Различение статистических гипотез

Определение.

Пусть X - выборки из неизвестного распределения $F_X \in \mathcal{F}$, где $\mathcal{F}$ - заданное множество априори возможных распределений выборки X. Выделим некоторое подмножество $\mathcal{F}_0 \subset \mathcal{F}, \mathcal{F}_1 = \mathcal{F} \backslash \mathcal{F}_0$.

- Гипотеза H_0 основная гипотеза $F_X \in \mathcal{F}_0$
- Гипотеза H_1 - альтернативная $F_X \in \mathcal{F}_1$

Определение.

Каждому критерию соответствует некоторое разбиение выборочного пространства $\mathfrak{X}$ на два взаимно дополнительных множества $\mathfrak{X}_0$ и $\mathfrak{X}_1$, где $\mathfrak{X}_0$ состоит из тех точек x, для которых H_0 принимается, а $\mathfrak{X}_1$ - из тех, для которых отвергается. Итак, критерий имеет вид:

$$H_0 \text{ отвергается } \iff X \in \mathfrak{X}_1$$

Определение.

Следуя любому критерию, мы можем принять правильно решение либо совершить одну из двух ошибок:

- ошибку 1 рода, отвергнув H_0, когда она верна.
- ошибку 2 рода, приняв H_0, когда она ложна.

Определение.

Функцией мощности критерия называется следующий функционал на множестве всех допустимых распределений $\mathcal{F}$ выборки X

$$W(F) = W(F, \mathfrak{X}_1) = P(X \in \mathfrak{X}_1 | F), \;\; F \in \mathcal{F}$$

Другими словами, W - это вероятность попадания значения выборки X в критическую область, когда F - ее истинное распределение.
Тогда вероятность ошибки 1 рода есть $W(F)$ при $F \in \mathcal{F}_0$, а 2 рода - $1 - W(F)$ при $F \in \mathcal{F}_1$

Описание критерия отношения правдоподобия

Функция отношения правдоподобия:
$l(x) = \dfrac{L(x, \theta_1)}{L(x, \theta_2)}$

Критическая область критерия Неймана-Пирсона:
$\mathfrak{X}^*_{1,\alpha} = \{x \in \mathfrak{X} : l(x) \geq c_\alpha\}$, где c_α: ошибка 1 рода равна α.
Наиболее мощный критерий с уровнем значимости α - параметрический критерий, минимизирующий ошибку 2 рода при заданной ошибке 1 рода.

По лемме Неймана-Пирсона: Критическая область $\mathfrak{X}^*_{1,\alpha}$ задает наиболее мощный критерий для гипотезы H_0 относительно альтернативы H_1 среди всех критериев с уровнем значимости α.

Вычисление функции отношения правдоподобия

$$l(x) = \frac{L(x, \theta_1)}{L(x, \theta_2)} = \frac{\prod_{i=1}^n \frac{-{\theta_1}^x}{\ln(1-\theta_1) x_i}}{\prod_{i=1}^n \frac{-{\theta_2}^x}{\ln(1-\theta_2) x_i}} = \prod_{i=1}^n \frac{\theta_1^{x_i}}{\theta_2^{x_i}} \cdot \frac{\ln(1-\theta_2)}{\ln(1-\theta_1)} =$$

$$= \frac{\ln^n(1-\theta_2)}{\ln^n(1-\theta_1)} \cdot \left(\frac{\theta_1}{\theta_2}\right)^{\sum xi}$$

Вычисление критической области

$$l(x) \geq c \iff \left(\frac{\theta_1}{\theta_2}\right)^{\sum x_i} \geq c \cdot \frac{\ln^n(1-\theta_1)}{\ln^n(1-\theta_2)} \iff \sum_{i=1}^n x_i \cdot \ln \frac{\theta_1}{\theta_2} \geq \ln c + \ln\left(\frac{\ln^n(1-\theta_1)}{\ln^n(1-\theta_2)}\right) \iff$$

$$\iff \sum_{i=1}^n x_i \geq \frac{\ln c + \ln\left(\frac{\ln^n(1-\theta_1)}{\ln^n(1-\theta_2)}\right)}{\ln \frac{\theta_1}{\theta_2}}$$

Пусть правая часть неравенства равна $t(c)$. Тогда $P(l(x) \geq c) = P\left(\sum x_i \geq t(c)\right)$

Рассмотрим асимптотический подход к различению гипотез. Выборка $X = (X_1, \dots, X_n)$ - независимые, одинаково распределенные случайные величины. $MX_i = \dfrac{-1}{\ln(1-\theta_1)} \dfrac{\theta_i}{1-\theta_i} = \mu_i$

$DX_i = -\theta_i \dfrac{\ln(1-\theta_i+\theta_i)}{(1-\theta_i)^2 \ln^2(1-\theta_i)} = \sigma_i^2$

Ниже, пока не будет написано иного $\mu_0 = \mu, \sigma_0 = \sigma$

Существуют конечные мат. ожидание и дисперсия $\longrightarrow$ выполняется ЦПТ. Тогда можем записать ЦПТ в форме Леви:

$$\sqrt{n}\frac{\overline{X}-\mu}{\sigma} \to N(0,1) \text{ при } \ n \to \infty$$

$$\sum_{i=1}^{n} X_i \sim N(n\mu_i, n\sigma^2)$$

Для определенности будем считать, что $\theta_2 < \theta_1$

Также известно, что если $\xi \sim N(\mu, \sigma^2)$, то $\eta = -\dfrac{\xi - \mu}{\sigma} \sim N(0,1)$.

В таком случае с.в. $-\dfrac{\sum_{i=1}^n X_i - n\mu_i}{\sqrt{n}\sigma_i} \sim N(0,1)$

$\alpha = P(l(x) \geq c_{\alpha)} = P\left(\sum_{i=1}^{n} x_i \geq t(c_\alpha)\right) = P\left(\dfrac{\sum x_i - n\eta}{\sqrt{n}\sigma} \geq \dfrac{t(c_\alpha) - n\mu}{\sqrt{n}\sigma}\right) = \Phi\left(-\dfrac{t_\alpha - n\mu}{\sqrt{n}\sigma}\right) = \Phi(-g_\alpha)$, где

$g_\alpha = g(t_\alpha) = g(t(c_\alpha)) = \dfrac{t(c_\alpha) - n\mu}{\sqrt{n}\sigma}$.

Так как $\Phi(-g)$ - непрерывная функция, то всегда найдем такое g_α. Таким образом, в данном случае критерий Неймана-Пирсона задается критической областью $\mathfrak{X}^*_{1,\alpha} = \{x : \dfrac{\sum x_i - n\mu}{\sqrt{n}\sigma} \geq g_\alpha\}$, $\Phi(-g_\alpha) = \alpha$

Из заданного нами g_α следует, что $t(c_\alpha) = n\mu + \sqrt{n}\sigma g_\alpha$.

$$\beta = P(l(x) < c_\alpha) = P\left(\sum x < t(c_\alpha)\right) = P\left(\frac{\sum x - n\mu}{\sqrt{n}\sigma} < \frac{t(c_\alpha) - n\mu}{\sqrt{n}\sigma}\right) =$$

$$P\left(\frac{\sum x - n\mu_1}{\sqrt{n}\sigma} < \frac{\mu_0 - \mu_1}{\sigma_1}\sqrt{n} + g_\alpha \frac{\sigma_0}{\sigma_1}\right) = \Phi\left(\frac{\mu_0 - \mu_1}{\sigma_1}\sqrt{n} + g_\alpha \frac{\sigma_0}{\sigma_1}\right)$$

Вычисление минимального количества материала

Заранее заданы вероятности ошибок α и β. Определим минимальное число наблюдений $n^* = n^*(\alpha, \beta) \to 0$ при $n \to \infty$. Получается, что n^* - наименьшее из n для которых $\beta(\alpha, n) \leq \beta$.

$\alpha = \Phi(-g_{\alpha)}, \beta = \Phi\left(\dfrac{\mu_0 - \mu_1}{\sigma_1}\sqrt{n} + g_\alpha \dfrac{\sigma_0}{\sigma_1}\right)$

Обозначим за квантили для α и β соответственно γ_α и γ_β.

$\gamma_\beta - g_\alpha \dfrac{\sigma_0}{\sigma_1} = \dfrac{\mu_0 - \mu_1}{\sigma_1}\sqrt{n}$

$n = \dfrac{(\sigma_1\gamma_\beta + \sigma_0\gamma_\alpha)^2}{(\mu_0 - \mu_1)^2}$

Так как n должно быть целым, то окгруляем серху.