Тема 5Б. Критерии проверки гипотез об однородности. Однофакторный дисперсионный анализ.

1. Критерий однородности Колмогорова-Смирнова.

Пусть совокупность наблюдений состоит из двух независимых выборок: $\xi^{(n)} = (\xi_1,..., \xi_n)$ – выборка из распределения $F_\xi(x)$, $\eta^{(m)} = (\eta_1,..., \eta_m)$ – выборка из распределения $F_\eta(x)$, $\xi^{(n)}$ и $\eta^{(m)}$ независимы. Основная гипотеза H_0 заключается в том, что выборки $\xi^{(n)}$ и $\eta^{(m)}$ однородны (являются выборками из одного и того же распределения), то есть $F_\xi(x) = F_\eta(x)$:

$$H_0: F_\xi(x) = F_\eta(x).$$

Альтернативная гипотеза напротив утверждает, что функции распределения $F_\xi(x)$ и $F_\eta(x)$ различны. Требуется составить критерий для проверки гипотезы H_0 против альтернативной гипотезы H_1.

Если функция $F(x) = F_\xi(x) = F_\eta(x)$ является непрерывной, то для проверки гипотезы H_0 может использоваться критерий однородности Колмогорова-Смирнова, статистика которого имеет вид:

$$D_{n,m}(\xi^{(n)}, \eta^{(m)}) = \sqrt{\frac{nm}{n+m}} \sup_{-\infty < x < \infty} \left| F^*_{\xi,n}(x; \xi^{(n)}) - F^*_{\eta,m}(x; \eta^{(m)}) \right|,$$

где $F^*_{\xi,n}(x; \xi^{(n)})$ – эмпирическая функция распределения выборки $\xi^{(n)}$ и $F^*_{\eta,m}(x; \eta^{(m)})$ – эмпирическая функция распределения выборки $\eta^{(m)}$. Если гипотеза H_0 не верна, то есть $F_\xi(x) \neq F_\eta(x)$, то функции эмпирического распределения $F^*_{\xi,n}(x; \xi^{(n)})$ и $F^*_{\eta,m}(x; \eta^{(m)})$ сходятся к различным функциям, поэтому точная верхняя грань модуля разности не стремится к нулю с увеличением n и m, а стремится к конечному числу отличному от нуля, которое затем умножается на возрастающую величину $\sqrt{\frac{nm}{n+m}}$. Отсюда следует, что в случае если гипотеза H_0 не верна статистика $D_{n,m}$ с большой вероятностью принимает «большие» значения, поэтому «большие» значения статистики $D_{n,m}$ свидетельствуют против гипотезы H_0 и в критическую область Γ_α гипотезы H_0 следует отнести «большие» значения статистики $D_{n,m}$:

$$\Gamma_\alpha = \{d : d = D_{n,m}(\xi^{(n)}, \eta^{(m)} \mid H_0) \geq h_\alpha\},$$

где пороговое значение h_α определяется из распределения статистики $D_{n,m}$, полученного при условии, что основная гипотеза верна, и заданного уровня значимости α. При больших n и m распределение статистики $D_{n,m}$ в том случае, когда основная гипотеза верна, может быть вычислено приближенно на основе теоремы Смирнова.

Теорема 5Б.1. (Смирнов)

Пусть $\xi^{(n)} = (\xi_1,..., \xi_n)$ и $\eta^{(m)} = (\eta_1,..., \eta_m)$ – независимые выборки из распределения $F(x)$, $F^*_{\xi,n}(x; \xi^{(n)})$ – функция эмпирического распределения выборки $\xi^{(n)}$, $F^*_{\eta,m}(x; \eta^{(m)})$ – функция эмпирического распределения выборки $\eta^{(m)}$. Если $F(x)$ – непрерывная функция, тогда для любого фиксированного $t > 0$:

$$\lim_{n,m \to \infty} P\left\{D_{n,m}(\xi^{(n)}, \eta^{(m)}) \leq t\right\} = K(t) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2t^2}.$$

Без доказательства.

Таким образом, при больших n и m для заданного уровня значимости α получим приближенное равенство:

$$\alpha = P\{D_{n,m}(\xi^{(n)}, \eta^{(m)}) \geq h_\alpha \mid F_\xi(x) = F_\eta(x)\} =$$

$$= 1 - P\{D_{n,m}(\xi^{(n)}, \eta^{(m)} \mid F_\xi(x) = F_\eta(x)) < h_\alpha\} \approx 1 - K(h_\alpha),$$

$$K(h_\alpha) \approx 1 - \alpha,$$

откуда численно определяется значение h_α.

2. Критерий Фишера.

Пусть совокупность наблюдений образована двумя независимыми выборками: $\xi^{(n)} = (\xi_1, ..., \xi_n)$ – выборка из нормального распределения $N(m_\xi, \sigma_\xi^2)$, $\eta^{(m)} = (\eta_1, ..., \eta_m)$ – выборка из нормального распределения $N(m_\eta, \sigma_\eta^2)$, $\xi^{(n)}$ и $\eta^{(m)}$ – независимы, параметры m_ξ, σ_ξ^2, m_η и σ_η^2 – неизвестны. Основная гипотеза H_0 заключается в том, что дисперсии σ_ξ^2 и σ_η^2 совпадают:

$$H_0: \sigma_\xi^2 = \sigma_\eta^2.$$

Альтернативная гипотеза H_1 заключается в том, что дисперсии σ_ξ^2 и σ_η^2 различны:

$$H_1: \sigma_\xi^2 = \sigma_\eta^2.$$

Требуется составить критерий проверки гипотезы H_0 против альтернативной гипотезы H_1.

Для проверки гипотезы H_0 используется критерий Фишера со статистикой:

$$f_{n,m}(\xi^{(n)}, \eta^{(m)}) = \frac{\tilde{\mu}_2(\xi^{(n)})}{\tilde{\mu}_2(\eta^{(m)})},$$

$$\tilde{\mu}_2(\xi^{(n)}) = \frac{1}{n-1}\sum_{i=1}^{n}(\xi_i - \hat{m}_1(\xi^{(n)}))^2,\ \hat{m}_1(\xi^{(n)}) = \frac{1}{n}\sum_{i=1}^{n}\xi_i, \qquad (5Б.1)$$

$$\tilde{\mu}_2(\eta^{(m)}) = \frac{1}{m-1}\sum_{j=1}^{m}(\eta_j - \hat{m}_1(\eta^{(m)}))^2,\ \hat{m}_1(\eta^{(m)}) = \frac{1}{m}\sum_{j=1}^{m}\eta_j$$

Можно показать, что статистика $f_{n,m}(\xi^{(n)}, \eta^{(m)})$ сходится по вероятности к отношению $\frac{\sigma_\xi^2}{\sigma_\eta^2}$ при одновременном возрастании n и m:

$$f_{n,m}(\xi^{(n)}, \eta^{(m)}) \xrightarrow{P} \frac{\sigma_\xi^2}{\sigma_\eta^2}, \text{ при } n, m \to \infty.$$

Если гипотеза H_0 не верна, то есть $\sigma_\xi \neq \sigma_\eta$, то значения статистики $f_{n,m}(\xi^{(n)}, \eta^{(m)})$ «концентрируются» в окрестности отношения $\frac{\sigma_\xi^2}{\sigma_\eta^2}$, которое больше 1, если $\sigma_\xi > \sigma_\eta$, либо меньше 1, если $\sigma_\xi < \sigma_\eta$. Другими словами, если гипотеза H_0 не верна, то значения статистики $f_{n,m}(\xi^{(n)}, \eta^{(m)})$ «концентрируются» за пределами окрестности 1.

Оказывается, что если основная гипотеза H_0 верна, то значения статистики $f_{n,m}(\xi^{(n)}, \eta^{(m)})$ напротив с большой вероятностью «сосредоточены» в окрестности 1 при достаточно больших n и m.

Утверждение 5Б.2.

Пусть $\xi^{(n)} = (\xi_1,..., \xi_n)$ – выборка из нормального распределения $N(m_\xi, \sigma^2)$, $\eta^{(m)} = (\eta_1,..., \eta_m)$ – выборка из нормального распределения $N(m_\eta, \sigma^2)$, $\xi^{(n)}$ и $\eta^{(m)}$ – независимы, тогда статистика $f_{n,m}(\xi^{(n)}, \eta^{(m)})$ (5Б.1) имеет распределение Фишера $F(n-1, m-1)$.

Доказательство:

Согласно теореме Фишера (теорема 4.5) случайная величина $\frac{(n-1)}{\sigma^2}\tilde{\mu}_2(\xi^{(n)})$ имеет распределение $\chi^2(n-1)$, а случайная величина $\frac{(m-1)}{\sigma^2}\tilde{\mu}_2(\eta^{(m)})$ имеет распределение $\chi^2(m-1)$, причем $\frac{(n-1)}{\sigma^2}\tilde{\mu}_2(\xi^{(n)})$ и $\frac{(m-1)}{\sigma^2}\tilde{\mu}_2(\eta^{(m)})$ независимы, поскольку выборки $\xi^{(n)}$ и $\eta^{(m)}$ независимы. Представим статистику $f_{n,m}(\xi^{(n)}, \eta^{(m)})$ в следующем виде:

$$f_{n,m}(\xi^{(n)}, \eta^{(m)}) = \frac{\tilde{\mu}_2(\xi^{(n)})}{\tilde{\mu}_2(\eta^{(m)})} = \frac{\frac{1}{\sigma^2}\tilde{\mu}_2(\xi^{(n)})}{\frac{1}{\sigma^2}\tilde{\mu}_2(\eta^{(m)})} = \frac{\frac{1}{(n-1)}\frac{(n-1)}{\sigma^2}\tilde{\mu}_2(\xi^{(n)})}{\frac{1}{(m-1)}\frac{(m-1)}{\sigma^2}\tilde{\mu}_2(\eta^{(m)})} = \frac{\frac{1}{(n-1)}\chi^2_{n-1}}{\frac{1}{(m-1)}\chi^2_{m-1}},$$

где χ^2_k обозначает случайную величину, имеющую распределение $\chi^2(k)$. Случайная величина $\frac{\frac{1}{(n-1)}\chi^2_{n-1}}{\frac{1}{(m-1)}\chi^2_{m-1}}$ с независимыми величинами χ^2_{n-1} и χ^2_{m-1} по определению имеет распределение Фишера $F(n-1, m-1)$.

Утверждение доказано.

Распределение Фишера $F(n, m)$ при одновременном возрастании n и m «концентрируется» в малой окрестности 1, поэтому если гипотеза H_0 верна, то значение статистики $f_{n,m}(\xi^{(n)}, \eta^{(m)})$ с большой вероятностью оказывается близким к 1.

Отсюда следует, что в качестве критической области Γ_α гипотезы H_0 следует выбирать те значения статистики $f_{n,m}(\xi^{(n)}, \eta^{(m)})$, которые оказываются за пределами окрестности 1:

$$\Gamma_\alpha = [0, h_{\alpha/2}) \cup (h_{1-\alpha/2}; \infty),$$

где в качестве пороговых значений $h_{\alpha/2}$ и $h_{1-\alpha/2}$ могут использоваться квантили распределения Фишера $F(n-1, m-1)$ уровней $\frac{\alpha}{2}$ и $1-\frac{\alpha}{2}$, где α – заданный уровень значимости.

3. Критерий Стьюдента.

Пусть совокупность наблюдений образована двумя независимыми выборками: $\xi^{(n)} = (\xi_1,..., \xi_n)$ – выборка из нормального распределения $N(m_\xi, \sigma^2)$, $\eta^{(m)} = (\eta_1,..., \eta_m)$ – выборка из нормального распределения $N(m_\eta, \sigma^2)$, $\xi^{(n)}$ и $\eta^{(m)}$ – независимы, параметры m_ξ, m_η и σ^2 – неизвестны. Основная гипотеза H_0 заключается в том, что математические ожидания m_ξ и m_η равны:

$$H_0: m_\xi = m_\eta.$$

Альтернативная гипотеза H_1 напротив заключается в том, что математические ожидания m_ξ и m_η различны:

$$H_1: m_\xi \neq m_\eta .$$

Требуется составить критерий для проверки гипотезы H_0 против альтернативной гипотезы H_1.

Для проверки гипотезы H_0 используется критерий Стьюдента со статисткой:

$$t_{n,m}(\xi^{(n)},\eta^{(m)}) = \sqrt{\frac{nm}{n+m}}\,\frac{\hat{m}_1(\xi^{(n)}) - \hat{m}_1(\eta^{(m)})}{\sqrt{\dfrac{(n-1)\tilde{\mu}_2(\xi^{(n)}) + (m-1)\tilde{\mu}_2(\eta^{(m)})}{n+m-2}}},$$

$$\tilde{\mu}_2(\xi^{(n)}) = \frac{1}{n-1}\sum_{i=1}^{n}(\xi_i - \hat{m}_1(\xi^{(n)}))^2,\ \hat{m}_1(\xi^{(n)}) = \frac{1}{n}\sum_{i=1}^{n}\xi_i , \qquad (5Б.2)$$

$$\tilde{\mu}_2(\eta^{(m)}) = \frac{1}{m-1}\sum_{j=1}^{m}(\eta_j - \hat{m}_1(\eta^{(m)}))^2,\ \hat{m}_1(\eta^{(m)}) = \frac{1}{m}\sum_{j=1}^{m}\eta_j .$$

Можно показать, что если гипотеза H_0 не верна, то есть $m_\xi \neq m_\eta$, то статистика $t_{n,m}(\xi^{(n)},\eta^{(m)})$ (5Б.2) неограниченна по вероятности. Действительно, выборочное среднее $\hat{m}_1(\xi^{(n)})$ сходится по вероятности к m_ξ, а выборочное среднее $\hat{m}_1(\eta^{(m)})$ – к m_η:

$$\hat{m}_1(\xi^{(n)}) \xrightarrow{P} m_\xi, \text{ при } n \to \infty ,$$

$$\hat{m}_1(\eta^{(m)}) \xrightarrow{P} m_\eta, \text{ при } m \to \infty ,$$

тогда в силу свойства сходимости по вероятности:

$$\hat{m}_1(\xi^{(n)}) - \hat{m}_1(\eta^{(m)}) \xrightarrow{P} m_\xi - m_\eta, \text{ при } n,m \to \infty .$$

Если гипотеза H_0 не верна и $m_\xi \neq m_\eta$, то $|m_\xi - m_\eta| \geq c > 0$.

Случайные величины $\tilde{\mu}_2(\xi^{(n)})$ и $\tilde{\mu}_2(\eta^{(m)})$ сходятся по вероятности к σ^2, поэтому знаменатель статистики $t_{n,m}(\xi^{(n)},\eta^{(m)})$:

$$\sqrt{\frac{(n-1)\tilde{\mu}_2(\xi^{(n)}) + (m-1)\tilde{\mu}_2(\eta^{(m)})}{n+m-2}} \xrightarrow{P} \sqrt{\frac{(n-1)\sigma^2 + (m-1)\sigma^2}{n+m-2}} \xrightarrow{P} \sigma, \text{ при } n,m \to \infty .$$

Таким образом, для всей статистики $t_{n,m}(\xi^{(n)},\eta^{(m)})$ имеет место сходимость:

$$t_{n,m}(\xi^{(n)},\eta^{(m)}) = \frac{\hat{m}_1(\xi^{(n)}) - \hat{m}_1(\eta^{(m)})}{\sqrt{\dfrac{(n-1)\tilde{\mu}_2(\xi^{(n)}) + (m-1)\tilde{\mu}_2(\eta^{(m)})}{n+m-2}}} \xrightarrow{P} \frac{m_\xi - m_\eta}{\sigma}, \text{ при } n,m \to \infty ,$$

$$\text{где } \frac{|m_\xi - m_\eta|}{\sigma} \geq \frac{c}{\sigma} > 0 .$$

Поскольку множитель $\sqrt{\dfrac{nm}{n+m}}$ возрастает с одновременным увеличением n и m, то статистика $t_{n,m}(\xi^{(n)},\eta^{(m)})$ с большой вероятностью принимает большие по модулю значения в том случае, когда гипотеза H_0 не верна. Отсюда, в критическую область Γ_α следует отнести «большие» значения модуля статистики $t_{n,m}(\xi^{(n)},\eta^{(m)})$:

$$\Gamma_\alpha = \{t : t = |t_{n,m}(\xi^{(n)},\eta^{(m)})| \geq h_\alpha\} = (-\infty, -h_\alpha) \cup (h_\alpha, \infty) ,$$

где пороговое значение h_α выбирается в соответствии с заданным уровнем значимости α и тем распределением статистики $t_{n,m}(\xi^{(n)},\eta^{(m)})$, которое она имеет при условии, что основная гипотеза H_0 верна.

Утверждение 5Б.3.

Пусть $\xi^{(n)}=(\xi_1,...,\xi_n)$ – выборка из нормального распределения $N(m,\sigma^2)$, $\eta^{(m)}=(\eta_1,...,\eta_m)$ – выборка из нормального распределения $N(m,\sigma^2)$, $\xi^{(n)}$ и $\eta^{(m)}$ – независимы, тогда статистика $t_{n,m}(\xi^{(n)},\eta^{(m)})$ (5Б.2) имеет распределение Стьюдента $T(n+m-2)$.

Доказательство:

Рассмотрим разность выборочных средних:

$$\hat{m}_1(\xi^{(n)})-\hat{m}_1(\eta^{(m)})=\hat{m}_1(\xi^{(n)})-m+m-\hat{m}_1(\eta^{(m)})=\hat{m}_1(\xi^{(n)})-m+m-\hat{m}_1(\eta^{(m)})=$$

$$=(\hat{m}_1(\xi^{(n)})-m)-(\hat{m}_1(\eta^{(m)})-m).$$

Отсюда, статистика $t_{n,m}(\xi^{(n)},\eta^{(m)})$ (5Б.2):

$$t_{n,m}(\xi^{(n)},\eta^{(m)})=\sqrt{\frac{nm}{n+m}}\frac{\hat{m}_1(\xi^{(n)})-\hat{m}_1(\eta^{(m)})}{\sqrt{\frac{(n-1)\tilde{\mu}_2(\xi^{(n)})+(m-1)\tilde{\mu}_2(\eta^{(m)})}{n+m-2}}}=$$

$$=\frac{\frac{(\hat{m}_1(\xi^{(n)})-m)-(\hat{m}_1(\eta^{(m)})-m)}{\sqrt{\frac{n+m}{nm}}\sigma}}{\sqrt{\frac{\frac{(n-1)\tilde{\mu}_2(\xi^{(n)})}{\sigma^2}+\frac{(m-1)\tilde{\mu}_2(\eta^{(m)})}{\sigma^2}}{(n+m-2)}}}. \qquad (5Б.3)$$

Поскольку $\xi^{(n)}$ – выборка из $N(m,\sigma^2)$, то случайная величина $\hat{m}_1(\xi^{(n)})$ имеет нормальное распределение (как сумма независимых в совокупности нормальных случайных величин):

$$\hat{m}_1(\xi^{(n)})=\frac{1}{n}\sum_{i=1}^{n}\xi_i \sim N\left(m,\frac{\sigma^2}{n}\right),$$

отсюда следует,

$$\hat{m}_1(\xi^{(n)})-m \sim N\left(0,\frac{\sigma^2}{n}\right).$$

Аналогично,

$$\hat{m}_1(\eta^{(m)})-m \sim N\left(0,\frac{\sigma^2}{m}\right).$$

Легко видеть, что

$$(\hat{m}_1(\xi^{(n)})-m)-(\hat{m}_1(\eta^{(m)})-m) \sim N\left(0,\frac{n+m}{nm}\sigma^2\right),$$

поскольку $\hat{m}_1(\xi^{(n)})$ и $\hat{m}_1(\eta^{(m)})$ независимые и нормальные случайные величины,

$$M[(\hat{m}_1(\xi^{(n)})-m)-(\hat{m}_1(\eta^{(m)})-m)]=M[\hat{m}_1(\xi^{(n)})-m]-M[\hat{m}_1(\eta^{(m)})-m]=0,$$

$$D[(\hat{m}_1(\xi^{(n)})-m)-(\hat{m}_1(\eta^{(m)})-m)]=D[\hat{m}_1(\xi^{(n)})-\hat{m}_1(\eta^{(m)})]=$$

$$=D[\hat{m}_1(\xi^{(n)})]-2\,\mathrm{cov}(\hat{m}_1(\xi^{(n)}),\hat{m}_1(\eta^{(m)}))+D[\hat{m}_1(\eta^{(m)})]=$$

$$= D[\hat{m}_1(\xi^{(n)})] + D[\hat{m}_1(\eta^{(m)})] = \frac{\sigma^2}{n} + \frac{\sigma^2}{m} = \frac{n+m}{nm}\sigma^2,$$

поскольку,

$$\operatorname{cov}(\hat{m}_1(\xi^{(n)}), \hat{m}_1(\eta^{(m)})) = M[(\hat{m}_1(\xi^{(n)}) - m)\cdot(\hat{m}_1(\eta^{(m)}) - m)] =$$

$$= M\left[\left(\frac{1}{n}\sum_{i=1}^{n}\xi_i - m\right)\left(\frac{1}{m}\sum_{j=1}^{m}\eta_j - m\right)\right] = M\left[\left(\frac{1}{n}\sum_{i=1}^{n}(\xi_i - m)\right)\left(\frac{1}{m}\sum_{j=1}^{m}(\eta_j - m)\right)\right] =$$

$$= \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} M[(\xi_i - m)(\eta_j - m)] = 0,$$

где $M[(\xi_i - m)(\eta_j - m)] = 0$ в силу независимости выборок $(\xi_1,\ldots,\xi_n)$ и $(\eta_1,\ldots,\eta_m)$.

Согласно теореме Фишера (теорема 4.5) случайная величина $\frac{(n-1)\tilde{\mu}_2(\xi^{(n)})}{\sigma^2}$ имеет распределение $\chi^2(n-1)$, а случайная величина $\frac{(m-1)\tilde{\mu}_2(\eta^{(m)})}{\sigma^2}$ – распределение $\chi^2(m-1)$, причем случайные величины $\frac{(n-1)\tilde{\mu}_2(\xi^{(n)})}{\sigma^2}$ и $\frac{(m-1)\tilde{\mu}_2(\eta^{(m)})}{\sigma^2}$ независимы, поэтому в силу свойства распределения хи-квадрат:

$$\frac{(n-1)\tilde{\mu}_2(\xi^{(n)})}{\sigma^2} + \frac{(m-1)\tilde{\mu}_2(\eta^{(m)})}{\sigma^2} \sim \chi^2(n-1+m-1).$$

Случайные величины $(\hat{m}_1(\xi^{(n)}) - m) - (\hat{m}_1(\eta^{(m)}) - m)$ и $\frac{(n-1)\tilde{\mu}_2(\xi^{(n)})}{\sigma^2} + \frac{(m-1)\tilde{\mu}_2(\eta^{(m)})}{\sigma^2}$ независимы: $\hat{m}_1(\xi^{(n)})$ и $\tilde{\mu}_2(\xi^{(n)})$ независимы по теореме Фишера (теорема 4.5), $\hat{m}_1(\xi^{(n)})$ и $\tilde{\mu}_2(\eta^{(m)})$ независимы, поскольку выборки $\xi^{(n)}$ и $\eta^{(m)}$ независимы, $\hat{m}_1(\eta^{(m)})$ и $\tilde{\mu}_2(\xi^{(n)})$ независимы, поскольку $\xi^{(n)}$ и $\eta^{(m)}$ независимы, и $\hat{m}_1(\eta^{(m)})$ и $\tilde{\mu}_2(\eta^{(m)})$ независимы по теореме Фишера (теорема 4.5).

Таким образом, в (5Б.3) случайная величина в числителе имеет распределение $N(0,1)$, случайная величина в знаменателе имеет распределение $\chi^2(n+m-2)$ и случайные величины числителя и знаменателя независимы, отсюда следует, что статистика $t_{n,m}(\xi^{(n)},\eta^{(m)})$ имеет распределение Стьюдента $T(n+m-2)$.

Утверждение доказано.

Из утверждения 5Б.3 следует, что в качестве порогового значения h_α критической области Γ_α следует использовать квантиль распределения Стьюдента $T(n+m-2)$ уровня $1-\alpha$, где α – заданный уровень значимости.

4. Однофакторный дисперсионный анализ.

Неформально задача дисперсионного анализа заключается в том, чтобы установить меру влияния факторов на наблюдаемый результат. Например, влияет ли качество сырья (первый фактор) и технология производства (второй фактор) на качество произведенного продукта (наблюдаемый результат), или, например, влияет ли уровень подготовки преподавателей (один фактор) на квалификацию выпускаемых специалистов (наблюдаемый результат).

В качестве исходных данных для анализа предоставляются несколько серий наблюдений, произведенных при различных значениях факторов. В процессе анализа фактически сравниваются не отдельные наблюдения, а средние значения наблюдений при различных значениях факторов. Если разброс средних значений оказывается большим, то

следовательно фактор имеет существенное влияние, если же разброс средних оказывается небольшим, то скорее всего фактор не имеет существенного влияния.

В общем случае, факторов может быть несколько, но если фактор один, то анализ называют однофакторным.

Пусть имеется один единственный фактор, который может иметь k различных уровней (значений), которые пронумерованы числами от 1 до k. При каждом значении фактора j производится n_j независимых наблюдений (измерений) результата, и результаты записываются – $(\xi_1^{(j)},..., \xi_{n_j}^{(j)})$.

Формально, будем считать, что заданы k выборок $\xi = (\xi^{(1)},..., \xi^{(k)})$:

$$\xi^{(1)} = (\xi_1^{(1)},..., \xi_{n_1}^{(1)}),$$

$$\dots,$$

$$\xi^{(k)} = (\xi_1^{(k)},..., \xi_{n_k}^{(k)}),$$

причем,

1) $\xi^{(j)}$ – выборка из нормального распределения $N(a_j, \sigma^2)$, $j = \overline{1,k}$;

2) $\xi^{(1)}$, …, $\xi^{(k)}$ совместно независимы;

3) числа a_j и σ неизвестны.

Требуется составить критерий для проверки основной гипотезы H_0, которая заключается в том, что:

$$H_0: \ a_1 = a_2 = ... = a_k.$$

против альтернативной гипотезы H_1, утверждающей, что величины a_i являются различными.

Фактически, гипотеза H_0 утверждает, что фактор не имеет никакого влияния на наблюдаемый результат (все наблюдения в среднем одинаковы).

Определим для каждой выборки $\xi^{(j)}$ выборочное среднее $\hat{m}_1(\xi^{(j)})$ и выборочную дисперсию $\hat{\mu}_2(\xi^{(j)})$:

$$\hat{m}_1(\xi^{(j)}) = \frac{1}{n_j}\sum_{i=1}^{n_j} \xi_i^{(j)},$$

$$\hat{\mu}_2(\xi^{(j)}) = \frac{1}{n_j}\sum_{i=1}^{n_j} (\xi_i^{(j)} - \hat{m}_1(\xi^{(j)}))^2.$$

Определим также общее выборочное среднее $\hat{m}_1(\xi)$ и общую выборочную дисперсию $\hat{\mu}_2(\xi)$ ($n = \sum_{j=1}^{k} n_j$):

$$\hat{m}_1(\xi) = \sum_{j=1}^{k} \frac{n_j}{n}\hat{m}_1(\xi^{(j)}) = \sum_{j=1}^{k} \frac{n_j}{n}\frac{1}{n_j}\sum_{i=1}^{n_j} \xi_i^{(j)} = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j} \xi_i^{(j)},$$

$$\hat{\mu}_2(\xi) = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j} (\xi_i^{(j)} - \hat{m}_1(\xi))^2.$$

Преобразуем общую выборочную дисперсию:

$$\hat{\mu}_2(\xi) = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j} (\xi_i^{(j)} - \hat{m}_1(\xi))^2 = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j} (\xi_i^{(j)} - \hat{m}_1(\xi^{(j)}) + \hat{m}_1(\xi^{(j)}) - \hat{m}_1(\xi))^2 =$$

$$= \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j} \left\{(\xi_i^{(j)} - \hat{m}_1(\xi^{(j)}))^2 + 2(\xi_i^{(j)} - \hat{m}_1(\xi^{(j)}))(\hat{m}_1(\xi^{(j)}) - \hat{m}_1(\xi)) + (\hat{m}_1(\xi^{(j)}) - \hat{m}_1(\xi))^2\right\} =$$

$$= \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(\xi_i^{(j)} - \hat{m}_1(\xi^{(j)}))^2 + 2\frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(\xi_i^{(j)} - \hat{m}_1(\xi^{(j)}))(\hat{m}_1(\xi^{(j)}) - \hat{m}_1(\xi)) + \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(\hat{m}_1(\xi^{(j)}) - \hat{m}_1(\xi))^2 =$$

$$= \frac{1}{n}\sum_{j=1}^{k} n_j \hat{\mu}(\xi^{(j)}) + 2\frac{1}{n}\sum_{j=1}^{k}(\hat{m}_1(\xi^{(j)}) - \hat{m}_1(\xi))\sum_{i=1}^{n_j}(\xi_i^{(j)} - \hat{m}_1(\xi^{(j)})) + \frac{1}{n}\sum_{j=1}^{k} n_j(\hat{m}_1(\xi^{(j)}) - \hat{m}_1(\xi))^2 =$$

$$= \sum_{j=1}^{k}\frac{n_j}{n}\hat{\mu}(\xi^{(j)}) + 2\frac{1}{n}\sum_{j=1}^{k}(\hat{m}_1(\xi^{(j)}) - \hat{m}_1(\xi))\left(\sum_{i=1}^{n_j}\xi_i^{(j)} - \sum_{i=1}^{n_j}\hat{m}_1(\xi^{(j)})\right) + \sum_{j=1}^{k}\frac{n_j}{n}(\hat{m}_1(\xi^{(j)}) - \hat{m}_1(\xi))^2 =$$

$$= \hat{s}^2(\xi) + 2\frac{1}{\sigma^2}\sum_{j=1}^{k}(\hat{m}_1(\xi^{(j)}) - \hat{m}_1(\xi))(n_j\hat{m}_1(\xi^{(j)}) - n_j\hat{m}_1(\xi^{(j)})) + \tilde{s}^2(\xi) =$$

$$= \hat{s}^2(\xi) + \tilde{s}^2(\xi).$$

Таким образом, имеет место, равенство:

$$\hat{\mu}_2(\xi) = \hat{s}^2(\xi) + \tilde{s}^2(\xi),$$

из которого следует, так называемое, *основное дисперсионное соотношение*:

$$\frac{n\hat{\mu}_2(\xi)}{\sigma^2} = \frac{n\hat{s}^2(\xi)}{\sigma^2} + \frac{n\tilde{s}^2(\xi)}{\sigma^2},$$

$$\hat{s}^2(\xi) = \sum_{j=1}^{k}\frac{n_j}{n}\hat{\mu}_2(\xi^{(j)}),$$

$$\tilde{s}^2(\xi) = \sum_{j=1}^{k}\frac{n_j}{n}(\hat{m}_1(\xi^{(j)}) - \hat{m}_1(\xi))^2.$$

в котором величина $\hat{s}^2(\xi)$ называется *внутригрупповой дисперсией*, а величина $\tilde{s}^2(\xi)$ – *межгрупповой дисперсией*.

Внутригрупповая дисперсия $\hat{s}^2(\xi)$ отражает «взвешенную, среднюю» выборочную дисперсию выборок: каждая выборочная дисперсия $\hat{\mu}_2(\xi^{(j)})$ входит в сумму с весом $\frac{n_j}{n}$, учитывающим объем выборки по отношению к суммарному объему всех выборок, и сумма весов $\sum_{j=1}^{k}\frac{n_j}{n} = 1$.

Межгрупповая дисперсия $\tilde{s}^2(\xi)$ отражает различие выборочных средних $\hat{m}_1(\xi^{(j)})$ между выборками $\xi^{(j)}$ (группами наблюдений $(\xi_1^{(j)},..., \xi_{n_j}^{(j)})$), причем каждое отклонение $(\hat{m}_1(\xi^{(j)}) - \hat{m}_1(\xi))^2$ входит в сумму с весом $\frac{n_j}{n}$, учитывающим объем выборки по отношению к суммарному объему всех выборок, и сумма всех весов $\sum_{j=1}^{k}\frac{n_j}{n} = 1$.

Дисперсионный анализ основан на сравнении этих двух дисперсий (отсюда и происходит название анализа – дисперсионный). Значения внутригрупповой дисперсии $\hat{s}^2(\xi)$ всегда являются величинами порядка дисперсии σ^2, независимо от того является верной основная гипотеза H_0 (величины a_j одинаковы) или альтернативная гипотеза H_1 (величины a_j существенно различаются). Значения межгрупповой дисперсии $\tilde{s}^2(\xi)$ напротив зависят от того, является или верной основная гипотеза или альтернативная. Если верна основная гипотеза, то значения межгрупповой дисперсии $\tilde{s}^2(\xi)$ сопоставимы с дисперсией σ^2 и значениями внутригрупповой дисперсии $\hat{s}^2(\xi)$. Если же верна

альтернативная гипотеза H_1, то значения межгрупповой дисперсии $\tilde{s}^2(\xi)$ становятся существенно больше величины дисперсии σ^2 и значений внутригрупповой дисперсии $\hat{s}^2(\xi)$. Таким образом, если величина межгрупповой дисперсии $\tilde{s}^2(\xi)$ существенно превосходит значение внутригрупповой дисперсий $\hat{s}^2(\xi)$, то вероятно основная гипотеза H_0 не верна, если же наоборот величины межгрупповой $\tilde{s}^2(\xi)$ внутригрупповой $\hat{s}^2(\xi)$ дисперсий различаются несущественно, то никаких оснований для отклонения основной гипотезы H_0 нет.

Преобразуем величину $\frac{n\hat{s}^2(\xi)}{\sigma^2}$:

$$\frac{n\hat{s}^2(\xi)}{\sigma^2} = \frac{n}{\sigma^2}\sum_{j=1}^{k}\frac{n_j}{n}\hat{\mu}_2(\xi^{(j)}) = \sum_{j=1}^{k}\frac{n_j\hat{\mu}_2(\xi^{(j)})}{\sigma^2},$$

по теореме Фишера (теорема 4.5), каждая статистика $\frac{n_j\hat{\mu}_2(\xi^{(j)})}{\sigma^2}$ имеет распределение $\chi^2(n_j-1)$.

Замечание 5Б.4.

В действительности, теорема Фишера была сформулирована для статистики $\frac{(n_j-1)\tilde{\mu}_2(\xi^{(j)})}{\sigma^2}$ с исправленной выборочной дисперсией $\tilde{\mu}_2(\xi^{(j)}) = \frac{1}{n_j-1}\sum_{i=1}^{n_j}(\xi_i^{(j)} - \hat{m}_1(\xi^{(j)}))^2$, тем не менее легко видеть, что $\frac{(n_j-1)\tilde{\mu}_2(\xi^{(j)})}{\sigma^2}$ и $\frac{n_j\hat{\mu}_2(\xi^{(j)})}{\sigma^2}$ это одна и та же статистика, действительно:

$$\frac{(n_j-1)\tilde{\mu}_2(\xi^{(j)})}{\sigma^2} = \frac{n_j-1}{\sigma^2}\frac{1}{n_j-1}\sum_{i=1}^{n_j}(\xi_i^{(j)} - \hat{m}_1(\xi^{(j)}))^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n_j}(\xi_i^{(j)} - \hat{m}_1(\xi^{(j)}))^2 =$$

$$= \frac{n_j}{\sigma^2}\frac{1}{n_j}\sum_{i=1}^{n_j}(\xi_i^{(j)} - \hat{m}_1(\xi^{(j)}))^2 = \frac{n_j\hat{\mu}_2(\xi^{(j)})}{\sigma^2}.$$

Поскольку все величины $\frac{n_j\hat{\mu}_2(\xi^{(j)})}{\sigma^2}$ независимы (все выборки $\xi^{(j)}$ независимы), то в силу свойства воспроизводимости при сложении распределения хи-квадрат случайная величина $\frac{n\hat{s}^2(\xi)}{\sigma^2}$ имеет распределение $\chi^2(n-k)$:

$$\frac{n\hat{s}^2(\xi)}{\sigma^2} = \sum_{j=1}^{k}\chi^2_{n_j-1} \sim \chi^2\left(\sum_{j=1}^{k}(n_j-1)\right) \sim \chi^2\left(\sum_{j=1}^{k}n_j - k\right) \sim \chi^2(n-k). \qquad (5Б.4)$$

Легко видеть, что межгрупповая дисперсия $\tilde{s}^2(\xi)$ и внутригрупповая дисперсия $\hat{s}^2(\xi)$ – независимые случайные величины. Действительно, случайная величина $\tilde{s}^2(\xi)$:

$$\tilde{s}^2(\xi) = \sum_{j=1}^{k}\frac{n_j}{n}(\hat{m}_1(\xi^{(j)}) - \hat{m}_1(\xi))^2 = \sum_{j=1}^{k}\frac{n_j}{n}\left(\hat{m}_1(\xi^{(j)}) - \sum_{j=1}^{k}\frac{n_j}{n}\hat{m}_1(\xi^{(j)})\right)^2$$

является функцией только выборочных средних $\hat{m}_1(\xi^{(j)})$ ($j=\overline{1,k}$). Случайная величина $\hat{s}^2(\xi)$:

$$\hat{s}^2(\xi) = \sum_{j=1}^{k}\frac{n_j}{n}\hat{\mu}_2(\xi^{(j)})$$

является функцией только выборочных дисперсий $\hat{\mu}_2(\xi^{(j)})$ ($j=\overline{1,k}$). Заметим, что при всех i и j случайные величины $\hat{m}_1(\xi^{(i)})$ и $\hat{\mu}_2(\xi^{(j)})$ независимы: если $i \neq j$, то случайные величины $\hat{m}_1(\xi^{(i)})$ и $\hat{\mu}_2(\xi^{(j)})$ независимы, поскольку выборки $\xi^{(i)}$ и $\xi^{(j)}$ независимы, если $i=j$, то $\hat{m}_1(\xi^{(j)})$ и $\hat{\mu}_2(\xi^{(j)})$ независимы, поскольку по теореме Фишера (теорема 4.5) случайные величины $\hat{m}_1(\xi^{(j)})$ и $\tilde{\mu}_2(\xi^{(j)})$ независимы и $\hat{\mu}_2(\xi^{(j)}) = \frac{n_j - 1}{n_j}\tilde{\mu}_2(\xi^{(j)})$. Таким образом, все слагаемые в $\tilde{s}^2(\xi)$ и все слагаемые в $\hat{s}^2(\xi)$ попарно независимы, поэтому $\tilde{s}^2(\xi)$ и $\hat{s}^2(\xi)$ независимы. Отсюда следует, что величины $\frac{n\hat{s}^2(\xi)}{\sigma^2}$ и $\frac{n\tilde{s}^2(\xi)}{\sigma^2}$ независимы.

Таким образом, всегда выполняется:

1) $\frac{n\hat{\mu}_2(\xi)}{\sigma^2} = \frac{n\hat{s}^2(\xi)}{\sigma^2} + \frac{n\tilde{s}^2(\xi)}{\sigma^2}$;

2) $\frac{n\hat{s}^2(\xi)}{\sigma^2} \sim \chi^2(n-k)$;

3) $\frac{n\hat{s}^2(\xi)}{\sigma^2}$ и $\frac{n\tilde{s}^2(\xi)}{\sigma^2}$ независимы.

Если гипотеза H_0 верна, то есть $a_1 = a_2 = \ldots = a_k$, тогда

$$\frac{n\hat{\mu}_2(\xi)}{\sigma^2} = \frac{1}{\sigma^2}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(\xi_i^{(j)} - \hat{m}_1(\xi))^2 = \frac{1}{\sigma^2}\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(\xi_i^{(j)} - \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j}\xi_i^{(j)}\right)^2,$$

где все случайные величины $\xi_i^{(j)}$ независимы и имеют одинаковое распределение $N(a_1, \sigma)$, то есть $\xi = (\xi^{(1)},\ldots,\xi^{(k)}) = (\xi_1^{(1)},\ldots,\xi_{n_1}^{(1)},\ldots,\xi_1^{(k)},\ldots,\xi_{n_k}^{(k)})$ – выборка объема n из нормального распределения $N(a_1, \sigma^2)$, тогда в силу теоремы Фишера (теоремы 4.5) с учетом замечания 5Б.4 статистика $\frac{n\hat{\mu}_2(\xi)}{\sigma^2}$ имеет распределение $\chi^2(n-1)$:

$$\frac{n\hat{\mu}_2(\xi)}{\sigma^2} \sim \chi^2(n-1).$$

Можно показать, что из условий:

1) $\frac{n\hat{\mu}_2(\xi)}{\sigma^2} = \frac{n\hat{s}^2(\xi)}{\sigma^2} + \frac{n\tilde{s}^2(\xi)}{\sigma^2}$,

2) $\frac{n\hat{\mu}_2(\xi)}{\sigma^2} \sim \chi^2(n-1)$,

3) $\frac{n\hat{s}^2}{\sigma^2} \sim \chi^2(n-k)$,

4) $\frac{n\hat{s}^2(\xi)}{\sigma^2}$ и $\frac{n\tilde{s}^2(\xi)}{\sigma^2}$ независимы,

в силу свойства распределения хи-квадрат, следует:

$$\frac{n\tilde{s}^2(\xi)}{\sigma^2} \sim \chi^2((n-1)-(n-k)).$$

Таким образом, всегда справедливо $\frac{n\hat{s}^2(\xi)}{\sigma^2} \sim \chi^2(n-k)$, $\frac{n\hat{s}^2(\xi)}{\sigma^2}$ и $\frac{n\tilde{s}^2(\xi)}{\sigma^2}$ независимы, а в случае если гипотеза H_0 верна, то $\frac{n\tilde{s}^2(\xi)}{\sigma^2} \sim \chi^2(k-1)$, поэтому если гипотеза H_0 верна, тогда статистика $T(\xi)$:

$$T(\xi) = \frac{\tilde{s}^2(\xi)}{k-1}\frac{n-k}{\hat{s}^2(\xi)}$$

имеет распределение Фишера $F(k-1, n-k)$:

$$T(\xi) = \frac{\tilde{s}^2(\xi)}{k-1}\frac{n-k}{\hat{s}^2(\xi)} = \frac{\frac{\tilde{s}^2(\xi)}{k-1}}{\frac{\hat{s}^2(\xi)}{n-k}} = \frac{\frac{n\tilde{s}^2(\xi)}{\sigma^2}\frac{1}{k-1}}{\frac{n\hat{s}^2(\xi)}{\sigma^2}\frac{1}{n-k}} = \frac{\frac{\chi^2_{k-1}}{k-1}}{\frac{\chi^2_{n-k}}{n-k}} \sim F(k-1, n-k).$$

Если же основная гипотеза H_0 не верна, и величины a_j существенно различаются, тогда значения межгрупповой дисперсии $\tilde{s}^2$ становятся существенно больше величины дисперсии σ^2 и значений внутригрупповой дисперсии $\hat{s}^2$. В этом случае значения статистики критерия $T(\xi)$ с «большой» вероятностью оказываются намного больше единицы. Отсюда следует, что «большие» значения статистики $T(\xi)$ свидетельствуют против основной гипотезы H_0 и в качестве критической области Γ_α следует выбирать области вида:

$$\Gamma_\alpha = \left\{ s : s = \frac{\tilde{s}^2(\xi)}{k-1}\frac{n-k}{\hat{s}^2(\xi)} \geq h_\alpha \right\} = (h_\alpha, \infty),$$

где h_α – квантиль распределения Фишера $F(k-1, n-k)$ уровня $1-\alpha$, и α – заданный уровень значимости.