1 Точечные оценки. Их свойства: состоятельность, несмещенность, эффективность.

Пусть имеется выборка объема n: $X = \begin{pmatrix} X_1 & \dots & X_n \end{pmatrix}$

Определение. Статистикой называется измеримая функция $\theta^* = \theta^*(X_1, \dots, X_n)$.

Пусть требуется найти значение параметра θ случайной величины X по данной выборке. Оценку будем считать с помощью некоторой статистики θ^*.

1.1 Свойства статистических оценок

1.1.1 Состоятельность

Определение. Статистика $\theta^* = \theta^*(X_1, \dots, X_n)$ называется **состоятельной оценкой** параметра θ, если:

$$\theta^* \xrightarrow[n\to\infty]{P} \theta$$

1.1.2 Несмещённость

Определение. Статистика $\theta^* = \theta^*(X_1, \dots, X_n)$ называется **несмещенной оценкой** параметра θ, если

$$\mathbb{E}\,\theta^* = \theta$$

Примечание. То есть с равной вероятностью можем ошибиться как в меньшую, так и в большую сторону. Нет систематической ошибки.

Определение. Статистика $\theta^* = \theta^*(X_1, \dots, X_n)$ называется **асимптотически несмещенной оценкой** параметра θ, если

$$\mathbb{E}\,\theta^* \xrightarrow[n\to\infty]{} \theta$$

Примечание. То есть при достаточно большом объеме выборки ошибка исчезает, но при малом она может существовать.

1.1.3 Эффективность

Определение. Оценка θ_1^* **не хуже** оценки θ_2^*, если

$$\mathbb{E}(\theta_1^* - \theta)^2 \le \mathbb{E}(\theta_2^* - \theta)^2$$

или, если обе оценки несмещенные,

$$\begin{gathered} \mathbb{E}(\theta_1^* - \theta)^2 \le \mathbb{E}(\theta_2^* - \theta)^2 \\ \mathbb{E}(\theta_1^* - \mathbb{E}\,\theta_1^*)^2 \le \mathbb{E}(\theta_2^* - \mathbb{E}\,\theta_2^*)^2 \\ \mathbb{D}\,\theta_1^* \le \mathbb{D}\,\theta_2^* \end{gathered}$$

Определение. Оценка θ^* называется **эффективной**, если она не хуже всех остальных оценок.

Теорема 1. Не существует эффективной оценки в классе всех возможных оценок.

Теорема 2. В классе несмещённых оценок существует эффективная оценка.

2 Точечные оценки моментов. Свойства оценок математического ожидания и дисперсии.

Определение. Выборочным средним $\overline{X_B}$ называется величина

$$\boxed{\overline{X_B} = \frac{1}{n}\sum_{i=1}^{n} X_i}$$

Определение. Выборочной дисперсией $\mathbb{D}_B$ называется величина

$$\boxed{\mathbb{D}_B = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X_B})^2}$$

Определение. Исправленной выборочной дисперсией S^2 называется величина

$$\boxed{S^2 = \frac{n}{n-1}\,\mathbb{D}_B}$$

или

$$\boxed{S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X_B})^2}$$

Определение. Выборочным средним квадратическим отклонением называется величина

$$\boxed{\sigma_B = \sqrt{\mathbb{D}_B}}$$

Определение. Исправленным выборочным средним квадратическим отклонением называется величина

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

Определение. Выборочным k-тым моментом называется величина

$$\boxed{\overline{X^k} = \frac{1}{n}\sum_{i=1}^{n} X_i^k}$$

Теорема 3. Выборочное среднее $\overline{X_B}$ является несмещенной состоятельной оценкой для математического ожидания, то есть:

1. $\mathbb{E}\,\overline{X_B} = \mathbb{E}\,X = a$ — несмещенность
2. $\overline{X_B} \xrightarrow[n\to\infty]{P} \mathbb{E}\,X$ — состоятельность

Proof.

1. $$\mathbb{E}\,\overline{X} = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\,X_i = \frac{1}{n}\cdot n\,\mathbb{E}\,X = \mathbb{E}\,X$$
2. $$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} \xrightarrow[n\to\infty]{P} \mathbb{E}\,X$$
   Это верно по закону больших чисел.

□

Теорема 4. Выборочный k-тый момент является несмещенной состоятельной оценкой для теоретического k-того момента, то есть:

1. $\mathbb{E}\,\overline{X^k} = \mathbb{E}\,X^k$
2. $\overline{X^k} \xrightarrow{P} \mathbb{E}\,X^k$

Proof. Следует из предыдущей теоремы, если в качестве случайной величины взять X^k. □

Теорема 5.

- $\mathbb{D}_B$ — смещённая состоятельная оценка дисперсии
- S^2 — несмещённая состоятельная оценка дисперсии

Proof.

$$\mathbb{D}_B = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \overline{X^2} - (\overline{X})^2$$

$$\begin{aligned}
\mathbb{E}\,\mathbb{D}_B &= \\
\mathbb{E}(\overline{X^2} - (\overline{X})^2) &= \\
\mathbb{E}\,\overline{X^2} - \mathbb{E}(\overline{X})^2 &\overset{\text{по 4}}{=} \\
\mathbb{E}\,X^2 - \mathbb{E}(\overline{X})^2 &= \\
\mathbb{E}\,X^2 - (\mathbb{D}\,\overline{X} + (\mathbb{E}\,\overline{X})^2) &=
\end{aligned}$$

$$\mathbb{E}\,X^2 - (\mathbb{E}\,\overline{X})^2 - \mathbb{D}\,\overline{X} =$$
$$(\mathbb{E}\,X^2 - (\mathbb{E}\,X)^2) - \mathbb{D}\,\overline{X} =$$
$$\mathbb{D}\,X - \underbrace{\mathbb{D}\,\overline{X}}_{\text{величина отклонения}} =$$
$$\mathbb{D}\,X - \mathbb{D}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) =$$
$$\mathbb{D}\,X - \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{D}\,X_i =$$
$$\mathbb{D}\,X - \frac{1}{n^2}\cdot n\,\mathbb{D}\,X =$$
$$\mathbb{D}\,X - \frac{1}{n}\,\mathbb{D}\,X =$$
$$\frac{n-1}{n}\,\mathbb{D}\,X \neq \mathbb{D}\,X$$

$$\mathbb{E}\,S^2 = \mathbb{E}\left(\frac{n}{n-1}\,\mathbb{D}_B\right) = \frac{n}{n-1}\cdot\frac{n-1}{n}\,\mathbb{D}\,X = \mathbb{D}\,X$$
$$\mathbb{D}_B = \overline{X^2} - (\overline{X})^2 \xrightarrow{P} \mathbb{E}\,X^2 - (\mathbb{E}\,X)^2 = \mathbb{D}\,X$$
$$S^2 = \frac{n}{n-1}\,\mathbb{D}_B \xrightarrow{P} \underbrace{\frac{n}{n-1}}_{\to 1}\,\mathbb{D}\,X$$

□

Примечание. $\mathbb{D}_B$ — асимптотически несмещённая оценка, т.к. при $n \to \infty$, $\frac{n-1}{n} \to 1$. Таким образом, при большой[1] выборке можно игнорировать смещённость.

3 Метод моментов. Пример.

Изобретен Карлом Пирсоном.

Пусть имеется выборка $(X_1 \dots X_n)$ неизвестного распределения, при этом известен тип[2] распределения. Пусть этот тип определяется k неизвестными параметрами $\theta_1 \dots \theta_k$. Теоретическое распределение задает теоретические k-тые моменты. Например, если распределение непрерывное, то оно задается плотностью $f(X, \theta_1 \dots \theta_k)$ и

$$m_k = \int_{-\infty}^{+\infty} X^k f(x, \theta_1 \dots \theta_k)dx = h_k(\theta_1 \dots \theta_k)$$

[1] $n \geq 100$, например.
[2] Нормальное, показательное и т.д.

Метод моментов состоит в следующем: вычисляем выборочные моменты и подставляем их в эти равенства вместо теоретических. В результате получаем систему уравнений:

$$\begin{cases} \overline{X} = h_1(\theta_1 \dots \theta_k) \\ \overline{X^2} = h_2(\theta_1 \dots \theta_k) \\ \vdots \\ \overline{X^k} = h_k(\theta_1 \dots \theta_k) \end{cases}$$

Решив эту систему, мы получим оценки на $\theta_1 \dots \theta_k$. Эти оценки будут состоятельными[3], но смещёнными.

Пример. Пусть $X \in U(a,b), a < b$. Обработав статданные, получили оценки первого и второго момента: $\overline{X} = 2.25; \overline{X^2} = 6.75$

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

$$\mathbb{E}\, X = \int_a^b x f(x) dx = \int_a^b \frac{x}{b-a} = \frac{1}{b-a} \cdot \frac{x^2}{2}\Big|_a^b = \frac{b^2-a^2}{2(b-a)} = \boxed{\frac{a+b}{2}}$$

$$\mathbb{E}\, X^2 = \int_a^b x^2 \cdot \frac{1}{b-a} dx = \frac{1}{b-a} \cdot \frac{x^3}{3}\Big|_a^b = \frac{b^3-a^3}{3(b-a)} = \boxed{\frac{a^2+ab+b^2}{3}}$$

$$\begin{cases} 2.25 = \frac{a+b}{2} \\ 6.75 = \frac{a^2+ab+b^2}{3} \end{cases}$$

$$\begin{cases} a+b = 4.5 \\ a^2+ab+b^2 = 20.25 \end{cases}$$

$$\begin{cases} a+b = 4.5 \\ ab = 0 \end{cases}$$

$$\begin{cases} a = 0 \\ b = 4.5 \end{cases}$$

□

[3] Если не придумывать специально плохие примеры

4 Метод максимального правдоподобия. Пример.

Метод максимального правдоподобия состоит в том, чтобы подобрать параметры таким образом, чтобы вероятность получения данной выборки была наибольшей. Если распределение дискретное, то вероятность выборки

$$P_\theta(X_1 = x_1, X_2 = x_2 \dots X_n = x_n) = P_\theta(X_1 = x_1)P_\theta(X_2 = x_2) \dots P_\theta(X_n = x_n)$$

Для непрерывной величины аналогично.

Поэтому исследуем такую функцию:

Определение. Функцией правдоподобия называется функция $L(\overline{X}, \theta)$, зависящая от выборки и неизвестных параметров, равная:

- В случае дискретного распределения:

$$P_\theta(X_1 = x_1)P_\theta(X_2 = x_2) \dots P_\theta(X_n = x_n)$$

- В случае абсолютно непрерывного распределения:

$$f_\theta(x_1)f_\theta(x_2) \dots f_\theta(x_n) = \prod_{i=1}^{n} f_\theta(x_i)$$

Эту функцию неудобно исследовать, поэтому мы используем следующую функцию:

Определение. Логарифмическая функция правдоподобия:

$$M(\overline{X}, \theta) = \ln L(\overline{X}, \theta)$$

Т.к. логарифм — строго возрастающая функция, экстремумы обычной и логарифмической функций правдоподобия совпадают.

Определение. Оценкой максимального правдоподобия $\hat{\theta}$ называется значение θ, при котором функция правдоподобия достигает наибольшего значения.

Пример. Пусть $X_1 \dots X_n$ — выборка неизвестного распределения Пуассона с параметром λ: $X \in \Pi_\lambda, \lambda > 0$

$$P(X = x_i) = \frac{\lambda^{x_i}}{x_i!}e^{-\lambda}$$

$$L(\overline{X}, \lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!}e^{-\lambda} = \frac{\lambda^{n \cdot \overline{X}}}{\prod_{i=1}^{n} x_i!}e^{-n\lambda}$$

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

$$\frac{\partial \ln L(\overline{X}, \lambda)}{\partial \lambda} = \frac{n\overline{X}}{\lambda} - n$$

Приравняем производную к нулю, чтобы найти точки экстремума:

$$\frac{n\overline{X}}{\lambda} - n = 0 \Rightarrow \lambda = \overline{X}$$

Таким образом, $\hat{\theta} = \overline{X}$ — ОМП.

Пример. Пусть $X_1 \dots X_n$ — выборка неизвестного нормального распределения: $X \in N(a, \sigma^2), a \in \mathbb{R}, \sigma > 0$

$$f_{a,\sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}}$$

$$L(\overline{X}, a, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-a)^2}{2\sigma^2}} = \frac{1}{\sigma^n \sqrt{2\pi}^n} e^{-\frac{\sum (x_i-a)^2}{2\sigma^2}}$$

$$\ln L(\overline{X}, a, \sigma^2) = n \ln \sigma - \frac{n}{2} \ln(2\pi) - \frac{1}{2\sigma^2} \sum (x_i - a)^2$$

Не дописано

Пример. Пусть $X_1 \dots X_n$ — выборка равномерного распределения вида $U(0, \theta)$

1. Метод моментов.
$$\mathbb{E} = \frac{a+b}{2} = \frac{\theta}{2} \Rightarrow \theta^* = 2\overline{X}$$

2. Метод максимального правдоподобия.

$$f_\theta(x) = \begin{cases} 0, & x < 0 \\ \frac{1}{\theta}, & x \in [0, \theta] \\ 0, & x > \theta \end{cases}$$

$$L(\overline{X}, \theta) = \prod_{i=1}^{n} f_\theta(x_i) = \begin{cases} 0, & \theta < \max x_i = X_{(n)} \\ \frac{1}{\theta^n}, & \theta \geq X_{(n)} \end{cases}$$

L достигает наибольшего значения при $\hat{\theta} = X_{(n)}$.

Примечание. ОМП состоятельны, часто эффективны, но могут быть смещенными.

5 Информация Фишера. Неравенство Рао-Крамера (без док–ва).

Пусть известно, что случайная величина $X \in \mathcal{F}_\theta$ — семейству распределений с θ.

Определение. Носителем семейства распределений $\mathcal{F}_\theta$ называется множество $C \subset \mathbb{R}$, такое что $\forall \theta \;\; P(X \in C) = 1$.

Обозначение.

$$f_\theta(x) = \begin{cases} f_\theta(x), & \text{если распределение абсолютно непрерывное} \\ P_\theta(X = x), & \text{если распределение дискретное} \end{cases}$$

Определение. Информацией Фишера называется величина

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

, если она существует.

Определение. Семейство распределений $\mathcal{F}_\theta$ называется **регулярным**, если:

1. Существует носитель C семейства $\mathcal{F}_\theta$, такой что $\forall x \in C$ функция $\ln f_\theta(x)$ непрерывно дифференцируема по θ.
2. $I(\theta)$ существует и непрерывна по θ.

Теорема 6 (неравенство Рао-Крамера)**.** Пусть $X_1 \dots X_n$ — выборка объема n из регулярного семейства распределений $\mathcal{F}_\theta$, θ^* — несмещенная оценка параметра θ, дисперсия которой ограничена на любом компакте в области θ.

Тогда

$$\boxed{\mathbb{D}\,\theta^* \geq \frac{1}{nI(\theta)}}$$

Следствие 6.1. Если при условиях выше $\mathbb{D}\,\theta^* = \frac{1}{nI(\theta)}$, то θ^* — эффективная оценка. Это не всегда достижимо.

6 Основные распределения математической статистики: хи–квадрат, Стьюдента, Фишера-Снедекора. Их свойства.

6.1 Нормальное распределение

$X \in N(a, \sigma^2)$:

$$\mathbb{E}\,X = a, \mathbb{D}\,X = \sigma^2$$

$N(0, 1)$ — стандартное нормальное распределение

6.2 Гамма-распределение

$X \in \Gamma_{\alpha,\lambda}$, если её плотность равна:

$$f(x) = \begin{cases} 0, & x \leq 0 \\ \frac{\alpha^\lambda}{\Gamma(\lambda)} x^{\lambda-1} e^{-\alpha x}, & x > 0 \end{cases}$$

Свойства.

1. $\mathbb{E}\,\xi = \frac{\lambda}{\alpha}, \mathbb{D}\,\xi = \frac{\lambda}{\alpha^2}$
2. Если $\xi_1 \in \Gamma_{\alpha,\lambda_1}, \xi_2 \in \Gamma_{\alpha,\lambda_2}$, то $\xi_1 + \xi_2 \in \Gamma_{\alpha,\lambda_1+\lambda_2}$
3. $\Gamma_{\alpha,1} = E_\alpha$ — показательное распределение.
4. Если $X_i \in E_\alpha$, то $\sum_{i=1}^n X_i \in \Gamma_{\alpha,n}$
5. Если $X \in N(0,1)$, то $X^2 \in \Gamma_{\frac{1}{2},\frac{1}{2}}$

Примечание. Гамма-распределение возникает в матстатистике как распределение квадрата стандартно нормально распределенной величины. Обобщим эту идею:

6.3 Распределение “хи-квадрат”

Определение. Распределением **хи-квадрат** с k степенями свободы называется распределение суммы k квадратов независимых стандартных нормальных величин.

$$\chi^2_k = X_1^2 + X_2^2 + \cdots + X_k^2, \quad X_i \in N(0,1)$$

Обозначение. $\chi^2 \in H_k$

Свойства.

1. $\chi^2_k \in \Gamma_{\frac{1}{2},\frac{k}{2}}$
2. $\chi^2_n + \chi^2_m = \chi^2_{n+m}$ — по определению
3. $\mathbb{E}\,\chi^2_k = \frac{\lambda}{\alpha} = \frac{\frac{k}{2}}{\frac{1}{2}} = k, \mathbb{D}\,\chi^2_k = \frac{\lambda}{\alpha^2} = \frac{\frac{k}{2}}{\left(\frac{1}{2}\right)^2} = 2k$

6.4 Распределение Стьюдента

Определение. Пусть случайные величины $X_0, X_1 \dots X_k$ — независимы и имеют стандартное нормальное распределение. Распределением **Стьюдента** с k степеней свободы называется распределение случайной величины

$$t_k = \frac{X_0}{\sqrt{\frac{1}{k}(X_1^2 + \cdots + X_k^2)}} = \frac{X_0}{\sqrt{\frac{1}{k}\chi^2_k}}$$

Свойства.

1. $\mathbb{E}\,t_k = 0$
2. $\mathbb{D}\,t_k = \frac{k}{k-2}$

6.5 Распределение Фишера-Снедекора

Определение. Распределение $F_{m,n}$ называется распределением **Фишера-Снедекора** *(или F-**распределением**)* со степенями свободы m и n называется распределение случайной величины

$$f_{m,n} = \frac{\frac{\chi^2_m}{m}}{\frac{\chi^2_n}{n}}$$

, где χ^2_n и χ^2_m — независимые случайные величины с распределением χ^2.

Свойства.

1. $\mathbb{E}\, f_{m,n} = \frac{n}{n-2}$
2. $\mathbb{D}\, f_{m,n} = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}$
3. $F_{m,n}(x) = P(f_{m,n} < x) = P\left(\frac{1}{f_{m,n}} > \frac{1}{x}\right) = P\left(f_{n,m} > \frac{1}{x}\right) = 1 - F_{n,m}\left(\frac{1}{x}\right)$

При $n, k, m \to \infty$ эти распределения слабо сходятся к нормальному. При $n > 30$ они достаточно близки.

7 Линейные преобразования нормальных выборок. Теорема об ортогональном преобразовании.

Пусть $\vec{X} = (X_1 \dots X_n)$, где $X_i \in N(0,1)$ и независимы. Будем рассматривать линейные комбинации этого вектора. Пусть A — невырожденная матрица размера $n \times n$. Рассмотрим случайный вектор $\vec{Y} = A\vec{X}$, где координаты случайного вектора $Y_i = a_{i1}X_1 + \dots + a_{in}X_n$. Будем исследовать, что из себя представляют Y_i и их совместное распределение.

Примечание. Если $\eta = a\xi + b$, то $f_\eta(\xi) = \frac{1}{|a|} f_\xi\left(\frac{\xi - b}{a}\right)$

Теорема 7. Пусть случайный вектор $\vec{X}$ имеет плотность распределения $f_{\vec{X}}(\vec{x})$ и A невырожденная матрица.

Тогда случайный вектор $\vec{Y} = A\vec{X} + \vec{b}$ имеет плотность

$$f_{\vec{Y}}(\vec{y}) = \frac{1}{|\det A|} \cdot f_{\vec{X}}(A^{-1}(\vec{y} - \vec{b}))$$

Примечание. $f_{\vec{X}}(\vec{x})$ — плотность $\vec{X}$, если $P(\vec{x} \in B) = \int \dots \int_B f_{\vec{X}}(\vec{x}) d\vec{x}$

Proof.

$$\begin{aligned} P(\vec{y} \in B) &= P(A\vec{x} + \vec{b} \in B) \\ &= P(\vec{x} \in A^{-1}(\vec{y} - \vec{b})) \end{aligned}$$

$$= \int \cdots \int_{A^{-1}(B-\vec{b})} f_{\vec{x}}(x) d\vec{x}$$

Сделаем замену $\vec{y} = A\vec{x} + \vec{b}$. Тогда $A^{-1}(B - \vec{b})$ перейдёт в B, $\vec{x}$ перейдёт в $A^{-1}(\vec{y} - \vec{b}), \vec{y} \in B$, $d\vec{x}$ перейдёт $|J| d\vec{y}$, где $J = |A^{-1}| = |A|^{-1}$

Итого:

$$= \int \cdots \int_B f(A^{-1}(\vec{y} - \vec{b})) \cdot \frac{1}{|\det A|} d\vec{y} \Rightarrow f_{\vec{Y}}(\vec{y}) = \frac{1}{|\det A|} f_{\vec{X}}(A^{-1}(\vec{y} - \vec{b}))$$

□

Определение. $A = C$ — **ортогональна**, т.е. $C^{\mathsf{T}} = C^{-1}, |\det C| = 1$

Теорема 8. Пусть дан случайный вектор $\vec{X} = (X_1 \dots X_n)$, где $\forall i \;\; X_i \in N(0,1)$ и X_i независимы, а C — ортогональная матрица.

Тогда координаты случайного вектора $\vec{Y} = C\vec{X}$ независимы и также имеют стандартное нормальное распределение.

Proof. Т.к. координаты $X_i \in N(0,1)$ и независимы, то плотность $\vec{X}$:

$$f_{\vec{X}}(\vec{x}) = \prod_{i=1}^{n} f_i(x_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{x_i^2}{2}} = \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}(x_1^2 + x_2^2 + \cdots + x_n^2)} = \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}||\vec{x}||^2}$$

По предыдущей теореме:

$$f_{\vec{Y}}(\vec{y}) = f_{\vec{X}}(C^{\mathsf{T}}\vec{y}) = \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}||C^{\mathsf{T}}\vec{y}||^2} = \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}||\vec{y}||^2} = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y_i^2} = \prod_{i=1}^{n} f_i(y_i)$$

Следовательно, $Y_i \in N(0,1)$ и независимы. □

8 Лемма Фишера.

Лемма 1 (Фишера)**.** Пусть случайный вектор $\vec{X}$ состоит из независимых стандартных нормальных случайных величин, $\vec{Y} = C\vec{X}$, где C — ортогональная матрица. Тогда $\forall k : 1 \le k \le n-1$ случайная величина

$$T(\vec{X}) = \left(\sum_{i=1}^{n} X_i^2 \right) - Y_1^2 - \cdots - Y_k^2$$

не зависит от случайного вектора $Y_1 \dots Y_k$ и имеет распределение H_{n-k}

Proof. Т.к. C ортогональна:

$$||\vec{Y}||^2 = ||C\vec{X}||^2 = ||\vec{X}||^2 = X_1^2 + \cdots + X_n^2 = Y_1^2 + \cdots + Y_n^2$$

Отсюда

$$T(\vec{X}) = \sum_{i=1}^{n} Y_i^2 - Y_1^2 - \cdots - Y_k^2 = Y_{k+1}^2 + \cdots + Y_n^2$$

$Y_{k+1} \dots Y_n$ — независимы, имеют стандартное нормальное распределение и $T(\vec{X}) \in H_{n-k}$ по определению распределения χ^2.

$T(\vec{X})$ не зависит от $Y_1 \dots Y_k$, т.к. $Y_{k+1} \dots Y_n$ по предыдущей лемме от них не зависят. □

9 Основная теорема о связи точечных оценок нормального распределения и основных распределений статистики.

Теорема 9 (основная)**.**

- $X_1 \dots X_k$ независимы и имеют нормальное распределение с параметрами a и σ^2
- $\overline{X}$ — выборочное среднее
- S^2 — исправленная выборочная дисперсия

Тогда имеют место следующие распределения:

1. $$\sqrt{n} \cdot \frac{\overline{X} - a}{\sigma} \in N(0,1)$$
2. $$\sum_{i=1}^{n} \left(\frac{X_i - a}{\sigma} \right)^2 \in H_n$$
3. $$\sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{\sigma} \right)^2 = \frac{(n-1)S^2}{\sigma^2} \in H_{n-1}$$
4. $$\sqrt{n} \cdot \frac{\overline{X} - a}{S} \in T_{n-1}$$
5. $\overline{X}$ и S^2 — независимые случайные величины

Proof.

1.

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

2. Верно, т.к. $\dfrac{X_i - a}{\sigma} \in N(0,1)$

3.

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

, где

$$z_i = \frac{X_i - a}{\sigma} \in N(0,1), \overline{z} = \frac{\sum_{i=1}^{n} z_i}{n} = \frac{\sum X_i - na}{\sigma n} = \frac{\overline{X} - a}{\sigma}$$

Поэтому можем считать, что $X_i \in N(0,1)$. Применим лемму Фишера.

$$T(\vec{X}) = \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 = n\,\mathbb{D}_B = n(\overline{X^2} - (\overline{X})^2) = \sum_{i=1}^{n} X_i^2 - n(\overline{X})^2 = \sum_{i=1}^{n} X_i^2 - Y_1^2$$

, где

$$Y_1^2 = n(\overline{X})^2 \quad Y_1 = \sqrt{n}\overline{X} = \frac{1}{\sqrt{n}}X_1 + \cdots + \frac{1}{\sqrt{n}}X_n$$

Так как длина[4] строки $\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}$ равна 1, можем[5] дополнить эту строку до ортогональной матрицы C. Тогда Y_1 — первая координата случайного вектора $\vec{Y} = C\vec{X}$ и по лемме Фишера $T(\vec{X}) \in H_{n-1}$

5. $T(\vec{X}) = \frac{(n-1)S^2}{\sigma^2}$ не зависит от $Y_1 = \sqrt{n}\overline{X} \Rightarrow S^2$ и $\overline{X}$ независимы.

4.

$$\sqrt{n}\frac{\overline{X} - a}{S} = \sqrt{n}\frac{\overline{X} - a}{\sigma} \cdot \frac{1}{\sqrt{\frac{(n-1)S^2}{\sigma^2} \cdot \frac{1}{n-1}}} = \frac{X_0}{\sqrt{\frac{\chi^2_{n-1}}{n-1}}} \in T_{n-1}$$, т.к.:

$X_0 \in N(0,1)$ по пункту 1, $\frac{(n-1)S^2}{\sigma^2} \in H_{n-1}$ по пункту 3 и X_0 не зависит от $\frac{(n-1)S^2}{\sigma^2}$ по пункту 5.

□

[4] То есть норма строки как вектора.

[5] По некой теореме.

10 Квантили распределений (оба определения). Функции для их вычисления в EXCEL.

Для простоты предполагаем, что все распределения непрерывные.

Определение (1). Число t_γ называется **квантилем**[6] уровня γ, если $F(t_\gamma) = \gamma$.

С точки зрения геометрии $P(X \in$ область слева от $t_\gamma) = \gamma$.

Примечание.

- Медиана — квантиль уровня $\frac{1}{2}$
- Квартили — квантили уровня $\frac{1}{4}, \frac{2}{4}, \frac{3}{4}$
- Децили — квантили уровня $\frac{1}{10}, \frac{2}{10}, \dots$

Примечание. Квантиль t_γ — значение обратной функции распределения: $t_\gamma = F^{-1}(\gamma)$

Определение (2 (альтернативное)). Число t_α называется **квантилем уровня значимости** α, если $F(t_\alpha) = 1 - \alpha$.

Примечание. $\alpha = 1 - \gamma$

10.1 Квантили основных распределений в Excel

1. НОРМ.СТ.ОБР.

$$F_0(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{z^2}{2}} dz$$

   Тогда $\dots(x + 0.5)$ — обратная функция функции Лапласа $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{z^2}{2}} dz$

2. (a) СТЬЮДЕНТ.ОБР. — обратная к функции распределения Стьюдента стандартной величины.

$$t_k = \frac{X_0}{\sqrt{\frac{1}{k}\chi_k^2}}$$

   (b) СТЬЮДЕНТ.ОБР.2Х

   Возвращает t_α, такое что $P(|X| > t_\alpha) = \alpha$. Отсюда $P(|X| < t_\alpha) = 1 - \alpha$ и применяем СТЬЮДЕНТ.ОБР.2Х$(1 - \alpha, k)$

3. (a) ХИ2.ОБР. — возвращает квантиль t_γ в первом смысле для распределения χ^2.

   (b) ХИ2.ОБР.ПХ — возвращает квантиль t_α

4. (a) F.ОБР. — возвращает квантиль t_γ F-распределения

[6] Или квантилью.

(b) F.ОБР.ПХ — возвращает квантиль t_α F-распределения

11 Интервальные оценки. Определения, смысл, терминология.

Недостаток точных оценок в том, что мы не знаем, насколько точная наша оценка.

Пусть требуется дать оценку неизвестного параметра θ.

Определение. Интервал $(\theta_\gamma^-, \theta_\gamma^+)$ называется **доверительным интервалом** для параметра θ надежности γ, если $P(\theta_\gamma^- < \theta < \theta_\gamma^+) = \gamma$

Примечание. Если θ — параметр дискретного распределения, то будет правильней написать $P(\theta_\gamma^- < \theta < \theta_\gamma^+) \geq \gamma$.

Примечание. Здесь случайные величины — интервальные оценки, а не θ. Поэтому более культурно говорить так: интервал $(\theta_\gamma^-, \theta_\gamma^+)$ накрывает неизвестный параметр θ с вероятностью γ.[7]

Примечание. В экономике γ берется 0.95, но можно брать и меньше — 0.9. Для чего-либо важного берется 0.99 или даже 0.999. Уровень надёжности выбирается в зависимости от решаемой задачи. Стандартные уровни: $0.9, 0.95, 0.99, 0.999$.

Кажется, тут недостаточно написано.

12 Доверительный интервал для математического ожидания нормального распределения при известном σ.

По пункту 1 теоремы 9:

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

, где $\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{z^2}{2}}dz$. Тогда t_γ — значение обратной к Φ в точке $\frac{\gamma}{2}$.

Осталось решить неравенство относительно a.

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

Итак, получили доверительный интервал для параметра a: $\left(\overline{X} - t_\gamma \cdot \frac{\sigma}{\sqrt{n}}, \overline{X} + t_\gamma \cdot \frac{\sigma}{\sqrt{n}}\right)$

[7] А не "θ попадает в интервал $(\theta_\gamma^-, \theta_\gamma^+)$ с вероятностью γ"

13 Доверительный интервал для математического ожидания нормального распределения при неизвестном σ.

По пункту 4 теоремы 9:

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

$F_{T_{n-1}}(t_\gamma) = \frac{1+\gamma}{2}$, т.е. t_γ — квантиль распределения Стьюдента T_{n-1} уровня $\frac{1+\gamma}{2}$.

Примечание. Если ξ — симметрично, то $P(|\xi| < t) = 2F(t) - 1$

Proof.

$$P(|\xi| < t) = 2P(0 < \xi < t) = 2(F(t) - F(0)) = 2F(t) - 1$$

□

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

14 Доверительный интервал для дисперсии нормального распределения при неизвестном a.

По пункту 2 теоремы 9 $\frac{(n-1)S^2}{\sigma^2} \in H_{n-1}$. Пусть χ_1^2 и χ_2^2 — квантили распределения H_{n-1} уровней $\frac{1-\gamma}{2}$ и $\frac{1+\gamma}{2}$[8]. Тогда:

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

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

Итак, доверительный интервал для параметра σ^2 надежности γ есть $\left(\frac{(n-1)S^2}{\chi_2^2}, \frac{(n-1)S^2}{\chi_1^2}\right)$, где χ_1^2 и χ_2^2 — квантили уровней $\frac{1-\gamma}{2}$ и $\frac{1+\gamma}{2}$. Следовательно, доверительный интервал для σ это $\left(\frac{\sqrt{n-1}S}{\chi_2}, \frac{\sqrt{n-1}S}{\chi_1}\right)$.

[8] На лекции было сказано $1-\frac{\gamma}{2}$ и $1+\frac{\gamma}{2}$, но это бред.

Этот интервал почти всегда не симметричен, можно его сделать симметричным, но мы этого делать не будем.

15 Доверительный интервал для дисперсии нормального распределения при известном a.

По пункту 3 теоремы $\frac{n\sigma^{2*}}{\sigma^2} \in H_n$, где $\sigma^{2*} = \frac{1}{n}\sum_{i=1}^{n}(X_i - a)^2$. Пусть χ_1^2 и χ_2^2 — квантили распределения H_n уровней $\frac{1-\gamma}{2}$ и $\frac{1+\gamma}{2}$ соответственно.

$$P\left(\chi_1^2 < \frac{n\sigma^{2*}}{\sigma^2} < \sigma_2^2\right) = F_{H_n}(\chi_2^2) - F_{H_n}(\chi_1^2) = \gamma$$

$$\chi_1^2 < \frac{n\sigma^{2*}}{\sigma^2} < \chi_2^2$$

$$\frac{n\sigma^{2*}}{\chi_2^2} < \sigma^2 < \frac{n\sigma^{2*}}{\chi_1^2}$$

Итак, доверительный интервал для σ^2 надежности γ это $\left(\frac{n\sigma^{2*}}{\chi_2^2}, \frac{n\sigma^{2*}}{\chi_1^2}\right)$, где χ_1^2 и χ_2^2 — квантили H_n уровней $\frac{1-\gamma}{2}$ и $\frac{1+\gamma}{2}$, $\sigma^{2*} = \frac{1}{n}\sum_{i=1}^{n}(X_i - a)^2$.

16 Проверка статистических гипотез. Определения, терминология. Уровень значимости и мощность критерия.

Определение. Гипотезой H называется предположение о распределении наблюдаемой случайной величины.

Определение. Гипотеза называется **простой**, если она однозначно определяет распределение, т.е. $H : \mathcal{F} = \mathcal{F}_1$, где $\mathcal{F}_1$ — распределение известного типа с известными параметрами.

Определение. Все остальные гипотезы называются **сложными**, т.к. они являются объединением конечного или бесконечного числа простых гипотез.

Определение (основная модель гипотез)**.** Гипотеза $H_1 = \overline{H_0}$ — конкурирующая *(альтернативная)* гипотеза, состоящая в том, что основная гипотеза H_0 неверна.

Примечание. С помощью статистических методов нельзя доказать гипотезу, можно только сказать, что она верна с некоторой уверенностью.

Основная гипотеза H_0 принимается или отклоняется с помощью статистики критерия K:

$$K(X_0 \dots X_n) \to \mathbb{R} = S \cup^{9} \overline{S} \to (H_0, H_1)$$

$$\begin{cases} H_0, & \text{если } K \in \overline{S} \\ H_1, & \text{если } K \in S \end{cases}$$

[9] Объединение на самом деле дизъюнктно.

Определение. Если точка находится на границе областей S и $\overline{S}$, она называется **критической**.

Определение. Ошибка I рода состоит в том, что нулевая гипотеза отвергается, когда она верна.

Определение. Ошибка II рода состоит в том, что отвергается альтернативная, когда она верна.

Определение. α — вероятность ошибки I рода, β — вероятность ошибки II рода

Пример. H_0 — деталь годная, H_1 — деталь бракованная.

Ошибка I рода — признать годную деталь бракованной.

Ошибка II рода — признать бракованную деталь годной.

Примечание. При росте выборки вероятности ошибок уменьшаются, при уменьшении вероятности одной ошибки другая вероятность увеличивается.

Определение. Критерий K называется **критерием асимптотического уровня** ε, если вероятность ошибки первого рода α стремится к ε при $n \to \infty$.

Определение. Критерий K для проверки гипотезы H_0 против альтернативы $H_1 = \overline{H_0}$ называется **состоятельным**, если вероятность ошибки II рода $\beta \to 0$ при $n \to \infty$.

Определение. Критерием **согласия** уровня ε называются состоятельные критерии асимптотического уровня ε.

17 Способы сравнения критериев проверки гипотез.

Пусть имеются критерии K_1 и K_2, $\alpha_1, \beta_1, \alpha_2, \beta_2$ — вероятности ошибок при соответствующих критериях, h_1 — потери в результате ошибке I рода, h_2 — потери в результате ошибки II рода.

Тогда рассмотрим способы сравнения критериев:

1. Минимакс: K_1 не хуже, чем K_2, если $\max(\alpha_1 h_1, \beta_1 h_2) \leq \max(\alpha_2 h_1, \beta_2 h_2)$
2. Критерий называется **баесовским**, если средние ожидаемые потери $U = \alpha h_1 + \beta h_2$ является минимальным.
3. Пусть ε — допустимый уровень ошибки I рода. Обозначим $K_\varepsilon \coloneqq \{K_i \mid \alpha_i \leq \varepsilon\}$.

   Определение. Критерий $K \in K_\varepsilon$ называется **наиболее мощным** критерием уровня ε, если $\beta \leq \beta_i \ \forall i$.

18 Построение критериев согласия (основные принципы).

В качестве критериев согласия берётся статистика $K(X_1 \dots X_n)$ со свойствами:

1. Если H_0 верна, то $K(X_1 \dots X_n) \rightrightarrows Z$ — известное распределение с известными параметрами.
2. Если H_0 не верна, то $K(X_1 \dots X_n) \xrightarrow{P} \infty$

Для заданного уровня значимости ε находим константу t_k, такую что $P(|Z| \geq t_k) = \alpha$. В результате получаем критерий согласия уровня значимости $\alpha = \varepsilon$:

$$\begin{cases} H_0, & |K| < t_k \\ H_1, & |K| \geq t_k \end{cases}$$

Теорема 10. Этот критерий является критерием согласия.

Proof.

1. K — критерий асимптотического уровня ε:

   Пусть H_0 верна. Тогда по построению $K \rightrightarrows Z$, т.е. $F_K(x) \to F_Z(x)$ и

   $$\begin{aligned} \alpha &= P(|K| \geq t_k \mid H_0) \\ &= 1 - P(|K| < t_k) \\ &= 1 - (F_K(t_k) - F_K(-t_k)) \\ &\xrightarrow[n\to\infty]{} 1 - (F_Z(t_k) - F_Z(-t_k)) \\ &= P(|Z| \geq t_k) \\ &= \varepsilon \end{aligned}$$

2. K — состоятельный критерий:

   Пусть H_1 верна. Тогда $K(X_1 \dots X_n) \xrightarrow{P} \infty$, т.е.

   $$\forall C \;\; P(|K| \geq C \mid H_1) \xrightarrow[n\to\infty]{P} 1 \Rightarrow \beta = P(|K| < t_k \mid H_1) \xrightarrow[n\to\infty]{P} 0$$

□

19 Гипотеза о среднем нормальной совокупности с известной дисперсией.

Пусть имеется выборка $(X_1 \dots X_n) \in X \in N(a, \sigma^2)$, причём второй параметр известен.[10]

$H_0 : a = a_0, H_1 : a \neq a_0$.

В качестве статистики критерия возьмём $\sqrt{n} \cdot \frac{\overline{X} - a_0}{\sigma}$. Проверим, что оно имеет требуемые свойства:

[10] Например, мы измеряем что-то инструментом заданной точности.

1. Если H_0 верна, т.е. $a = a_0$, то $\sqrt{n}\frac{\overline{X}-a_0}{\sigma} = \sqrt{n}\frac{\overline{X}-a}{\sigma} \in N(0,1)$
2. Если H_0 неверно, т.е. $a \neq a_0$, то $|K| \to \infty$:

$$|K| = \left|\sqrt{n}\frac{\overline{X}-a_0}{\sigma}\right| = \underbrace{\sqrt{n}}_{\to\infty}\left|\underbrace{\frac{\overline{X}-a}{\sigma}}_{\in N(0,1)} + \underbrace{\frac{a-a_0}{\sigma}}_{\neq 0}\right| \xrightarrow[n\to\infty]{P} \infty$$

Таким образом, этот критерий — критерий согласия. Для уровня значимости $\alpha = \varepsilon$ выберем C, такую что $\varepsilon = P(|K| \geq C) \Rightarrow P(|K| < C) = 1 - \varepsilon \Rightarrow 2\Phi(C) = 1 - \varepsilon \Rightarrow 2\Phi(C) = \frac{1-\varepsilon}{2}$. Тогда, C - это обратное значение Ф для $\frac{1-\varepsilon}{2}$

Итого:

$$\begin{cases} H_0, & |K| = \left|\sqrt{n}\frac{\overline{X}-a_0}{\sigma}\right| < C \\ H_1, & |K| = \left|\sqrt{n}\frac{\overline{X}-a_0}{\sigma}\right| \geq C \end{cases}$$

Заметим, что если мы решим это неравенство, то получим доверительный интервал для параметра a нормального распределения при известном σ.

20 Гипотеза о среднем нормальной совокупности с неизвестной дисперсией.

Аналогично можно проверять для неизвестного σ, тогда в критерии σ заменится на S.

В качестве статистики критерия возьмём $\sqrt{n} \cdot \frac{\overline{X}-a_0}{S}$. Проверим, что оно имеет требуемые свойства:

1. Если H_0 верна, т.е. $a = a_0$, то $\sqrt{n}\frac{\overline{X}-a_0}{S} = \sqrt{n}\frac{\overline{X}-a}{S} \in T_{n-1}$
2. Если H_0 неверно, т.е. $a \neq a_0$, то $|K| \to \infty$:

$$|K| = \left|\sqrt{n}\frac{\overline{X}-a_0}{S}\right| = \underbrace{\sqrt{n}}_{\to\infty}\left|\underbrace{\frac{\overline{X}-a}{S}}_{\in T_{n-1}} + \underbrace{\frac{a-a_0}{S}}_{\neq 0}\right| \xrightarrow[n\to\infty]{P} \infty$$

Таким образом, этот критерий — критерий согласия.

21 Доверительные интервалы как критерии гипотез о параметрах распределения.

Пусть имеется выборка $(X_1 \dots X_n)$ случайной величины $X \in \mathcal{F}_\theta$, где $\mathcal{F}_\theta$ — распределение известного типа с неизвестным параметром θ. Проверяется гипотеза: $H_0 : \theta = \theta_0$ против

$H_1 : \theta \neq \theta_0$. Пусть для θ построен доверительный интервал (θ^-, θ^+) надежности γ. Тогда следующий критерий является критерием согласия уровня $\alpha = 1 - \gamma$:

$$\begin{cases} H_0, & \theta_0 \in (\theta^-, \theta^+) \\ H_1, & \theta_0 \notin (\theta^-, \theta^+) \end{cases}$$

Proof.

$$\alpha = P(\theta_0 \notin (\theta^-, \theta^+) \mid X \in \mathcal{F}_\theta) = 1 - P(\theta_0 \in (\theta^-, \theta^+) \mid X \in \mathcal{F}_\theta) = 1 - \gamma = \alpha$$

Доказывать состоятельность критерия нужно в каждом случае отдельно. □

22 Критерий хи–квадрат для параметрической гипотезы.

Пусть дана выборка $(X_1 \dots X_n)$ неизвестного распределения $\mathcal{F}$. Проверяется основная сложная гипотеза $H_0 : \mathcal{F} \in \mathcal{F}_\theta$, т.е. $\mathcal{F}$ принадлежит классу распределений $\mathcal{F}_\theta$, параметризованное набором из m параметров: $\theta = (\theta_1 \dots \theta_m)$.

Пусть $\hat{\theta} = (\hat{\theta}_1 \dots \hat{\theta}_m)$ — оценка этих параметров методом максимального правдоподобия. Пусть выборка разбита на k интервалов $A_1 \dots A_k$, где $A_i = [a_{i-1}, a_i)$. Пусть n_i — соответствующие экспериментальные частоты попадания в интервал A_i, p_i — соответствующие теоретические вероятности попадания в эти интервалы при распределении $\mathcal{F}_{\hat{\theta}}$

Примечание. $p_i = \mathcal{F}_{\hat{\theta}}(a_i) - \mathcal{F}_{\hat{\theta}}(a_{i-1})$

Тогда $n'_i = np_i$ — теоретические частоты попадания в A_i.

В качестве статистики критерия берётся:

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

Теорема 11 (Фишера)**.** Если гипотеза $H_0 : \mathcal{F} \in \mathcal{F}_\theta$ верна, то

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

, т.е. K имеет распределение χ^2 с $k - m - 1$ степенями свободы, где k — число интервалов и m — число параметров, задающих распределение.

Proof. Использует многомерное нормальное распределение. □

Критерий используется следующим образом: для заданного уровня значимости α находим критическую точку t_k, такую что $P(\chi^2_{k-m-1} \geq t_k) = \alpha$. Тогда критерий имеет вид:

$$\begin{cases} H_0, & K < t_k \\ H_1, & K \geq t_k \end{cases}$$

Примечание. $t_k =$ ХИ2.ОБР.ПХ$(\alpha, k - m - 1)$

Примечание. Частота интервалов должна быть ≥ 5. Если нет, то объединяем соседние интервалы.

Примечание. Желательно выборку разбить на большое число равнонаполненных интервалов.

23 Критерий хи–квадрат для гипотезы о распределении.

Проверяется основная (простая) гипотеза $H_0 : \mathcal{F} = \mathcal{F}_\theta$, где $\mathcal{F}_\theta$ — распределение известного типа с известными параметрами, против $H_1 : \mathcal{F} \neq \mathcal{F}_\theta$. В качестве статистики берётся та же самая функция

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

Теорема 12 (Пирсона)**.** Если гипотеза H_0 верна, то

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

Дальше все аналогично с прошлым билетом.

24 Критерий Колмогорова для гипотезы о распределении.

Пусть имеется выборка $(X_1 \dots X_n)$ неизвестного распределения $\mathcal{F}$. Проверяется простая гипотеза $H_0 : \mathcal{F} = \mathcal{F}_1$ против $H_1 : \mathcal{F} \neq \mathcal{F}_1$. Пусть $F_1(x)$ — непрерывная функция распределения $\mathcal{F}_1$. Тогда применяем критерий:

$$K = \sqrt{n} \sup_x |F^*(x) - F_1(x)|$$

, где $F^*(x)$ — выборочная функция распределения.

Теорема 13. Если гипотеза H_0 верна, то

$$K = \sqrt{n} \sup_x |F^*(x) - F_1(x)| \xrightarrow{n \to \infty} \mathcal{K}$$

, где $\mathcal{K}$ — распределение Колмогорова с функцией распределения $F_{\mathcal{K}}(x) = \sum_{j=-\infty}^{\infty} (-1)^j e^{-2j^2x^2}$.

Для уровня значимости находим t_k и дальше как обычно.

В некоторых статистических пакетах это распределение есть, в Excel — нет. Исторически оно не распространилось.

Недостаток этого критерия в том, что он не применим в дискретном случае.

25 Критерий Колмогорова-Смирнова.

Также используется редко.

Пусть имеются две независимых выборки $(X_1 \dots X_n)$ и $(Y_1 \dots Y_m)$ объёмов n и m соответственно неизвестных непрерывных распределений $\mathcal{F}$ и $\mathcal{G}$. Проверяется гипотеза $H_0 : \mathcal{F} \neq \mathcal{G}$ против гипотезы $H_1 : \mathcal{F} = \mathcal{G}$. В качестве статистики берётся:

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

, где $F^*(x)$ и $G^*(x)$ — соответствующие выборочные функции распределения.

Теорема 14. Если гипотеза H_0 верна, то

$$K = \sqrt{\frac{nm}{n+m}} \sup_x |F^*(x) - G^*(x)| \underset{m\to\infty}{\overset{n\to\infty}{\rightrightarrows}} \mathcal{K}$$

Примечание. Чаще всего в случае нормальных распределений используются критерии Фишера и Стьюдента. Сначала применяем критерий Фишера и если он не отвергает основную гипотезу, то применяем критерий Стьюдента.

Ещё часто применяется ранговый критерий Уилкоксона-Манна-Уитни. Мы его не рассмотрим, но общая идея в следующем: рассматривается только одна выборка и если выборка составлялась не случайно, то порядок возрастания/убывания нарушен.

26 Критерий Фишера.

Пусть имеются две независимых выборки $(X_1 \dots X_n)$ и $(Y_1 \dots Y_m)$ объёмов n и m соответственно из нормальных распределений $N(a_1, \sigma_1^2)$ и $N(a_2, \sigma_2^2)$. Проверяется гипотеза $H_0 : \sigma_1 = \sigma_2$ против гипотезы $H_1 : \sigma_1 \neq \sigma_2$. В качестве статистики берётся:

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

, где S_x^2, S_y^2 — соответствующие исправленные дисперсии, причём $S_x^2 \geq S_y^2$

Теорема 15. Если H_0 верна, то $\frac{S_x^2}{S_y^2} \in F(n-1, m-1)$ — распределение Фишера с $n-1, m-1$ степенями свободы.

Proof. По пункту 3 основной теоремы $\frac{(n-1)S^2}{\sigma^2} \in H_{n-1}$ или $\frac{S^2}{\sigma^2} \in \frac{\chi^2_{n-1}}{n-1}$.

При $\sigma_1 = \sigma_2 = \sigma$:

$$\frac{S_x^2}{S_y^2} = \frac{S_x^2}{\sigma^2} \cdot \frac{\sigma^2}{S_y^2} = \frac{\chi^2_{n-1}}{n-1} \cdot \frac{m-1}{\chi^2_{m-1}} \stackrel{\text{def}}{=} F(n-1, m-1)$$

□

Критерий по статистике очевиден.

Примечание. При $H_1 : \sigma_1 \neq \sigma_2$, т.е. $\sigma_1 > \sigma_2, K = \frac{S_x^2}{S_y^2} \to \frac{\sigma_1^2}{\sigma_2^2} > 1$

При H_0 выполнено $K \to 1$.

27 Критерий Стьюдента.

Пусть имеются две независимых выборки $(X_1 \dots X_n)$ и $(Y_1 \dots Y_m)$ объёмов n и m соответственно из нормальных распределений $N(a_1, \sigma^2)$ и $N(a_2, \sigma^2)$ с одинаковой дисперсией σ^2. Проверяется гипотеза $H_0 : a_1 = a_2$ против гипотезы $H_1 : a_1 \neq a_2$.

Теорема 16. Случайная величина

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

, где T_{n+m-2} — распределение Стьюдента с $n + m - 2$ степенями свободы. Это не зависит от того, верна гипотеза или нет.

В качестве статистики берётся:

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

Из теоремы видим, что если H_0 верна, то $K \in T_{n+m-2}$, если нет, то $K \xrightarrow[n\to\infty]{m\to\infty} \infty$.

Критерий: пусть t_k — квантиль распределения Стьюдента $|T_{n+m-2}|$ уровня значимости α.

$$\begin{cases} H_0 : a_1 = a_2, & K < t_k \\ H_1 : a_1 \neq a_2, & K \geq t_k \end{cases}$$

28 Понятие статистической зависимости. Корреляционное облако и корреляционная таблица. Первоначальные выводы по ним.

Определение. Функциональная зависимость имеет место, когда две величины связаны жесткими законами природы.

Определение. Зависимость называется **статистической**, если изменение одной величины влияет на распределение другой. Если при этом изменяется среднее значение[11] другой случайной величины, то зависимость называется **корреляционной**. Если среднее значение *увеличивается* при увеличении первой случайной величины, то корреляция **прямая**, а если *уменьшается* — **обратная**.

28.1 Корреляционное облако

Пусть в ходе экспериментов получились значения случайных величин X и Y : $(X_i, Y_i), 1 \leq i \leq n$. Нанося эти точки на координатную плоскость XOY, получим корреляционное облако. По его виду можно сделать предположение о зависимости.

Пример. X, Y имеют нормальное распределение с одинаковыми параметрами.

- Если корреляционное облако имеет форму круга, то величины *независимы.*
- Если корреляционное облако имеет форму эллипса с большой осью параллельной прямой вида $y = kx + b, k > 0$, то скорее всего *зависимость прямая.*

28.2 Корреляционная таблица

Экспериментальные данные представляем в виде таблицы:

X_i \ Y_i	Y_1	Y_2	$\dots$	Y_m
X_1	n_{11}	n_{12}	$\dots$	n_{1m}
X_2	n_{21}	n_{22}	$\dots$	n_{2m}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
X_n	n_{n1}	n_{n2}	$\dots$	n_{nm}

Пример.

[11] Математическое ожидание.

X_i \ Y_i	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	50	

- n_x — частота значения x
- n_y — частота значения y
- $\overline{y_x}$ — условное среднее случайной величины y:

$$\overline{y_x} = \frac{1}{n_x} \sum_i n_{xy_i} y_i$$

В нашем примере условное матожидание $\overline{y_x}$ увеличивается при увеличении x, следовательно, скорее всего есть прямая корреляция.

Примечание. При большом числе данных удобнее составлять интервальную корреляционную таблицу и заменить интервалы на средние значения в этих интервалах.

29 Критерий хи–квадрат для проверки независимости.

Пусть выборка $(X_1, Y_1) \dots (X_n, Y_n)$ случайных величин X и Y представлена в виде интервальной корреляционной таблицы. Случайная величина X при этом разбита на k интервалов, а Y на m интервалов. Обозначим $v_{i\cdot}$ = число значений случайной величины X, попавших в i-тый интервал $[a_{i-1}, a_i), 1 \le i \le k$. Обозначим $v_{\cdot j}$ = число значений случайной величины Y, попавших в j-тый интервал $[b_{j-1}, b_j), 1 \le j \le k$. Обозначим v_{ij} = число точек (X, Y), попавших в $[a_{i-1}, a_i) \times [b_{i-1}, b_i)$.

X_i \ Y_i	$[b_0; b_1)$	$[b_1; b_2)$	$\dots$	$[b_{m-1}; b_m)$	$v_i = \sum_{j=1}^{n} v_{ij}$
$[a_0; a_1)$	v_{11}	v_{12}	$\dots$	v_{1m}	$v_{1\cdot}$
$[a_1; a_2)$	v_{21}	v_{22}	$\dots$	v_{2m}	$v_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	
$[a_{n-1}; a_n)$	v_{n1}	v_{n2}	$\dots$	v_{nm}	$v_{n\cdot}$
$v_j = \sum_{i=1}^{n} v_{ij}$	$v_{\cdot 1}$	$v_{\cdot 2}$	$\dots$	$v_{\cdot m}$	

По этой таблице проверяется основная гипотеза $H_0 : X$ и Y независимы против $H_1 : X$ и Y зависимы.

Вспомним определение независимых случайных величин:

$$P(X \in \mathfrak{B}_1, Y \in \mathfrak{B}_2) = P(X \in \mathfrak{B}_1) \cdot P(Y \in \mathfrak{B}_2)$$

Согласно этому определению, если гипотеза H_0 верна, то вероятность попадания пары (X, Y) в любой прямоугольник равна произведению теоретических вероятностей попасть случайным величинам в эти интервалы.

$$p_{ij} = P(X \in [a_{i-1}; a_i), Y \in [b_{i-1}; b_i)) = P(X \in [a_{i-1}; a_i)) \cdot P(Y \in [b_{i-1}; b_i)) = p_i \cdot p_y$$

По закону больших чисел при $n \to \infty$:

$$\frac{v_{i.}}{n} \xrightarrow{P} p_i \quad \frac{v_{.j}}{n} \xrightarrow{P} p_j \quad \frac{v_{ij}}{n} \xrightarrow{P} p_{ij}$$

Поэтому основанием для отклонения гипотезы служит заметная разница между величинами между $\frac{v_{ij}}{n}$ и $\frac{v_i}{n} \cdot \frac{v_j}{n}$, т.е. между v_{ij} и $\frac{v_i v_j}{n}$.

В качестве статистики критерия берётся функция:

$$K = n \sum_{i,j} \frac{\left(v_{ij} - \frac{v_i v_j}{n}\right)^2}{v_{i.} v_{.j}}$$

Теорема 17. Если гипотеза H_0 верна, то $K \rightrightarrows \chi^2_{(k-1)(m-1)}$

Получили критерий согласия: t_k — квантиль распределения $H_{(k-1)(m-1)}$

$$\begin{cases} H_0, & K < t_k \\ H_1, & K \geq t_k \end{cases}$$

30 Однофакторный дисперсионный анализ. Общая, межгрупповая и внутригрупповая дисперсии. Теорема о разложении дисперсии.

Предположим, что на случайную величину X *(признак–результат)* может влиять фактор Z *(признак–фактор)*. Z — не обязательно случайная величина.

Пример. Хотим проверить, как влияет температура на разложение ???. Проводим измерения при разных температурах, регулируя термостат. Тогда температура — не случайная величина, она управляема.

Пусть при различных k уровнях фактора Z получены k независимых выборок случайной величины X : $X^{(1)} = (X_1^{(1)} \dots X_{n_1}^{(1)}) \dots X^{(k)} = (X_1^{(k)} \dots X_{n_k}^{(k)})$. В общем случае размеры выборок могут быть различны.

30.0.1 Общая, межгрупповая и внутригрупповая дисперсия

$$\overline{X}^{(j)} = \frac{1}{n_j}\sum_{i=1}^{n_j} X_i^{(j)} \quad \mathbb{D}^{(j)} = \frac{1}{n_j}\sum_{i=1}^{n_j}(X_i^{(j)} - \overline{X}^{(j)})^2$$

Объединив все данные в одну общую выборку, получим выборку объёма $n = n_1 + \cdots + n_k$. Вычислим общее выборочное среднее как

$$\overline{X} = \frac{1}{n}\sum_{i,j} X_i^{(j)} = \frac{1}{n}\sum_{j=1}^{k}\overline{X}^{(j)}$$

и **общую выборочную дисперсию**:

$$D_o = \frac{1}{n}\sum_{i,j}(X_i^{(j)} - \overline{X})^2$$

Определение. Внутригрупповой *(остаточной)* дисперсией называется среднее[12] групповых дисперсий:

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

Определение. Межгрупповой *(факторной)* дисперсией или **дисперсией выборочных средних** называется величина

$$D = \frac{1}{n}\sum_{j=1}^{k}(\overline{X}^{(j)} - \overline{X})^2 n_j$$

Теорема 18 (о разложении дисперсий)**.** Общая дисперсия равна сумме межгрупповой и внутригрупповой дисперсий:

$$D_o = D + D$$

Proof. Неинтересное, алгебраическое. □

Смысл:

- Внутригрупповая дисперсия показывает средний разброс внутри выборок.
- Межгрупповая дисперсия показывает, насколько отличны выборочные средние при различных уровнях фактора. Таким образом, её величина в общей сумме отражает влияние фактора.

[12] взвешенное

31 Однофакторный дисперсионный анализ. Проверка гипотезы о влиянии фактора.

Предположим, что случайна величина X имеет нормальное распределение и фактор Z может влиять только на математическое ожидание, но не на дисперсию и тип распределения.

Может показаться, что ограничение слишком строгое, но в реальной жизни это условие выполняется часто.

Поэтому можно считать, что данные независимые выборки при разных уровнях Z также имеют нормальное распределение с одинаковым параметром σ^2:

$$X_i^{(j)} \in N(a_i, \sigma^2)$$

Проверяется основная гипотеза $H_0 : a_1 = a_2 = \cdots = a_k$, т.е. фактор Z не влияет на X. $H_1 : Z$ влияет на X. По пункту 3 основной теоремы:

$$\sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{\sigma} \right)^2 = \frac{nD}{\sigma^2} \in H_{n-1}$$

Отсюда для каждой из k выборок:

$$\frac{n_j \, \mathbb{D}^{(j)}}{\sigma} \in H_{n_j - 1}, 1 \le j \le k$$

Т.к. распределение χ^2 устойчиво по суммированию, то получаем:

$$\sum_{j=1}^{k} \frac{n_j \, \mathbb{D}^{(j)}}{\sigma} = \frac{\sum_{j=1}^{k} n_j \, \mathbb{D}^{(j)}}{\sigma^2} = \frac{nD}{\sigma^2} \in H_{n-k}$$

, т.к. $\sum_{j=1}^{k} (n_j - 1) = n - k$.

Все это выполнено вне зависимости от того, верна H_0 или нет. Пусть H_0 верна, тогда все выборки можно считать одной выборкой объёма и по тому же свойству:

$$\frac{nD_o}{\sigma^2} \in H_{n-1}$$

Согласно теореме о разложении дисперсии:

$$\underbrace{\frac{nD_o}{\sigma^2}}_{\in H_{n-1}} = \frac{nD}{\sigma^2} + \underbrace{\frac{nD}{\sigma^2}}_{\in H_{n-k}}$$

На записи все эти формулы видны примерно как “Ыаыаыаацпы”, так что за достоверность конспекта не отвечаю.

Следовательно, $\frac{nD}{\sigma^2} \in H_{k-1}$.[13]. В итоге при верной гипотезе H_0 мы получили $\frac{nD}{\sigma^2} \in H_{k-1}$, а $\frac{nD}{\sigma^2} \in H_{n-k}$. Тогда:

$$\frac{\frac{nD}{\sigma^2(k-1)}}{\frac{nD}{\sigma^2(n-k)}} = \frac{\frac{D}{k-1}}{\frac{D}{n-k}} \in F(k-1, n-k)$$

В результате имеем критерий $K = \frac{n-k}{n-1}\frac{D}{D}$, находим t_k — квантиль $F(k-1, n-k)$ уровня значимости α, искомое очевидно строится.

32 Математическая модель регрессии. Основные понятия и определения. Метод наименьших квадратов.

Требуется определить, как зависит наблюдаемая случайная величина от одной или нескольких других величин, причём необязательно все величины случайные. Хотелось бы построить математическую модель, которая будет с некоторой надежностью предсказывать значение искомой величины при известных прочих величинах.

Примечание. Предсказывать мы можем лишь среднее значение наблюдаемых случайных величин.

32.1 Математическая модель регрессии

Пусть случайная величина X зависит от случайной величины Z. Значение Z мы либо наблюдаем, либо задаём.

Определение. Регрессией X на Z называется функция $f(z) = \mathbb{E}(X \mid Z = z)$. Она показывает зависимость среднего значения X от значения Z.

Определение. Уравнение $X = f(Z)$ называется **уравнением регрессии**, а её график — **линией регрессии**.

Пусть при n экспериментах при значениях $z_1 \dots z_n$ случайной величины Z наблюдались значения $X_1 \dots X_n$ случайной величины X. Обозначим через ε_i разницу между экспериментальными и теоретическими значениями случайной величины X:

$$\varepsilon_i = X_i - \mathbb{E}(X \mid Z = z_i) = X_i - f(z_i)$$

Тогда $X_i = f(z_i) + \varepsilon_i, 1 \le i \le n$, и ε_i можно понимать, как ошибку эксперимента.

Примечание. Обычно можно считать, что ε_i — независимая одинаковая нормальная случайная величина: $\varepsilon_i \in N(0; \sigma^2)$. $a = 0$, т.к.

$$\mathbb{E}\,\varepsilon_i = \mathbb{E}(X_i) - \mathbb{E}(f(z_i)) = \mathbb{E}(X_i) - \mathbb{E}(X \mid Z = z_i) = 0$$

[13]Это неочевидный факт, но мы его доказывать не будем.

Примечание. В реальности ε_i могут быть зависимыми, это явление называется автокорреляция, особенно часто проявляется во временных рядах.

Цель: по данным значениям $z_1 \dots z_n$ и $X_1 \dots X_n$ как можно более точно оценить функцию $f(z)$.

Примечание. При этом предполагается (из теории), что $f(z)$ — функция определенного типа, параметры которой неизвестны.

32.1.1 Метод наименьших квадратов

Метод наименьших квадратов состоит в выборе параметров функции $f(z)$ таким образом, чтобы минимизировать сумму квадратов ошибок:

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (X_i - f(z_i))^2 \to \min$$

Определение. Пусть $\theta = (\theta_1 \dots \theta_k)$ — набор неизвестных параметров функции $f(z)$. Оценка $\hat{\theta}$, при которой достигается минимум суммы квадратов ошибок, называется **оценкой метода наименьших квадратов (ОМНК)**.

33 Вывод уравнения линейной парной регрессии. Геометрический смысл прямой регрессии.

Пусть имеется линейная регрессия $f(z) = a + bz$. Тогда $X_i = a + bz_i + \varepsilon_i, 1 \le i \le n$, найдём оценки неизвестных параметров a и b методом наименьших квадратов (МНК):

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

$$\begin{aligned} 0 = \frac{\partial}{\partial a} \sum_{i=1}^{n} \varepsilon_i^2 &= 2 \sum_{i=1}^{n} (X_i - a - bz_i) \cdot (-1) \\ &= -2 \sum_{i=1}^{n} X_i + 2 \sum_{i=1}^{n} a + 2 \sum_{i=1}^{n} bz_i \\ &= -2 \left(n\overline{X} - na - bn\overline{Z} \right) \end{aligned}$$

$$\begin{aligned} 0 = \frac{\partial}{\partial b} \sum_{i=1}^{n} \varepsilon_i^2 &= 2 \sum_{i=1}^{n} (X_i - a - bz_i) \cdot (-z_i) \\ &= -2 \left(n\overline{XZ} - an\overline{Z} - bn\overline{Z^2} \right) \end{aligned}$$

$$\begin{cases} n\overline{X} - na - bn\overline{Z} = 0 \\ n\overline{XZ} - na\overline{Z} - bn\overline{Z^2} = 0 \end{cases}$$

$$\begin{cases} \overline{X} - a - b\overline{Z} = 0 \\ \overline{XZ} - a\overline{Z} - b\overline{Z^2} = 0 \end{cases}$$

$$\begin{cases} a + b\overline{Z} = \overline{X} \\ a\overline{Z} + b\overline{Z^2} = \overline{XZ} \end{cases}$$

Система такого вида называется системой нормальных уравнений.

$$\begin{cases} a = \overline{X} - b\overline{Z} \\ (\overline{X} - b\overline{Z})\overline{Z} + b\overline{Z^2} = \overline{XZ} \end{cases}$$

$$\begin{cases} a = \overline{X} - b\overline{Z} \\ b(\overline{Z^2} - \overline{Z}^2) = \overline{XZ} - \overline{X} \cdot \overline{Z} \end{cases}$$

$$\begin{cases} a = \overline{X} - b\overline{Z} \\ b = \frac{\overline{XZ} - \overline{X} \cdot \overline{Z}}{\overline{Z^2} - \overline{Z}^2} \end{cases}$$

$$\begin{cases} a = \overline{X} - b\overline{Z} \\ b = \frac{\overline{XZ} - \overline{X} \cdot \overline{Z}}{\hat{\sigma}_z^2} \end{cases}$$

, где $\hat{\sigma}_z^2$ — выборочная дисперсия Z (неисправленная).

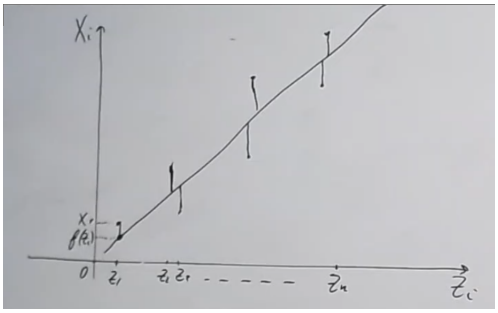
Запишем уравнение линейной регрессии в удобном виде.

$$\begin{aligned} \overline{X_z} &:= \mathbb{E}(\overline{X} \mid Z = z) = f(z) \\ \overline{X_z} &= a + bz \\ \overline{X_z} &= \overline{X} - b\overline{Z} + bz \\ \overline{X_z} - \overline{X} &= -b\overline{Z} + bz \\ \overline{X_z} - \overline{X} &= b(z - \overline{Z}) \\ \overline{X_z} - \overline{X} &= \frac{\overline{XZ} - \overline{X} \cdot \overline{Z}}{\hat{\sigma}_z^2}(z - \overline{Z}) \\ \overline{X_z} - \overline{X} &= \frac{\overline{XZ} - \overline{X} \cdot \overline{Z}}{\hat{\sigma}_z \hat{\sigma}_x} \cdot \frac{\hat{\sigma}_x}{\hat{\sigma}_z}(z - \overline{Z}) \\ \frac{\overline{X_z} - \overline{X}}{\hat{\sigma}_x} &= \frac{\overline{XZ} - \overline{X} \cdot \overline{Z}}{\hat{\sigma}_z \hat{\sigma}_x} \cdot \frac{z - \overline{Z}}{\hat{\sigma}_z} \\ \frac{\overline{X_z} - \overline{X}}{\hat{\sigma}_x} &= r\frac{z - \overline{Z}}{\hat{\sigma}_z} \end{aligned}$$

, где $r = \frac{\overline{XZ}-\overline{X}\cdot\overline{Z}}{\hat{\sigma}_z\hat{\sigma}_x}$ — **выборочный коэффициент линейной корреляции**, а само уравнение называется **выборочным уравнением линейной регрессии**.

Примечание. Прямая регрессии проходит через точку $\overline{Z}, \overline{X}$.

33.0.1 Геометрический смысл прямой линейной регрессии



Прямая регрессии строится таким образом, чтобы сумма квадратов длин отрезков от прямой до точек распределения была наименьшей.

34 Выборочный коэффициент линейной корреляции. Проверка гипотезы о его значимости.

Определение.

$$\boxed{r := \frac{\overline{XZ} - \overline{X}\cdot\overline{Z}}{\hat{\sigma}_z\hat{\sigma}_x}}$$

Примечание. Из прошлого семестра:

$$r_{\xi,\eta} = \frac{\mathbb{E}\,\xi\eta - \mathbb{E}\,\xi\cdot\mathbb{E}\,\eta}{\sigma_\xi\cdot\sigma_\eta}$$

Отсюда видим, что выборочный коэффициент линейной корреляции является оценкой теоретического коэффициента линейной корреляции, полученной по методу моментов. Поэтому выборочный коэффициент линейной корреляции характеризует силу линейной связи между двумя случайными величинами. Знак r показывает, является ли корреляция прямой или обратной.

$\lvert r\rvert$	Сила связи
$0.1 - 0.3$	слабая
$0.3 - 0.5$	умеренная
$0.5 - 0.7$	заметная
$0.7 - 0.9$	высокая
> 0.9	очень высокая

Figure 1: Шкала Чеддока

34.0.1 Проверка гипотезы о значимости выборочного коэффициента корреляции

Пусть двумерная случайная величина (Z, X) распределена нормально. По выборке объёма n вычислим r, а r — теоретический коэффициент линейной корреляции. Проверяется основная гипотеза $H_0 : r = 0$ против альтернативной гипотезы $H_1 : r \neq 0$, т.е. коэффициент r статистически значим.

Теорема 19. Если H_0 верна, то статистика

$$K = \frac{r\sqrt{n-2}}{\sqrt{1-r_2}} \in T_{n-2}$$

Критерий: пусть t_k — квантиль T_{n-2} уровня значимости α.

$$\begin{cases} H_0 : r = 0, & |K| < t_k \\ H_1 : r \neq 0, & |K| \geq t_k \end{cases}$$

В Excel $r = (Z, X)$

35 Выборочное корреляционное отношение, его свойства.

Пусть имеется k выборок случайной величины X, полученных при уровнях фактора $Z : Z_1 \dots Z_k$. Вычислены общая дисперсия D_o, межгрупповая дисперсия D и D — внутригрупповая дисперсия. По теореме о разложении дисперсий $D_o = D + D$. Обозначим:

- $\sigma_{\overline{X}_z} = \sqrt{D}$ — межгрупповое среднеквадратическое отклонение
- $\hat{\sigma}_X = \sqrt{D}$ — выборочное среднеквадратическое отклонение

Определение. Выборочным корреляционным отношением η_{XZ} случайной величины X на Z называется величина

$$\boxed{\eta_{XZ} = \frac{\sigma_{\overline{X}_z}}{\hat{\sigma}_X} = \sqrt{\frac{D}{D}}}$$

Свойства.

1. $0 \leq \eta \leq 1$

   Proof.

   $$D_o = D + D, 0 \leq D \leq D_o \Rightarrow 0 \leq \frac{D}{D_o} \leq 1 \Rightarrow 0 \leq \eta = \sqrt{\frac{D}{D_o}} \leq 1$$

   □

2. Если $\eta = 1$, то X функционально зависит от Z.

   Proof. $\eta = 1 \Rightarrow D = 0 \Rightarrow X = X(Z)$ — функция от переменной Z □

3. Если $\eta = 0$, то нет корреляционной зависимости.

   Proof. Если $\eta = 0$, то $D = 0$, следовательно значения $\overline{X}^{(i)}$ не зависят от Z_i. □

4. $\eta \geq |r|$
5. $\eta = |r| \Leftrightarrow$ когда имеет место точная линейная корреляционная зависимость, т.е. все точки экспериментальных данных лежат на одной прямой, которая совпадёт с прямой регрессии.

36 Свойства ошибок в модели линейной парной регрессии. Анализ дисперсии фактора–результата. Коэффициент детерминации, его свойства.

Пусть при n экспериментах получены точки $(Z_1, X_1) \dots (Z_n, X_n)$, при этом X — признак–результат, а Z — признак–фактор. Пусть $X = \alpha + \beta Z + e$ — теоретическая модель регрессии. Здесь α, β — неизвестные теоретические значения, e — неизвестная случайная величина, которая отражает:

- Влияние неучтённых факторов.
- Вероятность того, что модель нелинейная.
- Случайность.

Получим выборочное уравнение линейной регрессии:

$$\boxed{\hat{X} = a + bZ}$$

Тогда экспериментальные данные можно представить так:

$$X_i = \hat{X} + \varepsilon_i$$

Константы a и b можно рассматривать как точечные оценки α и β. Нас интересует, насколько качественны эти оценки и что можно сказать про распределение ошибки e.

Свойства ε_i:

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

   Proof.

   $$a = \overline{X} - b\overline{Z} \Rightarrow \overline{X} = a + b\overline{Z}$$
   $$\overline{\varepsilon_i} = \overline{X_i - \hat{X}} = \overline{X} - \overline{a + bZ_i} = \overline{X} - \overline{X} = 0$$

   □

2. $\operatorname{cov}(\hat{X}_i, \varepsilon_i) = 0$, т.е. эти величины не коррелируют.

   Proof.

   Примечание. Дальше будем обозначать выборочную дисперсию как $\mathbb{D}$, а не $\mathbb{D}_B$

   $$b = \frac{\overline{XZ} - \overline{X}\,\overline{Z}}{\hat{\sigma}_z^2} = \frac{\operatorname{cov}(X, Z)}{\mathbb{D}(Z)} \Rightarrow \operatorname{cov}(X, Z) = b\,\mathbb{D}(Z) \Rightarrow \operatorname{cov}(X, Z) - b\,\mathbb{D}(Z) = 0$$

   Т.к. $\hat{X} = a + bZ, \varepsilon = X - a - bZ$, то:

   $$\begin{aligned}
   \operatorname{cov}(\hat{X}, \varepsilon) &= \operatorname{cov}(a + bZ, X - a - bZ) \\
   &= \operatorname{cov}(bZ, x - bZ) \\
   &= \operatorname{cov}(bZ, X) - \operatorname{cov}(bZ, bZ) \\
   &= b\operatorname{cov}(Z, X) - b^2 \cdot \mathbb{D}(Z) \\
   &= b(\operatorname{cov}(Z, X) - b \cdot \mathbb{D}(Z)) \\
   &= 0
   \end{aligned}$$

   □

С помощью второго свойства можем провести анализ модели, анализируя дисперсии.

Обозначение.

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

Дисперсия расчётных значений:

$$\mathbb{D}(\hat{X}) = \frac{1}{n}\sum_{i=1}^{n}(\hat{X}_i - \overline{X})^2$$

$$\mathbb{D}(\varepsilon) = \frac{1}{n}\sum_{i=1}^{n}(\varepsilon_i - \bar{\varepsilon})^2 = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2$$

$$\begin{aligned} X &= \hat{X} + \varepsilon \\ \mathbb{D}\,X &= \mathbb{D}(\hat{X} + \varepsilon) \\ \mathbb{D}\,X &= \mathbb{D}\,\hat{X} + \mathbb{D}\,\varepsilon + 2\underbrace{\operatorname{cov}(\hat{x}, \varepsilon)}_{0} \end{aligned}$$

$$\boxed{\mathbb{D}\,X = \mathbb{D}\,\hat{X} + \mathbb{D}\,\varepsilon}$$

36.1 Коэффициент детерминации

Определение. $R^2 = \dfrac{\mathbb{D}\,\hat{X}}{\mathbb{D}\,X}$ **— коэффициент детерминации**.

Свойства.

1. $0 \le R^2 \le 1$
2. $R = 1 \Rightarrow$ ошибок нет и все точки экспериментальных данных лежат на найденной прямой.
3. $R = 0 \Rightarrow$ расчётное значение $\hat{X}$ постоянно, т.е. найденная прямая постоянна с уровнем $\overline{X}$ и $b = 0$.

Таким образом, чем ближе R^2 к 1, тем качественнее модель.

Поскольку R^2 — доля дисперсии расчётных значений, R^2 объясняет долю дисперсии экспериментальных данных. Соответственно $1 - R^2$ — доля необъясненной дисперсии.

37 Проверка гипотезы о значимости уравнения линейной регрессии. Связь между коэффициентом детерминации и коэффициентом линейной корреляции.

Пусть мы хотим проверить, является ли R^2 статистически значимым, т.е. проверяем основную гипотезу $H_0 : R_{\text{т}}^2 = 0$ против гипотезы $H_1 : R_{\text{т}} \neq 0$[14]

Теорема 20.

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

Пусть $t_{\text{кр.}}$ — квантиль $F(1, n-2)$ уровня значимости α. Тогда если $K < t_{\text{кр.}}$, то принимается гипотеза H_0, иначе H_1.

[14] $R_{\text{т}}$ — теоретический коэффициент детерминации.

Если взять корень из K, то полученная величина должна иметь распределение Стьюдента. Для похожей[15] величины мы это доказывали при проверке статистической значимости коэффициента линейной корреляции.

Проверить $R^2_{\text{T}} \neq 0$ — то же самое, что проверить гипотезу $b = 0$.

1. $\sqrt{R^2} = r_{\hat{X},X}$

   Proof.
   $$\operatorname{cov}(\hat{X}, X) = \operatorname{cov}(\hat{X}, \hat{X} + \varepsilon) = \underbrace{\operatorname{cov}(\hat{X}, \hat{X})}_{\mathbb{D}\,\hat{X}} + \underbrace{\operatorname{cov}(\hat{X}, \varepsilon)}_{0} = \mathbb{D}\,\hat{X}$$
   $$r_{\hat{X},X} \overset{\text{def}}{=} \frac{\operatorname{cov}(\hat{X}, X)}{\sqrt{\mathbb{D}(\hat{X}) \cdot \mathbb{D}(X)}} = \frac{\mathbb{D}(\hat{X})}{\sqrt{\mathbb{D}(\hat{X}) \cdot \mathbb{D}(X)}} = \sqrt{\frac{\mathbb{D}(\hat{X})}{\mathbb{D}(X)}} = \sqrt{R^2}$$
   □

2. $\sqrt{R^2} = r_{Z,X}$

   Proof.
   $$\operatorname{cov}(\hat{X}, X) = \operatorname{cov}(a + bZ, X) = b \operatorname{cov}(Z, X)$$
   $$\mathbb{D}(\hat{X}) = \mathbb{D}(a + bZ) = 0 + b^2\,\mathbb{D}(Z) = b^2\,\mathbb{D}(Z)$$
   $$r_{\hat{X},X} \overset{\text{def}}{=} \frac{\operatorname{cov}(\hat{X}, X)}{\sqrt{\mathbb{D}(\hat{X}) \cdot \mathbb{D}(X)}} = \frac{b \operatorname{cov}(Z, X)}{\sqrt{b^2\,\mathbb{D}(Z) \cdot \mathbb{D}(X)}} = \frac{\operatorname{cov}(Z, X)}{\sqrt{\mathbb{D}(Z) \cdot \mathbb{D}(X)}} \overset{\text{def}}{=} r_{Z,X}$$
   □

Таким образом, квадрат коэффициента линейной корреляции равен коэффициенту детерминации. Поэтому и результат проверки гипотез будет одинаковым. Этот факт верен только для модели парной регрессии.

38 Теорема Гаусса-Маркова.

Обсудим, насколько качественные оценки коэффициентов линейной регрессии мы находим методом наименьших квадратов.

Теорема 21 (Гаусса–Маркова)**.**

- $X = \alpha + \beta Z + e$
- $\hat{X} = a + bZ$
- e_i — независимые случайные величины, $\varepsilon_i \in N(0, \sigma^2)$

[15] Вместо R было r.

- Z_i и e_i независимы[16][17]

Тогда оценки a и b коэффициентов α и β будут:

1. Несмещёнными: $\mathbb{E}\, a = \alpha, \mathbb{E}\, b = \beta$
2. Состоятельными: $a \xrightarrow[n\to\infty]{P} \alpha$ и $b \xrightarrow[n\to\infty]{P} \beta$
3. Эффективными в классе линейных несмещенных оценок, т.е. их дисперсии в этом классе минимальны и при этом равны:

$$\mathbb{D}(a) = \frac{\overline{Z^2} \cdot \sigma^2}{n\,\mathbb{D}(Z)} \quad \mathbb{D}(b) = \frac{\sigma^2}{n\,\mathbb{D}(Z)}$$

Третье условие нужно для несмещённости, а четвертое — для состоятельности.

39 Стандартные ошибки коэффициентов регрессии. Их доверительные интервалы.

Дадим оценку дисперсии σ^2 величины ε:

$$\mathbb{D}(\varepsilon) = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2$$

$$\mathbb{E}(\mathbb{D}(\varepsilon)) = \frac{n-2}{n}\sigma^2$$

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

S^2 — несмещённая оценка σ^2, мы не будем это доказывать.

Определение. S **— стандартная ошибка регрессии**.

Если в $\mathbb{D}(a)$ и в $\mathbb{D}(b)$ подставим вместо σ^2 его оценку S^2, то получим **стандартные ошибки коэффициентов** a и b.

$$S_a^2 = \frac{\overline{Z^2} \cdot S^2}{n\,\mathbb{D}(Z)} \quad S_b^2 = \frac{S^2}{n\,\mathbb{D}(Z)}$$

Т.к. e_i имеет нормальное распределение, то и ε_i имеет нормальное распределение. Тогда пусть t_γ — квантиль двухстороннего распределения Стьюдента $|T_{n-2}|$. Тогда будут доверительные интервалы:

[16] Также используют условие Z_i — не случайная величина. Это условие слабее указанного.

[17] Обратное к этому условию называется автокорреляцией. Автокорреляция распространена во временных рядах.

- Для α: $(a - t_\gamma \cdot S_a; a + t_\gamma \cdot S_a)$
- Для β: $(b - t_\gamma \cdot S_b; a + t_\gamma \cdot S_b)$

40 Прогнозирование в модели линейной парной регрессии. Стандартная ошибка прогноза, доверительный интервал прогноза.

$X = \alpha + \beta Z + e$ — теоретическая неизвестная модель, $X = a + bZ$ — выборочная модель, которую нашли по экспериментальным данным.

Пусть требуется составить прогноз при $Z = Z_p$. Тогда в теории $X_p = \alpha + \beta Z_p + e$, на практике $\hat{X}_p = a + bZ_p$. Обозначим $\Delta_p = \hat{X}_p - X_p$.

Свойства.

1. $\mathbb{E}\,\Delta_p = 0$

   Proof.

   $$\begin{aligned}\mathbb{E}\,\Delta_p &= \mathbb{E}(\hat{X}_p - X_p)\\ &= \mathbb{E}(a + bZ_p) - \mathbb{E}(\alpha + \beta Z_p + e)\\ &= \mathbb{E}(a) + Z_p\,\mathbb{E}(b) - \alpha - \beta Z_p - \underbrace{\mathbb{E}\,e}_{0}\\ &= \alpha + \beta Z_p - \alpha - \beta Z_p\\ &= 0\end{aligned}$$

   □

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

   Proof. Несложно, но занудно. Поэтому не будем. □

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

   $$S\Delta_p = S\sqrt{1 + \frac{1}{n} + \frac{(Z_p - \overline{Z})^2}{n\,\mathbb{D}(Z)}}$$

   (a) $\mathbb{D}\,\Delta_p > \sigma^2$

   (b) При $n \to \infty, \mathbb{D}\,\Delta_p \to \sigma^2$.

   (c) Чем дальше Z_p от $\overline{Z}$, тем выше будет ошибка модели. Таким образом, качественно предсказывать далеко от среднего никакая модель не может.

Доверительный интервал прогноза будет $(\hat{X} - t_\gamma \cdot S\Delta_p, \hat{X} + t_\gamma \cdot S\Delta_p)$, где t_k — квантиль распределения Стьюдента T_{n-2} уровня γ. Это согласно моей интерпретации этого учебника, на лекции это не говорилось.

41 Общая модель линейной регрессии. Вывод нормального уравнения (свойство существования квадратного корня симметрической матрицы без доказательства).

Пусть признак–результат X зависит от k факторов $Z_1 \dots Z_k$. Рассматривается теоретическая модель линейной регрессии $\mathbb{E}(X \mid \vec{Z}) = \beta_1 Z_1 + \dots + \beta_k Z_k$

Обозначение.

$$\vec{Z} = \begin{pmatrix} Z_1 \\ \vdots \\ Z_k \end{pmatrix} \quad \vec{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$$

Пусть проведено $n \geq k$ экспериментов, $\vec{Z}^{(i)} = (Z_1^{(i)} \dots Z_k^{(i)})$ — значение фактора при i-том эксперименте (возможно заранее заданные). $\vec{X} = (X_1 \dots X_n)$ — полученные экспериментальные данные признака–результата X. Согласно модели:

$$\begin{cases} X_1 = \beta_1 Z_1^{(1)} + \dots + \beta_k Z_k^{(1)} + \varepsilon_1 \\ X_2 = \beta_1 Z_1^{(2)} + \dots + \beta_k Z_k^{(2)} + \varepsilon_2 \\ \vdots \\ X_n = \beta_1 Z_1^{(n)} + \dots + \beta_k Z_k^{(n)} + \varepsilon_n \end{cases}$$

, где ε_i — случайная теоретическая ошибка при i-том эксперименте.

Обозначение (вектор случайных ошибок).

$$\vec{\varepsilon} \coloneqq \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Обозначение (матрица плана).

$$Z_{k \times n} \coloneqq \begin{pmatrix} Z_1^{(1)} & Z_1^{(2)} & \dots & Z_1^{(n)} \\ Z_2^{(1)} & Z_2^{(2)} & \dots & Z_2^{(n)} \\ \vdots & \vdots & \ddots & \vdots \\ Z_k^{(1)} & Z_k^{(2)} & \dots & Z_k^{(n)} \end{pmatrix}$$

Тогда теоретическую модель можно записать в матричной форме:

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

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

Примечание. Заметим, что в данной модели мы не теряем свободный член a, т.к. можно считать, что $Z_1 = \mathbb{1}$ и ей соответствует строка из единиц.

41.1 Метод наименьших квадратов и нормальные уравнения

Будем считать, что выполнено два условия:

1. $\operatorname{rank} Z = k$, т.е. все строки матрицы плана линейно независимы.[18]
2. Случайные ошибки ε_i независимы и имеют одинаковое нормальное распределение с параметром $a = 0$.[19]

Обозначение. $A_{k\times k} \coloneqq ZZ^\mathsf{T}$

Свойства.

1. A — симметричная.
2. A — положительно определенная.
3. Существует вещественная симметрическая матрица $\sqrt{A}$, такая что $(\sqrt{A})^2 = A$.

Найдём оценку $\vec{B} = (b_1 \dots b_k)$, которая минимизирует функцию

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

Заметим, что $\left\|\vec{X} - Z^\mathsf{T}\vec{B}\right\|^2$ — квадрат расстояния от точки $\vec{X}$ до точки $Z^\mathsf{T}\vec{B}$, которая является точкой подпространства $\left\langle Z^\mathsf{T}\vec{t}\right\rangle$, где $\vec{t} \in \mathbb{R}^k$. Таким образом, искомое минимальное расстояние это расстояние до данного подпространства $\left\langle Z^\mathsf{T}\vec{t}\right\rangle$. Это расстояние получаем при условии, что вектор $\vec{X} - Z^\mathsf{T}\vec{B}$ будет ортогонален всем векторам подпространства, т.е. скалярное произведение $\left\langle Z^\mathsf{T}\vec{t}, \vec{X} - Z^\mathsf{T}\vec{B}\right\rangle = 0 \ \ \forall\vec{t}$

$$\left\langle Z^\mathsf{T}\vec{t}, \vec{X} - Z^\mathsf{T}\vec{B}\right\rangle = (Z^\mathsf{T}\vec{t})^\mathsf{T} \cdot (\vec{X} - Z^\mathsf{T}\vec{B})$$

[18] Это условие тривиально выполнить путём отброса линейно зависимых экспериментов.
[19] Это условие уже было в теореме Гаусса–Маркова

$$= \vec{t}^{\mathsf{T}} \cdot Z \cdot (\vec{X} - Z^{\mathsf{T}}\vec{B})$$
$$= \vec{t}^{\mathsf{T}} \cdot (Z\vec{X} - ZZ^{\mathsf{T}}\vec{B})$$

Т.к. скалярное произведение вектора со всеми другими векторами равно нулю тогда и только тогда, когда он является нулевым, то:

$$Z\vec{X} - ZZ^{\mathsf{T}}\vec{B} = 0$$
$$ZZ^{\mathsf{T}}\vec{B} = Z\vec{X}$$
$$A\vec{B} = Z\vec{X}$$

Это система из k линейных нормальных уравнений, из которой можно найти оценки $\vec{B}$ неизвестных параметров. По свойству 2 матрица невырожденная, поэтому эта система имеет единственное решение:

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

42 Свойства ОНМК в уравнении общей линейной регрессии.

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

   Proof.

$$\vec{B} - \vec{\beta} = A^{-1}Z\vec{X} - \vec{\beta}$$
$$= A^{-1}Z(Z^{\mathsf{T}}\vec{\beta} + \vec{\varepsilon}) - \vec{\beta}$$
$$= A^{-1}\underbrace{ZZ^{\mathsf{T}}}_{A}\vec{\beta} + A^{-1}Z\vec{\varepsilon} - \vec{\beta}$$
$$= A^{-1}Z\vec{\varepsilon}$$

   □

2. B — несмещённая оценка параметров $\vec{\beta}$

   Proof.

$$\mathbb{E}\,\vec{B} = \mathbb{E}(A^{-1}Z\vec{\varepsilon} + \vec{\beta})$$
$$= \mathbb{E}(A^{-1}Z\vec{\varepsilon}) + \mathbb{E}\,\vec{\beta}$$
$$= A^{-1}Z\underbrace{\mathbb{E}\,\vec{\varepsilon}}_{0} + \vec{\beta}$$
$$= \vec{\beta}$$

   □

Кроме того, если $\mathbb{E}\,\vec{\varepsilon} \neq 0$, то оценка смещённая.

3. $\mathbb{D}\,\vec{\varepsilon} = \begin{pmatrix} \sigma^2 & & \\ & \ddots & \\ & & \sigma^2 \end{pmatrix} = \sigma^2 E_n$ [20]

Proof. По условию 2 $\varepsilon_i \in N(0, \sigma^2)$ и независимы, следовательно, $d_{ij} = 0$ при $i \neq j$ и $d_{ii} = \sigma^2$ □

4. $\mathbb{D}(\sqrt{A} \cdot \vec{B}) = \begin{pmatrix} \sigma^2 & & \\ & \ddots & \\ & & \sigma^2 \end{pmatrix} = \sigma^2 E_k$

Proof.

$$\begin{aligned}
\mathbb{D}(\sqrt{A} \cdot \vec{B}) &= \mathbb{E}(\sqrt{A} \cdot \vec{B} - \mathbb{E}\,\sqrt{A} \cdot \vec{B})(\sqrt{A} \cdot \vec{B} - \mathbb{E}\,\sqrt{A} \cdot \vec{B})^{\mathsf{T}} \\
&= \mathbb{E}(\sqrt{A}(\vec{B} - \mathbb{E}\,\vec{B})) \cdot (\sqrt{A}(\vec{B} - \mathbb{E}\,\vec{B}))^{\mathsf{T}} \\
&= \mathbb{E}(\sqrt{A}(\vec{B} - \vec{\beta})) \cdot (\sqrt{A}(\vec{B} - \vec{\beta}))^{\mathsf{T}} \\
&= \mathbb{E}(\sqrt{A}A^{-1}Z\vec{\varepsilon}) \cdot (\sqrt{A}A^{-1}Z\vec{\varepsilon})^{\mathsf{T}} \\
&= \sqrt{A}A^{-1}Z\,\mathbb{E}(\vec{\varepsilon} \cdot \vec{\varepsilon}^{\mathsf{T}})Z^{\mathsf{T}}(A^{-1})^{\mathsf{T}}(\sqrt{A})^{\mathsf{T}} \\
&= \sqrt{A}A^{-1}Z\sigma^2 E_n Z^{\mathsf{T}}A^{-1}\sqrt{A} \\
&= \sigma^2\sqrt{A}A^{-1}\underbrace{ZZ^{\mathsf{T}}}_{A}A^{-1}\sqrt{A} \\
&= \sigma^2\sqrt{A}A^{-1}\sqrt{A} \\
&= \sigma^2\sqrt{A}\sqrt{A^{-1}}\sqrt{A^{-1}}\sqrt{A} \\
&= \sigma^2 E_k
\end{aligned}$$

□

Следствие 21.1. Координаты вектора $\sqrt{A} \cdot \vec{B}$ некоррелированны.

5. $\mathbb{D}\,\vec{B} = \sigma^2 A^{-1}$

Proof. По свойству 4 $\mathbb{D}(\sqrt{A} \cdot \vec{B}) = \sigma^2 E_k$. В силу билинейности ковариации:

$$\sigma^2 E_k = \mathbb{D}(\sqrt{A} \cdot \vec{B}) = (\sqrt{A})^2\,\mathbb{D}\,\vec{B} = A\,\mathbb{D}\,\vec{B}$$

□

[20] E_n — единичная матрица $n \times n$

Отсюда можем получить дисперсии оценок b_i по формулам:

$$\mathbb{D}\, b_i = \sigma^2 (A^{-1})_{ii}$$

Обозначим через $\hat{\sigma}^2 := \frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \frac{1}{n}\sum_{i=1}^{n} (\vec{X} - Z^{\mathsf{T}}\vec{B})^2$. Заметим, что $\hat{\sigma}^2$ — точная оценка неизвестного параметра σ^2.

43 Основная теорема об ОМНК (п.2 без доказательства).

Теорема 22. Пусть выполнены условия 1 и 2. Тогда:

1. Вектор $\frac{1}{\sigma}\sqrt{A}(\vec{B} - \vec{\beta})$ состоит из k независимых случайных величин со стандартным нормальным распределением.
2. $\frac{n\hat{\sigma}^2}{\sigma^2} \in H_{n-k}$ и не зависит от $\vec{B}$.
3. Исправленная оценка $S^2 = \frac{n\hat{\sigma}^2}{n-k} = \frac{1}{n-k}\sum_{i=1}^{k} \hat{\varepsilon}_i^2$ — несмещённая оценка для σ^2.

Proof.

1. Вектор $\sqrt{A}(\vec{B} - \vec{\beta}) = \sqrt{A}A^{-1}Z\vec{\varepsilon} = (\sqrt{A})^{-1}Z\vec{\varepsilon}$ является линейным преобразованием нормального вектора $\vec{\varepsilon}$, поэтому имеет нормальное совместное распределение, т.е. его координаты — нормальные случайные величины.

   По свойству 2 $\mathbb{E}(\sqrt{A}(\vec{B} - \vec{\beta})) = 0$, а его матрица ковариаций по свойству 4:

   $$\mathbb{D}(\sqrt{A}(\vec{B} - \vec{\beta})) = \mathbb{D}(\sqrt{A}\vec{B}) = \sigma^2 E_k$$

   Отсюда видим, что координаты данного вектора не коррелированны и, следовательно, независимы. Тогда $\mathbb{D}\left(\frac{1}{\sigma}\sqrt{A}\vec{B}\right) = E_k$
2. Нудно.
3. Т.к. $\mathbb{E}(\chi^2_{n-k}) = n - k$, то:

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

   Это смещённая оценка, следовательно, $S^2 = \dfrac{n\hat{\sigma}^2}{n-k}$

□

44 Мультиколлинеарность, ее неприятные последствия. Основные принципы отбора факторов в модель общей линейной регрессии.

Определение. Мультиколлинеарность — наличие заметной линейной связи между всеми или несколькими факторами.

Неприятные последствия:

1. Оценки параметров становятся ненадежными — имеют большие стандартные ошибки и малую значимость[21].
2. Небольшое изменение исходных данных существенно влияет на изменение оценок регрессии.
3. Трудно выявить изолированное влияние конкретного фактора на результат и физический (экономический) смысл этого влияния.

44.0.1 Начальный отбор факторов модели

Находим корреляционную матрицу, состоящую из коэффициентов линейной корреляции:

$$R = \begin{pmatrix} 1 & r_{X,Z_1} & r_{X,Z_2} & \dots & r_{X,Z_k} \\ r_{X,Z_1} & 1 & r_{Z_1,Z_2} & \dots & r_{Z_1,Z_k} \\ \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ r_{X,Z_k} & r_{Z_1,Z_k} & \dots & \dots & 1 \end{pmatrix}$$

Алгоритм:

1. Берём фактор, наиболее коррелирующий с X.
2. По очереди добавляем факторы со свойствами:
   (a) Корреляция с X как можно большая.
   (b) Корреляция с раннее введенными факторами как можно меньше.

Пример.

	X	Z_1	Z_2	Z_3
X	1	–	–	–
Z_1	0.85	1	–	–
Z_2	0.81	0.93	1	–
Z_3	−0.65	−0.43	−0.19	1

В модель вводится Z_1, Z_3

[21] Об этом — в конце лекции.

45 Стандартная ошибка общей линейной регрессии и стандартные ошибки коэффициентов регрессии. Проверка гипотезы о значимости отдельного коэффициента регрессии.

Пусть:

- $X = \beta_0 + \beta_1 Z_1 + \dots \beta_k Z_k + \varepsilon$ — теоретическая модель.
- $\hat{X} = b_0 + b_1 Z_1 + \dots + b_k Z_k$ — уравнение, полученное методом наименьших квадратов.

Как и раньше, считаем, что $\varepsilon \in N(0, \sigma^2)$.

Согласно пункту 3 теоремы 22 $S^2 = \frac{1}{n-k-1}$[22] $\sum_{i=1}^{k} \hat{\varepsilon}_i^2$ — несмещённая оценка для σ^2, где $\varepsilon_i = X_i - \hat{X}_i$.

Определение. S **— стандартная ошибка регрессии**.

Из свойства 5 имеем $\mathbb{D}\, b_i = \sigma^2 (A^{-1})_{ii}$, где $A = ZZ^{\mathsf{T}}$, Z — матрица плана. Соответственно $\hat{\mathbb{D}}\, b_i = S^2 (A^{-1})_{ii}$ — оценка $\mathbb{D}\, b_i$.

$S_{b_i} = S\sqrt{(A^{-1})_{ii}}$ **— стандартная ошибка коэффициента** b_i. Проверяется основная гипотеза $H_0 : \beta_i = 0$ против $H_1 : \beta_i \neq 0$.

Теорема 23. Если H_0 верна, то $T_i = \frac{b_i}{S_{b_i}} \in T_{n-k-1}$, где T_{n-k-1} — распределение Стьюдента с $n - k - 1$ степенями свободы, а S_{b_i} — стандартная ошибка коэффициента b_i.

Критерий: $t_{\text{кр}}$ — квантиль $|T_{n-k-1}|$ уровня значимости α. Тогда:

$$\begin{cases} H_0 : T_i < t_{\text{кр}} \Rightarrow Z_i \text{ можно выкинуть из модели} \\ H_1 : T_i \geq t_{\text{кр}} \Rightarrow Z_i \text{ можно оставить в модели} \end{cases}$$

При большой мультиколлинеарности может случиться так, что все коэффициенты статистически не значимы, а уравнение в целом значимо.

46 Уравнение регрессии в стандартных масштабах. Смысл стандартизованных коэффициентов и частных коэффициентов эластичности. Разложение влияния фактора на прямое и косвенное.

Так как факторы имеют различную природу и измеряются в различных единицах, то по коэффициентам уравнения МНК нельзя судить о силе влияния каждого фактора. Поэтому удобно стандартизовать исследуемые величины:

$$t_X := \frac{X - \overline{X}}{\sigma_X} \quad t_j := \frac{Z_j^{(i)} - \overline{Z_j}}{\sigma_{Z_j}}, 1 \leq j \leq k$$

[22] Из-за свободного члена.

Если в уравнении регрессии заменить величины стандартизованными величинами, то получим уравнение стандартных масштабов:

$$t_X = \gamma_1 t_1 + \dots \gamma_k t_k$$

Примечание. $\gamma_0 = 0$, т.к. $\underbrace{\overline{t_k}}_{=0} = \gamma_1 \underbrace{\overline{t_1}}_{=0} + \dots + \gamma_k \underbrace{\overline{t_k}}_{=0} + \gamma_0$

При этом система нормальных уравнений приобретает простой вид:

$$\begin{cases} \gamma_1 + r_{Z_1,Z_2}\gamma_2 + r_{Z_1,Z_3}\gamma_3 + \dots + r_{Z_1,Z_k}\gamma_k = r_{Z_1,X} \\ r_{Z_2,Z_1}\gamma_1 + \gamma_2 + r_{Z_2,Z_3}\gamma_3 + \dots + r_{Z_2,Z_k}\gamma_k = r_{Z_2,X} \\ \vdots \\ r_{Z_k,Z_1}\gamma_1 + r_{Z_k,Z_2}\gamma_2 + \dots + \gamma_k = r_{Z_k,X} \end{cases}$$

Или в матричной форме:

$$R\Gamma = R_X$$

, где R — матрица корреляции, $\Gamma = \begin{pmatrix} \gamma_1 \\ \vdots \\ \gamma_k \end{pmatrix}$, $R_X = \begin{pmatrix} r_{X,Z_1} \\ \vdots \\ r_{X,Z_k} \end{pmatrix}$

γ_i и b_i связаны соотношением:

$$\boxed{b_i = \gamma_i \cdot \frac{\sigma_X}{\sigma_{Z_i}}}$$

Частный случай — уравнение парной линейной регрессии:

$$\frac{X - \overline{X}}{\sigma_X} = r_{\text{в}} \cdot \frac{Z - \overline{Z}}{\sigma_Z}$$

В данном случае $\gamma_1 = r_{\text{в}}$

Смысл стандартизованных коэффициентов γ_i: γ_i показывает, на какую часть своего среднего отклонения σ_X изменится результат X при изменении фактора Z_i на величину своего среднего отклонения σ_{Z_i}.

При мультиколлинеарности факторы результата оказывают не только прямое воздействие на результат, но и косвенное через влияние других факторов. Стандартизованный коэффициент γ_i можно трактовать как показатель прямого влияния фактора Z_i на результат. Косвенное влияние этого фактора показывают остальные слагаемые: $\sum_{j \neq i} \gamma_j r_{Z_i,Z_j}$.

$$\overbrace{r_{Z_i,Z_1}\gamma_1 + r_{Z_i,Z_2}\gamma_2 + \dots}^{\text{косвенное влияние}} + \underbrace{\gamma_i}_{\text{прямое влияние}} + \overbrace{r_{Z_i,Z_{i+1}}\gamma_{i+1} + \dots + r_{Z_i,Z_k}\gamma_k}^{\text{косвенное влияние}} = r_{Z_i,X}$$

Примечание. Для измерения тесноты линейной связи между конкретным фактором и результатом при устранении других факторов есть понятие коэффициента частной корреляции.

46.1 Частные коэффициенты эластичности

Определение. Пусть имеется уравнение регрессии $X = f(Z_1 \dots Z_k)$. **Частными коэффициентами эластичности** называются величины

$$\boxed{Э_i = \frac{\partial f}{\partial Z_i} \cdot \frac{\overline{Z_i}}{\overline{X}}}$$

Смысл: $Э_i$ показывает, на сколько процентов от среднего изменится X при изменении Z_i на 1% от своего среднего уровня при фиксированных значениях других факторов. В случае линейной регрессии:

$$\boxed{Э_i = b_i \cdot \frac{\overline{Z_i}}{\overline{X}}}$$

Примечание. Коэффициенты эластичности и стандартизованные коэффициенты могут привести к противоположным выводам. Причины:

1. Вариация фактора очень велика.
2. Воздействие факторов разнонаправленно.
3. Мультиколлинеарность.

47 Коэффициенты детерминации и множественной корреляции, их свойства. Проверка гипотезы о значимости уравнения регрессии в целом.

Допустим, что дисперсию результата X можно разложить на объясненную моделью дисперсию и дисперсию остатков:

$$\mathbb{D}\, X = \mathbb{D}\, \hat{X} + \mathbb{D}\, \varepsilon$$

Определение. Коэффициентом детерминации R^2 называется:

$$\boxed{R^2 = \frac{\mathbb{D}\, \hat{X}}{\mathbb{D}\, X}}$$

или

$$\boxed{R^2 = 1 - \frac{\mathbb{D}\, \varepsilon}{\mathbb{D}\, X}}$$

Свойства.

1. $0 \leq R^2 \leq 1$
2. Если $R^2 = 1$, то $\mathbb{D}\,\varepsilon = 0$ и т.к. $\overline{\varepsilon} = 0$, то $\varepsilon_i = 0$ для всех i, т.е. все наблюдаемые точки лежат в гиперплоскости регрессии.
3. Если $R^2 = 0$, то $\mathbb{D}\,\hat{X} = 0, \forall i \;\; X_i = \overline{X}$ и $b_1 = \cdots = b_k = 0$.
4. Чем больше R^2, тем лучше модель.
5. В случае линейного уравнения регрессии:

$$R^2 = \sum_{i=1}^{k} \gamma_i r_{X,Z_i}$$

Определение. $R = \sqrt{R^2}$ **— коэффициент множественной корреляции**.

Примечание. При добавлении в модель нового фактора R^2 почти всегда увеличивается. Поэтому для выяснения необходимости введения новых факторов в модель используется **скорректированный коэффициент детерминации**:

$$\overline{R^2} = 1 - \frac{n-1}{n-k-1}\frac{\mathbb{D}\,\varepsilon}{\mathbb{D}\,X}$$

, где n — число экспериментов, k — число факторов в модели.

47.1 Проверка гипотезы о значимости уравнения регрессии в целом

Проверяется гипотеза $H_0 : R_T^2 = 0$ против гипотезы $H_1 : R_T^2 \neq 0$, т.е. уравнение статистически значимо.

Теорема 24. Если H_0 верна, то:

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

Критерий: $t_{\text{кр}}$ — квантиль $F(k, n-k-1)$ уровня значимости α. Тогда:

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

Примечание. По свойству 3 это эквивалентно проверке гипотезы $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$

48 Взвешенный МНК.

Пусть имеется уравнение множественной регрессии в векторной форме: $\vec{X} = Z^\mathsf{T}\vec{\beta} + \vec{\varepsilon}$, где $\vec{X} = (X_1 \dots X_k)^\mathsf{T}$ — наблюдаемые значения переменной, $Z = \left(Z_i^{(j)}\right)$ — матрица плана, $\vec{\beta} = (\beta_1 \dots \beta_k)^\mathsf{T}$ — ??? коэффициенты, $\vec{\varepsilon} = (\varepsilon_1 \dots \varepsilon_n)$ — ??? ошибки (случайные).

???, если[23]:

1. Строки Z — ??? независимые
2. $\varepsilon_i \in N(0, \sigma^2)$ — независимы

48.1 Некоррелированные наблюдения

Пусть $\mathbb{E}\,\varepsilon_i = 0, \mathbb{D}\,\varepsilon_i = \sigma^2 v_i$, но ошибки ε_i — некоррелированные, т.е. $\operatorname{cov}(\varepsilon_i, \varepsilon_j) = 0$ при $i \neq j$.

$\mathbb{D}\,\vec{\varepsilon} = \sigma^2 V$, где V — диагональная матрица с элементами v_i на главной диагонали.

Примечание. Логично придать наблюдениям с меньшим разбросом больший "вес". Делается это изменением масштаба следующим образом: положим $w_i = \frac{1}{v_i}, \tilde{X}_i = \sqrt{w_i} X_i$. Тогда получим новую модель $\vec{\tilde{X}} = \tilde{Z}^\mathsf{T}\vec{\beta} + \vec{\tilde{\varepsilon}}$, где $\tilde{Z}_i^{(j)} = \sqrt{w_j} Z_i^{(j)}, \tilde{\varepsilon}_i = \sqrt{w_i}\varepsilon_i$. При этом:

1. $\mathbb{E}\,\tilde{\varepsilon}_i = 0$
2. $\mathbb{D}\,\tilde{\varepsilon}_i = \mathbb{D}\sqrt{w_i}\varepsilon_i = \frac{1}{v_i}\mathbb{D}\,\varepsilon_i = \frac{1}{v_i} v_i \sigma^2 = \sigma^2$

Таким образом, получили прежнюю (классическую) ситуацию. Далее находим оценки $\vec{B} = (b_1 \dots b_k)^\mathsf{T}$ параметров $\beta = (\beta_1 \dots \beta_k)^\mathsf{T}$, как и раньше.

Пример.

1. Модель: $X = \beta_0 + \varepsilon$.

   $$\hat{\beta}_0 = \frac{\sum w_i Z_i X_i}{\sum w_i Z_i^2}$$

   В частности, при $Z_i = 1$, то $\hat{\beta}_0 = \frac{\sum w_i X_i}{\sum w_i}$ — взвешенное среднее.

2. Пропорция (прямая проходит через начало координат).

   Пусть X — потери тепла в квартире, Z — разность внутренней и наружней температуры. $X = \beta Z + \varepsilon$

   (a) Допустим, что $\mathbb{D}\,\varepsilon_i = cZ_i$. Тогда $w_i = \frac{\sigma^2}{cZ_i}$ и после вычислений получим

   $$\hat{\beta} = \frac{\sum X_i}{\sum Z_i} = \frac{\overline{X}}{\overline{Z}}$$

[23] Моя вольная интерпретация: это наши предположения.

(b) Допустим, что $\mathbb{D}\,\varepsilon_i = cZ_i^2$. Тогда

$$\hat{\beta} = \frac{1}{n}\sum_{i=1}^{n}\frac{X_i}{Z_i}$$

3. Повторные наблюдения.

   Пусть проводилось n серий из k_i экспериментов, $1 \le i \le n$. Пусть X_i — средний результат экспериментов этой серии, $\mathbb{E}\,X_{ij} = a, \mathbb{D}\,\varepsilon_{ij} = \sigma^2$. Тогда $\mathbb{D}\,\varepsilon_i = \frac{\sigma^2}{k_i}$. Берём $w_i = \frac{1}{k_i}$ и далее как раньше. То есть мы присваиваем больший вес сериям из большего числа экспериментов.

48.2 Коррелированные наблюдения

Пусть дисперсия ошибок $\mathbb{D}\,\varepsilon_i$ не только различны, но и зависимы между собой, т.е. $\operatorname{cov}(\varepsilon_i, \varepsilon_j) = \sigma^2 v_{ij}$. Пусть $V = (v_{ij})$. Тогда матрица ковариаций $\mathbb{D}\,\vec{\varepsilon} = \sigma^2 V$. Известно, что такая матрица симметричная и положительно определенная, следовательно, существует матрица $\sqrt{V}$. Итак, имеем модель $\vec{X} = Z^\mathsf{T}\vec{\beta} + \vec{\varepsilon}$. Умножим обе части слева на $\sqrt{V}^{-1}$, после чего получим новую модель:

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

, где $\vec{\tilde{X}} = \sqrt{V}^{-1}\vec{X}, \tilde{Z}^\mathsf{T} = \sqrt{V}^{-1}Z, \vec{\tilde{\varepsilon}} = \sqrt{V}^{-1}\vec{\varepsilon}$. Заметим, что:

1. $\mathbb{E}\,\vec{\varepsilon} = \mathbb{E}\,\sqrt{V}^{-1}\vec{\varepsilon} = \sqrt{V}^{-1}\,\mathbb{E}\,\vec{\varepsilon} = 0$
2. $\mathbb{D}\,\vec{\tilde{\varepsilon}} = \mathbb{E}\left(\left(\vec{\tilde{\varepsilon}} - \mathbb{E}\,\vec{\tilde{\varepsilon}}\right)\left(\vec{\tilde{\varepsilon}} - \mathbb{E}\,\vec{\tilde{\varepsilon}}\right)^\mathsf{T}\right) = \mathbb{E}\left(\vec{\tilde{\varepsilon}}\vec{\tilde{\varepsilon}}^\mathsf{T}\right) = \mathbb{E}\left(\sqrt{V}^{-1}\vec{\varepsilon}\left(\sqrt{V}^{-1}\vec{\varepsilon}\right)^\mathsf{T}\right) = \sigma^2 I$

Таким образом, мы получили стандартную ситуацию.

49 Приемы сведения нелинейных регрессий к линейным.

49.1 Модель очень быстрого роста

$$X = ae^{bZ}\cdot\varepsilon$$
$$\ln X = \ln a + bZ + \ln\varepsilon$$
$$X' = a' + bZ + \varepsilon'$$

49.2 Модель быстрого роста

$b > 1$

$$X = aZ^b\cdot\varepsilon$$
$$\ln X = \ln a + b\ln Z + \ln\varepsilon$$
$$X' = a' + bZ' + \varepsilon'$$

49.3 Модель медленного роста

$0 < b < 1$

$$X = aZ^b \cdot \varepsilon$$

Аналогично.

49.4 Модель очень медленнего роста

$$X = a + b \ln Z + \varepsilon$$
$$X = a + bZ' + \varepsilon$$

49.5 Модель медленной стабилизации

$$X = a + \frac{b}{Z} + \varepsilon$$
$$X = a + Z' + \varepsilon$$

49.6 Модель быстрой стабилизации

$$X = a + \frac{b}{e^Z} + \varepsilon$$
$$X = a + Z' + \varepsilon$$

49.7 Зависимость в виде полинома

$$X = \beta_0 + \beta_1 Z + \beta_2 Z^2 + \cdots + \beta_k Z^k$$

Сводим к линейной множественной регрессии: $Z_i = Z^i$

Примечание. Т.к. Z_i функционально зависимы, то обычно берут $k \leq 4$. Модель неустойчива.

50 Математические датчики случайных чисел.

Наиболее популярны мультипликативные датчики.

Задается начальное число k_0, множитель a и m — делитель (модуль). При этом $\gcd(k_0, m) = \gcd(a, m) = 1, 0 < a, k_0 < m$. Последовательность случайных чисел задается следующим образом:

$$k_n = ak_{n-1} \pmod m$$

Псевдослучайным числом тогда будет $\frac{k_n}{m}$.

Примечание. Эта процедура зациклится не более, чем за $m - 1$ итераций. Точнее, не более, чем за φ[24](m).

[24]Функция Эйлера.

50.0.1 Рекомендации

Для 32-х битных компьютеров:

- $m = 2^{31} - 1 = 2\ 147\ 483\ 647$
- $a = 630\ 360\ 016$ или $a = 764\ 261\ 123$
- k_0 — не важно.

50.0.2 Датчик Уичмана и Хилла

Одновременно запускаются три мультипликативных датчика с параметрами:

a	m
171	30 269
172	30 307
170	30 323

На n-том шаге получаем три числа: y'_n, y''_n, y'''_n. Тогда $y_n := \{y'_n + y''_n + y'''_n\}$ — дробная часть их суммы.

Преимущества:

1. Период $3 \cdot 10^{13}$, а у предыдущего $1 \cdot 10^9$.
2. Быстрее вычисляется.

51 Моделирование случайных величин методом обратной функции (включая дискретный случай).

Теорема 25. Пусть $F(x)$ — непрерывная строго возрастающая функция распределения. Если случайная величина θ имеет равномерное стандартное распределение, то случайная величина $\xi = F^{-1}(\theta)$ имеет функцию распределения $F(x)$.

Пример. Показательное распределение E_α.

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

$$y = 1 - e^{-x\alpha} \Leftrightarrow x = -\frac{\ln(1-y)}{\alpha} \in E_\alpha \Rightarrow x_i = -\frac{1}{\alpha}\ln \eta_i$$

Тогда x_i — значения E_α при $\eta_i \in U(0,1)$.

Пример. Нормальное распределение $N(a, \sigma^2)$. Достаточно смоделировать стандартное нормальное распределение $N(0,1)$.

$$F_0(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\alpha} e^{-\frac{z^2}{2}} dz$$
$$x_i = F_0^{-1}(\eta_i), \eta_i \in U(0,1)$$

- Достоинства: простота и универсальность
- Недостаток: неэффективность

Пусть имеется дискретное распределение, $p_k = P(\xi = c_k), k = 1, 2, \ldots$. $r_0 \coloneqq 0, r_m = \sum_{k=1}^{m} p_k$ — концы отрезков разбиения.

Пусть $y_i \in U(0,1)$ — псевдослучайное число. Если $y_i \in [r_{i-1}, r_i)$, то $x_i = c_i$. На самом деле это тот же метод обратной функции. В частности, если моделируем распределение Бернулли B_p, тогда $x_i = \begin{cases} 0, & y_i \in [0, 1-p) \\ 1, & y_i \in [1-p, 1] \end{cases}$

52 Моделирование нормальной случайной величины.

52.0.1 Нормальные случайные числа на основе ЦПТ

Пусть $\eta_i \in U(0,1), \mathbb{E}\,\eta_i = \frac{1}{2}, \mathbb{D}\,\eta_i = \frac{1}{12}, S_n \coloneqq \eta_1 + \cdots + \eta_n$. Согласно ЦПТ[25]:

$$\frac{S_n - na}{\sqrt{n\,\mathbb{D}\,\xi}} = \frac{S_n - \frac{n}{2}}{\sqrt{\frac{n}{12}}} \xrightarrow{n \to +\infty} Z \in N(0,1)$$

Уже при $n = 12$ распределение $S_n - 6$ неплохо приближается к $N(0,1)$.

52.0.2 Точное моделирование пары независимых случайных величин $N(0,1)$

Теорема 26. Пусть независимые случайные величины $\eta_1, \eta_2 \in U(0,1)$. Тогда следующие случайные величины X и Y независимы и $\in N(0,1)$:

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

Proof. Рассмотрим независимые случайные величины $X, Y \in N(0,1)$. Тогда плотность их совместного распределения это

$$f_{X,Y} = f_X(X) f_Y(Y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} = \frac{1}{2\pi} e^{-\frac{x^2+y^2}{2}}$$

Перейдём к полярным координатам, где радиус R, а угол — ϕ, тогда $X = R\cos\phi, Y = R\sin\phi, J = r$. Плотность в этих координатах:

$$f_{X,Y} = \underbrace{\frac{1}{2\pi}}_{\substack{f_\phi(\varphi) \\ 0 \le \varphi \le 2\pi}} \underbrace{e^{-\frac{r^2}{2}} r}_{\substack{f_R(r) \\ r \ge 0}}$$

[25] Центральной предельной теореме

Таким образом, R и ϕ — независимые случайные величины. Их функции распределения:

$$F_R(r) = \int_{-\infty}^{r} \rho e^{-\frac{\rho^2}{2}} d\rho = 1 - e^{-\frac{r^2}{2}}$$

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

$$F_R^{-1}(r) = \sqrt{-2\ln(1-y)} \quad F_\phi^{-1}(\varphi) = 2\pi y$$

Тогда $r_i := \sqrt{-2\ln\eta_1}, \varphi = 2\pi\eta_2$. Итого

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

□

53 Быстрый показательный датчик.

Теорема 27. Пусть независимые случайные величины $\eta_1 \dots \eta_{2n-1} \in U(0,1)$. Обозначим $\xi_1 \dots \xi_n$ за расставленные по возрастанию величины $\eta_{n+1} \dots \eta_{2n-1}, \xi_1 = 0, \xi_n = 1$. Тогда случайные величины

$$\mu_i = -\frac{1}{\alpha}(\xi_i - \xi_{i+1})\ln(\eta_1 \dots \eta_n), 1 \le i \le n$$

независимы и имеют показательное распределение с параметром α.

Proof. Не будем. □

Пример. При $n = 5$, пусть $\eta_5 > \eta_4$. Тогда:

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

Преимущество по сравнению с методом обратной функции — нужно вычислять только один логарифм. Минус — требуется сортировка. Способ наиболее эффективен при $n = 3$, при этом он примерно вдвое эффективнее метода обратной функции.

54 Моделирование дискретных случайных величин.

54.0.1 Биномиальное распределение

$$P(\xi = k) = C_n^k p^k q^{n-k} \quad 0 \le k \le n$$

Смысл: число успехов при n испытаниях. Будем на это опираться.

Берём n значений датчика $y_i, 1 \le i \le n$ и положим $z_i := \begin{cases} 0, & y_i \in [0, 1-p) \\ 1, & y_i \in [1-p, 1] \end{cases}$. Тогда $x = \sum_{i=1}^{n} z_i$ — смоделированное значение $\in B_{n,p}$

54.0.2 Геометрическое распределение

$$P(\xi = k) = (1-p)^{k-1} p$$

Смысл: номер первого успешного испытания.

Берём последовательные значения датчика y_i до тех пор, пока $y_i \in [1-p, p]$. Ответ — индекс последнего опыта.

54.0.3 Распределение Пуассона

$$\Pi_\lambda : P(\xi = k) = \frac{\lambda^k}{k!} e^{-\lambda}, k \ge 0$$

Теорема 28. Пусть $\mu_1, \mu_2 \cdots \in E_\lambda$ — независимые случайные величины. Положим $S_n = \mu_1 + \cdots + \mu_n, N = \max\{n \mid S_n \in [0,1]\}$. Тогда $N \in \Pi_\lambda$.

На основе теоремы и метода обратных функций получаем формулу для моделирования:

$$\begin{aligned} N &= \max\{k \mid S_k \in [0,1]\} \\ &= \max\{k \mid \sum_{j=1}^{k} \mu_j \le 1\} \end{aligned}$$

По методу обратных функций, если $y_j \in N(0,1)$:

$$\begin{aligned} &= \max\left\{k \,\middle|\, \sum_{j=1}^{k} \frac{-\ln(1-y_j)}{\lambda} \le 1\right\} \\ &= \max\left\{k \,\middle|\, \sum_{j=1}^{k} -\ln(1-y_j) \le \lambda\right\} \\ &= \max\left\{k \,\middle|\, \sum_{j=1}^{k} -\ln(y_j) \le \lambda\right\} \end{aligned}$$

$$= \max\left\{k \;\middle|\; \sum_{j=1}^{k} \ln(y_j) \geq -\lambda\right\}$$
$$= \max\left\{k \;\middle|\; e^{\sum_{j=1}^{k} \ln(y_j)} \geq e^{-\lambda}\right\}$$
$$= \max\left\{k \;\middle|\; \prod_{j=1}^{k} y_{ij} \geq e^{-\lambda}\right\}$$
$$= \min\left\{k+1 \;\middle|\; \prod_{j=1}^{k} y_j < e^{-\lambda}\right\}$$
$$= \min\left\{k \;\middle|\; \prod_{j=0}^{k} y_j < e^{-\lambda}\right\}$$

Добавив индекс j:

$$N_{ij} = \min\left\{k \;\middle|\; \prod_{j=0}^{k} y_{ij} < e^{-\lambda}\right\}$$

, где y_{ij} — псевдослучайные числа $\in N(0,1)$ i-той серии.

Возможно, ещё нужно рассказать про моделирование через обратную функцию, см. конец билета 51, стр. 55.

55 Метод Монте-Карло. Общая постановка, оценка погрешности.

Общая постановка метода: пусть требуется найти a и имеется случайная величина ξ, такая что $\mathbb{E}\,\xi = a$. Тогда согласно сильному закону больших чисел

$$\frac{\xi_1 + \cdots + \xi_n}{n} \xrightarrow[n\to\infty]{\text{п.н.}} a$$

Поэтому при достаточно больших n среднее выборочное $\overline{X}$ будет неплохой оценкой a.

Оценим погрешность вычислений. Пусть $\mathbb{D}\,\xi$ конечна, тогда по ЦПТ

$$\frac{S_n - na}{\sqrt{n\,\mathbb{D}\,\xi}} = \frac{n(\overline{X} - a)}{\sqrt{n\,\mathbb{D}\,\xi}} = \frac{\sqrt{n}(\overline{X} - a)}{\sqrt{\mathbb{D}\,\xi}} \xRightarrow{n\to\infty} Z \in N(0,1)$$

По правилу трёх сигм $P(|Z| < 3 \cdot 1) = 0,9973$. Тогда:

$$P\left(\frac{\sqrt{n}\,|\overline{X} - a|}{\sqrt{\mathbb{D}\,\xi}} < 3\right) \xrightarrow{n\to\infty} 0,9973$$

и поэтому можно считать, что $\frac{\sqrt{n}\left|\overline{X}-a\right|}{\sqrt{\mathbb{D}\,\xi}} < 3$ и, следовательно:

$$\left|\overline{X}-a\right| < \frac{3\sqrt{\mathbb{D}\,\xi}}{\sqrt{n}}$$

По возможности надо брать ξ с минимальной дисперсией.

56 Вычисление определенного и кратного интегралов методом Монте-Карло. Метод расслоенной выборки.

Отличается от метода прямоугольников тем, что в качестве узлов берутся случайные числа.

Пусть нужно посчитать $\int_0^1 \varphi(x)dx$, $\eta_i \in U(0,1)$. Тогда $\xi_i \coloneqq \varphi(\eta_i)$, плотность $f_\eta(y) = 1, y \in [0,1]$, 0 иначе.

$$\mathbb{E}\,\xi_1 = \int_{-\infty}^{+\infty} \varphi(y) f_\eta(y) dy = \int_0^1 \varphi(y)dy = I$$

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

Погрешность: $|I - I_n| \le \frac{3\sqrt{\mathbb{D}\,\xi_1}}{\sqrt{n}}$, где $\mathbb{D}\,\xi_1 = \int_0^1 \varphi^2(y)dy = I^2$

Недостатки:

1. Медленная сходимость (корень вместо квадрата или четвёртой степени).
2. Для оценки погрешности надо вычислить дисперсию.
3. Оценка справедлива лишь с вероятностью, пусть и близкой к единице.

В силу этого метод Монте-Карло не применяется для вычисления интегралов.

56.1 Вычисление кратных интегралов

По квадратурным формулам (при достаточно высокой кратности интеграла) погрешность $\le Cn^{-1+\varepsilon}$

Пусть требуется вычислить интеграл $\int \cdots \int_0^1 \varphi(x_1 \dots x_k)dx_1 \dots dx_k$. При методе Монте-Карло здесь достаточно набросать n случайных равномерно распределенных точек и взять среднее арифметическое значения φ в этих точках.

56.2 Метод Монте-Карло расслоенной выборкой

Пусть требуется вычислить интеграл $\int \cdots \int_0^1 \varphi(x_1 \dots x_k)dx_1 \dots dx_k$. Каждую из сторон k-мерного куба разобьем нам N равных частей. Тогда куб разобьется на N^k кубиков Δ_i и

в каждом из Δ_i возьмём k-мерную равномерно распределенную точку $\eta_i = (\eta_i^{(1)} \dots \eta_i^{(k)})$. Интеграл оценивается при помощи суммы

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

Погрешность: $|\tilde{I}_n - I| \leq C n^{-\frac{1}{2}-\frac{1}{k}}$

Пример. При $k = 1$ $|\tilde{I}_n - I| \leq C n^{-\frac{3}{2}}$