Семинар №12.
F-критерий, критерий χ^2 и критерий Колмогорова.

Проверка линейных гипотез в линейной гауссовской модели.

Пусть, как обычно, $X = Z\theta + \varepsilon$, $\varepsilon \sim N(0, \sigma^2 I_n)$, $\theta = (\theta_1, \dots, \theta_k)^T$, $k \le n$ — линейная гауссовская модель. Мы хотим построить критерий для проверки линейной гипотезы следующего вида:

$$H_0 : T\theta = \tau,$$

где T – матрица размера $m \times k$, $m \le k$, $rank(T) = m$. (Пример: $T = (\frac{1}{k}, \dots, \frac{1}{k})$, т.е. мы рассматриваем гипотезу, что $\frac{1}{k}\sum\limits_{i=1}^{k} \theta_i$ равняется некоему τ.)

Вопрос: как строить критерий?

Как мы уже знаем, $\widehat{\theta} = (Z^T Z)^{-1} Z^T X$ – оптимальная оценка θ, причём $\widehat{\theta} \sim N(\theta, \sigma^2 (Z^T Z)^{-1})$. Поскольку T – это линейное преобразование, то $\widehat{t} = T\widehat{\theta}$ – оптимальная оценка для $T\theta$. Тогда, в предположении верности гипотезы H_0,

$$\widehat{t} \sim N\left(\tau, \sigma^2 T (Z^T Z)^{-1} T^T\right).$$

Обозначим $B = T(Z^T Z)^{-1} T^T$.

Утверждение.

$$\frac{1}{\sigma^2}(\widehat{t} - \tau)^T B^{-1} (\widehat{t} - \tau) \sim \chi^2_m. \tag{1}$$

И правда, можно убедиться в том, что выражение выше – это сумма квадратов m независимых стандартных нормальных величин. Заметим также, что это выражение зависит только от $\widehat{\theta}$ (так как $\widehat{t} = T\widehat{\theta}$), значит, по теореме об ортогональном разложении гауссовского вектора, оно не зависит от $X - Z\widehat{\theta}$. С другой стороны, уже доказывалось ранее, что $\frac{1}{\sigma^2}\|X - Z\widehat{\theta}\|^2 \sim \chi^2_{n-k}$. Т.е., чтобы избавиться от $\frac{1}{\sigma^2}$ в выражении (1) (мы же хотим построить центральную статистику), можно разделить выражение (1) на $\frac{1}{\sigma^2}\|X - Z\widehat{\theta}\|^2$. Следовательно,

$$F_T = \frac{(\widehat{t} - \tau)^T B^{-1} (\widehat{t} - \tau)}{\|X - Z\widehat{\theta}\|^2} \frac{n-k}{m} \sim F_{m,n-k},$$

где $F_{m,n-k}$ – так называемое распределение Фишера, $F_{m,n-k} =^d \frac{n-k}{m} \frac{\chi^2_m}{\chi^2_{n-k}}$, где χ^2_m и χ^2_{n-k} независимы. Статистика F_T называется $F-$отношением.

Замечание 1 *Распределение Фишера также возникает в задаче проверки гипотезы о равенстве дисперсий двух нормальных выборок – в этом случае берётся отношение выборочных дисперсий.*

F-критерий проверки линейных гипотез

Пусть $u_{1-\alpha}$ — $(1-\alpha)$-квантиль $F_{m,n-k}$. Т.к. $F-$отношение всегда больше 0, то критерий будет такой:

$$\text{если } F_T = \frac{(\widehat{t} - \tau)^T D^{-1} (\widehat{t} - \tau)}{\|X - Z\widehat{\theta}\|^2} \frac{n-k}{m} > u_{1-\alpha},$$

то H_0 отвергается.

Замечание 2 *Матрица T и вектор τ зависят от конкретной задачи и выбираются нами. Например, есть у нас две независимые выборки: $X_1, \dots, X_n$ из $N(a, \sigma^2)$ и $Y_1, \dots, Y_m$ из $N(b, \sigma^2)$, хотим проверить гипотезу $H_0 : a = b$. Для этого приводим задачу к линейной гауссовской модели путём составления вектора $W = (X_1, \dots, X_n, Y_1, \dots, Y_m)^T$, получаем, что вектор параметров $\theta = (a, b)^T$. Тогда для $T = (1, -1)$ и $\tau = 0$ получаем $H_0 : a - b = 0$.*

Задача. Пусть $X_1, \dots, X_n \sim N(a_1, \sigma^2)$, $Y_1, \dots, Y_m$ – независимые выборки. Построить $F-$критерий для проверки гипотезы $H_0 : a_1 = a_2$.

Решение. Составим $W = (W_1, \dots, W_{n+m})^T = (X_1, \dots, X_n, Y_1, \dots, Y_m)^T$, тогда неизвестный параметр $\theta = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$, $Z = \begin{pmatrix} 1 \dots 1 0 \dots 0 \\ \underbrace{0 \dots 0}_{n} \underbrace{1 \dots 1}_{m} \end{pmatrix}^T$. Предположение гипотезы можно представить в виде $a_1 - a_2 = 0$, т.е. в равенстве $T\theta = \tau$ $T = (1, -1)$ и $\tau = 0$.

Посчитаем $B = T(Z^T Z)^{-1} T^T = (1, -1) \begin{pmatrix} \frac{1}{n} & 0 \\ 0 & \frac{1}{m} \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \frac{1}{n} + \frac{1}{m}$.

Посчитаем $\widehat{t} = T\widehat{\theta} = T(Z^T Z)^{-1} Z^T W = \overline{X} - \overline{Y}$.

Теперь можно написать $F-$отношение:

$$F_T = \frac{(\overline{X} - \overline{Y})(\frac{1}{n} + \frac{1}{m})(\overline{X} - \overline{Y})}{|W - Z\widehat{\theta}|^2} \cdot \frac{n + m - 2}{1},$$

где $|W - Z\widehat{\theta}|^2 = \sum\limits_{i=1}^{n} (X_i - \overline{X})^2 + \sum\limits_{j=1}^{m} (Y_i - \overline{Y})^2 = n s_X^2 + m s_Y^2$.

Отсюда получаем $F-$критерий: $\frac{(\overline{X} - \overline{Y})^2}{n s_X^2 + m s_Y^2} \cdot \frac{nm(n+m-2)}{n+m} > u_{1-\alpha}$, $u_{1-\alpha}$ – $(1-\alpha)-$квантиль $F_{1,n+m-2}$. ■

Критерий согласия хи-квадрат.

Пусть $X_1, \dots, X_n$ – выборка из схемы Бернулли с $m \geq 2$ исходами, т.е. $a_1, \dots, a_m$ – исходы и $P(X_i = a_j) = p_j$ для $j = 1, \dots, m$. Хотим проверить простую гипотезу

$$H_0 : p_j = p_j^0, j = 1 \dots m.$$

Определение. Статистикой хи-квадрат называется

$$\widehat{\chi} = \sum_{i=1}^{m} \frac{(\mu_j - n p_j^0)^2}{n p_j^0},$$

где $\mu_j = \sum\limits_{i=1}^{n} I\{X_i = a_j\}$ - число выпадений j-того исхода.

Теорема 1 *(К. Пирсон)*
В условиях гипотезы H_0

$$\widehat{\chi} \xrightarrow[n \to \infty]{d} \chi^2_{m-1}.$$

Критерий хи-квадрат: Если $\widehat{\chi} > u_{1-\alpha}$, где $u_{1-\alpha}$ – $(1-\alpha)$-квантиль χ^2_{m-1}, то отвергаем H_0.
Применимость критерия: $\forall i \; n p_i^0 \geq 5$.

Чем удобен и для чего нужен критерий хи-квадрат. Критерий хи-квадрат, как и критерий согласия Колмогорова, применяется для проверки гипотезы о равенстве распределения, из которого берётся наша выборка, какому-то определённому распределению. В отличие от критерия Колмогорова, критерий хи-квадрат не требует больших вычислений, но является менее "точным". Как работать с критерием хи-квадрат: область значений нашей выборки разбиваем на несколько интервалов, и смотрим, сколько членов выборки попало в каждый интервал. Это будут μ_i. Дальше находим вероятности этих интервалов для распределения, равенство которому нашего выборочного распределения мы проверяем, это будут p_i^0. Остаётся только вычислить статистику хи-квадрат, если

она будет меньше квантили хи-квадрат распределения, то принимаем гипотезу H_0, если больше, то отвергаем.

Задача. По статистике, собранной в психиатрической больнице в течение года, количество пациентов, поступивших в отделение интенсивной терапии, имело следующее распределение по дням недели:

ПН – 36, ВТ – 53, СР – 35, ЧТ – 26, ПТ – 30, СБ – 44, ВС – 28.

Согласуются ли эти данные с гипотезой о том, что попадание в отделение не зависит от дня недели на уровне значимости 0,95 и 0,99?

Решение. $m = 7$, $p_j^0 = \frac{1}{7}$, поскольку предполагаемое распределение является равномерным.
Считаем общее количество наблюдений: $n = 36 + 53 + 35 + 26 + 30 + 44 + 28 = 252$, таким образом, $np_j^0 = 36$ $\forall j = 1, \ldots, 7$.
Считаем статистику $\widehat{\chi} = 0 + 8,03 + 0,03 + 2,78 + 1 + 1,78 + 1,78 = 15,4$. Квантили распределения хи-квадрат с 6 степенями свободы такие: $u_{0,95} = 12,6$, $u_{0,99} = 16,8$. Таким образом, на уровне значимости $0,95$ мы гипотезу о равномерном поступлении пациентов отвергаем, а на уровне 0,99 – не отвергаем (но во втором случае вероятность ошибки второго рода, т.е. принятия основной гипотезы, если она неверна, существенно увеличивается – высокий уровень значимости не является целесообразным, если наблюдений в нашей модели не так много.) ■

Критерий согласия Колмогорова.

Определение. Если $X = (X_1, \ldots, X_n)$ – выборка растущего размера, то критерий S_n (фактически, последовательность критериев) называется состоятельным, если функция мощности $\beta(Q, S_n) \to 1$ при $n \to \infty$ для любого распределения Q из $\mathcal{P}_1$ (для любого распределения из альтернативной гипотезы).

Пусть имеется выборка из неизвестного распределения с непрерывной функцией распределения F. Мы хотим проверить гипотезу о том, что эта самая функция распределения есть некая известная нам функция распределения F_0, т.е. проверяем простую гипотезу $H_0 : F = F_0$.

Определение. Эмпирической функцией распределения выборки $X = (X_1, \ldots, X_n)$ называется следующая случайная функция: $F_n(x) = \frac{1}{n} \sum\limits_{i=1}^{n} I\{X_i \le x\}$.

Определение. Расстояние Колмогорова – это случайная величина $D_n = \sup\limits_x |F_n(x) - F(x)|$.

Теорема 2 *(Колмогоров)*
В условиях гипотезы H_0

$$\lim_{n\to\infty} P(\sqrt{n} D_n \le t) = \lim_{n\to\infty} P(\sqrt{n} \sup_x |F_n(x) - F_0(x)| \le t) = K(t),$$

где $K(t) = \sum\limits_{j=-\infty}^{+\infty} (-1)^j e^{-2j^2t^2}$ *– функция Колмогорова (она является функцией распределения).*

Критерий Колмогорова. Если $\sqrt{n} D_n > K_{1-\alpha}$, где $K_{1-\alpha}$ – $(1-\alpha)$-квантиль функции Колмогорова, то отвергаем гипотезу H_0.

Замечание. $K_{1-\alpha} \approx \sqrt{-\frac{1}{2} \ln \frac{\alpha}{2}}$, если $\alpha \approx 0$.