

Московский государственный технический университет
имени Н. Э. Баумана

Факультет: Фундаментальные науки

Кафедра: Математическое моделирование

Дисциплина: Математическая статистика

Теория к РК по модулю 2

Выполнила: Покасова А.И.
Группа: ИУ7-61

Москва, 2019 г.

Пусть X – случайная величина, закон распределения которой неизвестен (известен не полностью).

Определение 0.1. Статистической гипотезой называется любое утверждение относительно закона распределения случайной величины X .

Определение 0.2. Статистическая гипотеза называется простой, если она однозначно определяет закон распределения случайной величины X (однозначно задает функцию распределения случайной величины X как функцию своего аргумента). В противном случае статистическая гипотеза называется сложной.

Определение 0.3. Статистическая гипотеза называется параметрической, если она является утверждением относительно значений неизвестного параметра известного закона распределения.

1 Проверка статистических гипотез

Проверку статистической гипотезы обычно формулируют следующим образом:

1. Формулируют основную гипотезу H_0
2. Формулируют конкурирующую гипотезу H_1 . $H_0 \cap H_1 = \emptyset$, но, возможно, H_0 и H_1 не исчерпывают все возможные случаи.
3. На основании имеющейся выборки $\vec{x} \in \chi_n$ принимают решение об истинности H_0 и H_1 .

Определение 1.1. Правило, посредством которого принимается решение об истинности H_0 или H_1 называется статистическим критерием проверки гипотезы.

Задают критерий проверки статистической гипотезы обычно с помощью критического множества $W \in \chi_n$. При этом решающее правило имеет вид:

$$\vec{x} \in W \implies \begin{cases} \text{отклоняют } H_0 \\ \text{принимают } H_1 \end{cases} \quad \vec{x} \notin W \implies \begin{cases} \text{принимают } H_0 \\ \text{отклоняют } H_1 \end{cases}$$

Замечание. 1. Задать критерий проверки гипотез и задать критическое множество – одно и то же

2. При использовании любого критерия возможны ошибки двух видов:
 - (a) принять конкурирующую гипотезу при истинности основной гипотезы – ошибка первого рода: $P\{\vec{x} \in W | H_0\} = \alpha$
 - (b) принять основную гипотезу при истинности конкурирующей – ошибка второго рода: $P\{\vec{x} \notin W | H_1\} = \beta$

Определение 1.2. α называется уровнем значимости, а $1 - \beta$ – мощностью критерия.

Критерий Неймана-Пирсона

Пусть:

1. X – случайная величина
2. $F(x, \theta)$ – функция распределения случайной величины X (известны общий вид функции F , но она зависит от неизвестного параметра θ)

При построении критерия для проверки статистических гипотез, как правило, исходят из необходимости максимизации его мощности $1 - \beta$ (минимизация вероятности совершения ошибки второго рода) при фиксированном уровне значимости α критерия.

Введём в рассмотрение статистику:

$$\phi(\vec{X}) = \frac{L(\vec{X}; \theta_1)}{L(\vec{X}; \theta_0)},$$

где $L(\vec{X}; \theta)$ – функция правдоподобия.

Определение 1.3. Статистика $\phi(\vec{X})$ называется отношением правдоподобия.

Критическое множество должно иметь вид:

$$W = \{\vec{x} \in \chi_n : \phi(\vec{X}) \geq C_\phi\},$$

где константа C выбирается из условия

$$\alpha = P\{\phi(\vec{X}) \geq C_\phi | \theta = \theta_0\}$$

Пример. Пусть $X \sim N(m, \sigma^2)$, где m – неизвестно, σ^2 – известно.

Рассмотрим задачу проверки двух простых гипотез $H_0 = \{m = m_0\}$, $H_1 = \{m = m_1\}$, где $m_0 < m_1$.

В этом примере функция правдоподобия имеет вид:

$$L(X_1, \dots, X_n, m) = \left(\frac{1}{\sqrt{2\pi\sigma}}\right)^n e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - m)^2}$$

Тогда отношение правдоподобия:

$$\begin{aligned} \phi(\vec{X}) &= \frac{L(\vec{X}, m_1)}{L(\vec{X}, m_0)} = \frac{e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - m_1)^2}}{e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - m_0)^2}} = \\ e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n [(x_i - m_1)^2 - (x_i - m_0)^2]} &= e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n [x_i^2 - 2x_i m_1 + m_1^2 - x_i^2 + 2x_i m_0 - m_0^2]} = e^{\frac{m_1 - m_0}{\sigma^2} \sum_{i=1}^n x_i - \frac{n}{2\sigma^2} [m_1^2 - m_0^2]} \end{aligned} \quad (1)$$

Выше было показано, что критическое множество должно иметь вид

$$W = \{\vec{X} : \phi(\vec{X}) \geq C_\phi\},$$

где $C_\phi = \text{const}$ выбирается из условия

$$P\{\phi(\vec{X}) \geq C_\phi | H_0\} = \alpha$$

Условие

$$\begin{aligned} \phi(\vec{X}) \geq C_\phi &\Leftrightarrow \ln \phi(\vec{X}) \geq \ln C_\phi \Leftrightarrow |\text{см. (2)}| \Leftrightarrow \ln \left[e^{\frac{m_1 - m_0}{\sigma^2} \sum_{i=1}^n X_i - \frac{n}{2\sigma^2} (m_1^2 - m_0^2)} \right] \geq \\ \ln C_\phi &\Leftrightarrow \frac{m_1 - m_0}{\sigma^2} \sum_{i=1}^n X_i - \frac{n}{2\sigma^2} (m_1^2 - m_0^2) \geq \ln C_\phi \Leftrightarrow \frac{m_1^2 - m_0^2}{\sigma^2} \sum_{i=1}^n X_i \geq \\ &\ln C_\phi + \frac{n}{2\sigma^2 (m_1^2 - m_0^2)} \end{aligned}$$

$$\text{С учетом того, что } m_1 > m_0 \Leftrightarrow \sum_{i=1}^n X_i \geq \frac{\sigma^2}{m_1 - m_0} \left[\ln C_\phi - \frac{n}{2\sigma^2} [m_1^2 - m_0^2] \right], \quad C = \text{const}$$

Таким образом,

$$W = \{\vec{X} \in \chi_n : \sum_{i=1}^n X_i \geq C_\phi\},$$

где C выбирается из условия

$$\alpha = P\{\phi(\vec{X} \geq C_\phi | H_0)\} = P\{\sum_{i=1}^n X_i \geq C_\phi | m = m_0\}$$

Если истинна H_0 , т.е. $m = m_0$, то случайная величина $\sum_{i=1}^n X_i \sim N(nm_0, n\sigma^2)$ $|X_i \sim N(m_0, \sigma^2)$.

Таким образом, $\alpha = P\{\sum_{i=1}^n X_i \geq C | m = m_0\} = 1 - P\{\sum_{i=1}^n X_i \leq C | m = m_0\} = 1 - \Phi(\frac{C - nm_0}{\sqrt{n\sigma^2}})$, то есть $\Phi(\frac{C - nm_0}{\sqrt{n\sigma^2}}) = 1 - \alpha$.

Таким образом, $\frac{C - nm_0}{\sqrt{n\sigma^2}} = u_{1-\alpha}$, $C = \sigma u_{1-\alpha} \sqrt{n} + nm_0$.

Таким образом, критерий имеет вид

$$\sum_{i=1}^n X_i \geq \sigma u_{1-\alpha} \sqrt{n} + nm_0 \rightarrow \{\text{принять } H_1, \text{отклонить } H_0\}$$

$$\sum_{i=1}^n X_i < \sigma u_{1-\alpha} \sqrt{n} + nm_0 \{\text{принять } H_0, \text{отклонить } H_1\}$$

При этом вероятность совершения ошибки 1-го рода

$$p = P\{\vec{X} \notin W | H_1\} = P\{\sum_{i=1}^n X_i < C | m = m_0\} = |C = \sigma u_{1-\alpha} \sqrt{n} + nm_0, \text{прим } m = m_1 : \sum_{i=1}^n X_i \sim N(nm_1, n\sigma^2)|$$

$$\Phi\left(\frac{\sigma u_{1-\alpha} \sqrt{n} - n(m_1 - m_0)}{\sigma \sqrt{n}}\right) = \Phi\left(u_{1-\alpha} - \sqrt{n} \frac{m_1 - m_0}{n}\right)$$

Проверка сложных параметрических гипотез

Пусть

- X – случайная величина
- $F(x, \theta)$ – функция распределения случайной величины X (общий вид функции F известен, но F зависит от неизвестного параметра θ)

Рассмотрим задачу проверки двух сложных гипотез: $H_0 = \{\theta \in \Theta_0\}$ и $H_1 = \{\theta \in \Theta_1\}$, где $\theta_0 \cap \theta_1 = \emptyset$.

- $\theta_0 = \{\theta > \theta_0\}, \theta_1 = \{\theta < \theta_1\}$
- $\theta = \{\theta < \theta_0\}, \theta_1 = \{\theta > \theta_1\}$, где $\theta_0 < \theta_1$.

В этом случае критерий как и раньше задается с использованием критического множества W , а решающее правило имеет вид:

$$\begin{aligned} \vec{X} \in W &\rightarrow \{\text{принять } H_1, \text{отклонить } H_0\} \\ \vec{X} \notin W &\rightarrow \{\text{принять } H_0, \text{отклонить } H_1\} \end{aligned}$$

При этом ошибки первого и второго рода определяются как и раньше, но теперь их вероятности зависят от θ .

$$\alpha(\theta) = P\{\vec{X} \in W | \theta \in \Theta_0\},$$

$$\beta(\theta) = P\{\vec{X} \in \chi_n \setminus W | \theta \in \Theta_1\}.$$

Определение 1.4. Величина $\alpha = \sup_{\theta \in \Theta_0} \alpha(\theta)$ называется размером критерия.

Определение 1.5. Функция $M(\theta) = P\{\vec{X} \in W|\theta\} (*)$ называется функцией мощности критерия.

Замечание 1. 1) Условие (*), принятое в математической статистике удачней было бы записать в виде

$$M(t) = P\{\vec{X} \in W|\theta = t\},$$

то есть $M(\theta)$ – вероятность события $\{\vec{X} \in W\}$ при условии, что неизвестный параметр имеет значение θ .

2) Через функцию мощности можно выразить вероятности совершения ошибок первого и второго рода.

$$\alpha(\theta) = M(\theta), \quad \theta \in \Theta_0 \quad \beta(\theta) = 1 - M(\theta), \quad \theta \in \Theta_1.$$

Определение 1.6. Критерий, который при заданном размере α максимизирует функцию мощности одновременно по всем возможным критериям при всех $\theta \in \Theta_1$ называется равномерно наиболее мощным критерием.

Пример 1. Пусть $X \sim N(m, \sigma^2)$, где m – неизвестно, σ^2 – известна.

Рассмотрим задачу проверки гипотез $H_0 = \{m = m_0\}$ и $H_1 = \{m > m_0\}$.

1) Ранее была решена задача проверки двух параметрических гипотез $H_0 = \{m = m_0\}$ и $H_1 = \{m = m_1\}$, где $m_1 > m_0$. При этом критическое множество имеет вид:

$$W = \{\vec{X} \in \chi_n : \sum_{i=1}^n x_i \geq nm_0 + u_{1-\alpha}\sigma\sqrt{n}\} (*)$$

2) Так как построенное выше критическое множество не зависит от m_1 , то фактически этот критерий является равномерно наиболее мощным для проверки гипотез

$$H_0 = \{m = m_0\} \quad H_1 = \{m > m_0\}$$

Таким образом, для рассмотренной задачи критическое множество имеет вид (*).

Пример 2. Пусть $X \sim N(m, \sigma^2)$, где m – неизвестно, σ^2 – неизвестна.

Рассмотрим задачу проверки гипотез

$$H_0 = \{m = m_0\} \quad vs \quad H_1 = \{m > m_0\}$$

В этом примере целесообразно воспользоваться статистикой

$$T(\vec{X}) = \frac{\bar{X} - m_0}{S(\vec{X})} \sqrt{n} \sim (\text{при истинности } H_0) St(n-1)$$

Аналогично предыдущим примерам, критическое множество можно задать в виде

$$W = \{\vec{X} \in \chi_n : T(\vec{X}) \geq t_{1-\alpha}^{n-1}\},$$

где $t_{1-\alpha}^{n-1}$ – квантиль уровня $1 - \alpha$ распределения $St(n-1)$

Замечание 2. Пусть в условиях предыдущего примера требуется проверить следующие гипотезы:

- $H_0 = \{m = m_0\} \quad vs \quad H_1 = \{m < m_0\}$
- $H_0 = \{m = m_0\} \quad vs \quad H_1 = \{m \neq m_0\}$

Рассуждая аналогично, с использованием статистики

$$T(\vec{X}) = \frac{\bar{X} - m_0}{S(\vec{X})} \sqrt{n}$$

зададим критические множества в виде:

- $W = \{\vec{X} \in \chi_n : T(\vec{X}) \leq -t_{1-\alpha}^{n-1}\}$
- $W = \{\vec{X} \in \chi_n : |T(\vec{X})| \geq t_{1-\alpha/2}^{n-1}\}$

Пример 3. Пусть

- $X \sim N(m_1, \sigma_1^2)$
- $Y \sim N(m_2, \sigma_2^2)$
- m_1, m_2 – неизвестны, σ_1, σ_2 – известны

Рассмотрим задачу проверки следующих гипотез:

- $H_0 = \{m_1 = m_2\}$ vs $H_1 = \{m_1 > m_2\}$
- $H_0 = \{m_1 = m_2\}$ vs $H_1 = \{m_1 < m_2\}$
- $H_0 = \{m_1 = m_2\}$ vs $H_1 = \{m_1 \neq m_2\}$

Рассмотрим случайную величину $Z = X - Y$; $MZ = MX - MY$, поэтому сформулированные задачи эквивалентны задачам:

- $H_0 = \{m = 0\}$ vs $H_1 = \{m > 0\}$
- $H_0 = \{m = 0\}$ vs $H_2 = \{m < 0\}$
- $H_0 = \{m = 0\}$ vs $H_3 = \{m \neq 0\}$,

где $m = M[Z]$.

Рассмотрим статистику

$$T(\vec{X}, \vec{Y}) = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}},$$

где n_1 – объем выборки \vec{X} , n_2 – объем выборки \vec{Y} .

Закон распределения случайной величины T при истинности H_0 :

T является линейно наибольшей нормированной случайной величиной, следовательно T сама имеет нормальное распределение.

$$M[T] = \frac{1}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} (M\bar{X} - M\bar{Y}) = \text{при истинности } H_0 = 0$$

$$D[T] = \frac{1}{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} [D\bar{X} + D\bar{Y}] = \frac{1}{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \left[\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \right] = 1$$

Таким образом, при истинности H_0 статистика $T \sim N(0, 1)$. По этой причине критические множества в каждой из рассмотренных задач имеют вид:

- $W = \{(\vec{X}, \vec{Y}) \in \chi_n : T(\vec{x}, \vec{y}) \geq u_{1-\alpha}\}$
- $W = \{(\vec{X}, \vec{Y}) \in \chi_n : T(\vec{x}, \vec{y}) \leq -u_{1-\alpha}\}$
- $W = \{(\vec{X}, \vec{Y}) \in \chi_n : |T(\vec{x}, \vec{y})| \geq u_{1-\alpha/2}\}$

2 Критерии согласия

Определение 2.1. Первая задача математической статистики.

Дано: X – случайная величина, закон распределения которой неизвестен. Требуется найти закон распределения случайной величины X .

Определение 2.2. Вторая задача математической статистики.

Дано: X – случайная величина, закон распределения которой известен с точностью до вектора $\vec{\theta}$ неизвестных параметров.

Требуется оценить значение $\vec{\theta}$.

Решение первой задачи связано с проверкой основной гипотезы:

$$H_0 = \{F(t) \equiv F_0(t)\} = \{(\forall t \in \mathbb{R})(F(t) = F_0(t))\},$$

где

- $F(t)$ – функция распределения случайной величины X
- $F_0(t)$ – некоторая функция распределения

против конкурирующей гипотезы:

$$H_1 = \neg H_0 = \{(\exists t \in \mathbb{R})(F(t) \neq F_0(t))\}$$

Гипотеза может быть сложной и иметь вид:

$$H_0 = \{(\exists \vec{\theta}_0)(\forall t \in \mathbb{R})(F(t) = F_0(t, \vec{\theta}_0))\},$$

где

- $F(t)$ – функция распределения случайной величины X
- $F_0(t, \vec{\theta})$ – некоторая функция распределения, известная с точностью до вектора распределения θ

При этом конкурирующая гипотеза

$$H_1 = \neg H_0 = \{(\forall \vec{\theta})(\exists t \in \mathbb{R})(F(t) \neq F_0(t, \vec{\theta}))\}$$

Проверка основной гипотезы H_0 сводится к оценке величины

$$\Delta(F_n, F_0)$$

рассогласования эмпирической функции распределения и предполагаемой функции распределения F_0 .

Определение 2.3. Критерием согласия называется статистический критерий, предназначенный для проверки корректности гипотезы о том, что предполагаемый закон распределения $F_0(t, \vec{\theta})$ случайной величины X соответствует экспериментальным данным, представленным эмпирической функцией распределения $F_n(t)$.

2.1 Критерий Колмогорова

Для простой гипотезы:

Пусть:

- X – непрерывная случайная величина
- \vec{X} – случайная выборка из генеральной совокупности \vec{X} .

Рассмотрим задачу проверки гипотезы

$$H_0 = \{F(t) \equiv F_0(t)\} \quad vs \quad H_1 = \neg H_0$$

Замечание 3. Здесь $F_0(t)$ – полностью известная функция распределения, которая не зависит ни от каких неизвестных параметров. По этой причине H_0 – простая гипотеза.

Для решения этой задачи рассмотрим статистику

$$\Delta(\vec{X}),$$

реализации которой определяются соотношением

$$\Delta(\vec{X}) = \sup_{t \in \mathbb{R}} |F_n(t) - F_0(t)|,$$

где $F_n(t)$ – эмпирическая функция распределения, построенная по выборке \vec{x} .

Очевидно, что «малое» значение статистики Δ свидетельствуют об истинности H_0 , а «большие» – об истинности H_1 .

По этой причине критическое множество имеет вид

$$W = \{\vec{x} \in \chi_n : \Delta(\vec{X}) \geq \delta_{1-\alpha}\},$$

где $\delta_{1-\alpha}$ – квантиль уровня $1 - \alpha$ закона распределения случайной величины $\Delta(\vec{X})$.

При этом решающее правило имеет вид:

$$\begin{aligned} \vec{x} \in W &\rightarrow \text{принять } H_1, \text{ отклонить } H_0, \\ \vec{x} \in \bar{W} &\rightarrow \text{принять } H_0, \text{ отклонить } H_1 \end{aligned}$$

2.2 Критерий χ^2 для простой гипотезы

Пусть

- X – дискретная случайная величина
- X может принимать конечное множество значений a_1, \dots, a_n с неизвестными вероятностями p_1, \dots, p_l .

Требуется проверить основную гипотезу

$$H_0 = \{p_1 = p_1^0, \dots, p_l = p_l^0\},$$

где p_1^0, \dots, p_l^0 – некоторые известные значения,
против

$$H_1 = \neg H_0 = \{k \in \{1, \dots, l\} : p_k \neq p_k^0\}$$

Для решения этой задачи рассмотрим статистику

$$n_1(\vec{X}), \dots, n_l(\vec{X}),$$

где

$$n_k(\vec{x}) = \{\text{количество элементов выборки } \vec{x}, \text{ которые имеют значение } a_k\}$$

Замечание 4. Очевидно, что

$$n_1(\vec{X}) + \dots + n_l(\vec{X}) = n,$$

поэтому случайные величины $n_1(\vec{X}), \dots, n_l(\vec{X})$ – зависимы.

Теорема Пирсона. Пусть выполняются сделанные выше предположения.

Тогда при истинности H_0 последовательность случайных величин

$$\sum_{i=1}^l \frac{n_k(\vec{X} - np_k)^2}{np_k}$$

слабо сходится к случайной величине, имеющей распределение $\chi^2(l-1)$

Согласно этой теореме, при $n \rightarrow \infty$ случайная величина

$$\Delta(\vec{X}) = \sum_{i=1}^l \frac{(n_k(\vec{X}) - np_k^0)^2}{np_k} = n \sum_{i=1}^l \frac{\frac{n_k(\vec{X})}{n} - p_k^0}{p_k^0}$$

сходится к случайной величине, распределенной по закону $\chi^3(l-1)$.

Очевидно, что истинность основной гипотезы H_0 ассоциируется с малыми значениями статистики $\Delta(\vec{X})$, а истинность конкурирующей гипотезы H_1 – с «большими» положительными значениями. По этой причине критическое множество можно задать в следующем виде:

$$W = \{\vec{x} \in \chi_n : \Delta(\vec{x}) \geq h_{1-\alpha}^{l-1}\},$$

где $h_{1-\alpha}$ – квантиль уровня $1 - \alpha$ распределения $\chi^2(l-1)$

2.3 Критерий Колмогорова для сложной гипотезы

Требуется проверить гипотезу о принадлежности закону распределения случайной величины X заданному классу. По этой причине основная гипотеза H_0 будет сложной:

$$H_0 = \{(\exists \vec{\theta})(\forall t \in \mathbb{R})(F(t) = F_0(t, \vec{\theta}))\},$$

где

- $F(t)$ – теоретический (расово верный) закон распределения случайной величины X
- $F_0(t)$ – предполагаемый закон распределения случайной величины X
- $\vec{\theta}$ – вектор параметров закона F_0

Конкурирующая гипотеза $H_1 = \neg H_0$.

Для решения задачи:

1. построить точечную оценку $\hat{\vec{X}}$ для значения вектора параметров $\vec{\theta}$

2. использовать критерий Колмогорова для проверки простой гипотезы

$$H_0 = \{F(t) \equiv F_0(t, \hat{\vec{\theta}}(\vec{X}))\},$$

где $\hat{\vec{\theta}}$ – выборочное значение построенной оценки.

Недостаток: критерии перестают быть параметрическими, так как распределение модифицированной статистики

$$\Delta(\vec{X}) = \sup_{t \in \mathbb{R}} |F(t) - F_0(t, \hat{\vec{\theta}})|$$

зависит от выбранной точечной оценки, то есть от закона распределения случайной величины $\hat{\vec{\theta}}$.

Однако можно показать, что, если

1. $\hat{\vec{\theta}}$ – оценка максимального правдоподобия для вектора θ
2. Элементы $F_0(t, \vec{\theta})$ параметрического семейства получаются из какого-нибудь одного своего представителя с использованием преобразований сдвига и масштаба (вдоль оси $O t$), то есть

$$F_0(t, \vec{\theta}) = \tilde{F}_0\left(\frac{t - a}{b}\right),$$

где \tilde{F}_0 – какая-то фиксированная функция рассматриваемого семейства $F_0(t, \vec{\theta})$, а в-е значения которых зависят от значения $\vec{\theta}$ в левой части, то для использования критерия Колмогорова достаточно иметь только одну таблицу квантилей для каждого семейства.

2.4 Критерий χ^2 для сложной гипотезы

Пусть

1. X – дискретная случайная величина
2. X может принимать значения a_1, \dots, a_l с неизвестными вероятностями p_1, \dots, p_l
3. эти вероятности $p_k, k = \overline{1; l}$, зависят от неизвестных параметров $\vec{\theta}$, где $\vec{\theta} \in \Theta$, то есть в отличие от критерия для простой гипотезы теперь $p_k = p_k(\vec{\theta}), \theta \in \Theta, k = \overline{1; l}$

По этой причине основную гипотезу можно записать в виде:

$$H_0 = \{P\{X = a_k\} = p_{k_0}(\vec{\theta}), k = \overline{1; l}\}, \quad (2)$$

где $p_{k_0}(\vec{\theta})$ – известные функции, предполагаемые в зависимости вероятностей p_k от параметров $\vec{\theta}$.

$P\{X = a_k\} = p_k(\vec{\theta})$ – теоретические(расово верные) зависимости этих вероятностей от параметров; эти зависимости нам неизвестны.

Конкурирующую гипотезу выбирают такой: $H_1 = \neg H_0$

Для решения:

1. сначала строят оценку максимального правдоподобия для вектора $\vec{\theta} : \hat{\vec{\theta}}(\vec{X})$
2. вычисляют выборочное значение $\hat{\vec{\theta}}(\vec{x}) n_k(\vec{x})$

3. рассматривают статистику

$$\chi^2(\vec{X}) = \sum_{i=1}^l \frac{[n_k(\vec{X} - np_k(\vec{\theta}(\vec{X})))]^2}{np_{k_0}(\vec{\theta}(\vec{X}))} = n \sum_{i=1}^l \frac{[\frac{n_k(\vec{X})}{n} - p_{k_0}(\vec{\theta})(\vec{X})]^2}{p_{k_0}(\vec{\theta}(\vec{X}))},$$

которая в случае выполнения определенных условий гладкости функций $p_{k_0}(\vec{\theta})$ при $n \rightarrow \infty$ слабо сходится к случайной величине, имеющей распределение χ^{2l-r-1} , где r – размерность вектора θ .

4. поскольку при истинности основной гипотезы H_0 статистика $\chi^2(\vec{X})$ принимает «малые» значения, а при истинности H_1 – «большие» положительные значения, критическое множество можно записать в виде

$$W = \{\vec{x} : \chi^2(\vec{x}) \geq h_{1-\alpha}^{l-r-1}\},$$

где $h_{1-\alpha}^{l-r-1}$ – квантиль уровня $1 - \alpha$ распределения $\chi^2(l - r - 1)$

Замечание 5. О построении оценки максимального правдоподобия в рассматриваемом случае.

При истинности H_0 функция правдоподобия имеет следующий вид:

$$L(\vec{X}, \vec{\theta}) = \prod_{j=1}^n P\{X = X_j\} = \frac{n!}{n_1! \dots n_k!} \prod_{k=1}^l [p_{k_0}(\vec{\theta})]^{n_k(\vec{X})},$$

где $\sum_{k=1}^l n_k(\vec{X}) = n$.

Тогда уравнения правдоподобия

$$\frac{\partial \ln L}{\partial \theta_j} = 0, \quad j = \overline{1; r},$$

примут вид:

$$\sum_{k=1}^l \frac{n_k(\vec{X})}{p_{k_0}(\vec{\theta})} \cdot \frac{\partial p_{k_0}(\theta)}{\partial \theta_j} = 0, \quad j = \overline{1; r}$$

2.5 Критерий Смирнова

Пусть

1. X, Y – случайные величины
2. $F(t)$ – функция распределения случайной величины X , $G(t)$ – функция распределения случайной величины Y
3. \vec{X} – случайная выборка из генеральной совокупности X (объем n_1), \vec{Y} – случайная выборка из генеральной совокупности Y (объем n_2)

Требуется проверить гипотезу

$$H_0 = \{X, Y \text{ одинаково распределены}\} = \{(\forall t \in \mathbb{R})(F(t) = G(t))\} \quad \text{vs} \quad H_1 = \neg H_0$$

Если случайные величины X и Y непрерывны, то для решения этой задачи можно использовать статистику $\Delta(\vec{X}, \vec{Y})$, выборочное значение которой определяется формулой

$$\Delta(\vec{X}, \vec{Y}) = \sup_{t \in \mathbb{R}} |F_{n_1}(t) - G_{n_2}(t)|, \quad (3)$$

где $F_{n_1}(t), G_{n_2}(t)$ – эмпирические функции распределения, отвечающие выборкам \vec{x} и \vec{y} .

Если истинно H_0 , то в соответствии с теоремой о сходимости эмпирической функции распределения к теоретической функции распределения заключаем, что при достаточно больших n_1, n_2 значения статистики должны быть «малыми», а при истинности H_1 – «большими». По этой причине критическое множество можно задать в виде:

$$W = \{(\vec{x}, \vec{y}) : \Delta(\vec{X}, \vec{Y}) \geq \delta_{1-\alpha}\},$$

где $\alpha \in (0, 1)$ – заданный уровень значимости критерия, а $\delta_{1-\alpha}$ – квантиль уровня $1 - \alpha$ закона распределения статистики Δ при истинности H_0 .

Замечание 6. О законе распределения статистики $\Delta(\vec{X}, \vec{Y})$.

- Доказано, что при истинности H_0 закон распределения статистики Δ не зависит от $F(t)$ – теоретического закона распределения случайной величины X
- для небольших n_1, n_2 соответствующие распределения табулированы
- Смирнов доказал, что для $t > 0$

$$P\{\sqrt{\frac{n_1 n_2}{n_1 + n_2}} \Delta(\vec{X}, \vec{Y}) < t\} \rightarrow_{n_1 \rightarrow \infty, n_2 \rightarrow \infty} K(t),$$

$$\text{где } K(t) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2 t^2}$$

При достаточно больших n_1, n_2 можно считать, что случайная величина

$$A = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \Delta(\vec{X}, \vec{Y})$$

имеет своей функцией распределения $K(t), t > 0$.

3 Регрессионный анализ

Основные задачи регрессионного анализа – задачи, связанные с установкой стохастических зависимостей между случайной величиной Y и детерминированными X_1, \dots, X_p , носящими количественный характер.

Определение 3.1. Y стохастически зависит от X_1, \dots, X_p , если на изменение значений X_1, \dots, X_p реагирует изменением своего закона распределения.

Определение 3.2. Модель $\hat{\Phi}(t) = \theta_1 \Psi_1(t) + \dots + \theta_p \Psi_p(t)$, где $\Psi_j(t)$ – известная базовая функция, $\theta_j, j = 1, p$ – неизвестные параметры, называется линейной по параметрам, если каждый входящий в правую часть параметр входит линейно.

Определение 3.3. Оценкой, полученной методом наименьших квадратов (МНК - оценкой) вектора $\vec{\theta}$ называется такое его значение $\hat{\theta}$, которое дост.наим. значение функционалу $S(\vec{\theta})$, т.е. $S(\hat{\theta}) = \min_{\vec{\theta} \in \mathbb{R}^p} S(\vec{\theta})$, где $S(\hat{\theta})$ – мера близости аппроксимирующей функции $\hat{\Phi}$ и истинной Φ .

$$S(\vec{\theta}) = \sum_{i=1}^n (y_i - \hat{\Phi}(t_i))^2$$

Чем более удачно $\vec{\theta}$, тем меньше $S(\vec{\theta})$.

Теорема о вычислении МНК. Пусть $rg\Psi = p$, тогда $\hat{\theta} = (\Psi^T\Psi)^{-1}\Psi^T\vec{y}$.

Теорема о свойствах МНК-оценок. Пусть

1. $\varepsilon \sim N(0, \sigma^2)$
2. Реал. сл. величина ε в серии из n наблюдений незав.
3. $rg\Psi = p$
4. $\hat{\theta} = (\Psi^T\Psi)^{-1}\Psi^T\vec{y}$ – линейная оценка для θ

Тогда

1. $\hat{\theta}$ – несмещенная оценка θ
2. $\hat{\theta} \sim N(\vec{\theta}, \varepsilon)$ – нормальная случайная величина, где $\varepsilon = \sigma^2(\Psi^T\Psi)^{-1}$
3. Интервальная оценка уровня $1 - \alpha$ для параметра θ_j имеет вид

$$(\hat{\theta}_j - \Delta_j, \hat{\theta}_j + \Delta_j)$$

где

- $\Delta_j = t_{1-\alpha}^{n-p} \sqrt{\frac{d_j}{n-p} S(\hat{\theta})}$
- $t_{1-\alpha}^{n-p}$ – квантиль уровня $1 - \alpha$ $St(n - p)$
- d_j – j -й элемент главной диагонали матрицы $(\Psi^T\Psi)^{-1}$.

4 Примечания составителя

- Основная часть шпоры составлялась к 27.05.19, содержала значительное количество ошибок. В последней версии убраны старые ошибки, добавлены новые(верность данных уже особо не проверяется, т.к. РК я сдала, мне не актуально).
- Пункт «Пример» в секции «Критерий Неймана-Пирсона» относится к вопросу №3! Все остальное относится к вопросу №2.
Уже 2 человека наебнулись на этом.
Берегите себя и своих близких.
- «расово верный» – шутка, никогда не пишите так в РК(я серьезно, в лекциях этого нет, это все шутки составителя, т.е. меня)
- Шпоры распространяются под лицензией WTFPL, делайте с ними что хотите
- Если найдете ошибки в шпорах, создавайте issue в репозитории mathstat user-a-Euclidophren, или скачайте .tex файл и правьте сами(таки WTFPL), или напишите мне в телеграме(@neoisalie), я исправлю (или нет, идите нахуй).