Поиск ошибки второго рода и мощности критерия в однофакторном дисперсионном анализе (ANOVA)

Егор Личак

March 3, 2023

Определение 1: Пусть $X_1, X_2, ..., X_n$- независимые и одинаково распределенные случайные величины, причем $X_k \sim N(\mu_k, 1)$. Тогда случайная величина

$$\sum_{k=1}^{n} X_k^2 = \chi'^2(n, \lambda)$$

называется нецентральным распределением хи-квадрат, где n - число степеней свободы, а $\lambda = \sum_{k=1}^{n} \mu_k$ - параметр нецентральности.

Определение 2: Если $V_1 \sim \chi'^2(n, \lambda)$ - нецентральное распределение хи-квадрат с n степенями свободы и параметром нецентральности λ, $V_2 \sim \chi^2(m)$ - распределение хи-квадрат. Тогда случайная величина

$$\frac{V1/n}{V2/m} \sim \mathbb{F}'(n, m, \lambda)$$

называется нецентральным распределением Фишера с n, m степенями свободы и параметром нецнтральности λ.

Постановка задачи: Исходные данные состоят из $\sum_{j=1}^{k} n_j$ наблюдений x_{ij} по n наблюдений в j-ой выборке.

Ряды наблюдений (Treatments)

1	2	...	j	...	k
x_{11}	x_{12}	...	x_{1j}	...	x_{1k}
x_{21}	x_{22}	...	x_{2j}	...	x_{2k}
.	.	.	.	.	.
.	.	.	x_{ij}	.	.
.	.	.	.	.	.
...	...	...	x_{n_jj}	...	x_{n_jk}
.	.	.	.	.	.
.	.	.	x_{ij}	.	.
.	.	.	.	.	.
...	...	...	...	...	x_{n_kk}

Здесь x_{ij}- это i-ое наблюдение в j-ой выборке (j-ом ряду наблюдений).
Элементы x_{ij} можно считать реализацией случайных величин X_{ij}. Однафакторная модель предполагает, что случайные величины X_{ij} представимы в виде

$$X_{ij} = \mu_j + \varepsilon_{ij}, i = \overline{1, n_j}, j = \overline{1, k}$$

Здесь μ_j- неизвестный средний уровень фактора для j-ого ряда наблюдений, ε_{ij}- случайные ошибки.

Случаные выборки

$\overrightarrow{X_1}$	$\overrightarrow{X_2}$	...	$\overrightarrow{X_j}$	...	$\overrightarrow{X_k}$
1	2	...	j	...	k
X_{11}	X_{12}	...	X_{1j}	...	X_{1k}
X_{21}	X_{22}	...	X_{2j}	...	X_{2k}
.	.	.	.	.	.
.	.	.	X_{ij}	.	.
.	.	.	.	.	.
...	...	...	X_{n_jj}	...	X_{n_jk}
.	.	.	.	.	.
.	.	.	X_{ij}	.	.
.	.	.	.	.	.
...	...	...	...	...	X_{n_kk}

Здесь $\overline{X_j} = (X_{1j}, X_{2j}, ..., X_{n_jj})$- j-ая выборка объема n_j и таких выборок k штук.
Предположения:
п.1) Все случаные ошибки ε_{ij} независимы
п.2) Все ε_{ij} имеют одинаковое непрерывное (неизвестное) распределение.
Гипотеза однородности:
$H_0 : \mu_1 = \mu_2 = ... = \mu_k = \mu$
То есть гипотеза говорит об отсутствии различия в рядах наблюдений, то есть предполагается, что все ряды наблюдений (как и сами наблюдения) можно считать

одной выборкой из общей совокупности.
$H_1 : \exists i, j : \mu_i \neq \mu_j$
Определение 3: Статистика

$$SSE = n \cdot \overline{\sigma^2} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (X_{ij} - \overline{X_j})^2$$

называется внутригрупповой суммой квадратов или суммой квадратов отклонений внутри группы. Error Sum of Squares
Лемма 1: Вне зависимости от верности гипотез H_0 или H_1 случайная величина $\frac{SSE}{\sigma^2} \sim \chi^2(n-k)$.
Доказательство:
$\frac{SSE}{\sigma^2} = \frac{1}{\sigma^2} \sum_{j=1}^{k} (n_j - 1) \cdot S_j^2 = \sum_{j=1}^{k} \frac{(n_j-1) \cdot S_j^2}{\sigma^2}$, где $S_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (X_{ij} - \overline{X_j})^2$- исправленная выборочная дисперсия в j-ой выборке. Значимость или незначимость попарных разностей средних не влияет на эти статистики. Тогда, по следствию из теоремы Фишера, $\frac{(n_j-1) \cdot S_j^2}{\sigma^2} \sim \chi^2(n_j - 1)$.
Тогда $\sum_{j=1}^{k} \chi^2(n_j - 1) = \chi^2(\sum_{j=1}^{k} (n_j - 1)) = \chi^2(n-k)$ ч.т.д.
Определение 4:Статистика

$$SSTR = n\delta^2 = \sum_{j=1}^{k} (\overline{X_j} - \overline{X})^2 \cdot n_j$$

называется межгрупповой суммой квадратов или суммой квадратов между группами. Treatment Sum of Squares.
Лемма 2: Вне зависимости от верности гипотез H_0 или H_1 статистика $\frac{SSTR}{\sigma^2} \sim \chi'^2(l, \lambda)$- нецентральное распределение хи-квадрат с l степенями свободы и параметром нецентральности λ.
Доказательство: $\frac{SSTR}{\sigma^2} = \frac{1}{\sigma^2} \sum_{j=1}^{k} (\overline{X_j} - \overline{X})^2 \cdot n_j = \sum_{j=1}^{k} (\frac{\sqrt{n_j} \cdot (\overline{X_j} - \overline{X})}{\sigma})^2 = \sum_{j=1}^{k} Z_k^2$, где $Z_k = \frac{\sqrt{n_j} \cdot (\overline{X_j} - \overline{X})}{\sigma}$
$\overline{X_j} = \frac{1}{n_j} \sum_{i=1}^{n_j} X_{ij}$. X_{ij}- нормальные случайные величины, следовательно $\overline{X_j}$ имеет нормальное распределение. $\overline{X} = \frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n_j} X_{ij}$. Аналогично данная случайная величина распределена нормально. Разность случайных величин есть случайная величина, отсюда Z_k- нормальные случайные величины. Деление на σ и умножение на $\sqrt{n_j}$ означают, что параметр масштаба равен 1. Отсюда следует, что $\frac{SSTR}{\sigma}$ имеет нецентральное распределение хи-квадрат с неизвестными пока что параметрами l, λ. ч.т.д.
Определение 5: (One-Way Analysis of Variance F-Tests using Effect Size) Взвешенным средним назовем выражение, получаемое по следующей формуле.

$$\mu_w = \frac{1}{n}\sum_{j}^{k} n_j \cdot \mu_j$$

Утверждение 1: $E(\overline{X}) = \mu_w$

Доказательство: $E(\overline{X}) = E(\frac{1}{n}\sum_{j=1}^{k} n_j\overline{X_j}) = \frac{1}{n}(\sum_{j=1}^{k} n_j E(\overline{X_j})) = \frac{1}{n}\sum_{j=1}^{k} n_j \cdot \mu_j$ ч.т.д.

Лемма 3: $SSTR = \sum_{j=1}^{k}(\overline{X_j} - \mu_w)^2 \cdot n_j - n \cdot (\overline{X} - \mu_w)^2$

Доказательство: $SSTR = \sum_{j=1}^{k}(\overline{X_j} - \overline{X})^2 \cdot n_j = \sum_{j=1}^{k}[(\overline{X_j} - \mu_w) - (\overline{X} - \mu_w)]^2 \cdot n_j =$
$\sum_{j=1}^{k}(\overline{X_j}-\mu_w)^2\cdot n_j - 2\sum_{j=1}^{k}(\overline{X_j}-\mu_w)(\overline{X}-\mu_w)\cdot n_j + \sum_{j=1}^{k}(\overline{X}-\mu_w)^2\cdot n_j = \sum_{j=1}^{k}(\overline{X_j}-\mu_w)^2\cdot n_j - 2\cdot$
$(\overline{X}-\mu_w)\sum_{j=1}^{k}(\overline{X_j}-\mu_w)\cdot n_j + (\overline{X}-\mu_w)^2\sum_{j=1}^{k} n_j = \sum_{j=1}^{k}(\overline{X_j}-\mu_w)^2\cdot n_j - 2\cdot(\overline{X}-\mu_w)[\sum_{j=1}^{k}\overline{X_j}\cdot n_j -$
$\sum_{j=1}^{k}\mu_w\cdot n_j] + (\overline{X}-\mu_w)^2\cdot n = \sum_{j=1}^{k}(\overline{X_j}-\mu_w)^2\cdot n_j - 2\cdot(\overline{X}-\mu_w)(\overline{X}-\mu_w)\cdot n + (\overline{X}-\mu_w)^2\cdot n =$
$\sum_{j=1}^{k}(\overline{X_j}-\mu_w)^2\cdot n_j - 2\cdot(\overline{X}-\mu_w)^2\cdot n + (\overline{X}-\mu_w)^2\cdot n = \sum_{j=1}^{k}(\overline{X_j}-\mu_w)^2\cdot n_j - n\cdot(\overline{X}-\mu_w)^2.$

Лемма 4: $E(\frac{SSTR}{\sigma^2}) = (k-1) + \frac{\sum_{j=1}^{k}(\mu_j-\mu_w)^2 n_j}{\sigma^2}$ ч.т.д.
Доказательство: $E(\frac{SSTR}{\sigma^2}) = \frac{1}{\sigma^2}E(SSTR)$. Найдем $E(SSTR)$
$E(SSTR) = E[\sum_{j=1}^{k}(\overline{X_j}-\mu_w)^2\cdot n_j - n\cdot(\overline{X}-\mu_w)^2] = E(\sum_{j=1}^{k}(\overline{X_j}-\mu_w)^2\cdot n_j) - n\cdot E(\overline{X}-\mu_w)^2.$
С учетом утверждения 1 последнее слагамое- дисперсия, которая равна σ^2 для всех выборок. Отсюда
$E(SSTR) = E(\sum_{j=1}^{k}(\overline{X_j}-\mu_w)^2\cdot n_j) - n\cdot\frac{\sigma^2}{n} = \sum_{j=1}^{k}[Var(X_j-\mu_w) + (E(X_j-\mu_w))^2]n_j - \sigma^2 =$
$\sum_{j=1}^{k}[Var(\overline{X_j}) + (E(\overline{X_j}) - \mu_w)^2]n_j - \sigma^2 = \sum_{j=1}^{k}[\frac{\sigma^2}{n_j} + (\mu_j - \mu_w)^2]n_j - \sigma^2 = \sum_{j=1}^{k}[\sigma^2 + (\mu_j -$
$\mu_w)^2\cdot n_j] - \sigma^2 = k\cdot\sigma^2 - \sigma^2 + \sum_{j=1}^{k}(\mu_j-\mu_w)^2\cdot n_j = (k-1)\sigma^2 + \sum_{j=1}^{k}(\mu_j-\mu_w)^2\cdot n_j$. Тогда
$E(\frac{SSTR}{\sigma^2}) = (k-1) + \frac{\sum_{j=1}^{k}(\mu_j-\mu_w)^2\cdot n_j}{\sigma^2}$ ч.т.д.
Утверждение 2(Wikipedia): $E(\chi'^2(k,\lambda)) = k + \lambda$

Теорема 1: Если верна гипотеза H_1, то $\frac{SSTR}{\sigma^2} \sim \chi'^2(k-1,\lambda)$, где $\lambda = \frac{\sum_{j=1}^{k}(\mu_j-\mu_w)^2\cdot n_j}{\sigma^2}$.
Доказательство: По лемме 2 известно, что $\frac{SSTR}{\sigma^2}$ в общем случае имеет нецентральное хи-квадрат распределение $\chi'^2(l,\lambda)$. Тогда матожидание его равно $l+\lambda$. С учетом

леммы 4, получаем уравнение $l + \lambda = k - 1 + \frac{\sum_{j=1}^{k}(\mu_j - \mu_w)^2 \cdot n_j}{\sigma^2}(*)$. Известно, что если верна нулевая гипотеза, то число степеней свободы, не завсиящее от параметра нецентральности равно $k-1$. Отсюда вытекает, что $l = k - 1$. Из уравнения (*) тогда следует, что $\lambda = \frac{\sum_{j=1}^{k}(\mu_j - \mu_w)^2 \cdot n_j}{\sigma^2}$ ч.т.д.

Теорема 2: $\mathbb{F} = MSTR/MSE \sim \mathbb{F}'(k-1, n-k, \lambda)$- нецентральное распределение Фишера, где $\lambda = \frac{\sum_{j=1}^{k}(\mu_j - \mu_w)^2 \cdot n_j}{\sigma^2}$ - параметр нецентральности.

Доказательство: $\mathbb{F} = MSTR/MSE = \frac{SSTR/(k-1)}{SSE/(n-k)} = \frac{\chi'^2(k-1,\lambda)/(k-1)}{\chi^2(n-k)/(n-k)} = \mathbb{F}'(k-1, n-k, \lambda)$- по определению. ч.т.д.

Утверждение: Ошибка второго рода в однофакторном дисперсионном анализе равна $\beta(\overrightarrow{\mu}, \sigma) = F(f_\alpha(k-1, n-k))$,
где $F(\cdot)$- функция распределения нецентрального распределения Фишера с выведенымми выше параметрами, $f_\alpha(k-1, n-k)$- процентная точка центрального распределения Фишера с $k-1$ и $n-k$ степенями свободы. Мощность критерия равна
$W(\overrightarrow{\mu}, \sigma) = 1 - F(f_\alpha(k-1, n-k))$.

Доказательство: Критическая область правосторонняя и имеет вид: $K_\alpha = \{x_{ij} : \mathbb{F} > f_\alpha(k-1, n-k)\}$. Вероятность ошибки второго рода- вероятность непопадания значения статистики критерия в критическую область при условии верности гипотезы H_1. Таким образом, $\beta = \mathbb{P}(\mathbb{F} \leq f_\alpha(k-1, n-k)) = F(f_\alpha(k-1, n-k))$- по определению функции распределения.
$W = 1 - \beta = 1 - F(f_\alpha(k-1, n-k))$ ч.т.д.