ГЛАВА 3. ТЕОРИЯ ПРОВЕРОК СТАТИСТИЧЕСКИХ ГИПОТЕЗ

§3.1. Статистическая гипотеза

Статистическая гипотеза – гипотеза о законе распределения вероятностей.

Примеры:

1) случайная величина распределена по закону Пуассона;
2) дисперсии двух нормальных совокупностей равны между собой.

Выдвинутую статистическую гипотезу называют нулевой гипотезой и обозначают буквой H_0. Гипотезу, которая противоречит H_0 называют конкурирующей гипотезой и обозначают буквой H_1.

Пример: Пусть случайная величина ξ распределена по нормальному закону. Тогда в качестве нулевой гипотезы может выступать гипотеза вида $H_0 = \{M(\xi) = 10\}$. В этом случае конкурирующая гипотеза выглядит следующим образом $H_1 = \{M(\xi) \neq 10\}$.

Гипотезы бывают простыми и сложными. Простая гипотеза- гипотеза, которая однозначно определяет распределение вероятностей.

Пример: Пусть λ - параметр распределения Пуассона $P(m) = \frac{\lambda^m e^{-\lambda}}{m!}$. Тогда гипотеза $H_0 = \{\lambda = 1\}$ является простой гипотезой.

Сложная гипотеза- гипотеза, которая содержит конечное или бесконечное число простых гипотез.

Пример: пусть λ - параметр распределения Пуассона $P(m) = \frac{\lambda^m e^{-\lambda}}{m!}$.

Гипотеза $H_0 = \{\lambda > 1\}$ является сложной гипотезой, так как содержит бесконечное множество простых гипотез $H_{0i} = \{\lambda = b_i\}$, где b_i -любое число, большее 1.

§3.2. Статистический критерий

Нулевая гипотеза может быть правильной либо неправильной. Поэтому возникает необходимость ее проверки. Для этого используют статистический критерий.

Статистический критерий – правило, в соответствии с которым принимается либо отклоняется данная статистическая гипотеза. Это правило состоит в следующем:

1) Сначала выбирают случайную величину K, закон распределения которой известен. Такая случайная величина называется статистикой критерия;
2) Затем по выборке вычисляют наблюдаемое значение статистики $K_{набл}$;

3) Наконец находят критическую область D (на практике это делается с помощью соответствующих таблиц);
4) Если $K_{набл} \in D$, то гипотеза H_0 отклоняется, а если $K_{набл} \notin D$, то гипотеза принимается.

Обычно для статистики критерия используют специальные обозначения: 1) Z, V - если статистика K распределена по нормальному закону;

2) F, V^2 - если статистика K распределена по закону Фишера-Снедекора; 3) T - по закону Стьюдента; 4) χ^2 - по закону «хи-квадрат».

Критическая область, область принятия, критические точки

После выбора статистики критерия, множество всех ее возможных значений разбивают на два непересекающихся подмножества:

1) Одно из них содержит значения статистики критерия, при которых H_0 отклоняется;
2) Другое – значения, при которых H_0 принимается.

Первое подмножество называется критической областью. Второе подмножество называется областью принятия гипотезы. Обычно критическая область и область принятия гипотезы – это интервалы. Следовательно, существуют точки, которые их разделяют. Точки $K_{кр}$, отделяющие критическую область от области принятия, называются критическими точками. Различают одностороннюю и двустороннюю критическую область:

1) Интервал на числовой оси, определяемый неравенством $K > K_{кр}$, называется правосторонней критической областью;
2) Интервал на числовой оси, определяемый неравенством $K < K_{кр}$, называется левосторонней критической областью;
3) Интервалы, определяемые неравенствами $K < K_1, K > K_2 \quad (K_2 > K_1)$, называются двусторонней критической областью.

Отыскание правосторонней критической области

Итак, правосторонняя критическая область– интервал $K > K_{пр.кр}$. Требуется найти правую критическую точку $K_{пр.кр}$. Для этого задаются малой вероятностью – уровнем значимости α. Затем при условии справедливости H_0 находят $K_{пр.кр}$ из интегрального уравнения

$$P\left(K > K_{пр.кр}\right) = \int_{K_{пр.кр}}^{\infty} f_K(x)dx = \alpha \qquad (1)$$

Для каждой статистики критерия K имеется соответствующая таблица, по которой и находят правую критическую точку $K_{пр.кр}$, удовлетворяющую интегральному уравнению (1)

Отыскание левосторонней критической области

Левосторонняя критическая область – это интервал $K < K_{лев.кр}$. При условии справедливости H_0 левую критическую точку $K_{лев.кр}$ находят из интегрального уравнения

$$P\left(K < K_{лев.кр}\right) = \int_{-\infty}^{K_{лев.кр}} f_K(x)dx = \alpha \qquad (1)$$

Существующие таблицы составлены только для правых критических точек. Однако их можно использовать для нахождения и левых критических точек. Делается это следующим образом.

Так как события $\left(K < K_{лев.кр}\right), \left(K > K_{лев.кр}\right)$ противоположны, то

$$P\left(K < K_{лев.кр}\right) + P\left(K > K_{лев.кр}\right) = 1 \qquad (2)$$

Откуда, учитывая уравнение (1), получим, что

$$P\left(K > K_{лев.кр}\right) = 1 - \alpha \qquad (3)$$

Из равенства (3) следует, что $K_{лев.кр}$ можно найти как правую критическую точку, но по уровню значимости $(1-\alpha)$.

Отыскание двусторонней критической области

Двусторонняя критическая область – это интервалы $K < K_1, K > K_2$. При условии справедливости гипотезы H_0 критические точки K_1, K_2 находят из равенства

$$P\left(K < K_1\right) + P\left(K > K_2\right) = \alpha \qquad (1)$$

В общем случае вместо равенства (1) используют систему уравнений

$$\begin{cases} P\left(K < K_1\right) = \dfrac{\alpha}{2} \\ P\left(K > K_2\right) = \dfrac{\alpha}{2} \end{cases} \qquad (2)$$

Правую критическую точку K_2 находят по таблице по уровню значимости $\frac{\alpha}{2}$. Так же, как в предыдущем параграфе, левую критическую точку K_1 находят по таблице правых критических точек, но по уровню значимости $\left(1-\frac{\alpha}{2}\right)$.

Ошибки первого и второго рода

При статистической проверке статистических гипотез могут возникнуть ошибки двух родов:

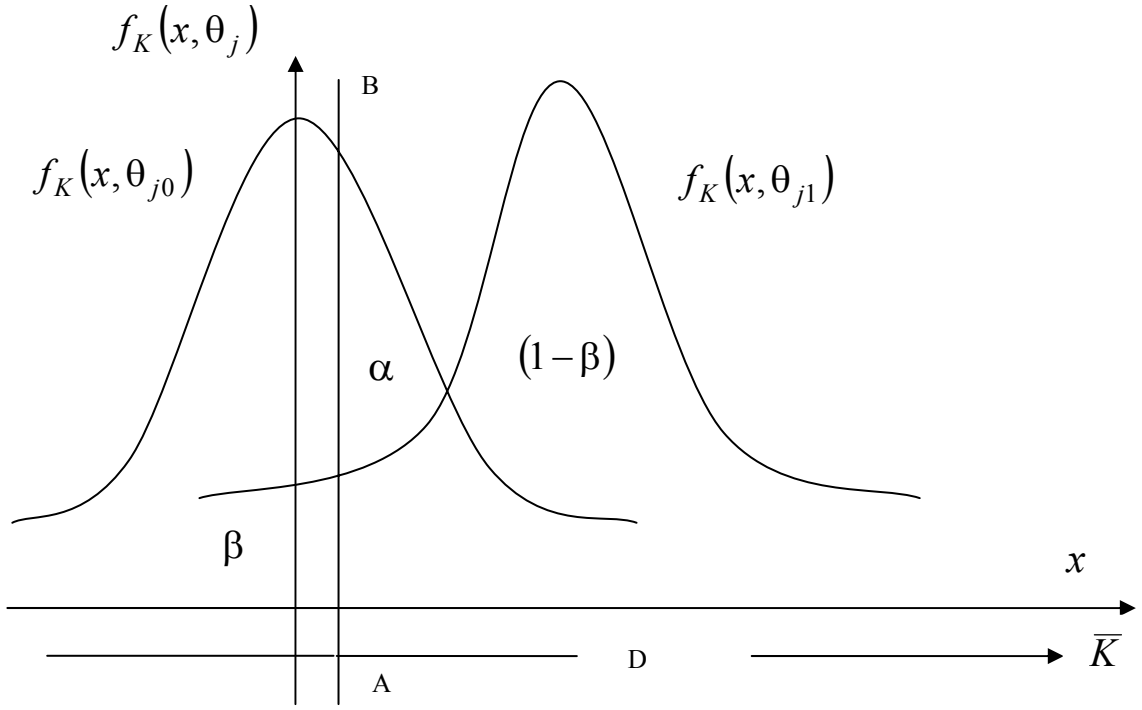
1) Ошибка первого рода состоит в том, что будет отклонена правильная гипотеза H_0;
2) Ошибка второго рода состоит в том, что будет принята ложная гипотеза H_0.

Вероятность совершить ошибку первого рода называют уровнем значимости и обозначают буквой α. Обычно выбирают $\alpha = 0.05, \alpha = 0.01$. Например, если $\alpha = 0.05$, то это означает, что в 5 случаях из 100 мы рискуем совершить ошибку первого рода. Вероятность совершить ошибку второго рода обозначают буквой β. Величину $(1-\beta)$ называют мощностью критерия. Другими словами, мощность критерия- это вероятность отклонить ложную гипотезу H_0. Значения α, β выбирают в зависимости от тяжести последствий ошибок 1 либо 2 рода. Например, если ошибки первого рода могут повлечь за собой более тяжелые последствия, чем ошибки второго рода, то следует выбрать возможно меньшее α.

Замечание1. При заданном объеме выборки уменьшить одновременно α, β невозможно.

Замечание2. Единственный способ одновременного уменьшения α и β состоит в увеличении объема выборки.

Дадим геометрическую иллюстрацию ошибок 1 и 2 рода. Для этого введем следующие обозначения: K – статистика критерия; $\overline{K}$ - наблюдаемые значения статистики критерия; $f_K(x, \theta_j)$- плотность вероятности статистики K $(j = 1,2...)$; $H_0 = \{\theta_1 = \theta_{10}, \theta_2 = \theta_{20}, ...\}$- простая нулевая гипотеза; $H_1 = \{\theta_1 = \theta_{11}, \theta_2 = \theta_{21}, ...\}$- конкурирующая гипотеза; $f_K(x, \theta_{j0})$ - плотность вероятности статистики при справедливости гипотезы H_0; $f_K(x, \theta_{j1})$ - плотность вероятности статистики критерия при справедливости гипотезы H_1; D – критическая область, AB–ее граница. Площади заштрихованных областей равны вероятностям $\alpha, \beta, (1-\beta)$ и иллюстрируют ошибки 1 и 2 рода.



§3.3. Проверка гипотезы $H_0 = \{\sigma^2 = \sigma_0^2\}$

Пусть σ^2 - неизвестная теоретическая дисперсия нормального распределения, σ_0^2 - предполагаемое значение теоретической дисперсии, S^2 - исправленная выборочная дисперсия. Выдвинутая гипотеза $H_0 = \{\sigma^2 = \sigma_0^2\}$ эквивалентна гипотезе $\widetilde{H}_0 = \{M(S^2) = \sigma_0^2\}$, так как $M(S^2) = \sigma^2$. Следовательно, для того, чтобы проверить гипотезу H_0, требуется установить, значимо или незначимо различаются S^2 и $\sigma_0{}^2$. На практике рассматриваемая гипотеза проверяется, если нужно проверить точность приборов, станков, методов исследования и устойчивость технологических процессов. Например, если допустимая величина характеристики рассеяния контролируемого размера деталей, изготовленных станком-автоматом, равна $\sigma_0{}^2$, а найденная по выборке дисперсия S^2 окажется значительно больше $\sigma_0{}^2$, то станок требует настройки. Проверка выдвинутой гипотезы $H_0 = \{\sigma^2 = \sigma_0^2\}$ основывается на следующем утверждении.

Утверждение 4. Если взаимно независимые случайные величины ξ_i одинаково распределены по нормальному закону, то тогда статистика критерия

$$\chi^2 = \frac{[n-1]S^2}{\sigma_0^2},$$

распределена по закону «хи-квадрат» с $k = n - 1$ степенями свободы (n – объем выборки) и имеет плотность вероятности

$$f_{\chi^2}(x) = \begin{cases} 0, & x \le 0 \\ C_k x^{\left(\frac{k}{2}-1\right)} \exp\left(-\frac{x}{2}\right), & x > 0 \end{cases}$$

где нормировочная константа $C_k = \dfrac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)}$.

Доказательство

Для этого представим статистику критерия следующим образом

$$\chi^2 = \sum_{i=1}^{n} U_i^2 ,$$

где $U_i = \dfrac{\xi_i - \overline{\xi}}{\sigma_0}$, случайные величины ξ_i распределены по нормальному закону

$$f_{\xi_i}(x_i) = \frac{1}{\sigma_0\sqrt{2\pi}} \exp\left[-\frac{(x_i - a)^2}{2\sigma_0^2}\right]. \qquad (1)$$

Найдем плотность вероятности для U_i, значения которой

$$u_i = \frac{x_i - \overline{x}}{\sigma_0}. \qquad (2)$$

Из теории вероятностей известно, что

$$f_{U_i}(u_i) = f_{\xi_i}(\sigma_0 u_i + \overline{x}) \cdot (x_i)'_{u_i}. \qquad (3)$$

Из выражения (2) следует, что

$$(x_i)' = \sigma_0, \qquad (4)$$

а согласно равенству (1)

$$f_{\xi_i}(\sigma_0 u_i + \overline{x}) = \frac{1}{\sigma_0\sqrt{2\pi}} \exp\left(-\frac{u_i^2}{2}\right). \qquad (5)$$

Подставим выражения (4-5) в равенство (3), получим

$$f_{U_i}(u_i) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u_i^2}{2}\right). \qquad (6)$$

Теперь найдем плотность вероятности $f_{Y_i}(y_i)$ для случайной величины $Y_i = U_i^2$. Так как функция $y = u^2$ немонотонна, то непосредственно применить формулу (3) нельзя. В таком случае следует сначала найти функцию распределения

$$F_{Y_i}(y_i) = P(Y_i < y_i) = P\left(U_i^2 < y_i\right) = P\left(-\sqrt{y_i} < U_i < \sqrt{y_i}\right) = \\ = F_{U_i}\left(\sqrt{y_i}\right) - F_{U_i}\left(-\sqrt{y_i}\right). \quad (7)$$

Дифференцируя выражение (7) по y_i, получим

$$f_{Y_i}(y_i) = F'_{Y_i}(y_i) = \frac{1}{2\sqrt{y_i}}\left[f_{U_i}\left(\sqrt{y_i}\right) + f_{U_i}\left(-\sqrt{y_i}\right)\right].$$

Так как, согласно выражению (6), функция f_{U_i} симметрична, то окончательно получим

$$f_{Y_i}(y_i) = \frac{1}{\sqrt{y_i}} f_{U_i}\left(\sqrt{y_i}\right) = \frac{1}{\sqrt{2\pi\, y_i}} \exp\left(-\frac{y_i}{2}\right), \; y > 0. \quad (8)$$

Найдем плотность вероятности для случайной величины $Y_1 + Y_2$, где случайные величины Y_1, Y_2 распределены по закону (8). Из теории вероятностей известно, что для независимых случайных величин

$$f_{Y_1+Y_2}(z) = \int_{-\infty}^{\infty} f_{Y_1}(y_1) f_{Y_2}(z - y_1)\, dy_1 = \\ = \frac{1}{2\pi}\int_0^z \frac{\exp\left(-\frac{y_1}{2}\right)}{\sqrt{y_1}} \cdot \frac{\exp\left(-\frac{z-y_1}{2}\right)}{\sqrt{z-y_1}} dy_1 = \\ = \left| t = \frac{y_1}{z} \right| = \frac{\exp\left(-\frac{z}{2}\right)}{2\pi}\int_0^1 t^{\left(\frac{1}{2}-1\right)} (1-t)^{(-0.5)} dt = \\ = \frac{\exp\left(-\frac{z}{2}\right)}{2\pi} \cdot \frac{\Gamma\left(\frac{1}{2}\right)\cdot\Gamma\left(\frac{1}{2}\right)}{\Gamma(1)} = \frac{\exp\left(-\frac{z}{2}\right)}{2}.$$

Аналогично

$$f_{Y_1+Y_2+\ldots+Y_{n-1}}(z) = C_k z^{\left(\frac{k}{2}-1\right)} \exp\left(-\frac{z}{2}\right), \; z > 0, k = n-1, \; C_k = \frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)},$$

что и требовалось показать.

В зависимости от вида конкурирующей гипотезы возможны три случая.

Случай 1. В этом случае конкурирующая гипотеза имеет вид $H_1 = \left\{\sigma^2 > \sigma_0^2\right\}$. Следовательно, критическая область является правосторонней. Правая критическая точка $\chi^2_{пр.кр}(\alpha, k)$ находится по таблице критических точек распределения «хи-квадрат» по входным данным α и k. Затем по выборке вычисляем наблюдаемое значение статистики. Если $\chi^2_{набл} < \chi^2_{пр.кр}$, то

гипотеза $H_0 = \{\sigma^2 = \sigma_0^2\}$ принимается. Если же $\chi^2_{набл} > \chi^2_{пр.кр}$, то гипотеза $H_0 = \{\sigma^2 = \sigma_0^2\}$ отклоняется.

Пример. По выборке объема n=13 вычислена выборочная дисперсия $S^2 = 14.6$. Выберем уровень значимости $\alpha = 0.01$. Пусть предполагаемое значение дисперсии равно $\sigma_0^2 = 12$. По таблице находим $\chi^2_{пр.кр}(0.01,12) = 26.2$

Вычислим наблюдаемое значение статистики $\chi^2_{набл} = \frac{(n-1)S^2}{\sigma_0^2} = 14.6$. Так как

$$\chi^2_{набл} < \chi^2_{пр.кр},$$

то гипотеза $H_0 = \{\sigma^2 = \sigma_0^2\}$ принимается (другими словами, отличие между S^2 и σ_0^2 незначимо).

Случай 2. В этом случае конкурирующая гипотеза имеет вид $H_1 = \{\sigma^2 < \sigma_0^2\}$. Следовательно, критическая область является левосторонней. Левая критическая точка $\chi^2_{лев.кр}$ находится по таблице правых критических точек распределения «хи-квадрат» по входным данным $1-\alpha$ и k. Затем по выборке вычисляем наблюдаемое значение статистики. Если $\chi^2_{набл} > \chi^2_{пр.кр}(1-\alpha,k)$, то гипотеза $H_0 = \{\sigma^2 = \sigma_0^2\}$ принимается. Если же $\chi^2_{набл} < \chi^2_{пр.кр}(1-\alpha,k)$ то гипотеза $H_0 = \{\sigma^2 = \sigma_0^2\}$ отклоняется

Случай 3. . В этом случае конкурирующая гипотеза имеет вид $H_1 = \{\sigma^2 \neq \sigma_0^2\}$. Следовательно, критическая область является двусторонней. По таблице критических точек распределения «хи-квадрат» по α и k находим критические точки $\chi^2_{лев.кр}\left(1-\frac{\alpha}{2},k\right), \chi^2_{пр.кр}\left(\frac{\alpha}{2},k\right)$ Затем по выборке вычисляем наблюдаемое значение статистики. Если $\chi^2_{лев.кр}\left(1-\frac{\alpha}{2},k\right) < \chi^2_{набл} < \chi^2_{пр.кр}\left(\frac{\alpha}{2},k\right)$, то гипотеза $H_0 = \{\sigma^2 = \sigma_0^2\}$ принимается. Если

$$\chi^2_{набл} < \chi^2_{лев.кр},\ \chi^2_{набл} > \chi^2_{пр.кр},$$

то гипотеза $H_0 = \{\sigma^2 = \sigma_0^2\}$ отклоняется.

Пример. По выборке объема n=13 вычислена выборочная дисперсия $S^2 = 10.3$. Выберем уровень значимости $\alpha = 0.02$. Пусть предполагаемое значение дисперсии равно $\sigma_0^2 = 12$. По таблице находим $\chi^2_{пр.кр}(0.01,12) = 26.2$, $\chi^2_{лев.кр}(0.99,12) = 3.57$ Вычислим наблюдаемое значение статистики $\chi^2_{набл} = \frac{(n-1)S^2}{\sigma_0^2} = 10.3$. Так как

$$\chi^2_{лев.кр}\left(1-\frac{\alpha}{2},k\right)<\chi^2_{набл}<\chi^2_{пр.кр}\left(\frac{\alpha}{2},k\right),$$

то гипотеза $H_0=\left\{\sigma^2=\sigma_0^2\right\}$ принимается (другими словами, отличие между S^2 и σ_0^2 незначимо).

§3.4. Проверка гипотезы $H_0=\{D(\xi)=D(\eta)\}$

Пусть ξ,η- случайные величины, распределенные нормально; S_ξ^2,S_η^2- исправленные выборочные дисперсии; $D(\xi),D(\eta)$- неизвестные теоретические дисперсии; $(x_1,x_2,....x_{n_1})$- выборка объема n_1 из генеральной совокупности ξ; $(y_1,y_2,....y_{n_2})$- выборка объема n_2 из генеральной совокупности η;

Так как

$$M\left(S_\xi^2\right)=D(\xi),M\left(S_\eta^2\right)=D(\eta),$$

то нулевая гипотеза $H_0=\{D(\xi)=D(\eta)\}$ эквивалентна гипотезе

$$\tilde{H}_0=\left\{\left(S_\xi^2\right)=\left(S_\eta^2\right)\right\}. \qquad (1)$$

Из выражения (1) видно, что для проверки $H_0=\{D(\xi)=D(\eta)\}$ требуется установить, значимо или незначимо различаются S_ξ^2 и S_η^2. Если окажется, что $H_0=\{D(\xi)=D(\eta)\}$ справедлива, то различие $D(\xi)$ и $D(\eta)$ незначимо. На практике выдвинутую гипотезу проверяют, если требуется сравнить точность приборов. Например, если различие S_ξ^2 и S_η^2 результатов измерений, выполненных двумя приборами, оказалось незначимым, то приборы имеют одинаковую точность.

Обоснованием первого пункта статистического критерия служит следующее утверждение. Пусть $S_\xi^2>S_\eta^2$. Для проверки гипотезы $H_0=\{D(\xi)=D(\eta)\}$ используется статистика Фишера-Снедекора

$$F=\frac{S_\xi^2}{S_\eta^2}.$$

Утверждение 5. При справедливости $H_0=\{D(\xi)=D(\eta)\}$ статистика

$$F=\frac{S_\xi^2}{S_\eta^2}$$

распределена по закону Фишера со степенями свободы

$$m_1=n_1-1,m_2=n_2-1$$

и имеет плотность вероятности

$$f_F(x)=\begin{cases}0, & x\le 0\\ C_m\dfrac{x^{\frac{m_1-2}{2}}}{\left(1+\dfrac{m_1}{m_2}x\right)^{\frac{m_1+m_2}{2}}}, & x>0,\end{cases}$$

где константа $C_m=\dfrac{\Gamma\left(\dfrac{m_1+m_2}{2}\right)}{\Gamma\left(\dfrac{m_1}{2}\right)\Gamma\left(\dfrac{m_2}{2}\right)}\left(\dfrac{m_1}{m_2}\right)^{\frac{m_1}{2}}$.

Доказательство

Пусть z- значения случайной величины F; x- значения случайной величины $\chi^2(m_1)=\dfrac{m_1S_\xi^2}{D(\xi)}$;

y-значения случайной величины $\chi^2(m_2)=\dfrac{m_2S_\eta^2}{D(\eta)}$. Ранее, мы показали, что $\chi^2(m_1)$распределена по закону «хи-квадрат» с числом степеней свободы $m_1=n_1-1$. Случайная величина $\chi^2(m_2)$ распределена по закону «хи-квадрат» с числом степеней свободы $m_2=n_2-1$. Эти величины имеют плотности вероятностей

$$f_{\chi^2\left(m_1\right)}(x)=\frac{x^{\frac{m_1}{2}-1}}{2^{\frac{m_1}{2}}\Gamma\left(\dfrac{m_1}{2}\right)}\exp\left(-\frac{x}{2}\right),\ x>0 \qquad (1)$$

$$f_{\chi^2\left(m_2\right)}(y)=\frac{y^{\frac{m_2}{2}-1}}{2^{\frac{m_2}{2}}\Gamma\left(\dfrac{m_2}{2}\right)}\exp\left(-\frac{y}{2}\right),\ y>0$$

Представим статистику критерия Фишера в виде $F=\dfrac{m_2}{m_1}\dfrac{\chi^2(m_1)}{\chi^2(m_2)}$.

При этом $z=\dfrac{m_2}{m_1}\dfrac{x}{y}$ и ,следовательно,

$$x = \frac{m_1}{m_2} zy, x'_z = \frac{m_1}{m_2} y. \qquad (2)$$

Из условия нормировки

$$\int_0^\infty f_F(z)dz = \int_0^\infty \int_0^\infty f_{\chi^2(m_1)}(x) f_{\chi^2(m_2)}(y)dx\,dy = 1.$$

с учетом равенств (1) и (2) следует, что

$$f_F(z) = \int_0^\infty x'_z f_{\chi^2(m_1)}(x) f_{\chi^2(m_2)}(y)dy = |(2)| =$$

$$= \frac{m_1}{m_2} \int_0^\infty y f_{\chi^2(m_1)}\left(\frac{m_1}{m_2} z\, y\right) f_{\chi^2(m_2)}(y)dy = |(1)| =$$

$$= \frac{\left(\frac{m_1}{m_2}\right)^{\frac{m_1}{2}} z^{\frac{m_1}{2}-1}}{2^{\frac{m_1+m_2}{2}} \Gamma\left(\frac{m_1}{2}\right)\Gamma\left(\frac{m_2}{2}\right)} \int_0^\infty y^{\frac{m_1+m_2}{2}-1} \exp\left[-\frac{y}{2}\left(1+\frac{m_1}{m_2} z\right)\right] dy =$$

$$= \left| t = \frac{y}{2}\left(1+\frac{m_1}{m_2} z\right) \right| = \frac{\left(\frac{m_1}{m_2}\right)^{\frac{m_1}{2}} z^{\frac{m_1}{2}-1}}{\Gamma\left(\frac{m_1}{2}\right)\Gamma\left(\frac{m_2}{2}\right)\left[1+\frac{m_1}{m_2} z\right]^{\frac{m_1+m_2}{2}}} \int_0^\infty t^{\left(\frac{m_1+m_2}{2}-1\right)} e^{-t}\, dt =$$

$$= \frac{\left(\frac{m_1}{m_2}\right)^{\frac{m_1}{2}} \Gamma\left(\frac{m_1+m_2}{2}\right)}{\Gamma\left(\frac{m_1}{2}\right)\Gamma\left(\frac{m_2}{2}\right)} z^{\frac{m_1}{2}-1} \left(1+\frac{m_1}{m_2} z\right)^{-\left(\frac{m_1+m_2}{2}\right)}$$

что и требовалось доказать.

При нахождении критических точек в зависимости от вида конкурирующей гипотезы возможны два случая.

Случай 1. В этом случае конкурирующая гипотеза имеет вид $H_1 = \{D(\xi) > D(\eta)\}$. Следовательно, критическая область является правосторонней. Правая критическая точка $F_{пр.кр}(\alpha, m_1, m_2)$ находится по таблице критических точек распределения Фишера. Затем по выборке вычисляется наблюдаемое значение статистики $F_{набл}$. Если окажется, что $F_{набл} < F_{пр.кр}$, то гипотеза H_0 принимается. Если окажется, что $F_{набл} > F_{пр.кр}$, то гипотеза H_0 отклоняется.

Случай 2. В этом случае конкурирующая гипотеза имеет вид $H_1 = \{D(\xi) \neq D(\eta)\}$. Следовательно, критическая область является двусторонней. Однако, так как по определению всегда $F_{набл} > 1$, то $F_{лев.кр}$ искать не нужно. Достаточно найти правую критическую точку, но по

уровню значимости $\frac{\alpha}{2}$ $F_{пр.кр}\left(\frac{\alpha}{2}, m_1, m_2\right)$. Затем по выборке находят наблюдаемое значение статистики Фишера $F_{набл}$. Если окажется, что $F_{набл} < F_{кр}$, то гипотеза H_0 принимается. Если же $F_{набл} > F_{кр}$, то гипотеза H_0 отклоняется.

§3.5. Проверка гипотезы $H_0 = \{M(\xi) = M(\eta)\}$ в случае, когда дисперсии известны

Пусть ξ, η - случайные величины, распределенные нормально; $D(\xi), D(\eta)$- известные теоретические дисперсии. Пусть

$$\overline{\xi} = \frac{1}{n}\sum_{i=1}^{n}\xi_i,\ \overline{\eta} = \frac{1}{m}\sum_{j=1}^{m}\eta_j,\ \overline{x} = \frac{1}{n}\sum_{i=1}^{n}x_i,\ \overline{y} = \frac{1}{m}\sum_{j=1}^{m}y_j,$$

ξ_i, η_j- взаимно независимые случайные величины, одинаково распределенные по нормальному закону, x_i, y_j-значения величин ξ_i, η_j, n - объем выборки из генеральной совокупности ξ, m-объем выборки из генеральной совокупности η.

В таком случае $M(\overline{\xi}) = M(\xi), M(\overline{\eta}) = M(\eta)$. Следовательно, гипотеза H_0 эквивалентна гипотезе $\tilde{H}_0 = \{(\overline{x}) = (\overline{y})\}$. Поэтому, чтобы проверить нулевую гипотезу H_0, требуется установить, значимо или незначимо различаются выборочные средние $\overline{x}$ и $\overline{y}$. На практике выдвинутая гипотеза проверяется, если нужно сравнить истинные средние размеры двух физических величин.

Проверка выдвинутой гипотезы основывается на следующем утверждении.

Утверждение 6. Если взаимно независимые случайные величины ξ_i, η_j одинаково распределены по нормальному закону

$$f_\xi(x) = \frac{1}{\sqrt{2\pi D(\xi)}}\exp\left(-\frac{(x-a)^2}{2D(\xi)}\right),\ f_\eta(y) = \frac{1}{\sqrt{2\pi D(\eta)}}\exp\left(-\frac{(y-b)^2}{2D(\eta)}\right)$$

то тогда при справедливости нулевой гипотезы статистика

$$Z = \frac{\overline{\xi} - \overline{\eta}}{\sigma(\overline{\xi} - \overline{\eta})}$$

распределена по стандартному нормальному закону $(a = 0, \sigma = 1)$, где

$$\sigma(\overline{\xi} - \overline{\eta}) = \sqrt{D(\overline{\xi} - \overline{\eta})} = \sqrt{D(\overline{\xi}) + D(\overline{\eta})} = \sqrt{\frac{D(\xi)}{n} + \frac{D(\eta)}{m}}.$$

Доказательство

Ранее было показано, что

$$f_{\overline{\xi}-a}(x)=\sqrt{\frac{n}{2\pi D(\xi)}}\exp\left(-\frac{nx^2}{2D(\xi)}\right),\ f_{\overline{\eta}-a}(y)=\sqrt{\frac{m}{2\pi D(\eta)}}\exp\left(-\frac{my^2}{2D(\eta)}\right)$$

Пусть $V=\overline{\xi}-\overline{\eta}$ а $v=x-y$ – возможные значения V. Тогда по закону композиции случайных величин с учетом справедливости нулевой гипотезы $(a=b)$

$$f_V(v)=\int_{-\infty}^{\infty}f_{\overline{\xi}-a}(x)f_{\overline{\eta}-a}(x-v)dx=\frac{\exp\left[-\frac{v^2}{2\left(\frac{D(\xi)}{n}+\frac{D(\eta)}{m}\right)}\right]}{\sqrt{2\pi\left(\frac{D(\xi)}{n}+\frac{D(\eta)}{m}\right)}}.$$

Отсюда окончательно находим ($f_Z(z)=\sigma f_V(\sigma z), z=\frac{v}{\sigma}, v'_z=\sigma$)

$$f_Z(z)=\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{z^2}{2}\right],$$

что и требовалось доказать.

В зависимости от вида конкурирующей гипотезы возможны три случая.

Случай 1. В этом случае конкурирующая гипотеза имеет вид

$$H_1=\{M(\xi)>M(\eta)\}.$$

Следовательно, критическая область является правосторонней. На практике такой случай имеет место, если введено усовершенствование технологического процесса, то естественно допустить, что оно приведет к увеличению выпуска продукции. Из $H_1=\{M(\xi)>M(\eta)\}$ следует, что критическая область является правосторонней, т.е. задается неравенством $Z>Z_{пр.кр}$. Критическую точку $Z_{пр.кр}$ находят при помощи функции Лапласа следующим образом. Из теории вероятностей известно, что

$$P(-\infty<Z<\infty)=\int_{-\infty}^{\infty}f_Z(x)dx=1. \qquad (1)$$

Так как $f_Z(x)$ симметрична относительно «0», то

$$P(0<Z<\infty)=\frac{1}{2}. \qquad (2)$$

Следовательно, если разбить интервал $(0,\infty)$ точкой $Z_{кр}$ на интервалы $(0,Z_{кр})$ и $(Z_{кр},\infty)$, то по теореме сложения вероятностей и с учетом того, что для непрерывных величин $P(Z=Z_{кр})=0$, имеем

$$P(0<Z<Z_{кр})+P(Z>Z_{кр})=\frac{1}{2}. \qquad (3)$$

Полагая $P(Z>Z_{кр})=\alpha$, из (3) получим

$$\Phi_0(Z_{кр})=\frac{1}{2}-\alpha. \qquad (4)$$

$Z_{кр}$, удовлетворяющее интегральному уравнению (4), находится по таблице функции Лапласа. Затем по выборке вычисляется

$$Z_{набл}=\frac{\bar{x}-\bar{y}}{\sqrt{\frac{D(\xi)}{n}+\frac{D(\eta)}{m}}}.$$

Если $Z_{набл}<Z_{кр}$, то гипотеза принимается; если $Z_{набл}>Z_{кр}$ - отклоняется.

Пример.

Дано: $n=10, m=10, \bar{x}=14.3, \bar{y}=12.2, D(\xi)=22, D(\eta)=18, \alpha=0.05$. По выборочным данным вычислим наблюдаемое значение статистики

$$Z_{набл}=\frac{14.3-12.2}{\sqrt{\frac{22}{10}+\frac{18}{10}}}=1.05$$

По таблице найдем $Z_{кр}=1.64$. Так как $Z_{набл}<Z_{кр}$, то гипотеза H_0 принимается. Другими словами $\bar{x}$ и $\bar{y}$ различаются незначимо.

Случай 2. В этом случае конкурирующая гипотеза имеет вид

$$H_1=\{M(\xi)<M(\eta)\}.$$

Следовательно, критическая область является левосторонней, т.е. задается неравенством $Z<Z_{лев.кр}$ **.** Левая критическая точка $Z_{лев.кр}$ находится как правая критическая точка, но по уровню значимости $(1-\alpha)$. Заменяя в (4) α на $1-\alpha$, получим

$$\Phi_0(Z_{лев.кр})=-\left(\frac{1}{2}-\alpha\right) \text{ или } \Phi_0(-Z_{лев.кр})=\left(\frac{1}{2}-\alpha\right). \qquad (5)$$

Сравнивая (4) и (5), заключаем $Z_{лев.кр}=-Z_{пр.кр}$. Затем по выборке вычисляем $Z_{набл}$. Если окажется, что $Z<Z_{лев.кр}=-Z_{пр.кр}$, то гипотеза отклоняется. Если $Z>Z_{лев.кр}=-Z_{пр.кр}$ - гипотеза принимается.

Пример.

Дано: $n=50, m=50, \bar{x}=142, \bar{y}=150, D(\xi)=28.2, D(\eta)=21.8, \alpha=0.01$. По выборочным данным вычислим наблюдаемое значение статистики

$$Z_{набл} = \frac{\bar{x} - \bar{y}}{\sigma(\bar{\xi} - \bar{\eta})} = \frac{142 - 150}{\sqrt{\frac{28.2}{50} + \frac{21.8}{50}}} = -8.$$

Из равенства $\Phi_0(Z_{пр.кр}) = \frac{1}{2} - \alpha = 0.49$ по таблице находим $Z_{пр.кр} = 2.33, Z_{лев.кр} = -2.33$. Так как $Z_{набл} < Z_{лев.кр}$, то нулевая гипотеза отклоняется. Другими словами, отличие $\bar{x}$ и $\bar{y}$ значимо.

Случай 3. В этом случае конкурирующая гипотеза имеет вид

$$H_1 = \{M(\xi) \neq M(\eta)\}.$$

Следовательно, критическая область является двусторонней, т.е. задается неравенствами $Z < Z_{лев.кр}$, $Z > Z_{пр.кр}$. Заменяя в (4) α на $\frac{\alpha}{2}$, получим

$$\Phi_0(Z_{пр.кр}) = \frac{1 - \alpha}{2}. \qquad (6)$$

Заменяя в (4) α на $1 - \frac{\alpha}{2}$, получим

$$\Phi_0(-Z_{лев.кр}) = \frac{1 - \alpha}{2}. \qquad (7)$$

Сравнивая (6) и (7), видим, что $Z_{лев.кр} = -Z_{пр.кр}$. Затем по выборке вычисляем $Z_{набл}$. Если

$-Z_{кр} < Z_{набл} < Z_{кр}$, то нулевая гипотеза принимается. Если же

$Z < -Z_{кр}, Z > Z_{кр}$, то нулевая гипотеза отклоняется.

Пример.

Дано: $n = 60, m = 50, \bar{x} = 1250, \bar{y} = 1275, D(\xi) = 120, D(\eta) = 100, \alpha = 0.01$.

Из $\Phi_0(Z_{кр}) = \frac{1 - \alpha}{2} = 0.495$

по таблице находим $Z_{пр.кр} = 2.58, Z_{лев.кр} = -2.58$. По выборке

$$Z_{набл} = \frac{\bar{x} - \bar{y}}{\sigma(\bar{\xi} - \bar{\eta})} = \frac{1250 - 1275}{\sqrt{\frac{120}{60} + \frac{100}{50}}} = -12.5.$$

Так как $Z_{набл} < Z_{лев.кр}$, то нулевая гипотеза отклоняется.

§3.6. Проверка гипотезы $H_0 = \{M(\xi) = M(\eta)\}$ для независимых величин

Пусть ξ,η - независимые случайные величины, распределенные нормально; $D(\xi), D(\eta)$ - неизвестные теоретические дисперсии, но известно, что они равны, n-объем выборки из генеральной совокупности ξ, m-объем выборки из генеральной совокупности η.

На практике такая гипотеза проверяется, например, если сравниваются средние размеры двух партий деталей, изготовленных на одном и том же станке. В таком случае естественно допустить, что дисперсии контролируемых размеров одинаковы, т.к детали изготовлены на одном и том же станке. Для проверки выдвинутой гипотезы используют следующее утверждение.

Утверждение 7. Если взаимно независимые случайные величины ξ_i, η_j одинаково распределены по нормальному закону, то тогда статистика критерия

$$T = \frac{\overline{\xi} - \overline{\eta}}{\sqrt{(n-1)S_\xi^2 + (m-1)S_\eta^2}} \sqrt{\frac{nm(n+m-2)}{n+m}},$$

распределена по закону Стьюдента с $k = n + m - 2$ степенями свободы и имеет плотность вероятности

$$f_T(x) = B_k \left[1 + \frac{x^2}{k}\right]^{-\frac{k+1}{2}}, \quad B_k = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{\pi k}\,\Gamma\left(\frac{k}{2}\right)},$$

здесь

$$\overline{\xi} = \frac{1}{n}\sum_{i=1}^{n}\xi_i, \overline{\eta} = \frac{1}{m}\sum_{j=1}^{m}\eta_j, \; S_\xi^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\xi_i - \overline{\xi}\right)^2, \; S_\eta^2 = \frac{1}{m-1}\sum_{j=1}^{m}\left(\eta_j - \overline{\eta}\right)^2.$$

Доказательство

Так как по условию утверждения $D(\xi) = D(\eta)$, то статистику T можно представить в следующем виде

$$T = \frac{U}{\mathrm{X}}\sqrt{\frac{mn\,(m+n-2)}{m+n}},$$

где $U = \frac{\overline{\xi} - \overline{\eta}}{\sigma}$, $\mathrm{X} = \sqrt{\mathrm{X}^2(n) + \mathrm{X}^2(m)}$, $\mathrm{X}^2(n) = \frac{(n-1)S_\xi^2}{\sigma^2}$,

$\mathrm{X}^2(m) = \frac{(m-1)S_\eta^2}{\sigma^2}$. Согласно **«Утверждению 1»**

$$f_{\frac{\xi}{\sigma}}(x)=\sqrt{\frac{n}{2\pi}}\exp\left(-\frac{n\,x^2}{2}\right),$$

$$f_{\frac{\eta}{\sigma}}(y)=\sqrt{\frac{m}{2\pi}}\exp\left(-\frac{m\,y^2}{2}\right).$$

Тогда в силу закона композиции

$$f_U(u)=\int\limits_{-\infty}^{\infty} f_{\frac{\xi}{\sigma}}(x)f_{\frac{\eta}{\sigma}}(x-u)dx=\sqrt{\frac{mn}{2\pi(n+m)}}\exp\left[-\left(\frac{mn}{m+n}\right)\frac{u^2}{2}\right].$$

Согласно **«Утверждению 4»**

$$f_{\mathrm{X}^2(n)}(x)=C_n\,x^{\frac{(n-1)}{2}-1}\exp\left(-\frac{x}{2}\right),\ x>0,$$

$$f_{\mathrm{X}^2(m)}(y)=C_m\,y^{\frac{(m-1)}{2}-1}\exp\left(-\frac{y}{2}\right),\ y>0\,.$$

Тогда плотность вероятности величины $\mathrm{X}^2\equiv\mathrm{X}^2(n)+\mathrm{X}^2(m)$ по закону композиции принимает вид

$$f_{\mathrm{X}^2}[\chi]=C_nC_m\int f_{\mathrm{X}^2(n)}(x)\,f_{\mathrm{X}^2(m)}(\chi-x)\,dx=\frac{\chi^{\frac{k}{2}-1}\exp\left(-\frac{\chi}{2}\right)}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)}.$$

По теореме о законе распределения функции одной величины

$$f_{\mathrm{X}}(t)=\frac{2t\left(t^2\right)^{\frac{k}{2}-1}\exp\left(-\frac{t^2}{2}\right)}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)}.$$

Из условия нормировки

$$\int\limits_0^{\infty} f_{\frac{U}{\mathrm{X}}}(z)\,dz=1$$

следует, что

$$f_{\frac{U}{\mathrm{X}}}(z)=\int\limits_0^{\infty} f_U(z\,t)f_{\mathrm{X}}(t)(x)'_z\,dt=\frac{\sqrt{mn}}{\sqrt{\pi}\sqrt{mn}}\frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)}\left(1+\frac{t^2\,mn}{m+n}\right)^{-\frac{k+1}{2}}.$$

Отсюда по теореме о законе распределения функции одной величины находим

$$f_T(w)=\frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{\pi k}\,\Gamma\left(\frac{k}{2}\right)}\left[1+\frac{w^2}{k}\right]^{-\frac{k+1}{2}},$$

что и требовалось показать.

В зависимости от вида конкурирующей гипотезы возможны три случая.

Случай 1. В этом случае конкурирующая гипотеза имеет вид $H_1=\{M(\xi)>M(\eta)\}$. Следовательно, критическая область является правосторонней. Правая критическая точка $t_{пр.кр}(\alpha,k)$, удовлетворяющая интегральному уравнению

$$P\left(T>t_{пр.кр}\right)=B_k\int\limits_{t_{пр.кр}}^{\infty}\left(1+\frac{x^2}{k}\right)^{-\frac{1+k}{2}}dx=\alpha.$$

находится по таблице критических точек распределения Стьюдента. Затем по выборке вычисляют наблюдаемое значение статистики $T_{набл}$. Если окажется, что $T_{набл}<t_{пр.кр}$, то нулевая гипотеза принимается. Если окажется, что $T_{набл}>t_{пр.кр}$, то нулевая гипотеза отклоняется.

Случай 2. В этом случае конкурирующая гипотеза имеет вид $H_1=\{M(\xi)<M(\eta)\}$. Следовательно, критическая область является левосторонней. Так как распределение Стьюдента является симметричным от относительно нуля, то левая критическая точка $t_{лев.кр}(\alpha,k)=-t_{пр.кр}$. Затем по выборке вычисляют наблюдаемое значение статистики $T_{набл}$. Если окажется, что $T_{набл}<t_{лев.кр}$, то нулевая гипотеза отклоняется. Если окажется, что $T_{набл}>t_{лев.кр}$, то нулевая гипотеза принимается.

Случай 3. В этом случае конкурирующая гипотеза имеет вид $H_1=\{M(\xi)\neq M(\eta)\}$. Следовательно, критическая область является двусторонней. Критические точки, удовлетворяющие системе интегральных уравнений

$$P\left(T<t_{лев.кр}\right)=B_k\int\limits_{-t_{пр.кр}}^{\infty}\left(1+\frac{x^2}{k}\right)^{-\frac{1+k}{2}}dx=\frac{\alpha}{2},$$

$$P\left(T>t_{пр.кр}\right)=B_k\int\limits_{t_{пр.кр}}^{\infty}\left(1+\frac{x^2}{k}\right)^{-\frac{1+k}{2}}dx=\frac{\alpha}{2}.$$

находится по таблице критических точек распределения Стьюдента. Затем по выборке вычисляют наблюдаемое значение статистики $T_{набл}$. Если окажется,

что $T_{набл} < t_{лев.кр}, T_{набл} > t_{пр.кр}$ то нулевая гипотеза отклоняется. Если окажется, что $t_{пр.кр} < T_{набл} < t_{лев.кр}$, то нулевая гипотеза принимается.

Пример.Дано: $n = 5,\ m = 5,\ S_\xi^2 = 0.25,\ \overline{x} = 3.3,\ \overline{y} = 2.48,\ S_\eta^2 = 0.108,\ \alpha = 0.05$

Проверим гипотезу $H_0 = \{M(\xi) = M(\eta)\}$ против конкурирующей гипотезы $H_1 = \{M(\xi) \neq M(\eta)\}$. Так как $S_\xi^2 \neq S_\eta^2$, то следует проверить прежде гипотезу

$$\tilde{H}_0 = \{D(\xi) = D(\eta)\}$$

по критерию Фишера:

$F_{набл} = \dfrac{0.25}{0.108} = 2.31,\ F_{кр}(0.05,4,4) = 6.39 \Rightarrow F_{набл} < F_{кр}$ - гипотеза $\tilde{H}_0 = \{D(\xi) = D(\eta)\}$ принимается.

По выборке вычислим наблюдаемое значение статистики $T_{набл} = \dfrac{\overline{x} - \overline{y}}{\sqrt{(n-1)S_\xi^2 + (m-1)S_\eta^2}} \sqrt{\dfrac{n\,m(n+m-2)}{n+m}} = 3.23$. По таблице найдем $t_{кр}(0.025,8) = 2.73$. Так как $T_{набл} > t_{кр}$, то $H_0 = \{M(\xi) = M(\eta)\}$ отклоняется.

§3.7. Проверка гипотезы $H_0 = \{M(\xi) = M(\eta)\}$ для зависимых величин

Пусть ξ, η - зависимые случайные величины, распределенные нормально; $D(\xi), D(\eta)$ - неизвестные теоретические дисперсии. Выборки имеют одинаковый объем n.

На практике такая гипотеза проверяется, например, если требуется сравнить два метода исследования, осуществленных одной лабораторией, или если исследование проведено одним и тем же методом двумя различными лабораториями.

Для проверки такой гипотезы используют статистику

$$T = \frac{\overline{D}\sqrt{n}}{S_d},$$

которая распределена по закону Стьюдента с числом степеней свободы $k = n - 1$, где

$$\overline{D} = \overline{\xi} - \overline{\eta} = \frac{1}{n}\sum_{i=1}^{n} D_i,\ D_i = \xi_i - \eta_i,\ S_d = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(D_i - \overline{D}\right)^2}.$$

Обычно используют конкурирующую гипотезу вида $H_1 = \{M(\xi) \neq M(\eta)\}$. Следовательно, критическая область является двусторонней. Правая критическая точка $t_{пр.кр}\left(\dfrac{\alpha}{2}, k\right)$ находится по таблице критических точек распределения Стьюдента. Так как распределение Стьюдента симметрично относительно «0», то $t_{лев.кр} = -t_{пр.кр}$. Затем по выборке вычисляем наблюдаемое значение статистики $T_{набл}$. Если окажется,

что $t_{лев.кр} < T_{набл} < t_{пр.кр}$, то гипотеза принимается. Если же $T_{набл} < t_{лев.кр}, T_{набл} > t_{пр.кр}$ - гипотеза отклоняется.

Пример. Двумя приборами измерены 5 деталей и получены следующие результаты:

$$\begin{cases} x_1 = 6,\ x_2 = 7,\ x_3 = 8,\ x_4 = 5,\ x_5 = 7 \\ y_1 = 7,\ y_2 = 6,\ y_3 = 8,\ y_4 = 7,\ y_5 = 8, \quad \alpha = 0.05 \end{cases}$$

Тогда

$D_1 = 6-7 = -1,\ D_2 = 7-6 = 1,\ D_3 = 8-8 = 0,\ D_4 = 5-7 = -2,\ D_5 = 7-8 = -1$ и, следовательно, $\overline{D} = -0.6$, $S_d = \sqrt{1.3}$, $T_{набл} = -1.18$. По таблице критических точек распределения Стьюдента находим

$$t_{пр.кр}(0.025,4) = 2.78,\ t_{лев.кр}(0.025,4) = -2.78.$$

Так как $t_{лев.кр} < T_{набл} < t_{пр.кр}$, то гипотеза $H_0 = \{M(\xi) = M(\eta)\}$ принимается.

§3.8. Проверка гипотезы $H_0 = \{M(\xi) = a_0\}$ в случае известной дисперсии

Пусть ξ - случайная величина, распределенная нормально; $D(\xi)$ - известная дисперсия; $M(\xi)$ - неизвестное математическое ожидание; a_0 - предполагаемое значение математического ожидания; Пусть $\overline{\xi} = \frac{1}{n}\sum_{i=1}^{n}\xi_i$ - выборочное среднее; в соответствии с классической вероятностной моделью ξ_i - взаимно независимые одинаково распределенные по нормальному закону случайные величины.

В таком случае $M(\overline{\xi}) = M(\xi)$. Поэтому нулевая гипотеза эквивалентна гипотезе $\widetilde{H}_0 = \{(\overline{\xi}) = a_0\}$. Следовательно, для проверки нулевой гипотезы требуется установить, значимо различие $\overline{\xi}$ и a_0 или нет.

На практике такая гипотеза проверяется, если требуется сравнить средний размер $\overline{x}$ партии деталей, изготовленных станком-автоматом, с проектным размером a_0.

Проверка гипотезы $H_0 = \{M(\xi) = a_0\}$ основывается на утверждении.

Утверждение 8. Если взаимно независимые случайные величины ξ_i, одинаково распределены по нормальному закону, то тогда статистика критерия

$$V = \frac{\overline{\xi} - a_0}{\sigma(\overline{\xi})},$$

где $\sigma(\bar{\xi}) = \sqrt{D(\bar{\xi})} = \sqrt{\frac{D(\xi)}{n}}$, распределена по стандартному нормальному закону

$$f_V(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right).$$

Доказательство

Для этого представим статистику в виде $V = \sum_{i=1}^{n} U_i$, где случайная величина $U_i = \frac{\xi_i - a_0}{\sqrt{nD(\xi)}}$, а ее значение

$$u_i = \frac{x_i - a_0}{\sqrt{nD(\xi)}}. \quad (1)$$

По постановке задачи предполагается, что

$$f_{\xi_i}(x_i) = \frac{1}{\sqrt{2\pi D}} \exp\left(-\frac{(x_i - a_0)^2}{2D}\right). \quad (2)$$

Найдем $f_{U_i}(u_i)$. Для этого из (1) получим $x_i = \sqrt{nD}u_i + a_0$. Тогда производная

$$(x_i)'_u = \sqrt{nD}. \quad (3)$$

Из теории вероятностей известно, что

$$f_{U_i}(u_i) = f_{\xi_i}\left(\sqrt{nD}u_i + a_0\right) \cdot (x_i)'_u. \quad (4)$$

Учитывая (2) и (3) из (4) получим

$$f_{U_i}(u_i) = \sqrt{\frac{n}{2\pi}} \exp\left(-\frac{n u_i^2}{2}\right). \quad (5)$$

Теперь найдем $f_{U_1+U_2}(z)$. Из теории вероятностей известно, что

$$f_{U_1+U_2}(z) = \int_{-\infty}^{\infty} f_{U_1}(u_1) f_{U_2}(z - u_1) du_1. \quad (6)$$

Подставляя (5) в (6), получим при n=2

$$f_{U_1+U_2}(z) = \frac{1}{\pi} \int_{-\infty}^{\infty} \exp\left(-\frac{u_1^2}{2 \cdot \frac{1}{2}}\right) \cdot \exp\left(-\frac{(z-u_1)^2}{2 \cdot \frac{1}{2}}\right) du_1 = \frac{1}{\pi} \exp\left(-\frac{z^2}{2}\right) \int_{-\infty}^{\infty} \exp\left(-\frac{(u_1 - A)^2}{2\left(\frac{1}{2}\right)^2}\right) du_1 =$$

$$= \frac{1}{\pi} \exp\left(-\frac{z^2}{2}\right)\left(\sqrt{2\pi} \cdot \frac{1}{2}\right) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right).$$

Аналогично $f_{U_1+U_2+U_3}(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$ и т.д. $f_V = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$, что и требовалось показать.

В зависимости от вида конкурирующей гипотезы возможны три случая.

Случай 1. В этом случае конкурирующая гипотеза имеет вид $H_1 = \{M(\xi) > a_0\}$. Следовательно, критическая область в этом случае является правосторонней. Правая критическая точка $V_{пр.кр}$, удовлетворяющая уравнению

$$\Phi_0(V_{пр.кр}) = \frac{1}{2} - \alpha \qquad (7)$$

находится по таблице функции Лапласа $\Phi_0(x)$. Затем по выборке вычисляют наблюдаемое значение статистики $V_{набл}$. Если $V_{набл} < V_{пр.кр}$, то нулевая гипотеза принимается. Если $V_{набл} > V_{пр.кр}$,то нулевая гипотеза отклоняется.

Случай 2. В этом случае конкурирующая гипотеза имеет вид $H_1 = \{M(\xi) < a_0\}$. Следовательно, критическая область в этом случае является левосторонней. Правая критическая точка $V_{пр.кр}$, удовлетворяющая уравнению (7), находится по таблице. Затем находят левую критическую точку $V_{лев.кр} = -V_{пр.кр}$. Затем по выборке вычисляют наблюдаемое значение статистики $V_{набл}$. Если $V_{набл} < V_{лев.кр}$, то нулевая гипотеза отклоняется. Если $V_{набл} > V_{лев.кр}$,то нулевая гипотеза принимается.

Случай 3. В этом случае конкурирующая гипотеза имеет вид $H_1 = \{M(\xi) \neq a_0\}$. Следовательно, критическая область в этом случае является двусторонней. Вспомогательная правая критическая точка $V_{пр.кр}$ находится из уравнения (7), в котором в соответствии с общими правилами построения критических областей вместо α следует подставить $\frac{\alpha}{2}$

$$\Phi_0(V_{пр.кр}) = \frac{1-\alpha}{2}.$$

Затем находят левую критическую точку $V_{лев.кр} = -V_{пр.кр}$. После этого по выборке вычисляют наблюдаемое значение статистики $V_{набл}$. Если $V_{лев.кр} < V_{набл} < V_{пр.кр}$, то нулевая гипотеза принимается.

Если $V_{набл} > V_{пр.кр}, V_{набл} < V_{лев.кр}$,то нулевая гипотеза отклоняется.

Пример. Дано: $n = 36, \sigma = 0.36, \bar{x} = 21.6, \alpha = 0.05, a_0 = 21$. Проверим гипотезу

$$H_0 = \{M(\xi) = 21\}$$

против конкурирующей гипотезы $H_1 = \{M(\xi) \neq 21\}$. Запишем уравнение $\Phi_0(V_{пр.кр}) = \frac{1-\alpha}{2} = 0.475$. По таблице находим $V_{пр.кр}$=1.96, $V_{лев.кр} = -1.96$. По выборке находим наблюдаемое значение статистики $V_{набл} = 10$. Так как $V_{набл} > V_{пр.кр}$, то нулевая гипотеза $H_0 = \{M(\xi) = 21\}$ отклоняется.

Пример. Данные те же. Проверим гипотезу $H_0 = \{M(\xi) = 21\}$ против конкурирующей гипотезы $H_1 = \{M(\xi) > 21\}$. В этом случае уравнение, содержащее функцию Лапласа, имеет другой вид

$$\Phi_0\left(V_{пр.кр}\right)=\frac{1}{2}-\alpha=0.45.$$

По таблице находим $V_{пр.кр}$=1.65. Так как $V_{набл}>V_{пр.кр}$, то нулевая гипотеза $H_0=\{M(\xi)=21\}$ отклоняется.

§3.9. Проверка гипотезы $H_0=\{M(\xi)=a_0\}$ в случае неизвестной дисперсии

Пусть ξ - случайная величина, распределенная нормально; $D(\xi)$ - неизвестная дисперсия; $M(\xi)$ - неизвестное математическое ожидание; a_0 - предполагаемое значение математического ожидания; $\overline{\xi}=\frac{1}{n}\sum_{i=1}^{n}\xi_i$ - выборочное среднее; в соответствии с классической вероятностной моделью ξ_i - взаимно независимые одинаково распределенные по нормальному закону случайные величины.

Проверки гипотезы $H_0=\{M(\xi)=a_0\}$ основывается на утверждении.

Утверждение 9. Если взаимно независимые случайные величины ξ_i,η_j одинаково распределены по нормальному закону, то тогда статистика критерия

$$T=\frac{\overline{\xi}-a_0}{S}\sqrt{n},$$

где $S=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\xi_i-\overline{\xi}\right)^2}$, распределена по закону Стьюдента с числом степеней свободы $k=n-1$.

Доказательство

Представим статистику критерия в виде

$$T=\frac{U}{S},$$

где $U=\sqrt{n}\left(\overline{\xi}-a_0\right)$. Пусть t- значения T,u-значения U,z-значения S. Тогда $t=\frac{u}{z}$, так что

$$u=tz,\ u'_t=z. \qquad (1)$$

Из условия нормировки

$$\int f_T(t)dt=\iint f_U(u)f_S(z)du\,dz=\iint f_U(t\cdot z)f_S(z)u'_t\,dz\,dt=1$$

и с учетом равенства (1) следует, что

$$f_T(t)=\iint f_U(t\cdot z)f_S(z)z\,dz\ . \qquad (2)$$

Пусть x-значения величина $\overline{\xi}$. Тогда $u=\sqrt{n}(x-a_0)$ и

$$x = \frac{u}{\sqrt{n}} + a_0,\ x'_u = \frac{1}{\sqrt{n}}. \qquad (3)$$

С учетом **«Утверждения 1»** и соотношения (3) по теореме о законе распределения функции одной величины находим

$$f_U(u) = \frac{1}{\sqrt{n}} f_{\bar{\xi}}\left[\frac{u}{\sqrt{n}} + a_0\right] = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{u^2}{2\sigma^2}\right). \qquad (4)$$

Аналогично

$$f_\chi(y) = 2y f_{\chi^2}\left(y^2\right) = 2C_n \left(y^2\right)^{\frac{k+1}{2}-1} \exp\left(-\frac{y^2}{2}\right),$$

$$C_n = \frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)},$$

где $\chi = \sqrt{\chi^2}$, χ^2-случайная величина, распределенная по закону «хи-квадрат» с $k = n-1$ степенями свободы. Так как $S = \frac{\sqrt{n-1}}{\sigma}\chi$, то

$$f_S(z) = \frac{\sqrt{n-1}}{\sigma} f_\chi\left(\frac{\sqrt{n-1}}{\sigma} z\right) =$$

$$= 2C_n \left(\frac{n-1}{\sigma^2}\right)^{\frac{k}{2}} \left(z^2\right)^{\frac{k+1}{2}-1} \exp\left(-\frac{z^2(n-1)}{2\sigma^2}\right). \qquad (5)$$

Подставляя выражения (4) и (5) в формулу (2),
получим

$$f_T(t) = \frac{2^{\frac{k}{2}} C_n}{\sqrt{\pi k}} \left[1 + \frac{t^2}{k}\right]^{-\frac{k+1}{2}} \int_0^\infty v^{\frac{k+1}{2}-1} e^{-v} dv =$$

$$= \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{\pi k}\,\Gamma\left(\frac{k}{2}\right)} \left[1 + \frac{t^2}{k}\right]^{-\frac{k+1}{2}}, t > 0,$$

что и требовалось показать.

В зависимости от вида конкурирующей гипотезы возможны три случая.

Случай 1. В этом случае конкурирующая гипотеза имеет вид $H_1 = \{M(\xi) > a_0\}$. Следовательно, критическая область в этом случае является правосторонней. Правую критическую точку $t_{пр.кр}(\alpha, k)$ находят по

таблице критических точек распределения Стьюдента. По выборке вычисляют $T_{набл}$. Если $T_{набл} < t_{пр.кр}$, то нулевая гипотеза принимается. Если $T_{набл} > t_{пр.кр}$,то нулевая гипотеза отклоняется.

Случай 2. В этом случае конкурирующая гипотеза имеет вид $H_1 = \{M(\xi) < a_0\}$. Следовательно, критическая область в этом случае является левосторонняя. Левую критическую точку, используя равенство $t_{лев.кр} = -t_{пр.кр}$, находят по таблице правых критических точек распределения Стьюдента. По выборке вычисляют $T_{набл}$. Если $T_{набл} < t_{лев.кр}$, то нулевая гипотеза отклоняется. Если $T_{набл} > t_{лев.кр}$,то нулевая гипотеза принимается.

Случай 3. В этом случае конкурирующая гипотеза имеет вид $H_1 = \{M(\xi) \neq a_0\}$. Следовательно, критическая область в этом случае является двусторонней. По таблице критических точек распределения Стьюдента находим критические точки $t_{пр.кр}\left(\frac{\alpha}{2}, k\right)$, $t_{лев.кр} = -t_{пр.кр}$. Если $t_{лев.кр} < T_{набл} < t_{пр.кр}$, то нулевая гипотеза принимается.

Если $T_{набл} > t_{пр.кр}, T_{набл} < t_{лев.кр}$,то нулевая гипотеза отклоняется.

Пример. Дано: $n = 20, \bar{x} = 16, S = 4.5, a_0 = 15, \alpha = 0.05, H_1 = \{M(\xi) \neq 15\}$. По таблице находим $t_{пр.кр}(0.025,19) = 2.09, t_{лев.кр}(0.025,19) = -2.09$. По выборке находим $T_{набл} = 0.99$. Так как $t_{лев.кр} < T_{набл} < t_{пр.кр}$, то гипотеза $H_0 = \{M(\xi) = 15\}$ принимается.

§3.10. Критерий Бартлета

Критерий Бартлета используется для проверки гипотезы об однородности дисперсий

$$H_0 = \{D(\xi_1) = D(\xi_2) = ... = D(\xi_l)\}$$

Пусть $\xi_1, \xi_2, ...\xi_l$ - случайные величины, распределенные нормально; $D(\xi_1), ..., D(\xi_l)$- теоретические дисперсии; $S^2_{\xi_1}, ..., S^2_{\xi_l}$ - выборочные дисперсии. Для проверки нулевой гипотезы используют статистику Бартлета

$$B = \frac{V}{C},$$

где

$$V = 2.303\left[k \lg\left(\overline{S^2}\right) - \sum_{i=1}^{l} k_i \lg\left(S_i^2\right)\right],$$

$$C = 1 + \frac{1}{3(l-1)}\left[\sum_{i=1}^{l} \frac{1}{k_i} - \frac{1}{k}\right] > 1,$$

n_i - объемы выборок, $k_i = n_i - 1$ - числа степеней свободы; $\overline{S^2} = \frac{\sum_{i=1}^{l} k_i S_i^2}{k}, k = \sum_{i=1}^{l} k_i$.

Статистика В приближенно распределена по закону «хи-квадрат» с $m = l - 1$ степенями свободы, если $n_i > 3$.

Замечание. Распределение статистики В очень чувствительно к отклонению распределения величин ξ_l от нормального закона. Поэтому к выводам, полученным по этому критерию, надо относиться осторожно.

Обычно используют правостороннюю критическую область. Правую критическую точку $\chi^2_{пр.кр}(\alpha, l-1)$ находят по таблице критических точек распределения «хи-квадрат». Затем по выборке вычисляют наблюдаемое значение статистики.

Если $B_{набл} < B_{кр}$, то гипотеза $H_0 = \{D(\xi_1) = D(\xi_2) = ... = D(\xi_l)\}$ принимается.

Если $B_{набл} > B_{кр}$, то гипотеза $H_0 = \{D(\xi_1) = D(\xi_2) = ... = D(\xi_l)\}$ отклоняется.

Пример.

Дано: $n_1 = 10, n_2 = 13, n_3 = 15, n_4 = 16, S_1^2 = 0.25, S_2^2 = 0.40, S_3^2 = 0.36, S_4^2 = 0.46, \alpha = 0.05$

Составим расчетную таблицу

1	2	3	4	5	6	7	8
i	n_i	k_i	S_i^2	$k_i S_i^2$	$\lg(S_i^2)$	$k_i \lg(S_i^2)$	$\frac{1}{k_i}$
1	10	9	0.25	2.25	- 0.6021	- 5.4189	
2	13	12	0.40	4.80	- 0.3979	- 4.7748	
3	15	14	0.36	5.04	- 0.4437	- 6.2118	
4	16	15	0.46	6.90	- 0.3372	- 5.0580	
Σ		$k = 50$		18.99		- 21.4635	

$\overline{S^2} = 0.3798, V_{набл} = 1.02, \quad \chi^2_{пр.кр}(0.05, 3) = 7.8$. Так как $V_{набл} < \chi^2_{кр}$, то подавно и $B_{набл} < \chi^2_{кр}$, поскольку всегда $C > 1$. Следовательно, гипотеза принимается.

§3.11. Критерий Кочрена

Критерий Кочрена используется для проверки гипотезы об однородности дисперсий

$$H_0 = \{D(\xi_1) = D(\xi_2) = ... = D(\xi_l)\}$$

в случае, когда объемы выборок одинаковы. Для проверки такой гипотезы используют статистику

$$G = \frac{S^2_{\max}}{\sum\limits_{i=1}^{l} S_i^2}$$

которая распределена по закону Кочрена с числом степеней свободы $k = n - 1$, где n - объем выборки; S_i^2 - выборочные дисперсии, $S^2_{\max}$ - наибольшая из них.

На практике обычно используют правостороннюю критическую область. Критическую точку $G_{пр.кр}(\alpha, k, l)$ находят по таблице критических точек распределения Кочрена. Затем по данным наблюдений находят $G_{набл}$. Если $G_{набл} < G_{пр.кр}$, то нулевая гипотеза $H_0 = \{D(\xi_1) = D(\xi_2) = ... = D(\xi_l)\}$ принимается. Если $G_{набл} > G_{пр.кр}$, то нулевая гипотеза $H_0 = \{D(\xi_1) = D(\xi_2) = ... = D(\xi_l)\}$ отклоняется.

Пример.Дано: $n = 17, S_1^2 = 0.26, S_2^2 = 0.36,\ S_3^2 = 0.40,\ S_4^2 = 0.42,\ \alpha = 0.05$.
По выборке вычислим наблюдаемое значение статистики

$$G_{набл} = \frac{0.42}{1.44} = 0.29 \ .$$

По таблице находим критическую точку $G_{кр}(0.05, 16, 4) = 0.4366$. Так как $G_{набл} < G_{пр.кр}$, гипотеза $H_0 = \{D(\xi_1) = D(\xi_2) = ... = D(\xi_l)\}$ принимается.
Замечание. Так как $H_0 = \{D(\xi_1) = D(\xi_2) = ... = D(\xi_l)\}$ принята, то в качестве оценки общей теоретической дисперсии можно принять

$$D(\xi) = \frac{0.26 + 0.36 + 0.40 + 0.42}{4} = 0.36$$

§3.12. Проверка гипотезы $H_0 = \{P = p_0\}$

Пусть P – вероятность появления события A в каждом из n независимых испытаний, которая неизвестна. Пусть p_0 -предполагаемое значение вероятности P; $W = \frac{\xi}{m}$, ξ - случайная величина числа появлений события A; $\frac{m}{n}$ - значения W. Для проверки гипотезы $H_0 = \{P = p_0\}$ используют статистику

$$V = \frac{W - p_0}{\sqrt{p_0 q_0}}\sqrt{n},\ q_0 = 1 - p_0$$

Проверка гипотезы основывается на следующем утверждении.

Утверждение10. Случайная величина V при справедливости гипотезы $H_0 = \{P = p_0\}$ распределена по стандартному нормальному закону с плотностью вероятности

$$f_V = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)$$

Доказательство

Пусть y – значение V; x- значение ξ. Преобразуем $V = \frac{\xi - np_0}{\sqrt{np_0 q_0}}$. Тогда $y = \frac{x - np_0}{\sqrt{np_0 q_0}}$ или $x = \sqrt{np_0 q_0}\,y + np_0$. Отсюда находим $x'_y = \sqrt{np_0 q_0}$.

Представим случайную величину ξ в виде $\xi = \sum_{i=1}^{n}\xi_i$, где случайные величины ξ_i распределены по закону

ξ_i	0	1
p_i	q_0	p_0

Из таблицы следует

$$M(\xi_i) = p_0 \qquad (1)$$

$$D(\xi_i) = p_0 q_0 \qquad (2)$$

Тогда

$$M(\xi) = \sum_{i=1}^{n} M(\xi_i) = np_0 \qquad (3)$$

$$D(\xi) = \sum_{i=1}^{n} D(\xi_i) = np_0 q_0 \qquad (4)$$

По теореме Ляпунова случайная величина ξ имеет асимптотически нормальное распределение

$$f_\xi(x) = \frac{1}{\sqrt{2\pi D(\xi)}}\exp\left[-\frac{(x - M(\xi))^2}{2D(\xi)}\right] \qquad (5)$$

Из теории вероятностей известно, что

$$f_V(y) = x'_y f_\xi\left(\sqrt{np_0 q_0}\,y + np_0\right) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{y^2}{2}\right),$$

что и требовалось доказать.

В зависимости от вида конкурирующей гипотезы возможны три случая.

Случай 1. В этом случае конкурирующая гипотеза имеет вид $H_0 = \{P > p_0\}$. Следовательно, критическая область является правосторонней. Правая критическая $V_{пр.кр}$ точка находится по равенству

$$\Phi_0\left(V_{пр.кр}\right) = \frac{1}{2} - \alpha \qquad (6)$$

с помощью таблицы функции Лапласа. Затем по выборке находим наблюдаемое значение статистики $V_{набл}$. Если окажется, что $V_{набл} > V_{пр.кр}$, то гипотеза $H_0 = \{P = p_0\}$ отклоняется. Если же $V_{набл} < V_{пр.кр}$, то гипотеза $H_0 = \{P = p_0\}$ принимается.

Случай 2. В этом случае конкурирующая гипотеза имеет вид $H_0 = \{P < p_0\}$. Следовательно, критическая область является левосторонней. Правая критическая $V_{пр.кр}$ точка находится по равенству (6), а левая критическая точка равна $V_{лев.кр} = -V_{пр.кр}$. Затем по выборке находим наблюдаемое значение статистики $V_{набл}$. Если окажется, что $V_{набл} > V_{лев.кр}$, то гипотеза $H_0 = \{P = p_0\}$ принимается. Если же $V_{набл} < V_{лев.кр}$, то гипотеза $H_0 = \{P = p_0\}$ отклоняется.

Случай 3. В этом случае конкурирующая гипотеза имеет вид $H_0 = \{P \neq p_0\}$. Следовательно, критическая область является двусторонней. Правая критическая точка находится по равенству

$$\Phi_0\left(V_{пр.кр}\right) = \frac{1-\alpha}{2}, \qquad (7)$$

левая же критическая точка равна $V_{лев.кр} = -V_{пр.кр}$. Затем по выборке находим наблюдаемое значение статистики $V_{набл}$.Если окажется, что $V_{лев.кр} < V_{набл} < V_{пр.кр}$, то гипотеза $H_0 = \{P = p_0\}$ принимается. Если же $V_{набл} < V_{лев.кр}, V_{набл} > V_{пр.кр}$, то гипотеза $H_0 = \{P = p_0\}$ отклоняется.

Пример. Дано: $n = 100, \frac{m}{n} = 0.08, \alpha = 0.05, p_0 = 0.12$. Проверим гипотезу $H_0 = \{P = 0.12\}$ против конкурирующей гипотезы $H_1 = \{P \neq p_0\}$. Вычислим наблюдаемое значение статистики

$$V_{набл} = \frac{(0.08 - 0.12)}{\sqrt{0.12 \cdot 0.88}}\sqrt{100} = -1.23.$$

По таблице находим

$$V_{пр.кр} = 1.96, V_{лев.кр} = -1.96.$$

Так как $-1.96 < V_{набл} < 1.96$, то гипотеза $H_0 = \{P = 0.12\}$ принимается.

§3.13. Проверка гипотезы $H_0 = \{\rho = 0\}$

Пусть (ξ, η)- двумерная случайная величина, распределенная нормально; ρ - теоретический коэффициент корреляции; r - выборочный коэффициент корреляции; n - объем выборки.

Для проверки гипотезы $H_0 = \{\rho = 0\}$ используют статистику

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}},$$

которая распределена по закону Стьюдента с числом степеней свободы $k = n - 2$. Обычно используют двустороннюю критическую область, т.е. проверяют гипотезу $H_0 = \{\rho = 0\}$ против конкурирующей $H_1 = \{\rho \neq 0\}$. По таблице критических точек распределения Стьюдента находят правую критическую точку $t_{пр.кр}\left(\frac{\alpha}{2}, k\right)$. При этом левая критическая точка равна

$$t_{лев.кр} = -t_{лев.кр}.$$

Если окажется, что $t_{лев.кр} < T_{набл} < t_{пр.кр}$, то гипотеза принимается. Если же $T_{набл} < t_{лев.кр}, T_{набл} > t_{пр.кр}$, то гипотеза $H_0 = \{\rho = 0\}$ отклоняется.

Пример. Дано: $n = 122,\ r = 0.4,\ \alpha = 0.05$.

Тогда

$$T_{набл} = \frac{0.4\sqrt{122-2}}{\sqrt{1-(0.4)^2}} = 4.78.$$

По таблице находим $t_{пр.кр}(0.025, 120) = 1.98$. Так как $T_{набл} > 1.98$, то $H_0 = \{\rho = 0\}$ отклоняется.

§3.14. Критерий Пирсона

Пусть ξ - случайная величина; $F_\xi(x)$ - ее неизвестная функция распределения; $F_0(x)$- предполагаемая известная функция распределения. Критерий согласия Пирсона используется для проверки гипотезы

$$H_0 = \{F_\xi(x) = F_0(x)\}$$

Для проверки такой гипотезы используется статистика вида

$$\chi^2 = \sum_{i=1}^{r} \frac{[n_i - \nu_i]^2}{\nu_i},$$

где r - число интервалов (x_i, x_{i+1}), покрывающих числовую ось (если ξ - непрерывная величина), либо r - число вариантов (если ξ - дискретная величина); n_i - эмпирические частоты; $\nu_i = np_i$ - теоретические частоты; n - объем выборки; вероятности p_i вычисляются по заданному закону распределения.

Пирсон доказал, что в случае справедливости гипотезы $H_0 = \{F_\xi(x) = F_0(x)\}$ при $n \to \infty$ статистика $\chi^2 = \sum_{i=1}^{r} \frac{[n_i - \nu_i]^2}{\nu_i}$ распределена по закону «хи-квадрат» с числом степеней свободы $k = r - 1 - l$, где l - число неизвестных параметров $F_0(x)$.

Для проверки гипотезы $H_0 = \{F_\xi(x) = F_0(x)\}$ обычно используют правостороннюю критическую область. Критическую точку $\chi^2_{кр}(\alpha, k)$ находят по таблице критических точек распределения «хи-квадрат» по входным данным α, k. Если окажется, что $\chi^2_{набл} > \chi^2_{кр}$, то гипотеза $H_0 = \{F(x) = F_0(x)\}$ отклоняется. Если же $\chi^2_{набл} < \chi^2_{кр}$, то гипотеза принимается.

Замечание. Отметим, что на практике в случае непрерывных величин интервалы (x_i, x_{i+1}) выбирают так, чтобы теоретические частоты были не очень малыми, например $\nu_i \geq 7$.

Пример1. Через равные промежутки времени в тонком слое раствора золота регистрировалось число частиц золота, попадавших в поле зрения микроскопа. В результате наблюдений было получено следующее статистическое распределение

x_i	0	1	2	3	4	5	6	7
n_i	112	168	130	68	32	5	1	1

Здесь x_i - число частиц золота; n_i - число интервалов времени, в течение которых в поле зрения попало ровно x_i частиц; объем выборки равен $n = 517$. Проверим согласие эмпирических данных с законом Пуассона, приняв за уровень значимости $\alpha = 0.05$. Распределение Пуассона имеет вид

$$P(\xi = x_i) = \frac{\lambda^{x_i} \exp(-\lambda)}{(x_i)!}, \qquad (1)$$

где λ - параметр, который неизвестен. Известно, что точечной оценкой этого параметра является выборочное среднее. Найдем выборочное среднее $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} n_i x_i = 1.54$. Применим в качестве оценки $\lambda = 1.54$. Найдем p_i при $\lambda = 1.54$, используя формулу(1):

$$p_0 = P(\xi = 0) = \frac{\lambda^0 e^{-\lambda}}{0!} = 0.2144,\ p_1 = P(\xi = 1) = \frac{\lambda^1 e^{-\lambda}}{1!} = 0.3301, p_2 = 0.2542$$

$$p_3 = 0.1305,\ p_4 = 0.0502,\ p_5 = 0.0155, p_6 = 0.0040,\ p_7 = 0.0009$$

В соответствии с замечанием объединим малочисленные частоты (5+1+1) и соответствующие им теоретические вероятности сложим (0.0155+0.0040+0.0009=0.0204). В результате объединения получим следующую таблицу

x_i	0	1	2	3	4	5
n_i	112	168	130	68	32	7
p_i	0.2144	0.3301	0.2542	0.1305	0.0502	0.0204

Используя данные последней таблицы, найдем наблюдаемое значение статистики

$$\chi^2_{набл} = \sum_{i=1}^{5} \frac{[n_i - \nu_i]^2}{\nu_i} = 2.8$$

По таблице критических точек распределения «хи-квадрат» найдем критическое значение статистики $\chi^2_{кр}(0.05, 6) = 12.6$. Так как $\chi^2_{набл} < \chi^2_{кр}$, то гипотеза $H_0 = \{F(x) = F_0(x)\}$ принимается.

Пример2. Пусть дан интервальный статистический ряд

Δx_i	4-6	6-8	8-10	10-12	12-14	14-16	16-18	18-20	20-22	Всего
n_i	15	26	25	30	26	21	24	20	13	200

Проверим согласие эмпирических данных с нормальным законом при уровне значимости $\alpha = 0.05$, т.е. проверим гипотезу $H_0 = \{f(x) = f_0(x)\}$, где $f(x)$ - неизвестная плотность вероятности; $f_0(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-a)^2}{2\sigma^2}\right]$ - предполагаемая плотность вероятности. Параметры a, σ неизвестны. Оценим эти параметры. Для этого вычислим выборочное среднее и выборочное среднее квадратичное отклонение

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} n_i x_i = 12.63,\ S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} n_i (x_i - \bar{x})^2} = 4.695$$

Следовательно, $a \approx \bar{x} = 12.630, \sigma \approx S = 4.695$. Перейдем к величинам $Z_i = \frac{x_i - \bar{x}}{S}$. Вычислим концы интервалов (Z_i, Z_{i+1}). При этом будем предполагать, что $Z_1 = -\infty, Z_9 = +\infty$. Составим расчетную таблицу

i	x_i	x_{i+1}	$x_i - \bar{x}$	$x_{i+1} - \bar{x}$	Z_i	Z_{i+1}
1	4	6	-8.63	-6.63	$-\infty$	-1.41
2	6	8	-6.63	- 4.63	-1.41	-0.99
3	8	10	-4.63	-2.63	-0.99	-0.156
4	10	12	-2.63	-0.63	-0.156	-0.13
5	12	14	-0.63	1.37	-0.13	0.29
6	14	16	1.37	3.37	0.29	0.72
7	16	18	3.37	5.37	0.72	1.14
8	18	20	5.37	7.37	1.14	1.57
9	20	22	7.37	9.37	1.57	$+\infty$

Вычислим теоретические вероятности p_i по равенству

$$p_i = P(x_i < \xi < x_{i+1}) = P\left(Z_i < \frac{\xi - \bar{x}}{S} < Z_{i+1}\right) =$$

$$= \int_{Z_i}^{Z_{i+1}} f_0(z)dz = \Phi_0(Z_{i+1}) - \Phi_0(Z_i).$$

и найдем теоретические частоты $\nu_i = np_i$. Расчетные данные представим в виде таблицы

i	Z_i	Z_{i+1}	$\Phi_0(Z_i)$	$\Phi_0(Z_{i+1})$	p_i	ν_i
1	$-\infty$	-1.41	-0.5	-0.4207	0.0793	15.86
2	-1.41	-0.99	-0.4207	-0.3389	0.0818	16.36
3	-0.99	-0.156	-0.3389	-0.2123	0.1266	25.32
4	-0.156	-0.13	-0.2123	-0.0517	0.1606	32.16
5	-0.13	0.29	-0.0517	0.1141	0.1658	33.16
6	0.29	0.72	0.1141	0.2642	0.1501	30.02
7	0.72	1.14	0.2642	0.3729	0.1087	21.74
8	1.14	1.57	0.3729	0.4418	0.0689	13.78
9	1.57	$+\infty$	0.4418	0.5	0.0582	11.64
Проверка					$\sum p_i = 1$	$\sum \nu_i = 200$

Окончательно имеем

n_i	15	26	25	30	26	21	24	20	13
ν_i	15.86	16.36	25.32	32.16	33.16	30.02	21.74	13.78	11.64

По данным последней таблицы вычислим наблюдаемое значение статистики $\chi^2_{набл} = \sum_{i=1}^{9} \frac{(n_i - \nu_i)^2}{\nu_i} = 13.35$. Так как заданное нормальное распределение содержит два неизвестных параметра, то число степеней свободы $k = 9 - 1 - 2 = 6$. По таблице критических точек распределения «хи-квадрат» находим критическое значение статистики $\chi^2_{кр}(0.05, 6) = 12.59$. Поскольку $\chi^2_{набл} > \chi^2_{кр}$, то гипотеза отклоняется. Если же $\alpha = 0.01$, то $\chi^2_{кр}(0.01, 6) = 16.81$ и, следовательно, гипотеза принимается.

§3.15. Критерий Колмогорова

Критерий Колмогорова используется для проверки гипотезы

$$H_0 = \{F(x) = F_0(x)\}$$

Пусть ξ - непрерывная случайная величина; $F_\xi(x)$- неизвестная функция распределения ξ; $F_0(x)$ - известная предполагаемая функция распределения; $(x_1, x_2, ..x_n)$ - выборка объема n из генеральной совокупности; $(Z_1, Z_2, ..Z_k)$- вариационный ряд, соответствующий выборке.

Для проверки гипотезы используют статистику Колмогорова

$$K = \max_{-\infty < x < \infty} |F_n(x) - F_0(x)|,$$

где $F_n(x)$- выборочная функция распределения

$$F_n(x) = \begin{cases} 0, \ x \le Z_1 \\ \dfrac{m_j}{n}, \ Z_j < x \le Z_{j+1}, \ m_j = n_1 + n_2 + ..n_j, \ j = 1,2.....(k-1) \\ 1, \ x > Z_k \end{cases}$$

Для определения наблюдаемого значения статистики находят величины

$$K^{(1)} = \max_{j_j = 1,..k} \left(\frac{m_j}{n} - F_0(Z_j) \right)$$

$$K^{(2)} = \max_{j = 2.,..k} \left(F_0(Z_j) - \frac{m_{j-1}}{n} \right)$$

Тогда наблюдаемое значение статистики равно

$$K_{набл} = \max\left(K^{(1)}, K^{(2)}\right)$$

При выборе двусторонней критической области гипотеза отвергается, если

$$K_{набл} > K_{кр}(n, \alpha)$$

Критические значения $K_{кр}$ находят по таблице критических точек распределения Колмогорова. При выборе односторонней критической области гипотеза отвергается, если

$$K^{(1)}{}_{набл} > K_{кр}(n,2\alpha)$$

Пример. Требуется проверить на уровне значимости $\alpha = 0.05$ гипотезу $H_0 = \{F(x) = F_0(x)\}$ против конкурирующей гипотезы $H_1 = \{F_\xi(x) \neq F_0(x)\}$, здесь $F_0(x) = x$ - равномерное распределение на отрезке $[0,1]$. Ниже представлены данные выборки.

m_j	Z_j	$\frac{m_j}{n}$	$\frac{m_{j-1}}{n}$	$\frac{m_j}{n} - F_0(Z_j)$	$F_0(Z_j) - \frac{m_{j-1}}{n}$
1	0.0834	0.1	0.0	0.0166	0.0834
2	0.1174	0.2	0.1	0.0826	0.0174
3	0.1794	0.3	0.2	0.1206	- 0.0206
4	0.3094	0.4	0.3	0.0906	0.0094
5	0.5424	0.5	0.4	- 0.0424	**0.1424**
6	0.6288	0.6	0.5	- 0.0288	0.1288
7	0.6606	0.7	0.6	0.0394	0.0606
8	0.6917	0.8	0.7	0.1083	- 0.0083
9	0.7410	0.9	0.8	**0.1590**	- 0.0590
10	0.9401	1.0	0.9	0.0599	0.0401

Итак, имеем $K^{(1)}_{набл} = 0.1590,\ K^{(2)}_{набл} = 0.1424$, и, следовательно, $K_{набл} = \max(0.1590, 0.1424) = 0.1590$ По таблице находим $K_{кр}(0.05,10) = 0.4025$. Так как $K_{набл} < K_{кр}(n,\alpha)$, то гипотеза принимается.

§ 3.16. Критерий Манна-Уитни

Пусть ξ, η-непрерывные случайные величины. $F_\xi(x), F_\eta(y)$-неизвестные функции распределения этих величин. Пусть $(x_1,...x_{n_1})$-выборка объема n_1 из генеральной совокупности ξ; $(y_1,...y_{n_2})$- выборка объема n_2 из генеральной совокупности η; $n_1 > n_2$.

Критерий Манна-Уитни используется для проверки гипотезы

$$H_0 = \{F_\xi(x) = F_\eta(y)\}.$$

Для проверки такой гипотезы используется статистика

$$U = \min[U_1, U_2],$$

где

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2},\ U_2 = R_2 - \frac{n_2(n_2+1)}{2},\ R_1 = \sum_{i=1}^{n_1} r(x_i),\ R_2 = \sum_{j=1}^{n_2} r(y_j),$$

$r(x_i)$-ранг выборочного значения x_i; $r(y_j)$-ранг выборочного значения y_j.

В зависимости от вида конкурирующей гипотезы возможны 3 случая. В случае 1

$$H_1 = \{F_\xi(x) > F_\eta(y)\},$$

в случае 2

$$H_1 = \{F_\xi(x) < F_\eta(y)\},$$

в случае 3

$$H_1 = \{F_\xi(x) \neq F_\eta(y)\}.$$

Пусть $\widetilde{U}(n_1, n_2, \alpha)$-критическое значение статистики в случае односторонней критической области (случаи 1,2). Тогда при $U_{набл} < \widetilde{U}_{кр}$ нулевая гипотеза отклоняется, а в противном случае принимается. Пусть $U_{кр}(n_1, n_2, \alpha)$-критическое значение в случае двусторонней критической области (случай 3). Если $U_{набл} < U_{кр}$, то нулевая гипотеза отклоняется, а в противном случае–принимается, $\widetilde{U}(n_1, n_2, \alpha) = U(n_1, n_2, 2\alpha)$.

Пример. Даны выборки:

x_i	2.19	2.26	2.28	1.90	2.03	2.08	2.00	2.04	2.32	2.02	2.24	2.35

y_j	1.98	2.31	2.25	2.07	1.89	2.13	2.22	2.01	1.86	1.95	1.84

Проверим гипотезу $H_0 = \{F_\xi(x) = F_\eta(y)\}$ против гипотезы

$$H_1 = \{F_\xi(x) > F_\eta(y)\}$$

при уровне значимости $\alpha = 0.01$.

Для этого построим общий вариационный ряд:

1.84	1
1.86	2
1.89	3
1.90	4
1.95	5
1.98	6

2.00	7
2.01	8
2.02	9
2.03	10
2.04	11
2.07	12
2.08	13
2.13	14
2.19	15
2.22	16
2.24	17
2.25	18
2.26	19
2.28	20
2.31	21
2.32	22
2.35	23

Из общего вариационного ряда найдем ранги соответствующих выборочных значений:

x_i	2.19	2.26	2.28	1.90	2.03	2.08	2.00	2.04	2.32	2.02	2.24	2.35
$r(x_i)$	15	19	20	4	10	13	7	11	22	9	17	23

y_j	1.98	2.31	2.25	2.07	1.89	2.13	2.22	2.01	1.86	1.95	1.84
$r(x_j)$	6	21	18	12	3	14	16	8	2	5	1

Отсюда находим $R_1 = \sum_{i=1}^{12} r(x_i) = 170,\ R_2 = \sum_{j=1}^{11} r(y_j) = 106,$

$$U_1 = 170 - \frac{12 \cdot 13}{2} = 92,\ U_2 = 106 - \frac{11 \cdot 12}{2} 40.$$

Тогда $U_{набл} = \min(U_1, U_2) = 40.$ Из таблицы критических точек распределения Манна-Уитни находим $\tilde{U}_{кр}(12,11,0.01) = 24.$ Так как $U_{набл} = 40 > \tilde{U}_{кр} = 24$, то нулевая гипотеза принимается.

§3.17. Критерий Томпсона

В выборке иногда присутствуют такие результаты наблюдений, которые сильно отличаются от других значений. В этой связи возникает задача выявления и устранения таких результатов наблюдения. Для решения такой задачи применяют критерии исключения резко выделяющихся результатов

наблюдений. Критерий Томпсона является одним из таких. Пусть ξ - случайная величина, распределенная по нормальному закону $N(a,\sigma^2)$; Здесь a - математическое ожидание ξ; σ^2- дисперсия ξ; (x_i) - выборка объема n из генеральной совокупности ξ. В соответствии с классической вероятностной моделью сопоставим $x_i \to \xi_i$. Пусть $H_0 = \{\xi_i \in N(a,\sigma^2)\}$- гипотеза, состоящая в том, что все случайные величины ξ_i распределены по нормальному закону с одинаковыми параметрами (a,σ^2). Пусть $H_j = \{\xi_j \in N(a+\Delta,\sigma^2)\}$- гипотеза, состоящая в том, что при фиксированном j из множества $\{1,2...n\}$ ξ_j распределена по нормальному закону с той же дисперсией σ^2, но с другим математическим ожиданием $(a+\Delta)$. Критерий Томпсона используется для проверки гипотезы $H_0 = \{\xi_i \in N(a,\sigma^2)\}$ при конкурирующей гипотезе $H_j = \{\xi_j \in N(a+\Delta,\sigma^2)\}$. Для проверки такой гипотезы используется статистика Томпсона

$$T_i = \frac{\xi_i - \overline{\xi}}{\widetilde{S}},$$

$i = 1,2....n$, $\overline{\xi} = \frac{1}{n}\sum_{i=1}^{n}\xi_i$, $\widetilde{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(\xi_i - \overline{\xi})^2$. Обычно применяется двусторонняя критическая область. Критические точки $Z(k,\alpha)$ находятся по формуле

$$Z_{кр}(k,\alpha) = \frac{t(k,\alpha/2)}{\sqrt{k + t^2(k,\alpha/2)}}\sqrt{k+1}, \qquad (1)$$

где $k = n-2$, $t(k,\alpha/2)$- критические точки распределения Стьюдента. Если окажется, что $|T_{набл}| < Z_{кр}(k,\alpha)$, то гипотеза принимается. Если же $|T_{набл}| > Z_{кр}(k,\alpha)$, то гипотеза отклоняется.

Пример. Пусть дана выборка объема 9: 0.13,0.11, 0.10, 0.07, 0.12, 0.30, 0.17, 0.09, 0.15; уровень значимости $\alpha = 0.05$. По выборке найдем точечные оценки $\overline{x} = 0.138$, $\widetilde{S} = 0.0636$. Проверим, не имеется ли среди выборочных значений резко выделяющихся значений. Вначале проверим гипотезу относительно $x_{\max} = 0.30$. Вычислим наблюдаемое значение статистики Томпсона $T_{набл} = \frac{0.30 - 0.138}{0.0636} = 2.547$. По таблице Стьюдента $t_{кр}(0.025,7) = 2.36$ и по формуле (1) находим критическое значение $Z_{кр}(7,0.05) = 1.885$. Так как $T_{набл} > 1.885$, то гипотеза отклоняется, т.е. значение 0.30 значительно отклоняется от остальных наблюдаемых значений и его следует исключить. Далее проверим гипотезу относительно значения $x_{\min} = 0.07$. В этом случае $T_{набл} = \frac{0.07 - 0.138}{0.0636} = -1.07$. Так как $|T_{набл}| < 1.885$, то гипотеза принимается, т.е. нет основания для исключения значения 0.07.

Аналогичное заключение можно сделать относительно всех остальных значений в выборке.

§3.18. Однофакторный дисперсионный анализ

Пусть $\xi_1, \xi_2, ...\xi_n$ - случайные величины, распределенные нормально; дисперсии $D(\xi_i)$-неизвестны, но известно, что они равны; математические ожидания $M(\xi_i)$- неизвестны, но могут быть различны.

Возникает задача при заданном уровне значимости по выборочным средним проверить гипотезу

$$H_0 = \{M(\xi_1) = M(\xi_2) = ... = M(\xi_n)\}$$

Для проверки выдвинутой гипотезы используют метод дисперсионного анализа. На практике дисперсионный анализ применяют, чтобы установить, оказывает ли существенное влияние на изучаемую величину X качественный фактор F, который имеет p уровней $F_1, F_2, ...F_p$. Например, если требуется выяснить, оказывает ли влияние удобрение на урожай, то фактор F–удобрение, а его уровни $F_1, F_2, ...F_p$ -виды удобрений, X- урожай.

Основная идея дисперсионного анализа состоит в сравнении факторной дисперсии $S^2_{фак}$, порождаемой воздействием фактораF, и остаточной дисперсии $S^2_{ост}$, обусловленной случайными причинами.

Замечание1. Если уже установлено, что F существенно влияет на X, и требуется выяснить, какой из уровней оказывает существенное воздействие, то дополнительно производят попарное сравнение средних по критерию Стьюдента.

В более общих случаях исследуют воздействие нескольких факторов. Мы же ограничимся случаем, когда на X воздействует только один фактор, который имеет p уровней.

Общая, факторная и остаточная суммы

Пусть X – нормально распределенный признак, F -фактор, который имеет p уровней; q – число наблюдений на каждом уровне; x_{ij} - наблюдаемые значения X, где i – номер испытания (i=1,2,3….q); j-номер уровня (j=1,2….p);pq–число x_{ij}; $\bar{x}_j = \frac{1}{q}\sum_{i=1}^{q} x_{ij}$ -групповые средние; $\bar{x} = \frac{1}{pq}\sum_{j=1}^{p}\sum_{i=1}^{q} x_{ij}$ - общее среднее.

Допустим, что F оказывает существенное влияние на X. Следовательно, будут различаться $\bar{x}_j$, причем они тем больше рассеяны вокруг $\bar{x}$, чем больше воздействие фактора F. Отсюда следует, что для оценки воздействия фактора F целесообразно составить факторную сумму вида

$$S_{фак} = q \sum_{j=1}^{p} \left(\bar{x}_j - \bar{x}\right)^2 \qquad (1)$$

Так как на X, кроме F, воздействуют и случайные причины, то все x_{ij}, принадлежащие j группе, вообще говоря, различны и, значит, рассеяны вокруг $\bar{x}_j$. Отсюда следует, что для оценки влияния случайных причин целесообразно составить остаточную сумму вида

$$S_{ост} = \sum_{j=1}^{p} \sum_{i=1}^{q} \left(x_{ij} - \bar{x}_j\right)^2 \qquad (2)$$

Если все x_{ij} рассматривать как единую совокупность, то x_{ij} будут различны из-за воздействия F и случайных причин. Для оценки воздействия F и случайных причин целесообразно составить общую сумму вида

$$S_{общ} = \sum_{j=1}^{p} \sum_{i=1}^{q} \left(x_{ij} - \bar{x}\right)^2 \qquad (3)$$

При вычислении $S_{фак}$, $S_{общ}$, $S_{ост}$ удобно использовать следующие формулы

$$S_{общ} = \sum_{j=1}^{p} S_j - \frac{\left[\sum_{j=1}^{p} T_j\right]^2}{pq} \qquad (4)$$

$$S_{фак} = \frac{1}{q} \cdot \sum_{j=1}^{p} T_j{}^2 - \frac{\left[\sum_{j=1}^{p} T_j\right]^2}{pq} \qquad (5)$$

$$S_{ост} = \sum_{j=1}^{p} S_j - \frac{\left[\sum_{j=1}^{p} T_j{}^2\right]}{q} \qquad (6)$$

где $S_j = \sum_{i=1}^{q} x_{ij}^2$, $T_j = \sum_{i=1}^{q} x_{ij}$.

Выведем первую из них:

$$S_{общ} = \sum_{j=1}^{p} \sum_{i=1}^{q} \left(x_{ij} - \bar{x}\right)^2 = \sum_{j=1}^{p} \left(\sum_{i=1}^{q} x_{ij}{}^2 \right) + pq\bar{x}^2 - 2\bar{x} \sum_{j=1}^{p} \sum_{i=1}^{q} x_{ij} = \sum_{j=1}^{p} S_j - pq\bar{x}^2 = \sum_{j=1}^{p} S_j - \frac{1}{pq}\left[\sum_{j=1}^{p} T_j\right]^2$$

Выведем вторую формулу

$$S_{фак} = q \sum_{j=1}^{p} \left(\bar{x}_j - \bar{x}\right)^2 = q \sum_{j=1}^{p} x_j{}^2 + qp\bar{x}^2 - 2q\bar{x} \sum_{j=1}^{p} \bar{x}_j = \frac{1}{q} \sum_{j=1}^{p} T_j{}^2 - qp\bar{x}^2 = \frac{1}{q} \sum_{j=1}^{p} T_j{}^2 - \frac{1}{pq}\left[\sum_{j=1}^{p} T_j\right]^2$$

Выведем формулу

$$S_{общ} = S_{фак} + S_{ост}.$$

$$S_{общ} = \sum_{j=1}^{p}\sum_{i=1}^{q}\left(x_{ij} - \bar{x}\right)^2 = \sum_{j=1}^{p}\sum_{i=1}^{q}\left[\left(x_{ij} - \bar{x}_j\right) + \left(\bar{x}_j - \bar{x}\right)\right]^2 = \sum_{j=i=1}^{p}\sum^{q}\left(x_{ij} - \bar{x}_j\right)^2 +$$

$$+ \sum_{j=1}^{p}\sum_{i=1}^{q}\left(\bar{x}_j - \bar{x}\right)^2 + 2\sum_{j=1}^{p}\left(\bar{x}_j - \bar{x}\right)\sum_{i=1}^{q}\left(x_{ij} - \bar{x}_j\right) =$$

$$= S_{ост} + q\sum_{j=1}^{p}\left(\bar{x}_j - \bar{x}\right)^2 = S_{ост} + S_{фак}$$

Вычтем из формулы (4) формулу (5), получим формулу (6).
Что и требовалось показать.

Общая, факторная и остаточная дисперсии

Общей дисперсией называется величина

$$S^2{}_{общ} = \frac{1}{pq-1}S_{общ}$$

Факторной дисперсией называют величину

$$S^2{}_{фак} = \frac{1}{p-1}S_{фак}$$

Остаточной дисперсией называют величину

$$S^2{}_{ост} = \frac{1}{p(q-1)}S_{ост}$$

Проверка гипотезы $H_0 = \{M(\xi_1) = M(\xi_2) = ... = M(\xi_n)\}$ методом дисперсионного анализа основывается на следующем утверждении.

Утверждение11. Необходимым и достаточным условием справедливости гипотезы H_0 является справедливость гипотезы $\widetilde{H}_0 = \{S^2_{\text{фак}} = S^2_{\text{ост}}\}$.

Доказательство

I. *Необходимость*. Покажем, что если гипотеза о равенстве средних H_0 справедлива, т.е.

$$M(\xi_1) = M(\xi_2) = ... = M(\xi_n) = a,$$

то дисперсии $S^2_{общ}, S^2_{ост}, S^2_{фак}$ являются несмещенными оценками теоретической дисперсии $D(\xi)$.

В соответствии с классической вероятностной моделью совершим переход $x_{ij} \to \xi_{ij}, X \to \xi, \bar{x} \to \bar{\xi}, \bar{x}_j \to \bar{\xi}_j$. Тогда

$$M\left(S^2_{общ}\right)=\frac{1}{pq-1}M\left[\sum_{i,j}\left(\xi_{ij}-\bar{\xi}\right)^2\right]=\frac{1}{pq-1}M\left[\sum_{i,j}\left[\left(\xi_{ij}-a\right)+\left(a-\bar{\xi}\right)\right]^2\right]=$$

$$=\frac{1}{pq-1}M\left\{\sum_{i,j}\left(\xi_{ij}-a\right)^2+\sum_{i,j}\left(a-\bar{\xi}\right)^2+2\sum_{i,j}\left(\xi_{ij}-a\right)\left(a-\bar{\xi}\right)\right\}=$$

$$=\frac{1}{pq-1}M\left\{\sum_{i,j}\left(\xi_{ij}-a\right)^2+pq\left(a-\bar{\xi}\right)^2-2pq\left(a-\bar{\xi}\right)^2\right\}=$$

$$=\frac{1}{pq-1}\left\{\sum_{i,j}M\left(\xi_{ij}-a\right)^2-pqM\left(a-\bar{\xi}\right)^2\right\}=\frac{1}{pq-1}\left\{\sum_{i,j}D\left(\xi_{ij}\right)-pqD\left(\bar{\xi}\right)\right\}=$$

$$=\frac{1}{pq-1}\left\{pqD\left(\xi\right)-D\left(\xi\right)\right\}=D\left(\xi\right)$$

Таким образом $M\left(S^2_{общ}\right)=D\left(\xi\right)$, что и требовалось показать. Аналогично можно показать, что $S^2_{фак}, S^2_{ост}$ являются несмещенными оценками теоретической дисперсии**.** Так как дисперсии равны по условию задачи, то $M\left(S^2_{фак}\right)=M\left(S^2_{ост}\right)$. Следовательно, $S^2_{фак}, S^2_{ост}$ различаются незначимо. Поэтому если сравнивать эти оценки по критерию Фишера, то очевидно, критерий укажет, что гипотезу $\widetilde{H}_0$ следует принять.

Отметим, что если $\widetilde{H}_0$ ложна, то ложна и гипотеза H_0. Действительно, если бы H_0 была бы правильна, то согласно предыдущему $\widetilde{H}_0$ была бы правильна, что противоречит условию о том, что $\widetilde{H}_0$ ложна.

II**.** *Достаточность*. Если $\widetilde{H}_0$ правильна, то тогда $S^2_{фак}=S^2_{ост}$. Поэтому

$$\frac{S_{фак}}{p-1}=\frac{S_{ост}}{p(q-1)} \qquad (1)$$

Так как $S_{общ}=S_{фак}+S_{ост}$, то

$$S_{фак}=S_{общ}-S_{ост} \qquad (2)$$

Подставим (2) в (1), получим

$$\frac{S_{общ}}{pq-1}=\frac{S_{ост}}{p(q-1)} \qquad (3)$$

По определению

$$\begin{cases} S_{общ}=\sum\limits_{j=1}^{p}\sum\limits_{i=1}^{q}\left(x_{ij}-\bar{x}\right)^2 \\ S_{ост}=\sum\limits_{j=1}^{p}\sum\limits_{i=1}^{q}\left(x_{ij}-\bar{x}_j\right)^2 \end{cases} \qquad (4)$$

Учитывая (4) из (3) находим, что

$$\frac{\left(x_{ij}-\bar{x}\right)^2}{pq-1}-\frac{\left(x_{ij}-\bar{x}_j\right)^2}{p(q-1)}=0$$

Откуда заключаем, $x_{ij} = \bar{x} = \bar{x}_j$, т.е. групповые средние $\bar{x}_j$ равны, что и требовалось доказать.

Из доказанного утверждения следует, что для того, чтобы проверить гипотезу H_0 о равенстве групповых средних нормальных случайных величин с одинаковыми дисперсиями, достаточно проверить по критерию Фишера гипотезу $\widetilde{H}_0$.

Замечание1. Если окажется, что $S^2_{фак} < S^2_{ост}$, то уже отсюда следует справедливость H_0 и, значит, нет надобности прибегать к критерию Фишера.

Замечание 2. Если нет уверенности в справедливости предположения о равенстве дисперсий, то это предположение следует проверить предварительно по критерию Кочрена.

В заключение рассмотрим случай неодинакового числа испытаний на различных уровнях. Пусть q_1 - число испытаний на первом уровне F_1; q_2 - число испытаний на втором уровне F_2.. ; q_p - число испытаний на уровне F_p. В этом случае

$$S_{общ} = \sum_{j=1}^{p} P_j - \frac{1}{n}\left[\sum_{j=1}^{p} R_j\right]^2,$$

$$P_j = \sum_{i=1}^{q_j} x_{ij}^2,\ R_j = \sum_{i=1}^{q_j} x_{ij},\ n = \sum_{j=1}^{p} q_j$$

$$S_{фак} = \sum_{j=1}^{p} \frac{R_j^2}{q_j} - \frac{1}{n}\left[\sum_{j=1}^{p} R_j\right]^2$$

Соответственно $S_{ост} = S_{общ} - S_{фак}$, $S^2_{фак} = \dfrac{S_{фак}}{p-1}$, $S^2_{ост} = \dfrac{S_{ост}}{n-p}$.

Пример. Произведено по 4 испытания на каждом из 3 уровней. Результаты испытаний приведены в таблице.

Номер испытания i	F_j		
	F_1	F_2	F_3
1	51	52	42
2	52	54	44
3	56	56	50
4	57	58	52
$\bar{x}_j$	54	55	47

Методом дисперсионного анализа при уровне значимости $\alpha = 0.05$ проверим гипотезу о равенстве $\bar{x}_j$.

Решение. Перейдем к условным вариантам $y_{ij} = x_{ij} - 52$. Составим расчетную таблицу в условных вариантах

Номер Испытания i	F_j						Σ
	F_1		F_2		F_3		
	y_{i1}	y_{i1}^2	y_{i2}	y_{i2}^2	y_{i3}	y_{i3}^2	
1 2 3 4	-1 0 4 5	1 0 16 25	0 2 4 6	0 4 16 36	-10 -8 -2 0	100 64 4 0	
$S_j = \sum_{i=1}^{4} y_{ij}^2$		42		56		168	$\sum_{j=1}^{3} S_j = 266$
T_j	8		12		-20		$\sum_{j=1}^{3} T_j = 0$
T_j^2	64		144		400		$\sum_{j=1}^{3} T_j^2 = 608$

Пользуясь таблицей и учитывая, что p=3, q=4, найдем

$$S_{общ} = \sum_{j=1}^{3} S_j - \frac{1}{pq}\left[\sum_{j=1}^{3} T_j\right]^2 = 266 - 0 = 266$$

$$S_{факт} = \frac{1}{q}\sum_{j=1}^{3} T_j^2 - \frac{1}{pq}\left[\sum_{j=1}^{3} T_j\right]^2 = \frac{608}{4} - 0 = 152$$

Тогда $S_{ост} = 266 - 152 = 114$. Найдем $S_{фак}^2 = \frac{S_{фак}}{p-1} = \frac{152}{3-1} = 76$,

$$S_{ост}^2 = \frac{S_{ост}}{p(q-1)} = \frac{114}{3(4-1)} = \frac{114}{9} = 12.67$$

Сравним $S_{фак}^2$ и $S_{ост}^2$ по критерию Фишера . Для этого найдем наблюдаемое значение статистики критерия

$$F_{набл} = \frac{S_{фак}^2}{S_{ост}^2} = \frac{72}{12.67} = 6$$

Число степеней свободы числителя равно $k_1 = p - 1 = 3 - 1 = 2$; а знаменателя $k_2 = p(q-1) = 9$. Тогда по таблице $F_{кр}(0.05, 2, 9) = 4.26$. Так как $F_{набл} > F_{кр}$, то H_0 отвергаем. Другими словами, групповые средние $\bar{x}_j$ «в целом» различаются значимо. Если требуется сравнить средние попарно, то следует воспользоваться критерием Стьюдента.

§3.19. Дискриминантный анализ

Основные понятия и алгоритм метода

Дискриминантный анализ используется для классификации объектов. Исходными данными метода являются обучающие выборки, принадлежащими различным классам объектов. В качестве примера рассмотрим классификацию на 2 класса. Обычно предполагается, что математические ожидания случайных векторов каждого из классов различны, а дисперсии и ковариационные матрицы одинаковы, распределения предполагают подчиняющимися нормальному закону. Таким образом, имеем 2 плотности распределения $f_1(\bar{x}), f_2(\bar{x})$. В рассмотрение вводится некоторый новый объект с измеренной реализацией случайного вектора $\bar{z}$. Считается, что если $f_1(\bar{z}) > f_2(\bar{z})$, то рассматриваемый объект $\bar{z}$ будет принадлежать первому классу, и второму классу, если $f_1(\bar{z}) < f_2(\bar{z})$. Плотность вероятности каждого класса можно представить в виде

$$f_1(\bar{x}) = (2\pi)^{-\frac{k}{2}}|K_1|^{-\frac{1}{2}}exp\left[-\frac{1}{2}\left(\bar{x}-\bar{m}^{(1)}\right)^T K_{(1)}^{-1}\left(\bar{x}-\bar{m}^{(1)}\right)\right],$$

$$f_2(\bar{x}) = (2\pi)^{-\frac{k}{2}}|K_2|^{-\frac{1}{2}}exp\left[-\frac{1}{2}\left(\bar{x}-\bar{m}^{(2)}\right)^T K_{(2)}^{-1}\left(\bar{x}-\bar{m}^{(2)}\right)\right],$$

где $\bar{x}^T = (x_1, x_2, \dots x_k)$, $\bar{m}^{(i)}$ – вектор математических ожиданий размерности k, K_i – ковариационная матрица, $K_{(i)}^{-1}$ – обратная ковариационная матрица, предполагается, что $K_1 = K_2$.

Первое неравенство $f_1(\bar{x}) > f_2(\bar{x})$ с учетом допущения $K_1 = K_2$ соответствует неравенству между квадратичными формами

$$-\left(\bar{x}-\bar{m}^{(1)}\right)^T K_{(1)}^{-1}\left(\bar{x}-\bar{m}^{(1)}\right) > -\left(\bar{x}-\bar{m}^{(2)}\right)^T K_{(2)}^{-1}\left(\bar{x}-\bar{m}^{(2)}\right),$$

из которого с учетом

$$\bar{m}^{(1)T}K^{-1}\bar{x} = \bar{x}^T K^{-1}\bar{m}^{(1)} \equiv \left(\bar{x}, K^{-1}\bar{m}^{(1)}\right),$$

$$\bar{m}^{(1)T}K^{-1}\bar{m}^{(1)} = \bar{m}^{(1)}K^{-1}\bar{m}^{(1)T} \equiv \left(\bar{m}^{(1)}, K^{-1}\bar{m}^{(1)}\right),$$

находим

$$\left(\bar{x}, K^{-1}\bar{m}^{(1)}\right) - \frac{1}{2}\left(\bar{m}^{(1)}, K^{-1}\bar{m}^{(1)}\right) > \left(\bar{x}, K^{-1}\bar{m}^{(2)}\right) - \frac{1}{2}\left(\bar{m}^{(2)}, K^{-1}\bar{m}^{(2)}\right).$$

Функция

$$h_i = \left(\bar{x}, K^{-1}\bar{m}^{(i)}\right) - \frac{1}{2}\left(\bar{m}^{(i)}, K^{-1}\bar{m}^{(i)}\right)$$

называется линейной дискриминантной функцией. Эту функцию можно представить в виде

$$h_i = (\bar{x}, a_i) + b_i,$$

где $a_i = K^{-1}\bar{m}^{(i)}$ – вектор коэффициентов дискриминантной функции. На практике обычно исходят из определения знака разности $h_1 - h_2$, которую и рассматривают как дискриминантную функцию, так что если $h_1 - h_2 > 0$, то точку z относят к 1 классу. При этом дискриминантная функция принимает вид

$$U(\bar{x}) = (\bar{x}, \tilde{a}) + (b_1 - b_2),$$

где $\tilde{a} = \left(\bar{m}^{(1)}, K^{-1}(\bar{m}^{(1)} - \bar{m}^{(2)})\right)$. Величина $(b_1 - b_2)$ дает постоянное смещение дискриминантной функции, одинаковое как для обучающих выборок, так и для исследуемого объекта Z. Поэтому ее можно не учитывать.

Таким образом, алгоритм метода дискриминантного анализа выглядит следующим образом:

1. На основании имеющихся обучающих выборок двух классов

$$X = \begin{pmatrix} x_{11}\ x_{12}..x_{1k} \\ \dots\dots\dots\dots \\ x_{n_1 1}\ x_{n_1 2}\ x_{n_1 k} \end{pmatrix}, Y = \begin{pmatrix} y_{11}\ y_{12}..y_{1k} \\ \dots\dots\dots\dots \\ y_{n_2 1}\ y_{n_2 2}\ y_{n_2 k} \end{pmatrix},$$

где k – число измеряемых параметров, $n_{1,2}$ – объемы выборок по каждому из параметров, определяют выборочные средние значения

$$\bar{X} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \bar{x}_3 \\ \dots \\ \bar{x}_k \end{pmatrix}, \bar{Y} = \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \bar{y}_3 \\ \dots \\ \bar{y}_k \end{pmatrix},$$

где $\bar{x}_j = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{ij}$, $\bar{y}_j = \frac{1}{n_2}\sum_{i=1}^{n_2} y_{ij}$.

2. Затем вычисляют оценки ковариационных матриц

$$S_X = \{s_{ij}\}_X, S_Y = \{s_{ij}\}_Y.$$

Например,

$$\{s_{ij}\}_X = \frac{1}{n_1}\sum_{j=1}^{n_1}(x_{ji} - \bar{x}_i)(x_{jl} - \bar{x}_l) = \overline{x_i x_l} - \bar{x}_i\,\bar{x}_l.$$

3. Рассчитывают несмещенную оценку суммарной ковариационной матрицы

$$\hat{S} = \frac{1}{n_1 + n_2 - 2}[n_1 S_X + n_2 S_Y].$$

4. Определяют матрицу $\hat{S}^{-1}$, обратную к матрице $\hat{S}$.
5. Вычисляют вектор оценок дискриминантной функции

$$a = \hat{S}^{-1}(\bar{X} - \bar{Y}).$$

6. Рассчитывают оценки векторов значений дискриминантной функции для матриц исходных данных

$$\hat{U}_X = Xa, \hat{U}_Y = Ya.$$

7. Вычисляют средние значения оценок дискриминантной функции

$$\bar{u}_X = \frac{1}{n_1}\sum_{i=1}^{n_1} u_{X_i},\ \bar{u}_Y = \frac{1}{n_2}\sum_{i=1}^{n_2} u_{Y_i}.$$

8. Определяют границу между оценками дискриминантной функции двух классов как среднее арифметическое

$$\tilde{C} = \frac{1}{2}(\bar{u}_X + \bar{u}_Y).$$

9. Находят дискриминантную функцию для наблюдаемого объекта, заданного вектором измерений $z = (z_1, z_2, \ldots . z_k)$:

$$\hat{U}_Z = Za.$$

10. Проводят сравнение: если $\hat{U}_Z > \tilde{C}$, то наблюдаемый объект Z относится к первому классу X, а если $\hat{U}_Z < \tilde{C}$, то – ко второму классу Y.

Пример дисриминантного анализа

Из списка заводов, деятельность которых была оценена экспертами, выбраны 5 заводов (класс А), имеющих повышенную финансовую эффективность, и 5 заводов (класс В), имеющих недостаточную финансовую эффективность. С помощью дискриминантного анализа требуется оценить финансовую эффективность трех предприятий D1,D2,D3 и отнести каждый из них к одному из двух классов А, В. Деятельность каждого предприятия оценивается по трем показателям α(выручка от реализации),β (рентабельность продаж),γ (оборачиваемость активов). Исходные данные для предприятий приведены в таблицах.

Класс А	α	β	γ
Завод А1	1.7774	0.289	1.053
Завод А2	9.0798	0.124	1.347
Завод А3	1.0634	0.199	2.252
Завод А4	11.5849	0.284	1.665
Завод А5	0.6805	0.256	0.901

Класс В	α	β	γ
Завод В1	0.9335	0.221	1.285
Завод В2	0.0733	0.050	1.313
Завод В3	3.3637	0.215	1.064
Завод В4	3.3490	0.213	0.385
Завод В5	3.2397	0.176	0.577

Исследуемые предприятия D	α	β	γ
Завод D1	10.0796	0.136	1.073
Завод D2	2.0616	0.150	1.382
Завод D3	5.8125	0.234	1.828

Запишем исходные данные в виде матриц обучающих выборок X и Y с объемами выборок $n_1 = n_2 = 5$:

$$X = \begin{pmatrix} 1.7774 & 0.289 & 1.053 \\ 9.0798 & 0.124 & 1.347 \\ 1.0634 & 0.199 & 2.252 \\ 11.5849 & 0.284 & 1.665 \\ 0.6805 & 0.256 & 0.901 \end{pmatrix}, Y = \begin{pmatrix} 0.9335 & 0.221 & 1.285 \\ 0.0733 & 0.050 & 1.313 \\ 3.3637 & 0.215 & 1.064 \\ 3.3490 & 0.213 & 0.385 \\ 3.2397 & 0.176 & 0.577 \end{pmatrix}.$$

Предполагается, что данные выборки взяты из генеральных совокупностей, имеющих трехмерный нормальный закон распределения с неизвестными, но равными ковариационными матрицами.

Данные об исследуемых предприятиях «D» составляют матрицу новых наблюдений:

$$Z = \begin{pmatrix} 10.0796 & 0.136 & 1.073 \\ 2.0616 & 0.150 & 1.382 \\ 5.8125 & 0.234 & 1.828 \end{pmatrix}.$$

Целью дискриминантного анализа является отнесение строк матрицы Z либо к X, либо к Y.

Для построения дискриминантной функции найдем оценки математических ожиданий и ковариационной матрицы.

1. Найдем векторы средних:

$$\bar{X} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \bar{x}_3 \end{pmatrix}, \bar{Y} = \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \bar{y}_3 \end{pmatrix},$$

где $\bar{x}_j = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{ij}$, $\bar{y}_j = \frac{1}{n_2}\sum_{i=1}^{n_2} y_{ij}$. Например,

$$\bar{x}_1 = \frac{1}{5}(1.7774 + 9.0798 + 1.0634 + 11.5849 + 0.6805).$$

Таким образом,

$$\bar{X} = \begin{pmatrix} 4.8372 \\ 0.2304 \\ 1.4436 \end{pmatrix}, \bar{Y} = \begin{pmatrix} 2.19184 \\ 0.175 \\ 0.7248 \end{pmatrix}.$$

2. Определим оценки ковариационных матриц

$$S_X = \{s_{ij}\}_X = \frac{1}{n_1}\tilde{X}_i\tilde{X}_j,$$

$$S_Y = \{s_{ij}\}_Y = \frac{1}{n_2}\tilde{Y}_i\tilde{Y}_j,$$

где $\tilde{X}_j = \{x_{ij} - \bar{x}_j\}$ – столбцы матрицы X, $\tilde{Y}_j = \{y_{ij} - \bar{y}_j\}$ – столбцы матрицы Y. Например,

$$\{s_{12}\}_X = \frac{1}{5}[(1.7774 - 4.8372)(0.289 - 0.2304) + (9.0798 - 4.8372)(0.124 - 0.2304) + (1.0634 - 4.8372)(0.199 - 0.2304) + (11.5849 - 4.8372)(0.284 - 0.2304) + (0.6805 - 4.8372)(1.4436 - 0.2304)].$$

Таким образом,

$$S_X = \begin{pmatrix} 20.8826 & -0.0514 & 0.2968 \\ -0.0514 & 0.0039 & -0.0080 \\ 0.2968 & -0.0080 & 0.2318 \end{pmatrix},$$

$$S_Y = \begin{pmatrix} 0.9249 & -0.0005 & -0.2732 \\ -0.0005 & 0.0008 & -0.0023 \\ -0.2732 & -0.0023 & 0.1539 \end{pmatrix}.$$

3. Получим несмещенную оценку суммарной ковариационной матрицы:

$$\hat{S} = \frac{1}{n_1 + n_2 - 2}[n_1 S_X + n_2 S_Y] = \begin{pmatrix} 13.6292 & -0.0324 & 0.0147 \\ -0.0324 & 0.0029 & -0.0064 \\ 0.0147 & -0.0064 & 0.2410 \end{pmatrix}.$$

4. Найдем обратную матрицу $\hat{S}^{-1} = \frac{A_{ij}}{det(S)}$ к матрице $\hat{S}$.

Например, алгебраическое дополнение

$$A_{12} = (-1)^{1+2}\begin{vmatrix} -0.0324 & 0.0147 \\ -0.0064 & 0.2410 \end{vmatrix} = 0.0077,$$

а определитель $\det(S) = 0.0087$. Таким образом, получаем обратную матрицу

$$\hat{S}^{-1} = \begin{pmatrix} 0.0755 & 0.8924 & 0.0192 \\ 0.8924 & 379.604 & 10.0934 \\ 0.0192 & 10.0934 & 4.4172 \end{pmatrix}.$$

5. Найдем вектор оценок коэффициентов дискриминации:

$$a = \hat{S}^{-1}(\bar{X} - \bar{Y}) = \begin{pmatrix} 0.2110 \\ 20.5410 \\ 2.6312 \end{pmatrix}.$$

6. Вычислим оценки дискриминантной функции:

$$\widehat{U}_X = Xa = \begin{pmatrix} 9.0820 \\ 8.6800 \\ 10.2374 \\ 12.6586 \\ 7.7728 \end{pmatrix}, \widehat{U}_Y = Ya = \begin{pmatrix} 8.1176 \\ 7.1524 \\ 7.9256 \\ 6.0948 \\ 5.8169 \end{pmatrix}.$$

7. Определим средние значения оценок дискриминантной

функции:

$$\bar{u}_X = \frac{1}{n_1}\sum_{i=1}^{n_1} u_{X_i} = 9.5516, \bar{u}_Y = \frac{1}{n_2}\sum_{i=1}^{n_2} u_{Y_i} = = 7.0215.$$

8. Найдем граничную константу

$$\tilde{C} = \frac{1}{2}(\bar{u}_X + \bar{u}_Y) = 8.2865.$$

9. Найдем средние значения оценок дискриминантной функции

для предприятий «D»:

$$\bar{u}_1 = Z_1 a = 7.7433,$$
$$\bar{u}_2 = Z_2 a = 1.8661,$$
$$\bar{u}_3 = Z_3 a = 10.8427,$$

где Z_1, Z_2, Z_3 – строки матрицы Z. Сравнивая полученные значения с константой $\tilde{C}$, делаем вывод, что к классу A можно отнести только завод D3 ($\bar{u}_3 > \tilde{C}$).

§3.20. Теория ошибок

Классификация ошибок

Теория ошибок используется при изучении таких явлений, которые не являются вероятностно случайными, а вероятностным закономерностям подчиняются лишь методы исследования этих явлений.

Дело в том, что любое измерение физической величины–это сравнение ее с другой величиной того же рода, принятой за единицу, так что при выбранной системе единиц результаты измерений выражаются определенными числами. Из опыта известно, что при достаточно точных измерениях одной и той же величины численные значения отдельных измерений отличаются друг от друга, и, следовательно, содержат ошибки.

Различают три основных вида ошибок:

1) систематические ошибки; 2)грубые ошибки;
3) случайные ошибки.

Систематические ошибки постоянно либо преувеличивают, либо преуменьшают результаты измерений и происходят от таких причин, как неправильная установка прибора, влияние окружающей среды и т. д.

Оценка систематических ошибок производится с помощью методов, выходящих за пределы математической статистики.

Грубые ошибки возникают в результате просчета, неправильного чтения показаний измерительного прибора и т. д.

Случайные ошибки происходят от случайных причин, действующих при каждом измерении непредвиденным образом то в сторону уменьшения, то в сторону увеличения результатов.

Теория ошибок занимается изучением лишь грубых и случайных ошибок. Теория ошибок решает следующие задачи:

1) нахождение законов распределения случайных ошибок;
2) определение оценок неизвестных величин по результатам измерений $\left(\overline{x}, S^2, S\right)$;
3) установление погрешностей таких оценок;

4) устранение грубых ошибок.

Распределение случайных ошибок

Пусть в результате n независимых измерений истинного значения a некоторой неизвестной величины получены значения x_i. Разности $\Delta_i = x_i - a$ называют истинными ошибками. Так как a неизвестно, то и истинные ошибки $\Delta_i = x_i - a$ тоже неизвестны.

В соответствии с классической вероятностной моделью выборочным значениям x_i сопоставляют случайные величины ξ_i, и, значит, истинным ошибкам сопоставляют случайные величины $\delta_i = \xi_i - a$, где ξ_i, δ_i-взаимно независимые случайные величины, имеющие одинаковое распределение, $M(\xi_i) = a$. Независимость измерений трактуется как взаимная независимость случайных величин $\delta_i = \xi_i - a$. При этом общее математическое ожидание

$$M(\delta_1) = M(\delta_2) = ... = M(\delta_n) \equiv b$$

называют систематической ошибкой, а разности $Z_i = \delta_i - b$-случайными ошибками. Опыт показывает, что случайные ошибки часто распределены по нормальному закону

$$f_{Z_i}(z) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(z-a)^2}{2\sigma^2}\right).$$

Причины этого вскрыты центральной теоремой Ляпунова.

Оценки величин измерений

В качестве оценки истинного значения a величины X берут выборочное среднее

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i ,$$

а разности $\widetilde{\Delta} = x_i - \overline{x}$ называют кажущимися ошибками (они уже известны). Выбор $\overline{x}$ в качестве оценки для a основан на теореме Чебышева, согласно которой

$$\overline{x} \underset{n\to\infty}{\overset{вер}{\to}} M(\overline{\xi}) = M(\xi),$$

где математическое ожидание $M(\xi)$ трактуется как истинное значение a величины X .Так как

$$M(\overline{\xi}) = \frac{1}{n} M\left(\sum_{i=1}^{n} \xi_i \right) = a ,$$

то $M(\overline{\xi} - a) = 0$, т.е. оценка $\overline{x}$ лишена систематической ошибки.

Известно, что

$$D(\overline{\xi}) = \frac{\sigma^2}{n},$$

где σ^2-дисперсия отдельного измерения. Если дисперсия σ^2 отдельных измерений заранее неизвестна, то для ее оценки пользуются величиной

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2 .$$

Погрешности точечных оценок $\overline{x}, S$

Если случайные ошибки Z_i имеют нормальное распределение, то отношение

$$T = \frac{\overline{\xi} - a}{S} \sqrt{n}$$

подчиняется распределению Стьюдента с $k = n - 1$ степенями свободы. Это используется для оценки погрешности приближенного равенства $a \approx \overline{x}$, которая равна

$$|\overline{x} - a| = t \frac{S}{\sqrt{n}},$$

где коэффициент Стьюдента t находится по соответствующей таблице.

Величина

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$

при тех же предположениях имеет распределение «хи-квадрат» с $k = n - 1$ степенями свободы. Это позволяет оценить погрешность приближенного

равенства $\sigma \approx S$. Относительная погрешность этого приближенного соотношения равна

$$\frac{|S-\sigma|}{S} = q,$$

где число q находится по соответствующей таблице.

Четвертая задача устранения грубых ошибок решается, например, по критерию Томпсона.