исходных данных может привести к существенному изменению оценок регрессии.
3. Трудно выявить изолированное влияние конкретного фактора на результат и физический/экономический/т.д. смысл этого влияния.

Начальный отбор факторов в уравнение регрессии

Построим корреляционную матрицу, состоящую из коэффициентов линейной корреляции.

$$P = \begin{pmatrix} 1 & \rho_{x,z_1} & \rho_{x,z_2} & \cdots & \rho_{x,z_m} \\ \rho_{z_1,x} & 1 & \rho_{z_1,z_2} & \cdots & \rho_{z_1,z_m} \\ \rho_{z_2,x} & \rho_{z_2,z_1} & 1 & \cdots & \rho_{z_2,z_m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{z_m,x} & \rho_{z_m,z_1} & \rho_{z_m,z_2} & \cdots & 1 \end{pmatrix}$$

Выбираем фактор, наиболее коррелирующий с X. Далее добавляем в модель факторы, которые с одной стороны имеют бóльшую корреляцию с результатом X, а с другой стороны наименее коррелированны с факторами, которые уже включены в модель.

Пример 1.12.2. Пусть дана корреляционная таблица.

	X	Z_1	Z_2	Z_3
X	1	0.81	0.85	−0.65
Z_1	0.81	1	0.93	−0.38
Z_2	0.85	0.93	1	−0.28
Z_3	−0.65	−0.38	−0.28	1

Согласно алгоритму естественно включить фактор Z_2, т.к. он имеет наибольшую корреляцию с X. Далее логично включить фактор Z_3, т.к. Z_1 сильно коррелирован с уже включенным фактором Z_2. Итого выбираем факторы Z_2 и Z_3.

Анализ уравнения линейной регрессии

Пусть после отсева осталось k факторов. Теоретическая модель регрессии имеет вид

$$X = \beta_0 + \beta_1 z_1 + \ldots + \beta_k z_k + \varepsilon$$

где ε — случайный член, отражающий влияние неучтенных факторов, возможную нелинейность модели, случай и т.д. Методом наименьших квадратов построили модель

$$\widehat{X} = b_0 + b_1 z_1 + \ldots + b_k z_k + \widehat{\varepsilon}$$

Предполагаем, что $\forall i \,\Big|\, \varepsilon_i \in N\left(0; \sigma^2\right)$ и независимы. Согласно пункту 3 в 1.11.9 получили несмещенную оценку для σ^2 в виде

$$S^2 = \frac{1}{n-k-1} \sum_{i=1}^{n} \widehat{\varepsilon}_i^2$$

Def 1.12.3. S это стандартная ошибка регрессии.

Из 1.11.8 получили $\mathbb{D}\left(b_i\right) = \sigma^2 \left(A^{-1}\right)_{i,i}$, где $A = ZZ^T$, а Z это матрица плана. Тогда

$$S_{b_i} = S\sqrt{\left(A^{-1}\right)_{i,i}}$$

Это стандартные ошибки коэффициентов регрессии b_i.

Уравнение регрессии в стандартных масштабах

Замечание 1.12.4. При обычном уравнении регрессии по коэффициентам b_i нельзя судить о силе влияния фактора на результат X, т.к. факторы могут быть разной природы и иметь различные единицы измерения.

Стандартизация данных. Пусть имеется выборка $(X_1, \ldots, X_n)$ случайной величины X. Заменяем ее выборкой $t_x \colon X_i \to \frac{X_i - \overline{x}}{\widehat{\sigma}_x}$, которую можно считать выборкой случайной величины $t_x = \frac{X - \mathbb{E}(X)}{\sigma_X}$, не имеющей единиц измерения.

Замечание 1.12.5. Очевидно, что $\overline{t_x} = 0$ и $\widehat{\mathbb{D}}_{t_x} = 1$.

Lm 1.12.6.

$$\widehat{\rho}_{X,Y} = \widehat{\rho}_{t_x,t_y} = \overline{t_x t_y}$$

□

$$\widehat{\rho}_{X,Y} = \frac{\operatorname{cov}^*(X,Y)}{\widehat{\sigma}_x \widehat{\sigma}_y} = \frac{\frac{1}{n} - \sum_{i=1}^{n} \left(x_i - \overline{x}\left(y_i - \overline{y}\right)\right)}{\widehat{\sigma}_x \widehat{\sigma}_y} = \frac{1}{n} \sum_{i=1}^{n} \frac{x_i - \overline{x}}{\widehat{\sigma}_x} \frac{y_i - \overline{y}}{\widehat{\sigma}_y} = \frac{1}{n} \sum_{i=1}^{n} t_x t_y = \overline{t_x t_y}$$

Равенство $\widehat{\rho}_{t_x,t_y} = \overline{t_x t_y}$ будет выполняться, т.к. $\overline{t_x} = \overline{t_y} = 0$ и $\widehat{\mathbb{D}}_{t_x} = \widehat{\mathbb{D}}_{t_y} = 1$ в силу стандартизованности этих величин. ■

Все значения исследуемых признаков и результата стандартизируем

$$t_j = \frac{z_j^{(i)} - \overline{z}_j}{\widehat{\sigma}_{z_j}} \qquad 1 \leqslant i \leqslant k$$

$$t_x = \frac{x_i - \overline{x}}{\widehat{\sigma}_x}$$

При линейной модели регрессии можно все величины в уравнении регрессии заменить на стандартизованные, т.к. все операции при стандартизации линейные. В результате получим так называемое уравнение в стандартных масштабах.

$$t_x = \gamma_1 t_1 + \ldots + \gamma_k t_k$$

Замечание 1.12.7. Заметим, что свободного члена нет, т.к. $\overline{t}_x = \overline{t}_1 = \ldots = \overline{t}_k = 0$.

Lm 1.12.8. При стандартизации система нормальных уравнений приобретает более простой вид.

$$\begin{cases} \gamma_1 + \rho_{z_1,z_2}\gamma_2 + \ldots + \rho_{z_1,z_k}\gamma_k = \rho_{z_1,x} \\ \rho_{z_2,z_1}\gamma_1 + \gamma_2 + \ldots + \rho_{z_2,z_k}\gamma_k = \rho_{z_2,x} \\ \ldots \\ \rho_{z_k,z_1}\gamma_1 + \rho_{z_k,z_2}\gamma_2 + \ldots + \gamma_k = \rho_{z_k,x} \end{cases}$$

В матричной форме это можно записать как $P\Gamma = P_x$, где P это матрица корреляций между факторами, $\Gamma = (\gamma_1, \ldots, \gamma_k)^T$ и $P_x = (\rho_{x,z_1}, \ldots, \rho_{x,z_k})^T$.

□ Действительно, нормальное уравнение регрессии имело вид $AB = Z\vec{X}$, где $A = ZZ^T$, а Z это матрица плана. Допустим, что все данные стандартизированны, тогда i-тый элемент столбца $Z\vec{X}$ имеет вид

$$\left(Z_1^{(i)}, \ldots, Z_n^{(i)}\right) \cdot \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} = n\overline{Z^{(i)}X} = n\rho_{z_i,x}$$

При этом элемент $a_{i,j}$ матрицы A будет равен

$$a_{i,j} = \left(Z_1^{(i)}, \ldots, Z_n^{(i)}\right) \cdot \begin{pmatrix} Z_1^{(j)} \\ \vdots \\ Z_n^{(j)} \end{pmatrix} = \begin{cases} n\overline{z_i z_j} = n\rho_{z_i,z_j} & i \neq j \\ n\overline{z_i^2} = n\mathbb{D}(z_i) = n & i = j \end{cases}$$

После сокращения n в левой и правой части получаем искомую систему уравнений. ■

Переход от уравнения в стандартных масштабах к обычному уравнению регрессии и обратно можно выполнить по следующим формулам.

$$b_i = \gamma_i \frac{\sigma_x}{\sigma_{z_i}} \qquad b_0 = \overline{X} - \sum_{i=1}^{n} b_i \overline{z_i}$$

Замечание 1.12.9. В случае парной линейной регрессии уравнение в стандартных масштабах имеет вид $t_x = \rho t_z$.

Смысл стандартизованных коэффициентов

Коэффициент γ_i на какую часть своего среднеквадратического отклонения σ_x изменится результат X при изменении фактора Z_i на величину своего среднеквадратического отклонения σ_{z_i} (при неизменных значениях других факторов). При мультиколлинеарности факторы оказывают не только прямое влияние на результат, но и косвенное (через влияние на другие факторы). Стандартизованный коэффициент γ_i можно трактовать как показатель прямого влияния, а остальные слагаемые в уравнении как результат косвенного влияния.

$$\rho_{z_i,z_1}\gamma_1 + \rho_{z_i,z_2}\gamma_2 + \ldots + \rho_{z_i,z_{i-1}}\gamma_{i-1} + \gamma_i + \rho_{z_i,z_{i+1}}\gamma_{i+1} + \ldots + \rho_{z_i,z_k}\gamma_k = \rho_{z_i,x}$$

Результат $\rho_{z_i,x}$ показывает величину полного влияния, а γ_i можно грубо трактовать как величину прямого влияния. Все остальное можно трактовать как косвенное влияние.

Замечание 1.12.10. Для измерения тесноты линейной связи между фактором и результатом при устранении влияния остальных факторов есть понятия коэффициентов частной корреляции. Например, при $k = 2$

$$\rho_{x,z_1|z_2} = \frac{\rho_{x,z_1} - \rho_{x,z_2}\rho_{z_1,z_2}}{\sqrt{\left(1-\rho_{z_1,z_2}^2\right)\left(1-\rho_{x,z_2}^2\right)}} \qquad \rho_{x,z_2|z_1} = \frac{\rho_{x,z_2} - \rho_{x,z_1}\rho_{z_1,z_2}}{\sqrt{\left(1-\rho_{z_1,z_2}^2\right)\left(1-\rho_{x,z_1}^2\right)}}$$

Коэффициенты детерминации и множественной корреляции

Допустим, что как и в случае парной линейной регрессии дисперсию результата X можно разложить на объясненную и необъясненную составляющую, т.е.

$$\mathbb{D}(X) = \mathbb{D}\left(\widehat{X}\right) + \mathbb{D}(\widehat{\varepsilon})$$

где $\mathbb{D}\left(\widehat{X}\right)$ это дисперсия расчетных значений по построенной модели МНК, а $\mathbb{D}(\widehat{\varepsilon})$ это дисперсия экспериментальных ошибок.

Def 1.12.11. Коэффициентом детерминации R^2 называется величина

$$R^2 = 1 - \frac{\mathbb{D}(\widehat{\varepsilon})}{\mathbb{D}(X)}$$

Замечание 1.12.12. Ясно, что $0 \leqslant R^2 \leqslant 1$, причем чем больше R^2, тем лучше качество модели. Если $R^2 = 1$, то $\mathbb{D}(\widehat{\varepsilon}) = 0$, т.е. $\widehat{\varepsilon} \equiv 0$, значит все экспериментальные данные легли на гиперплоскость регрессии. Если $R^2 = 0$, то $\mathbb{D}\left(\widehat{X}\right) = 0$, т.е. $\widehat{X} = \overline{X}$, значит $b_0 = \overline{X}, b_1 = \ldots = b_k = 0$. Такая модель ничего не объясняет.

Замечание 1.12.13. В случае линейного уравнения регрессии $R^2 = \sum_{i=1}^{k} \gamma_i \rho_{x,z_i}$, где γ_i это стандартизованные коэффициенты.

Def 1.12.14. R называется коэффициентом множественной корреляции.

Замечание 1.12.15. При добавлении в модель новых факторов R^2 всегда вырастет, однако не всегда эти факторы следует вводить в модель, поэтому для выяснения того, следует ли это делать, существует скорректированный коэффициент детерминации $\overline{R^2}$.

$$\overline{R^2} = 1 - \frac{n-1}{n-k-1} \cdot \frac{\mathbb{D}(\widehat{\varepsilon})}{\mathbb{D}(X)}$$

где n это число экспериментов, а k это число факторов в модели.

F-тест: проверка гипотезы о значимости уравнения регрессии в целом

Проверяется основная гипотеза H_0 о том, что $R^2_{\text{т}} = 0$ (т.е. уравнение в целом не значимо), против альтернативной H_1 о том, что $R^2_{\text{т}} \neq 0$.

Теорема 1.12.16. Если нулевая гипотеза верна, то

$$F = \frac{R^2}{1-R^2} \cdot \frac{n-k-1}{k} \in F_{k,n-k-1}$$

Пусть $t_{\text{кр}}$ это квантиль распределения $F_{k,n-k-1}$ уровня значимости α, тогда

$$\begin{cases} H_0, & F < t_{\text{кр}} \\ H_1, & F \geqslant t_{\text{кр}} \end{cases}$$

T-тест: проверка гипотезы о значимости отдельного коэффициента регрессии

Проверяется основная гипотеза H_0 о том, что $\beta_i = 0$, против альтернативной H_1 о том, что $\beta_i \neq 0$.

Теорема 1.12.17. Если нулевая гипотеза верна, то

$$T_i = \frac{b_i}{S_{b_i}} \in T_{n-k-1}$$

Пусть $t_{\text{кр}}$ это квантиль распределения $|T_{n-k-1}|$ уровня значимости α, тогда

$$\begin{cases} H_0, & |T_i| < t_{\text{кр}} \\ H_1, & |T_i| \geqslant t_{\text{кр}} \end{cases}$$

Замечание 1.12.18. Данный критерий применяется для отсева несущественных факторов.

Замечание 1.12.19. При мультиколлинеарности может оказаться так, что все коэффициенты по отдельности статистически не значимы, в то время как модель в целом имеет высокую значимость.

1.13. Лекция 24.05.02.

Нюансы регрессионного анализа

Пусть имеется линейное уравнение регрессии в матричной форме $\vec{X} = Z^T\vec{\beta} + \vec{\varepsilon}$, где $\vec{X} = (X_1, \ldots, X_n)$ это столбец результатов n экспериментов, $Z_{k\times n}$ это матрица плана, $\vec{\beta} = (\beta_1, \ldots, \beta_k)^T$ это столбец неизвестных (теоретических) коэффициентов регрессии и $\vec{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ это вектор теоретических ошибок, возникающих из-за возможной нелинейности модели, невключенных факторов, случая и ошибок измерения.
Ранее предполагали, что строки матрицы плана Z независимы, т.е. $\operatorname{rank} Z = k$ (на практике почти всегда выполнено), а также $\forall i \,\big|\, \varepsilon_i \in N(0;\sigma)$ и независимы. На практике часто бывают ситуации, когда ошибки коррелированны между собой и дисперсия зависит от эксперимента.

Взвешенный метод наименьших квадратов

Пусть ошибки ε_i некоррелированны, имеют нормальное распределение с первым параметром 0, но их дисперсия зависит от эксперимента, т.е.

$$\operatorname{cov}(\varepsilon_i, \varepsilon_j) = 0 \qquad \overline{\varepsilon}_i = 0 \qquad \mathbb{D}(\varepsilon_i) = \sigma^2 v_i$$
$$\mathbb{D}\left(\vec{\varepsilon}\right) = \sigma^2 \operatorname{diag}(v_1, \ldots, v_n) = \sigma^2 V$$

В этом случае оценки по методу наименьших квадратов неэффективные. Логично придать наблюдениям с меньшей дисперсией больший вес. Обозначим $w_i = \frac{1}{v_i}$ и каждое уравнение из системы $\vec{X} = Z^T\vec{\beta} + \vec{\varepsilon}$ умножим на $\sqrt{w_i}$. Тогда

$$\widetilde{X}_i = \sqrt{w_i}X_i \qquad \widetilde{Z}_i^{(j)} = \sqrt{w_j}Z_i^{(j)} \qquad \widetilde{\varepsilon}_i = \sqrt{w_i}\varepsilon_i$$

При этом

$$\widetilde{\varepsilon}_i = 0 \qquad \mathbb{D}\left(\widetilde{\varepsilon}_i\right) = w_i \mathbb{D}\left(\varepsilon_i\right) = \frac{1}{v_i} \cdot \sigma^2 v_i = \sigma^2$$

Таким образом $\mathbb{D}\left(\vec{\widetilde{\varepsilon}}\right) = \sigma^2 E$, т.е. получили классическую ситуацию, и при применении к этим данным метода наименьших квадратов получаем эффективные оценки неизвестных коэффициентов регрессии.
Рассмотрим несколько приложений этой модели.

I. Модель $\vec{X} = \beta_0 Z + \vec{\varepsilon}$

$$\vec{X} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \qquad \vec{\beta} = \beta_0 \qquad \vec{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \qquad Z^T = \begin{pmatrix} z_1 \\ \vdots \\ z_n \end{pmatrix}$$

Подставим эти данные в нормальные уравнения $A\vec{B} = Z\vec{X}$, где

$$Z = (z_1, \ldots, z_n) \qquad A = ZZ^T = z_1^2 + \ldots + z_n^2 \qquad Z\vec{X} = (z_1, \ldots, z_n) \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = z_1 x_1 + \ldots + z_n x_n$$

Таким образом нормальные уравнения приобретают вид

$$\sum_{i=1}^n z_i^2 \beta_0 = \sum_{i=1}^n z_i x_i$$
$$\widehat{\beta}_0 = \frac{\sum_{i=1}^n z_i x_i}{\sum_{i=1}^n z_i^2}$$

При взвешенном методе наименьших квадратов получаем оценку

$$\widetilde{\beta}_0 = \frac{\sum_{i=1}^n w_i z_i x_i}{\sum_{i=1}^n w_i z_i^2}$$

Пример 1.13.1. Рассмотрим модель $X = \beta_0 + \varepsilon$, т.е. $Z \equiv (1, \ldots, 1)$. Тогда

$$\widetilde{\beta}_0 = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$$

Это можно трактовать как результат измерений скоропортящегося измерительного инструмента.

Пример 1.13.2. Пусть проводится n серий по k_i измерений (повторные измерения). Обозначим $\overline{x}_i$ это среднее арифметическое измерений в i-ой серии, а ε_i — ошибка среднего измерения в i-ой серии. Тогда дисперсия $\mathbb{D}\left(\varepsilon_i\right) = \frac{\sigma^2}{k_i}$, где σ^2 это дисперсия ошибки при одном измерении. Тогда $w_i = k_i$ и получаем формулу

$$\widetilde{\beta}_0 = \frac{\sum k_i x_i}{\sum k_i}$$

Пример 1.13.3. Пусть X это потери тепла в квартире. Логично предположить, что основной фактор это разница температур снаружи и внутри. Рассмотрим уравнение $X = \beta Z + \varepsilon$, где Z это разница температур. Логично предположить, что с возрастанием Z возрастает дисперсия ошибки. Рассмотрим две ситуации.

I. $\mathbb{D}\left(\varepsilon_i\right) = cz_i$

$$\mathbb{D}\left(\varepsilon_i\right) = \sigma^2 \cdot \frac{cz_i}{\sigma^2} \Longrightarrow w_i = \frac{\sigma^2}{cz_i} \Longrightarrow \widetilde{\beta}_0 = \frac{\sum \frac{\sigma^2}{cz_i} z_i x_i}{\sum \frac{\sigma^2}{cz_i} z_i^2} = \frac{\sum x_i}{\sum z_i} = \frac{\overline{x}}{\overline{z}}$$

II. $\mathbb{D}\left(\varepsilon_i\right) = cz_i^2$

$$\mathbb{D}\left(\varepsilon_i\right) = \sigma^2 \cdot \frac{cz_i^2}{\sigma^2} \Longrightarrow w_i = \frac{\sigma^2}{cz_i^2} \Longrightarrow \widetilde{\beta}_0 = \frac{\sum \frac{\sigma^2}{cz_i^2} z_i x_i}{\sum \frac{\sigma^2}{cz_i^2} z_i^2} = \frac{\sum \frac{x_i}{z_i}}{\sum 1} = \overline{\left(\frac{x_i}{z_i}\right)}$$

Коррелированные наблюдения

Пусть дисперсии ошибок $\mathbb{D}(\varepsilon_i)$ различны и ошибки ε_i коррелированны между собой, причем эта корреляция известна $\operatorname{cov}(\varepsilon_i, \varepsilon_j) = \sigma^2 v_{i,j}$. Тогда матрица ковариаций ошибок будет равна $\mathbb{D}\left(\vec{\varepsilon}\right) = \sigma^2 V$, где $V = (v_{i,j})$. Т.к. матрица V симметричная и положительно определенная, то существует $\sqrt{V}$. Умножим обе части модели на $\left(\sqrt{V}\right)^{-1}$ слева. В результате получаем новую модель

$$\vec{\widetilde{X}} = \widetilde{Z}^T \vec{\beta} + \vec{\widetilde{\varepsilon}}$$

$$\vec{\widetilde{X}} = \sqrt{V^{-1}}\vec{X} \qquad \widetilde{Z}^T = \sqrt{V} Z^T \qquad \vec{\widetilde{\varepsilon}} = \sqrt{V}\vec{\varepsilon}$$

Заметим, что $\widetilde{\overline{\varepsilon}}_i = 0$ и

$$\mathbb{D}\left(\vec{\widetilde{\varepsilon}}\right) = \mathbb{D}\left(\sqrt{V^{-1}}\vec{\varepsilon}\right) = \sqrt{V^{-1}}\mathbb{D}(\varepsilon)\left(\sqrt{V^{-1}}\right)^T = \sqrt{V^{-1}}\sigma^2 V\sqrt{V^{-1}} = \sigma^2 E$$

Таким образом получили стандартную ситуацию теоремы 1.10.15 и при этих данных получаем эффективные оценки неизвестных коэффициентов регрессии.

Нелинейная регрессия

Помимо общего метода наименьших квадратов многие нелинейные зависимости могут быть сведены к линейным при помощи простых приемов. Рассмотрим несколько случаев.

I. $X = \alpha + \beta f(Z) + \varepsilon$

Пусть $f(Z)$ это известная функция, например $f(Z) = \ln Z$ (быстрозамедляющийся процесс). Вычислим новые данные $Z'_i = f(Z_i)$ и получим стандартную модель парной линейной регрессии $X = \alpha + \beta Z' + \varepsilon$.

II. Степенная $X = \alpha Z^\beta \varepsilon$

Прологарифмируем обе части и получим

$$\ln X = \ln\alpha + \beta \ln Z + \ln\varepsilon$$
$$X' = \alpha' + \beta Z' + \varepsilon'$$

Далее по найденным α' и β' находим исходные α и β.

III. Показательная $X = \alpha e^{\beta Z}\varepsilon$

Прологарифмируем обе части и получим

$$\ln X = \ln\alpha + \beta Z + \ln\varepsilon$$
$$X' = \alpha' + \beta Z + \varepsilon'$$

IV. Полиномиальная $X = \alpha + \beta_1 Z + \ldots + \beta_k Z^k + \varepsilon$

Введем новые переменные, положим $u_i = Z^i$, тогда

$$X = \alpha + \beta_1 u_1 + \ldots + \beta_k u_k + \varepsilon$$

Полученное уравнение можно рассматривать как уравнение общей (множественной) регрессии.

Замечание 1.13.4. На практике обычно $k \leqslant 4$ во избежание мультиколлинеарности.

Замечание 1.13.5. При выборе между несколькими моделями выбираем ту, где коэффициент детерминации больше (т.е. меньше дисперсия экспериментальных ошибок).

Замечание 1.13.6. Построение даже удачной регрессионной модели не означает выявления причинно-следственной связи. Одна из причин заключается в том, что не учтен скрытый фактор.

Пример 1.13.7. Строилась модель точности бомбометания. Были выбраны факторы Z_1 — высота, Z_2 — ветер и Z_3 — количество истребителей противника. В итоге была получена модель

$$X = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 + \varepsilon$$

где коэффициент β_3 оказался больше нуля, что противоречит логике. Это произошло из-за того, что в модель не был включен фактор Z_4 — облачность.

Замечание 1.13.8. Пусть эксперимент управляем и матрицу плана можно выбирать (почти) произвольно. Наиболее эффективные оценки получаются в том случае, когда строки матрицы плана ортогональны.

Замечание 1.13.9. Иногда вместо метода наименьших квадратов используется метод главных осей. Допустим, что есть корреляционное облако в виде эллипса. Согласно этому методу уравнение регрессии строим так, чтобы прямая совпадала с главной осью эллипса.

1.14. Лекция 24.05.16.

Датчики случайных чисел

Def 1.14.1. Члены последовательности $y_1, \ldots, y_n, \ldots$, которые можно рассматривать как экспериментальные значения данной случайной величины, называются псевдослучайными числами, а устройства или алгоритмы для их получения — датчиками случайных чисел.

I. Физические датчики

Теорема 1.14.2. Случайная величина $\eta \in U(0;1)$ тогда и только тогда, когда разряды ξ_i ее двоичной записи $\eta = \sum_{i=1}^{\infty} 2^{-i}\xi_i$ — схема Бернулли с вероятностью $p = 0.5$.

Замечание 1.14.3. Недостатки физических датчиков:

1. Необходимо оборудование.
2. При повторном опыте нельзя получить ту же самую последовательность (а иногда это надо).

II. Таблицы случайных чисел

Пусть имеется таблица чисел (обычно двузначных) — результат работы некоторого датчика случайных чисел. Далее наугад выбирались строка и столбец и начиная с этого места брались «случайные» числа. При необходимости увеличения точности дописывали числа из соседнего столбца.

Замечание 1.14.4. Недостатки:

1. Требует много оперативной памяти.
2. Слишком большая предсказуемость.

III. Математические датчики

Обычно это рекуррентные последовательности вида $y_n = f(y_{n-1})$. Основным считается мультипликативный датчик, который задается так: k_0 — начальное число, с которого начинается генерация, a — это множитель и m это модуль. Причем модуль выбирается так, что множитель a и k_0 взаимно просты с ним. В итоге последовательность псевдослучайных чисел строится по формулам

$$\begin{cases} k_n = k_{n-1} \cdot a \pmod m \\ y_n = \frac{k_n}{m} \in (0;1) \end{cases}$$

Замечание 1.14.5. Рекомендации (старые, для 32-ух битных компьютеров):

1. $m = 2^{31} - 1$
2. $a = 630360016$ или $a = 764261123$
3. k_0 практически неважно.

IV. Датчик Уичмана и Хилла (1982)

Одновременно запускаются три мультипликативных датчика с параметрами $a_1 = 171$, $m_1 = 30269$, $a_2 = 172$, $m_2 = 30307$, $a_3 = 170$, $m_3 = 30323$. На i-ом шаге генерируется три псевдослучайных числа y_n', y_n'', y_n'''. Тогда $y_n = \{y_n' + y_n'' + y_n'''\}$ (дробная часть от суммы). Преимущества:

1. Он быстрее предыдущего.
2. Период этого датчика примерно $3 \cdot 10^{13}$, а предыдущего $2 \cdot 10^9$.

Моделирование случайных величин

В начале рассмотрим непрерывное распределение. Будем использовать метод обратной функции (квантильное преобразование).

Теорема 1.14.6. Пусть $F(x)$ — функция распределения абсолютно непрерывной случайной величины. Если $\eta \in U(0;1)$, то случайная величина $\xi = F^{-1}(\eta)$ имеет функцию распределения $F(x)$.

Пример 1.14.7 (Показательное распределение). Пусть $\xi \in E_\alpha$, тогда

$$F(x) = \begin{cases} 0, & x < 0 \\ 1 - e^{-\alpha x}, & x \geqslant 0 \end{cases}$$

Значит получаем

$$y = 1 - e^{-\alpha x} \Longrightarrow x = -\frac{1}{\alpha}\ln(1-y) = F^{-1}(y)$$

Тогда согласно теореме, если y_i это значение датчика, то $x_i = F^{-1}(y) \in E_\alpha$.

Пример 1.14.8 (Нормальное распределение). Пусть $\xi \in N\left(a;\sigma^2\right)$. Известно, что если $\xi \in N(0;1)$, то $(\sigma\xi + a) \in N(a;\sigma)$, поэтому достаточно уметь моделировать стандартное нормальное распределение.

$$F_0(x) = \frac{1}{2\sqrt{\pi}}\int_{-\infty}^{x}\exp\left(-\frac{z^2}{2}\right)\mathrm{d}z$$

Таким образом $x_i = F_0^{-1}(y_i) \in N(0;1)$, если $y_i \in U(0;1)$.

Замечание 1.14.9. Преимущество этого метода заключается в простоте и универсальности, а недостаток заключается в не очень высокой эффективности (раньше было так, сейчас уже возможно иначе).

Нормальные случайные числа

I. На основе центральной предельной теоремы

Пусть $\eta_i \in U(0;1)$, тогда $\mathbb{E}(\eta_i) = a = 0.5$ и $\mathbb{D}(\eta_i) = \frac{1}{12}$. По центральной предельной теореме

$$\frac{S_n - na}{\sqrt{n\mathbb{D}(\xi)}} = \frac{S_n - \frac{n}{2}}{\sqrt{\frac{n}{12}}} \rightrightarrows N(0;1)$$

Уже при $n = 12$ получаются довольно хорошие результаты.

II. Точное моделирование пары случайных величин с распределением $N(0;1)$

Теорема 1.14.10. Пусть $\eta_1, \eta_2 \in N(0;1)$ и независимы. Тогда следующие величины $X, Y \in N(0;1)$ и независимы.

$$X = \sqrt{-2\ln\eta_1}\cos(2\pi\eta_2) \qquad Y = \sqrt{-2\ln\eta_1}\sin(2\pi\eta_2)$$

□ Пусть $X, Y \in N(0;1)$ и независимы. Тогда плотность совместного распределения

$$f_{X,Y}(x,y) = \frac{1}{2\sqrt{\pi}}\exp\left(-\frac{x^2}{2}\right)\frac{1}{2\sqrt{\pi}}\exp\left(-\frac{y^2}{2}\right) = \frac{1}{2\pi}\exp\left(-\frac{1}{2}\left(x^2+y^2\right)\right)$$

При переходе к полярным координатам плотность данного распределения примет вид

$$f_{X,Y}(\rho,\varphi) = \frac{1}{2\pi}\exp\left(-\frac{1}{2}\rho^2\right)\rho = \underbrace{\tfrac{1}{2\pi}}_{f_\varphi}\cdot\underbrace{\exp\left(-\tfrac{\rho^2}{2}\right)\rho}_{f_\rho}$$

Найдем их функции распределения.

$$F_\varphi(x) = \frac{x}{2\pi}$$

$$F_\rho(x) = \int_0^\rho \rho\exp\left(-\frac{\rho^2}{2}\right)\mathrm{d}\rho = -\exp\left(-\frac{\rho^2}{2}\right)\Big|_0^\rho = 1 - \exp\left(-\frac{\rho^2}{2}\right)$$

Отсюда методом обратной функции получаем формулы для моделирования случайных величин φ и ρ.

$$y = \frac{x}{2\pi} \Longrightarrow x = 2\pi y \qquad y \in U(0;1)$$

$$y = 1 - \exp\left(-\frac{\rho^2}{2}\right) \Longrightarrow \rho = \sqrt{-2\ln(1-y)} \qquad y \in U(0;1)$$

Т.к. $y \in U(0;1)$, то $1 - y \in U(0;1)$. В итоге получаем

$$X = \rho\cos\varphi = \sqrt{-2\ln\eta_1}\cos(2\pi\eta_2) \qquad Y = \rho\sin\varphi = \sqrt{-2\ln\eta_1}\sin(2\pi\eta_2)$$

■

Быстрый показательный датчик

Теорема 1.14.11. Пусть случайные величины $\eta_1, \ldots, \eta_{2n-1} \in U(0;1)$. Обозначим $\xi_1, \ldots, \xi_{n-1}$ это расставленные в порядке возрастания значения $\eta_{n+1}, \ldots, \eta_{2n-1}$. Отдельно положим $\xi_0 = 0$ и $\xi_n = 1$. Тогда μ_i независимы и имеют показательное распределение с параметром α.

$$\mu_i = -\frac{1}{\alpha}(\xi_i - \xi_{i-1})\ln(\eta_1 \cdot \ldots \cdot \eta_n) \qquad 1 \leqslant i \leqslant n$$

Пример 1.14.12. При $n = 3$ получаем $\eta_1, \eta_2, \eta_3, \eta_4, \eta_5$. Пусть $\eta_4 < \eta_5$ (в противном случае переобозначим их), тогда

$$\mu_1 = -\frac{1}{\alpha}\eta_4 \ln(\eta_1\eta_2\eta_3)$$
$$\mu_2 = -\frac{1}{\alpha}(\eta_5 - \eta_4) \ln(\eta_1\eta_2\eta_3)$$
$$\mu_3 = -\frac{1}{\alpha}(1 - \eta_5) \ln(\eta_1\eta_2\eta_3)$$

Замечание 1.14.13. При этом методе экономим время при вычислении логарифма, но теряем на сортировке. Алгоритм оптимален при $n = 3$. В этом случае он примерно в два раза быстрее метода обратной функции.

Моделирование дискретных распределений

Пусть имеется дискретное распределение, которое задается парами x_k и $p_k = \mathsf{P}(\xi = c_k)$, где $k = 1, 2, \ldots$. Разобьем отрезок $[0; 1]$ на отрезки длины p_k. Обозначим $\eta_m = \sum_{k=1}^{m} p_k$ и $\eta_0 = 0$ — границы отрезков. Пусть $y_i \in [0; 1]$ это случайное число. Если $y_i \in [\rho_{j-1}; \rho_j)$, то полагаем, что $x_i = c_j$.

I. Распределение Бернулли

Пусть $\xi \in B_p$, тогда получаем два отрезка, т.е.

$$\begin{cases} y_i \in [0; 1-p) \Longrightarrow x_i = 0 \\ y_i \in [1-p; 1] \Longrightarrow x_i = 1 \end{cases}$$

II. Биномиальное распределение

Пусть $\xi \in B_{n,p}$, т.е. $\mathsf{P}(\xi = k) = C_n^k p^k q^{1-k}$, где $k = 0, \ldots, n$. Исходя из смысла биномиального распределения $\xi = \xi_1 + \ldots + \xi_n$, где $\xi_i \in B_p$. Далее берем n значений датчика $y_i \in U(0; 1)$ и

$$\begin{cases} y_i \in [0; 1-p) \Longrightarrow z_i = 0 \\ y_i \in [1-p; 1] \Longrightarrow z_i = 1 \end{cases} \quad \Longrightarrow x_k = \sum_{i=1}^{n} z_i$$

III. Геометрическое распределение

Пусть $\xi \in G_p$, т.е. $\mathsf{P}(\xi = k) = q^{k-1}p$, где $k = 1, 2, \ldots$. Исходя из смысла геометрического распределения ξ это номер первого успешного испытания, поэтому берем значения датчика $y_i \in U(0; 1)$ до тех пор, пока оно не попадет в интервал $[1-p; 1]$. Далее полагаем $x_k = i$.

IV. Распределение Пуассона

Пусть $\xi \in \Pi_\lambda$, т.е. $\mathsf{P}(\xi = k) = \frac{\lambda^k}{k!}e^{-\lambda}$, где $k = 0, 1, \ldots$.

Теорема 1.14.14. Пусть $\mu_1, \mu_2, \ldots$ — независимые случайные величины, имеющие показательное распределение с параметром λ. Положим $S_n = \mu_1 + \ldots + \mu_n$ и $N = \max n \,\big|\, S_n \in [0; 1]$. Тогда $N \in \Pi_\lambda$.

На основе данной теоремы и метода обратной функции получаем формулу для моделирования распределение Пуассона.

$$n_i = \min\left\{ k \,\Big|\, \prod_{j=1}^{k} y_{i,j} < e^{-\lambda} \right\}$$

где $y_{i,j}$ это значение датчика в i-ой серии.

1.15. Лекция 24.05.23.

Метод Монте—Карло

Цель метода состоит в том, чтобы находить неизвестные значения изучаемой величины при помощи «разыгрывания» некоторой случайной величины.
Общая постановка метода. Пусть требуется найти неизвестное число a и имеется случайная величина ξ такая, что $\mathbb{E}(\xi) = a$. Тогда согласно закону больших чисел

$$\frac{\xi_1 + \ldots + \xi_n}{n} \xrightarrow{\text{п.н.}} a$$

Следовательно, при достаточно больших n среднее выборочное $\overline{x} \approx a$.
Оценка погрешности. Пусть $\mathbb{D}(\xi) < \infty$, тогда согласно центральной предельной теореме

$$\frac{S_n - na}{\sqrt{n\mathbb{D}(\xi_1)}} = \frac{n(\overline{x} - a)}{\sqrt{n\mathbb{D}(\xi_1)}} \rightrightarrows z \in N(0;1)$$

По правилу трех сигм $\mathsf{P}(|z| < 3) \approx 0.9973$. Следовательно, при больших n можно считать, что

$$\left|\frac{n(\overline{x} - a)}{\sqrt{n\mathbb{D}(\xi_1)}}\right| < 3 \Longrightarrow |(\overline{x} - a)| < 3\frac{\sqrt{\mathbb{D}(\xi_1)}}{\sqrt{n}}$$

Отсюда видим, что сходимость есть, но она достаточно медленная, порядка $\frac{1}{\sqrt{n}}$, поэтому на практике не удается получить очень точные оценки для a.

Вычисление определенных интегралов

$$\int_a^b \varphi(x)\mathrm{d}x = \lim_{\Delta x_i \to 0} \sum_{i=1}^{n} \varphi(c_i)\Delta x_i$$

где Δx_i это длины интервалов разбиения, а c_i — точки внутри интервалов. На этом определении основаны так называемые квадратурные формулы.

I. Формула прямоугольников

Разобьем отрезок $[a;b]$ на n равных частей длины $\Delta x_i = \frac{b-a}{n}$. Обозначим x_i — середину i-ого интервала. Тогда

$$I = \int_a^b \varphi(x)\mathrm{d}x \approx \frac{b-a}{n}\sum_{i=1}^{n}\varphi(x_i) = I_n$$

Можно показать, что $|I - I_n| \leqslant \frac{M_1}{n^2}$, где $M_1 = const$.

II. Формула трапеций

Разобьем отрезок $[a;b]$ на n равных частей $[x_i;x_{i+1}]$ длины $\Delta x_i = \frac{b-a}{n}$. Обозначим $y_i = \varphi(x_i)$, где $0 \leqslant i \leqslant n$. Тогда

$$I = \int_a^b \varphi(x)\mathrm{d}x \approx \frac{b-a}{2n}\left(y_0 + y_n + 2\left(y_1 + \ldots + y_{n-1}\right)\right) = I_n$$

Можно показать, что $|I - I_n| \leqslant \frac{M_2}{n^2}$, где $M_2 = const$, но M_2 меньше, чем M_1, примерно в два раза.

III. Формула Симпсона (формула парабол)

Разобьем отрезок $[a;b]$ на $n = 2m$ равных отрезков $[x_i;x_{i+1}]$ длины $\Delta x_i = \frac{b-a}{n}$. Обозначим $y_i = \varphi(x_i)$, где $0 \leqslant i \leqslant n$. Тогда

$$I = \int_a^b \varphi(x)\mathrm{d}x \approx \frac{b-a}{3n}\Big(y_0 + y_n + 4\left(y_1 + y_3 + \ldots + y_{n-1}\right) + 2\left(y_2 + y_4 + \ldots + y_{n-2}\right)\Big) = I_n$$

Можно показать, что $|I - I_n| \leqslant \frac{M}{n^4}$, где $M = const$.

IV. Метод Монте—Карло

В качестве узлов берутся псевдослучайные числа. Пусть $I = \int_0^1 \varphi(x)\mathrm{d}x$. Ясно, что $[0;1] \to [a;b]$ при помощи линейной замены. Обозначим $\eta_i \in U(0;1)$ — значение датчика случайных чисел, $f_{\eta_i}(x) \equiv 1, x \in [0;1]$. Пусть $\xi_i = \varphi(\eta_i)$, тогда

$$\mathbb{E}(\xi_i) = \int_{-\infty}^{\infty} \varphi(x)\cdot f_{\eta_i}(x)\mathrm{d}x = \int_0^1 \varphi(x)\mathrm{d}x = I$$

По методу Монте—Карло получаем, что

$$I \approx \widehat{I}_n = \frac{1}{n}\sum_{i=1}^{n}\varphi(\eta_i)$$

где η_i это значения датчика случайных чисел. Погрешность вычислений не будет превосходить

$$|I - I_n| \leqslant \frac{3\sqrt{\mathbb{D}(\xi_1)}}{\sqrt{n}} \qquad \mathbb{D}(\xi_1) = \int_0^1 \varphi^2(x)\mathrm{d}x - I^2$$

Замечание 1.15.1. Скорость сходимости намного хуже, чем в квадратурных формулах. При этом для оценки погрешности потребуется вычислить (или оценить сверху) дополнительный интеграл, поэтому метод Монте—Карло не используется для вычисления определенных интегралов.

Кратные интегралы

Замечание 1.15.2. При вычислении k-кратных интегралов число узлов сетки возрастает как n^k и аналог метода прямоугольников будет довольно трудоемким. Таким образом метод Монте—Карло становится уместным, т.к. он не зависит от размерности.

Для вычисления интеграла

$$I = \int_0^1 \cdots \int_0^1 \varphi\left(x_1, \dots x_k\right) \mathrm{d}x_1 \dots \mathrm{d}x_k$$

достаточно случайно бросить n точек в данный k-мерный единичный куб. Для этого последовательные k значений датчика берем в качестве координат соответствующей случайной точки. Вычисляем значение подынтегральной функции в этих случайных точках и получаем

$$I \approx I_n = \frac{1}{n} \sum_{i=1}^{n} \varphi\left(x_{i_1}, \dots, x_{i_k}\right)$$

При этом скорость сходимости остается равной $\frac{1}{\sqrt{n}}$. В частности, если функция $\varphi = I_D$ это индикатор области D, то с помощью метода Монте—Карло можно приблизительно найти объем области D.

Пример 1.15.3. Пусть необходимо оценить площадь четверти единичной окружности. Генерируем последовательность псевдослучайных чисел и разбиваем их на пары $\langle x_{i_1}, x_{i_2} \rangle$, которые будут координатами точек.

$$I_D = \begin{cases} 0, & x_{i_1}^2 + x_{i_2}^2 > 1 \\ 1, & x_{i_1}^2 + x_{i_2}^2 \leqslant 1 \end{cases}$$

Итого $S = \frac{n_D}{n}$, где n_D это число точек, попавших в область D.

Метод расслоенной выборки

Пусть имеется k-мерный интеграл

$$I = \int_0^1 \cdots \int_0^1 \varphi\left(x_1, \dots x_k\right) \mathrm{d}x_1 \dots \mathrm{d}x_k$$

Каждую из сторон k-мерного куба разобьем N равных частей, тогда куб разобьется на $n = N^k$ маленьких кубиков Δ_i со стороной $\frac{1}{N}$. В каждом из этих кубиков возьмем случайную точку, построенную с помощью датчика случайных чисел, $\eta_i = \left(\eta_i^1, \dots, \eta_i^k\right) \in \Delta_i$, где $1 \leqslant i \leqslant n$. Интеграл оценивается при помощи суммы

$$I_n = \frac{1}{n} \sum_{i=1}^{n} \varphi\left(\eta_i\right)$$

При этом методе погрешность будет лучше, а именно

$$|I - I_n| \leqslant C \cdot \frac{1}{n^{\frac{1}{2} + \frac{1}{k}}} \qquad C = const$$

Замечание 1.15.4. В самых лучших квадратурных формулах для k-кратных интегралов погрешность составляет

$$\left|I - \widehat{I}\right| \leqslant C \cdot \frac{1}{n^{1+\varepsilon}} \qquad C = const$$

Равномерность по Вейлю

Def 1.15.5. Числовая последовательность $x_1, \dots, x_n$, где $x_i \in [0; 1]$ называется равномерной по Вейлю, если частота попадания точек x_i на любой отрезок $[a; b]$ стремится к его длине $b - a$ при $n \to \infty$.

Замечание 1.15.6. Ясно, что значения равномерного стандартного распределения или датчика случайных чисел будут равномерными по Вейлю.

Теорема 1.15.7. Пусть α это иррациональное число, тогда последовательность $x_n = \{n\alpha\}$ является равномерной по Вейлю.

Замечание 1.15.8. Если последовательность x_n равномерная по Вейлю, то

$$\frac{1}{n} \sum_{i=1}^{n} \varphi(x_i) \to \int_0^1 \varphi(x) \mathrm{d}x$$

но возможно что

$$\frac{1}{n}\sum_{i=1}^{n}\psi(x_{i_1},x_{i_2})\to C\neq\int_0^1\int_0^1\psi(x,y)\mathrm{d}x\mathrm{d}y$$

Def 1.15.9. Числовая последовательность $x_1,\dots,x_n,\dots$ называется вполне равномерной по Вейлю, если для произвольного числа k частота попадания k-мерных точек $\left\langle x_i^{(1)},\dots,x_i^{(k)}\right\rangle$ в любой параллелепипед внутри единичного куба стремится к его объему.

Замечание 1.15.10. Хороший датчик случайных чисел должен выдавать вполне равномерную последовательность. Мультипликативный датчик не является вполне равномерным.

Пример 1.15.11 (Парадокс первой цифры). Найти вероятность того, что число 2^n начинается с цифры 7.
Решение: Пусть 2^n начинается с цифры m. Это означает, что

$$\begin{aligned} m\cdot 10^l\leqslant 2^n<(m+1)\cdot 10^l\\ \log_{10}m+l\leqslant n\log_{10}2<\log_{10}(m+1)+l\in[l;l+1)\\ \log_{10}m\leqslant\{n\log_{10}2\}<\log_{10}(m+1)\end{aligned}$$

Согласно теореме 1.15.7 последовательность $\{n\log_{10}2\}$ является равномерной по Вейлю, следовательно вероятность того, что число начнется с цифры m стремится к длине интервала, т.е. к $\log_{10}(m+1)-\log_{10}m=\log_{10}\left(1+\frac{1}{m}\right)$. Итого при $m=7$ получаем $\log_{10}\left(1+\frac{1}{7}\right)\approx 0.058$.

1.16. Лекция 24.05.30.

Энтропия

Пусть случайная величина ξ это результат некоторого эксперимента с N исходами $A_1,\dots,A_N$. Вероятности этих исходов — $p_1,\dots,p_N$.

Def 1.16.1. Энтропией эксперимента называется величина

$$H(\xi)=-\sum_{i=1}^{N}p_i\log_2 p_i$$

При $p_i=0$ соответствующее слагаемое полагаем равным нулю.

Замечание 1.16.2. $H(\xi)\geqslant 0$, т.к. $\log_2 p_i<0$, то каждое слагаемое будет неположительным. Минус перед знаком суммы дает общую неотрицательность энтропии.

Замечание 1.16.3. $H(\xi)=0$ тогда и только тогда, когда $\exists i$ такое, что $p_i=1$ и $p_k=0$ при $k\neq i$. В этом случае результат эксперимента не случаен, а полностью предопределен (нет неопределенности результата).

Lm 1.16.4. Максимум энтропии, равный $\log_2 N$, достигается при $p_1=\ldots=p_N=\frac{1}{n}$.

□ Рассмотрим дискретную случайную величину η, значениями которой будут $p_1,\dots,p_N$, а вероятности этих значений все будут равны $\frac{1}{N}$. Рассмотрим функцию $\varphi(x)=x\log_2 x$ и найдем ее вторую производную.

$$\varphi''(x)=\left(\log_2 x+x\cdot\frac{1}{x\ln 2}\right)'=\frac{1}{x\ln 2}$$

Это больше нуля при $x>0$, значит эта функция выпукла вниз. Тогда по неравенству Йенсена получаем

$$\begin{aligned}\varphi\left(\mathbb{E}\left(\eta\right)\right)&\leqslant\mathbb{E}\left(\varphi(\eta)\right)\\ \varphi\left(\mathbb{E}\left(\eta\right)\right)=\varphi\left(\sum_{i=1}^{N}p_i\frac{1}{N}\right)&=\varphi\left(\frac{1}{N}\right)=-\frac{1}{N}\log_2 N\\ \mathbb{E}\left(\varphi(\eta)\right)=\sum_{i=1}^{N}\frac{1}{N}p_i\log_2 p_i=\frac{1}{N}\sum_{i=1}^{N}p_i\log_2 p_i&=\frac{1}{N}H(\xi)\\ -\frac{1}{N}\log_2 N&\leqslant\frac{1}{N}H(\xi)\\ \log_2 N&\geqslant H(\xi)\end{aligned}$$

■

Замечание 1.16.5. Энтропию можно рассматривать как количественную характеристику неопределенности эксперимента. Если $H=0$, то результат предопределен, а если $H=H_{\max}=\log_2 N$, то все исходы равновероятны и ни одному из них нельзя отдать предпочтение.

Пример 1.16.6. Пусть $\xi \in B_p$, тогда

$$H(\xi) = -(1-p)\log_2(1-p) - p\log_2 p$$

При $p = 0.5$ имеем

$$H(\xi) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$$

Пример 1.16.7. Перед испытуемым зажигалось n лампочек, и он должен был быстро указать на загоревшуюся лампочку. Оказалось, что самое длинное время реакции было в том случае, когда лампочки загорались с одинаковой частотой. Если частоты были разными, то время реакции было прямопропорционально энтропии эксперимента.

Кодирование информации

Пусть требуется передать сообщение из N символов. Пусть длины кодовых слов равны, тогда для каждого символа понадобится $\log_2 N$ бит, а для всего сообщения потребуется $N\log_2 N$ бит. Для больших по объему сообщений можно сократить число бит, учитывая что разные символы возникают с разной частотой. Пусть $p_1, \ldots, p_N$ — вероятности соответствующих N символов, тогда при длинном сообщении i-ый символ встречается $v_i = Np_i$ раз.

Def 1.16.8. Назовем сообщение типичным, если $\forall i\colon\ |v_i - Np_i| < \delta$.

Обозначим $M_{N,\delta}$ — число типичных сообщений.

Теорема 1.16.9. (Макмиллана (частный случай))

$$\frac{1}{N}\log_2 M_{N,\delta} \xrightarrow{n\to\infty} H$$

Отсюда число типичных сообщений $M_{N,\delta} \leqslant 2^{N(H+\varepsilon)}$, где $\varepsilon > 0$ — мало. Обозначим $H_0 = \log_2 N$, тогда с вероятностью, близкой к единице, можно сократить длину сообщения с коэффициентом сжатия $\gamma = \frac{H}{H_0}$. Если символы появляются независимо, то большее сжатие невозможно, а если использовать их зависимость, то коэффициент сжатия можно существенно уменьшить. Пусть γ_∞ — данный коэффициент сжатия. Для русского языка $\gamma \approx 0.87$, а $\gamma_\infty \approx 0.24$ (для литературного языка) и $\gamma_\infty \approx 0.17$ (для деловых сообщений).

Def 1.16.10. Величина $1 - \gamma_\infty$ называется избыточностью языка.

Энтропия абсолютно непрерывных распределений

Def 1.16.11. Пусть ξ это абсолютно непрерывная случайная величина с плотностью $f(x)$. Энтропией случайной величины ξ называется величина

$$H(\xi) = -\int_{-\infty}^{\infty} f(x)\log_2 f(x)\mathrm{d}x$$

Теорема 1.16.12. Следующие распределения имеют наибольшую энтропию:

1. Если $\xi \in [0;1]$, то $\xi \in U(0;1)$.
2. Если $\xi \in [0;+\infty)$ и $\mathbb{E}(\xi) = 1$, то $\xi \in E_1$.
3. Если $\xi \in (-\infty;+\infty)$, $\mathbb{E}(\xi) = 0$ и $\mathbb{D}(\xi) = 1$, то $\xi \in N(0;1)$.

Задача о разорении игрока

Играют два игрока. Вероятность выигрыша первого игрока — p, а второго — $1-p$. Ставка в каждой игре это одна единица. Капитал первого игрока равен k, а второго — $M-k$. Игра прекращается, когда один из игроков потеряет свой капитал. Найти вероятность разорения первого игрока.
Выигрыш первого игрока $S_n = z_1 + \ldots + z_n$, где $\mathsf{P}(z_i = 1) = p$ и $\mathsf{P}(z_i = -1) = q$. Пусть r_k это вероятность разорения при стартовом капитале k, тогда $r_k = pr_{k+1} + qr_{k-1}$. Это линейное однородное рекуррентное уравнение. Сначала рассмотрим случай $p \neq \frac{1}{2}$. Составим и решим характеристическое уравнение.

$$px^2 - x + (1-p) = 0 \Longrightarrow x_{1,2} = \frac{1 \pm \sqrt{1-4p(1-p)}}{2p} = \frac{1 \pm (1-2p)}{2p} = \begin{cases} \frac{1-p}{p} = \lambda \\ 1 \end{cases} \Longrightarrow r_k = c_1 1^k + c_2\lambda^k$$

Это общее решение. Чтобы найти константы подставим начальные (в данном случае их лучше назвать граничными) условия.

$$\begin{cases} r_0 = 1 = c_1 + c_2 \\ r_M = 0 = c_1 + c_2\lambda^M \end{cases} \Longleftrightarrow \begin{cases} c_1 = 1 - c_2 \\ 1 - c_2 + c_2\lambda^M = 0 \end{cases} \Longleftrightarrow \begin{cases} c_2 = \frac{1}{1-\lambda^M} \\ c_1 = \frac{-\lambda^M}{1-\lambda^M} \end{cases}$$

Таким образом получаем итоговое решение

$$r_k = \frac{-\lambda^M}{1-\lambda^M} + \frac{\lambda^k}{1-\lambda^M} = \frac{\lambda^k - \lambda^M}{1-\lambda^M}$$

Далее рассмотрим ситуацию $M \to \infty$. Возможны два случая.

1. $p < \frac{1}{2}$, тогда $\lambda > 1$ и $\lambda^M \to \infty$, значит

$$r_k = \frac{\lambda^k - \lambda^M}{1-\lambda^M} = \frac{\lambda^{k-M} - 1}{\lambda^{-M} - 1} = 1$$

   Таким образом игрок гарантированно разоряется.

2. $p > \frac{1}{2}$, тогда $\lambda < 1$ и $\lambda^M \to 0$, значит

$$r_k = \frac{\lambda^k - \lambda^M}{1-\lambda^M} = \lambda^k = \left(\frac{q}{p}\right)^k$$

Вернемся к решению характеристического уравнения и рассмотрим случай $p = \frac{1}{2}$. Тогда $x_1 = x_2 = 1$ и общее решение будет иметь вид

$$r_k = c_1 1^k + c_2 k \cdot 1^k = c_1 + c_2 k$$

Для нахождения констант подставим начальные условия.

$$\begin{cases} r_0 = 1 = c_1 \\ r_M = 0 = c_1 + c_2 M \end{cases} \iff \begin{cases} c_1 = 1 \\ c_2 = -\frac{1}{M} \end{cases} \implies r_k = 1 - \frac{k}{M}$$

При $M \to \infty$ получаем, что $r_k \to 1$. Значит вне зависимости от стартового капитала при бесконечной игре первый игрок рано или поздно разорится.