Дисперсионный анализ 2

Н.П. Алексеева,

СПбГУ, математико-механический факультет

2020 г.

Многомерный дисперсионный анализ

Векторы $Y_i = (y_{i1}, \ldots, y_{in})^T$, $i = 1, \ldots, p$, соответствуют n независимым измерениям p переменных с одномерными линейными моделями

$$Y_i = X\beta_i + e_i, \quad \operatorname{cov}(Y_i) = \mathbb{E}(Y_i - \mathbb{E}Y_i)(Y_j - \mathbb{E}Y_j)^T = \sigma_{ii}I, \tag{1}$$

где X – матрица плана ранга $r \le m < n$, $p \le n - r$, $r < n$.
$\beta_i = (\beta_{i1}, \ldots, \beta_{im})^T$ – вектор параметров, специфичных для каждой переменной. Взятые вместе p линейных моделей (1) составляют многомерную обобщенную линейную модель: $Y = X\beta + e$, где

$$Y = \begin{bmatrix} y_{11} & y_{21} & \ldots & y_{p1} \\ y_{12} & y_{22} & \ldots & y_{p2} \\ \ldots & \ldots & \ldots & \ldots \\ y_{1n} & y_{2n} & \ldots & y_{pn} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_{11} & \beta_{21} & \ldots & \beta_{p1} \\ \beta_{12} & \beta_{22} & \ldots & \beta_{p2} \\ \ldots & \ldots & \ldots & \ldots \\ \beta_{1m} & \beta_{2m} & \ldots & \beta_{pm} \end{bmatrix} \tag{2}$$

e – матрица n на p, строки которой составляют случайную выборку из $\mathcal{N}(0, \Sigma)$, $\Sigma = \{\sigma_{ij}\}$.

Оценкой параметров $\beta_i,\ i = 1, \ldots, p$, служит МНК-оценка по одной переменной Y_i вида

$$\hat{\beta}_i = (X^T X)^- (X^T Y_i), \quad i = 1, \ldots, p\,.$$

Для вычисления несмещенных оценок для σ_{ii} и σ_{ij} используются остаточная сумма квадратов

$$R_0^2(i,i) = Y_i^T Y_i - Y_i^T X \beta_i, \quad \hat{\sigma}_{ii} = \frac{R_0^2(i,i)}{n-r}$$

и остаточная сумма произведений

$$R_0^2(i,j) = Y_i^T Y_j - Y_i^T X \beta_j, \quad \hat{\sigma}_{ij} = \frac{R_0^2(i,j)}{n-r}\,.$$

$$\mathbb{R}_0 = \begin{bmatrix} R_0^2(1,1) & \ldots & R_0^2(1,p) \\ \ldots & \ldots & \ldots \\ R_0^2(p,1) & \ldots & R_0^2(p,p) \end{bmatrix}$$

называется матрицей остаточных сумм и произведений.

Требуется одновременно проверить гипотезы о том, что

$$\mathrm{H}^{\mathrm{T}}\beta_{\mathrm{i}} = \theta_{\mathrm{i}},\ \mathrm{i} = 1, \ldots, \mathrm{p}\,, \tag{3}$$

где матрица H^{T} размерности s на m имеет ранг $\mathrm{s} \leq \mathrm{r}$.
Если эта гипотеза выполняется, то получается модель с ограничениями, на основе которой можно найти оценки β_{i}^{*} и матрицу остаточных сумм и произведений $\mathbb{R}_1$.
Матрица $\mathbb{R}_1 - \mathrm{R}_0$ называется матрицей квадратов и произведений, обусловленных отклонением от гипотезы.

Таким образом, отклонение от гипотезы может быть определено сравнением матриц $\mathbb{R}_1 - \mathbb{R}_0$ и $\mathbb{R}_0$.

Сведем задачу $H^T\beta_i = \theta_i,\ i = 1, \ldots, p$, к одномерному случаю,

$$\beta = L_1\beta_1 + \ldots + L_p\beta_p, \quad \xi = L_1\xi_1 + \ldots + L_p\xi_p,$$

рассмотрев гипотезу $H^T\beta = \xi$ и вместо множества случайных величин одну $Y = L_1Y_1 + \ldots + L_pY_p$.

$$R_0^2 = \min(Y - X\beta)^T(Y - X\beta), \quad R_1^2 = \min_{H^T\beta=\xi}(Y - X\beta)^T(Y - X\beta).$$

При $H^T\beta = \xi$ статистики $R_1^2 - R_0^2$ и R_0^2 распределены независимо как хи-квадрат распределения $\sigma_L^2\chi^2(s)$ и $\sigma_L^2\chi^2(n - r)$.

$$L^T\mathbb{R}_0L = R_0^2, \quad L^T\mathbb{R}_1L = R_1^2, \quad L^T(R_1 - R_0)L = R_1^2 - R_0^2, \text{например,}$$

$$R_0^2 = (Y - X\beta)^T(Y - X\beta) = \left[\sum L_i(Y_i - X\beta_i)\right]^T \left[\sum L_i(Y_i - X\beta_i)\right] =$$

$$= \sum L_iL_j(Y_i - X\beta_i)^T(Y_i - X\beta_i) = \sum L_iL_jR_0^2(i, j) = L^T\mathbb{R}_0L\,.$$

При справедливости $H^T\beta_i = \theta_i,\ i = 1, \ldots, p$, для каждого L верна гипотеза $H^T\beta = \xi$ и подходящими являются статистики

$$F = \frac{n - r}{s} \cdot \frac{R_1^2 - R_0^2}{R_0^2} \quad \text{или} \quad B = \frac{R_0^2}{R_1^2} = \frac{1}{1 + \frac{sF}{n-r}} = \frac{L^T\ \mathbb{R}_0\ L}{L^T\ \mathbb{R}_1\ L},$$

последняя имеет бета-распределение. Чем больше F и значимость, тем меньше B.

Выберем L так, чтобы статистика B имела наименьшее значение (F наибольшее), тем самым пытаясь опровергнуть гипотезу. Тогда критерий сводится к виду:

$$\lambda = \min_{\mathrm{L}} \mathrm{B} = \min_{\mathrm{L}} \frac{\mathrm{L}^{\mathrm{T}}\ \mathbb{R}_0\ \mathrm{L}}{\mathrm{L}^{\mathrm{T}}\ \mathbb{R}_1\ \mathrm{L}},$$

или λ является наименьшим корнем характеристического уравнения

$$|\mathbb{R}_0 - \lambda\mathbb{R}_1| = 0\,, \quad \text{так как} \quad \mathbb{R}_0\mathrm{L} = \lambda\mathbb{R}_1\mathrm{L} \iff \mathrm{L}^{\mathrm{T}}\mathbb{R}_0\mathrm{L} = \lambda\mathrm{L}^{\mathrm{T}}\mathbb{R}_1\mathrm{L} \iff \lambda = \frac{\mathrm{L}^{\mathrm{T}}\ \mathbb{R}_0\ \mathrm{L}}{\mathrm{L}^{\mathrm{T}}\ \mathbb{R}_1\ \mathrm{L}}\,.$$

Один из возможных критериев на собственных числах - Λ-критерий Уилкса:

$$\Lambda = \lambda_1 \cdot \ldots \cdot \lambda_{\mathrm{p}} = \frac{|\mathbb{R}_0|}{|\mathbb{R}_1|}\,.$$

Двуфакторный дисперсионный анализ с повторениями

Исследуется уровень адреналина у мышей, содержащихся в разных условиях до периода полового созревания, и выкормленных по-разному. Таким образом, имеются два фактора:

- A - условия вскармливания (без матери и с матерью),
- B - условия содержания (в изолированных, смежных и общих клетках).

Пусть x_{ijk} обозначает значение переменной X (уровень адреналина), полученное при k-м повторении эксперимента в ячейке ij, $i = 1, \ldots, I$, $j = 1, \ldots, J$, $k = 1, \ldots, K$.

Двуфакторная модель с фиксированными эффектами

Модель с фиксированными эффектами имеет вид:

$$x_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}, \tag{4}$$

$i = 1, \ldots, I,\ j = 1, \ldots, J,\ k = 1, \ldots, K,$

- μ – генеральное среднее,
- α_i - дифференциальный эффект фактора A,
- β_j – дифференциальный эффект фактора B.
- γ_{ij} – взаимодействие факторов учитывает дифференциальный эффект комбинаций i-го уровня фактора A и j-го уровня фактора B, если он не выражается суммой $\alpha_i + \beta_j + \mu$.
- ошибки e_{ijk} предполагаются независимыми и нормально распределенными $\mathcal{N}(0, \sigma)$.

При ограничениях на параметры

$$\sum_{i=1}^{I} \alpha_i = 0, \quad \sum_{j=1}^{J} \beta_j = 0,$$

$$\sum_{i=1}^{I} \gamma_{ij} = 0, \quad j = 1, 2, \ldots, J;$$

$$\sum_{j=1}^{J} \gamma_{ij} = 0, \quad i = 1, 2, \ldots, I,$$

оценками параметров модели $x_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$ являются
$\hat{\mu} = \bar{x}$ – общее среднее,
$\hat{\alpha}_i = \bar{x}_{i\cdot} - \bar{x}$, где $\bar{x}_{i\cdot}$ – среднее по i-му значению фактора A;
$\hat{\beta}_j = \bar{x}_{\cdot j} - \bar{x}$, где $\bar{x}_{\cdot j}$ – среднее по j-му значению фактора B;
$\widehat{\gamma}_{ij} = \bar{x}_{ij\cdot} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{x}$, где $\bar{x}_{ij\cdot}$ – среднее в ячейке ij.
Заметим, что $\widehat{\gamma}_{ij} = (\bar{x}_{ij\cdot} - \bar{x}) - (\bar{x}_{i\cdot} - \bar{x}) - (\bar{x}_{\cdot j} - \bar{x}) = \bar{x}_{ij\cdot} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{x}$.

Уровень адреналина при разном вскармливании и содержании

вскармл. с матерью 3.35 ± 0.15			вскармл. без матери 4.48 ± 0.15		
отд. ξ_1	сообщ. ξ_2	общ. ξ_3	отд. ξ_1	сообщ. ξ_2	общ. ξ_3
1.9	4	3.2	3.3	6.3	4.6
2.3	4.6	2.6	4	7.2	4.8
2.2	5.7	2.2	5	4.6	4.6
2	5.7	2.6	3.2	7.2	4.4
2.7	4.8	3.2	2.4	3.8	4.5
2.8	4.8	2.5	3.6	4.4	4.2
2.4	5.4	3	3	4.8	4.4
2.7	3.8	3.3	3	5.8	4.3
$\bar{x}_{11} = 2.38$	$\bar{x}_{12} = 4.85$	$\bar{x}_{13} = 2.83$	$\bar{x}_{21} = 3.44$	$\bar{x}_{22} = 5.51$	$\bar{x}_{23} = 4.48$

отд. ξ_1	сообщ. ξ_2	общ. ξ_3
$\bar{x}_{\cdot 1} = 2.91$	$\bar{x}_{\cdot 2} = 5.18$	$\bar{x}_{\cdot 3} = 3.65$

Имеем вектор Y размерности IJK, матрицу плана размерности IJK на $r = 1 + (I - 1) + (J - 1) + (I - 1)(J - 1) = IJ$. Например, при $I = 3$ и $J = 4$ она имеет вид:

Y	μ	α_1	α_2	β_1	β_2	β_3	γ_{11}	γ_{12}	γ_{13}	γ_{21}	γ_{22}	γ_{23}
y_{11*}	1	1	0	1	0	0	1	0	0	0	0	0
y_{12*}	1	1	0	0	1	0	0	1	0	0	0	0
y_{13*}	1	1	0	0	0	1	0	0	1	0	0	0
y_{14*}	1	1	0	−1	−1	−1	−1	−1	−1	0	0	0
y_{21*}	1	0	1	1	0	0	0	0	0	1	0	0
y_{22*}	1	0	1	0	1	0	0	0	0	0	1	0
y_{23*}	1	0	1	0	0	1	0	0	0	0	0	1
y_{24*}	1	0	1	−1	−1	−1	0	0	0	−1	−1	−1
y_{31*}	1	−1	−1	1	0	0	−1	0	0	−1	0	0
y_{32*}	1	−1	−1	0	1	0	0	−1	0	0	−1	0
y_{33*}	1	−1	−1	0	0	1	0	0	−1	0	0	−1
y_{34*}	1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1

Вектор параметров $\beta = (\mu, \alpha_1, \alpha_{I-1}, \beta_1, \ldots, \beta_{J-1}, \gamma_{11}, \ldots, \gamma_{I-1,J-1})^T$.

Общее число наблюдений равно IJK.

Остаточная сумма квадратов основной модели равна

$$Q_R = R_0^2 = \sum_{i,j,k} (x_{ijk} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j - \hat{\gamma}_{ij})^2$$

имеет число степеней свободы $IJK - IJ$. Для проверки гипотезы $H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_I = 0$ используем $H^T\beta = 0$, матрицу H^T размерности $I - 1$ на $r = IJ$ ранга $I - 1$, вида $H^T = \left[0_{I-1,1}|I_{I-1,I-1}|0_{I-1,IJ-I)}\right]$. Остаточная сумма квадратов усеченной модели равна

$$R_1^2 = \sum_{i,j,k} (x_{ijk} - \hat{\mu} - \hat{\beta}_j - \hat{\gamma}_{ij})^2 = R_0^2 + Q_A,$$

$$\text{где } Q_A = \sum_{i,j,k} (\hat{\alpha}_i)^2 = JK \sum_{i=1}^{I} (\bar{x}_{i\cdot} - \bar{x})^2$$

Отсюда получаем статистику критерия

$$F = \frac{IJK - IJ}{I - 1} \cdot \frac{R_1^2 - R_0^2}{R_0^2} = \frac{IJK - IJ}{I - 1} \cdot \frac{Q_A}{Q_R} \sim F(I - 1, IJK - IJ).$$

Аналогично строятся остальные статистики.

Таблица двуфакторного дисперсионного анализа

источник дисперсии	сумма квадратов	степени свободы	средний квадрат
фактор A	$Q_A = JK\sum_{i=1}^{I}(\bar{x}_{i.} - \bar{x})^2$	$\nu_A = I - 1$	$MQ_A = \frac{Q_A}{\nu_A}$
фактор B	$Q_B = IK\sum_{j=1}^{J}(\bar{x}_{.j} - \bar{x})^2$	$\nu_B = J - 1$	$MQ_B = \frac{Q_B}{\nu_B}$
взаимо-действие AB	$Q_{AB} = K\sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{x}_{ij.} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x})^2$	$\nu_{AB} = (I-1)(J-1)$	$MQ_{AB} = \frac{Q_{AB}}{\nu_{AB}}$
остаток (ошибка)	$Q_R = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(x_{ijk} - \bar{x}_{ij.})^2$	$\nu_R = IJ(K-1)$	$MQ_R = \frac{Q_R}{\nu_R}$
полная	$Q_T = \sum_{i=1}^{I}\sum_{j=1}^{J}s\sum_{k=1}^{K}(x_{ijk} - \bar{x})^2$	$\nu_T = IJK - 1$	

- Для проверки гипотезы об отсутствии эффекта взаимодействия

  H_0 : все $\gamma_{ij} = 0$ используется статистика $F = \dfrac{MQ_{AB}}{MQ_R} \sim F(\nu_{AB}, \nu_R)$.

- Для проверки гипотезы об отсутствии эффекта A-фактора

  H_0 : все $\alpha_i = 0$ используется отношение $F = \dfrac{MQ_A}{MQ_R} \sim F(\nu_A, \nu_R)$.

- Для проверки гипотезы об отсутствии эффекта B-фактора

  H_0 : все $\beta_j = 0$ берется отношение $F = \dfrac{MQ_B}{MQ_R} \sim F(\nu_B, \nu_R)$.

Если пренебречь фактором взаимодействия, или недостаточно наблюдений для достаточного представительства в каждой ячейке ij, тогда суммы квадратов Q_{AB} и Q_R могут складываться в новую сумму Q_R^T с соответственным сложением числа степеней свободы.

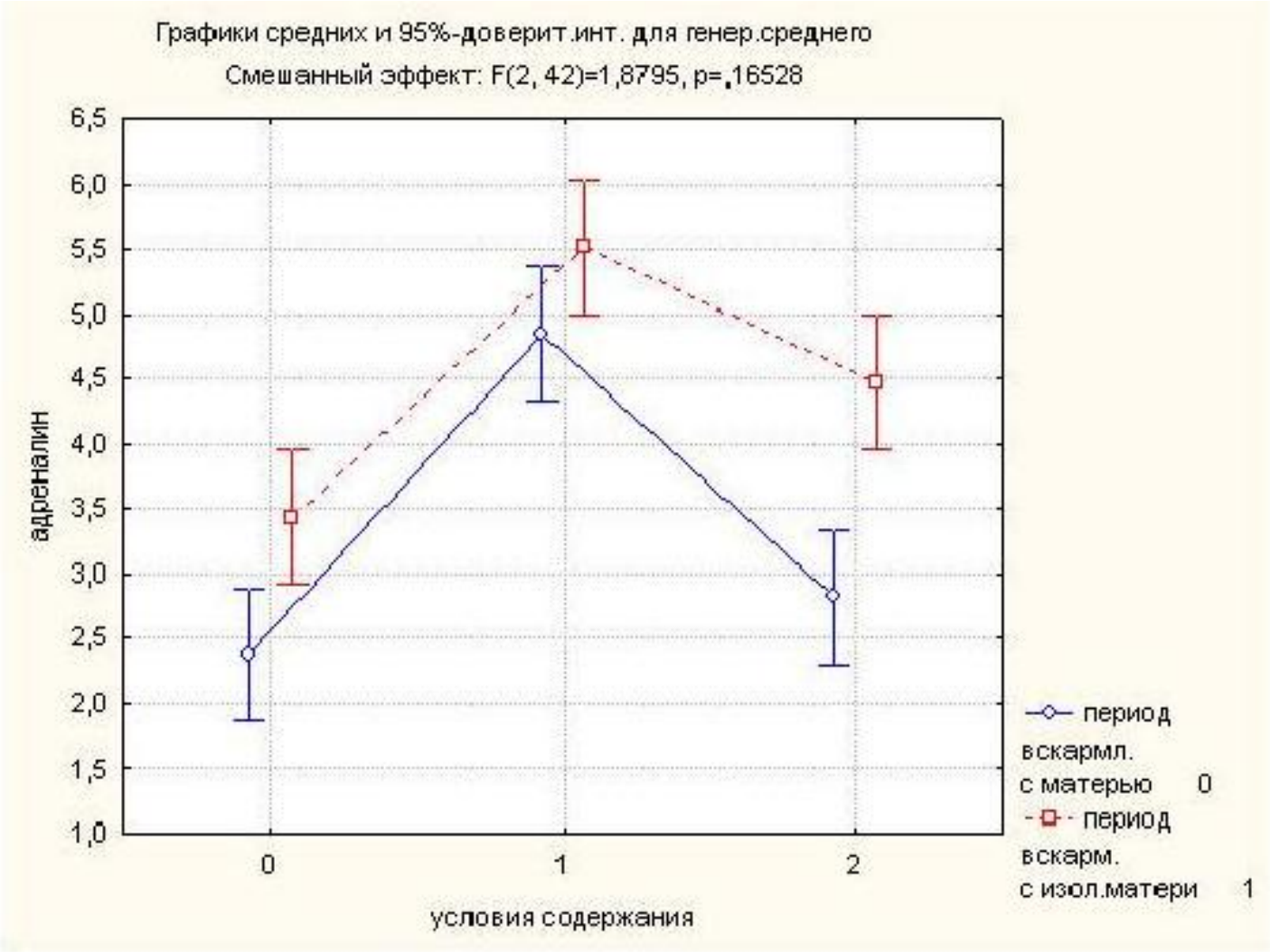


Figure: Иллюстрация отсутствия эффекта взаимодействия факторов вскармливания и содержания.

Проверка гипотез

- H_0 : все $\alpha_i = 0$. $p = 0.000003$ – следовательно, влияние фактора вскармливания значимо для уровня адреналина, средний уровень адреналина 3.35 ± 0.15 в группе, до конца вскармливания находящейся с матерью, значимо ниже среднего уровня адреналина 4.48 ± 0.15 в группе, изолированной от матери.
- H_0 : все $\beta_j = 0$. $p < 0.000001$ – следовательно, влияние фактора условия содержания значимо для уровня адреналина: средние 2.91 ± 0.18, 5.18 ± 0.18 и 3.65 ± 0.18 в группах, отличающихся условиями содержания, значимо различаются.
- H_0 : все $\gamma_{ij} = 0$. $p = 0.16$ – следовательно, фактор взаимодействия не значим.

Модель со случайными эффектами

$$x_{ij} = \mu + a_i + e_{ij}, \quad i = 1, \ldots, r, \ j = 1, \ldots, n_i,$$

- a_i распределены по $\mathcal{N}(0, \sigma_a)$,
- e_{ij} распределены по $\mathcal{N}(0, \sigma)$,
- все a_i и e_{ij} в совокупности независимы.

Гипотеза $H_0 : \sigma_a^2 = 0$ означает, что фактор не вносит значимого вклада в дисперсию.

Для проверки гипотезы $H_0 : \sigma_a^2 = 0$ используются выражения для математических ожиданий средних квадратов $\frac{Q_1}{r-1}$ и $\frac{Q_2}{n-r}$. Непосредственными вычислениями можно убедиться, что

$$E\left(\frac{Q_2}{n-r}\right) = \sigma^2, \quad E\left(\frac{Q_1}{r-1}\right) = \sigma^2 + k\sigma_a^2, \quad \text{где}$$

$$k = \frac{1}{r-1}\left(n - \frac{1}{n}\sum_{i=1}^{r} n_i^2\right). \tag{5}$$

Если используется одинаковое число повторений в каждой подпопуляции, то есть $n_1 = \ldots = n_r = m$, то $k = m$.

Оценка дисперсии σ_a^2

$$\widehat{\sigma}_a^2 = \frac{1}{k}\left(\frac{Q_1}{r-1} - \frac{Q_2}{n-r}\right).$$

Используем выражение $\bar{x}_i = \mu + a_i + \bar{\epsilon}_i$ для вычисления EQ_2:

$$Q_2 = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2 = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(\mu + a_i + \epsilon_{ij} - \mu - a_i - \bar{\epsilon}_i)^2 =$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{n_i}\epsilon_{ij}^2 - \sum_{i=1}^{r} n_i\bar{\epsilon}_i^2, \quad \sum_{i=1}^{r} n_i = n,$$

$$EQ_2 = \sum_{i=1}^{r}\sum_{j=1}^{n_i} E\epsilon_{ij}^2 - \sum_{i=1}^{r} n_i E\bar{\epsilon}_i^2 = n\sigma^2 - \sum_{i=1}^{r} n_i\frac{\sigma^2}{n_i} = (n-r)\sigma^2\,.$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_i} x_{ij} = \frac{1}{n}\sum_{i=1}^{r} n_i\bar{x}_i = \mu + \frac{1}{n}\sum_{i=1}^{r} n_i a_i + \bar{\epsilon},$$

$$Q_1 = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(\bar{x}_i - \bar{x})^2 = \sum_{i=1}^{r}\sum_{j=1}^{n_i}\left((a_i + \bar{\epsilon}_i) - \left(\frac{1}{n}\sum_{i=1}^{r} n_i a_i + \bar{\epsilon}\right)\right)^2 =$$

$$= \sum_{i=1}^{r} n_i(a_i + \bar{\epsilon}_i)^2 - n\left(\frac{1}{n}\sum_{i=1}^{r} n_i a_i + \bar{\epsilon}\right)^2;$$

$$EQ_1 = \sum_{i=1}^{r} n_i E(a_i + \bar{\epsilon}_i)^2 - nE\left(\frac{1}{n}\sum_{i=1}^{r} n_i a_i + \bar{\epsilon}\right)^2 =$$

$$= \sum_{i=1}^{r} n_i D(a_i + \bar{\epsilon}_i) - nD\left(\frac{1}{n}\sum_{i=1}^{r} n_i a_i + \bar{\epsilon}_i\right) =$$

$$= \sum_{i=1}^{r} n_i\left(\sigma_a^2 + \frac{\sigma^2}{n_i}\right) - n\left(\sigma_a^2\frac{1}{n^2}\sum_{i=1}^{r} n_i^2 + \frac{\sigma^2}{n}\right) =$$

$$= (r-1)\sigma^2 + \sigma_a^2\left(n - \frac{1}{n}\sum_{i=1}^{r} n_i^2\right) = (r-1)\left(\sigma^2 + k\sigma_a^2\right),$$

Модель с фиксированными эффектами. Проверка гипотезы

При справедливости гипотезы $H_0 : \sigma_a^2 = 0$ статистики $\frac{Q_2}{n-r}$ и $\frac{Q_1}{r-1}$ имеют одинаковые средние σ^2, поэтому можно построить отношение Фишера

$$F = \frac{n-r}{r-1} \cdot \frac{Q_1}{Q_2} \sim F(r-1, n-r).$$

Значимость критерия равна $\alpha_* = P\{F > F_*\}$, где F_* наблюдаемое значение статистики.

Двуфакторная модель со случайными эффектами

Модель со случайными эффектами имеет вид:

$$x_{ijk} = \mu + a_i + b_j + (ab)_{ij} + e_{ijk}, \tag{6}$$

где $i = 1, \ldots, I,\ j = 1, \ldots, J,\ k = 1, \ldots, K$,

- μ - генеральное среднее,
- случайные дифференциальные эффекты A-фактора a_i независимы и распределены по $\mathcal{N}(0; \sigma_a)$,
- случайные дифференциальные эффекты B-фактора b_j независимы и распределены по $\mathcal{N}(0; \sigma_b)$,
- эффекты взаимодействия $(ab)_{ij}$ независимы и распределены по $\mathcal{N}(0; \sigma_{ab})$.
- Величины e_{ijk} независимы и распределены по $\mathcal{N}(0; \sigma)$.
- Все величины $a_i, b_j, (ab)_{ij}, e_{ijk}$ независимы в совокупности.

Двухфакторная модель со случайными эффектами. Проверка гипотез

источник дисперсии	средний квадрат	математическое ожидание
A	$\mathrm{MQ_A}$	$\sigma^2 + \mathrm{K}\sigma^2_{\mathrm{ab}} + \mathrm{JK}\sigma^2_{\mathrm{a}}$
B	$\mathrm{MQ_B}$	$\sigma^2 + \mathrm{K}\sigma^2_{\mathrm{ab}} + \mathrm{IK}\sigma^2_{\mathrm{b}}$
AB	$\mathrm{MQ_{AB}}$	$\sigma^2 + \mathrm{K}\sigma^2_{\mathrm{ab}}$
R	$\mathrm{MQ_R}$	σ^2

Отсюда получаем выражения для оценок дисперсий:

$$\sigma^2_{\mathrm{a}} = \frac{\mathrm{MQ_A} - \mathrm{MQ_{AB}}}{\mathrm{JK}}, \quad \sigma^2_{\mathrm{b}} = \frac{\mathrm{MQ_B} - \mathrm{MQ_{AB}}}{\mathrm{IK}}, \quad \sigma^2_{\mathrm{ab}} = \frac{\mathrm{MQ_{AB}} - \mathrm{MQ_R}}{\mathrm{K}}.$$

Для проверки гипотезы $\mathrm{H}_0 : \sigma^2_{\mathrm{ab}} = 0$ используется отношение $\mathrm{F} = \frac{\mathrm{MQ_{AB}}}{\mathrm{MQ_R}}$, для $\mathrm{H}_0 : \sigma^2_{\mathrm{a}} = 0$ используется отношение $\mathrm{F} = \frac{\mathrm{MQ_A}}{\mathrm{MQ_{AB}}}$, для $\mathrm{H}_0 : \sigma^2_{\mathrm{b}} = 0$ отношение $\mathrm{F} = \frac{\mathrm{MQ_B}}{\mathrm{MQ_{AB}}}$ с соответствующими степенями свободы.

Пример двухфакторной модели со случайными эффектами

- При проверке гипотезы $H_0 : \sigma^2_{ab} = 0$ об отсутствии взаимодействия получаем тот же критерий, что и в случае модели с фиксированными эффектами. $p = 0.16$, взаимодействие незначимо.
- При проверке гипотезы $H_0 : \sigma^2_{a} = 0$ об отсутствии эффекта условий вскармливания получаем значимость $p = 0.059$.
- При проверке гипотезы $H_0 : \sigma^2_{b} = 0$ получаем значимость $p = 0.044$.

Значимости отличаются от значимостей в модели с фиксированными эффектами, сохраняется только соотношение между ними: фактор условия содержания более значим, чем фактор вскармливания.

Двуфакторная модель со смешанными эффектами

$$x_{ijk} = \mu + \alpha_i + b_j + (\alpha b)_{ij} + e_{ijk}, \tag{7}$$

где $i = 1, \ldots, I,\ j = 1, \ldots, J,\ k = 1, \ldots, K,$

- μ - генеральное среднее,
- α_i – i-й дифференциальный эффект A-фактора,
- b_j – случайные эффекты B-фактора независимы и нормально распределены $\mathcal{N}(0; \sigma_b)$, эффекты взаимодействия
- $(\alpha b)_{ij}$ независимы и нормально распределены $\mathcal{N}(0; \sigma_{ab})$.
- Величины e_{ijk} независимы распределены по $\mathcal{N}(0; \sigma)$.
- Все величины b_j, $(\alpha b)_{ij}$, e_{ijk} независимы в совокупности.

Двухфакторная модель со смешанными эффектами. Проверка гипотез

источник дисперсии	средний квадрат	математическое ожидание
A	MQ_A	$\sigma^2 + K\sigma^2_{ab} + \frac{JK\sum_{i=1}^{I}\alpha_i^2}{I-1}$
B	MQ_B	$\sigma^2 + K\sigma^2_{ab} + IK\sigma^2_b$
AB	MQ_{AB}	$\sigma^2 + K\sigma^2_{ab}$
R	MQ_R	σ^2

Для проверки гипотезы
$H_0 : \sigma^2_{ab} = 0$ используется отношение $F = \frac{MQ_{AB}}{MQ_R}$, для
$H_0 : \sigma^2_b = 0$ используется отношение $F = \frac{MQ_B}{MQ_{AB}}$, для
H_0 : все $\alpha_i = 0$ отношение $F = \frac{MQ_A}{MQ_{AB}}$ с соответствующими степенями свободы.

Заметим, что в случае двуфакторного плана результаты дисперсионного анализа в случае хотя бы одного случайного фактора совпадают.

Двуфакторная модель с группировкой

ПРИМЕР. Пусть имеется фактор диеты A с $r = 4$ уровнями. Зависимой переменной является дважды ($K = 2$) измеряемое количество выдыхаемого азота x_{ijk}, $i = 1, \ldots, I$; $j = 1, \ldots, J$, $k = 1, \ldots, K$. Фактор B является фактором испытуемого, $J = 4$.

Данные для двуфакторного плана с группировкой

	диета 1		диета 2		диета 3		диета 4	
B	1	2	1	2	1	2	1	2
1	4.079	4.859	4.368	5.668	4.169	5.709	4.928	5.608
2	3.541	5.047	3.752	5.848	4.416	5.666	4.941	5.291
3	3.298	4.679	3.802	4.844	4.123	5.059	4.674	5.038
4	2.871	4.648	3.578	5.393	4.403	4.496	4.905	5.208

Модель при случайном факторе A модель имеет вид:

$$x_{ijk} = \mu + a_i + b_{j(i)} + e_{ijk},$$

где μ - генеральное среднее, величины a_i независимы и распределены по $\mathcal{N}(0; \sigma_a)$, величины $b_{j(i)}$ - независимы и распределены по $\mathcal{N}(0; \sigma_{b(a)})$. Величины e_{ijk} независимы распределены по $\mathcal{N}(0; \sigma)$. Все величины $a_i, b_{j(i)}, e_{ijk}$ независимы в совокупности. $\hat{b}_{j(i)} = \bar{x}_{ij\cdot} - \bar{x}_{i\cdot\cdot}$.

Таблица двуфакторного плана с группировкой

источник дисперсии	сумма квадратов	степени свободы	средний квадрат
фактор A	$Q_A = JK\sum_{i=1}^{I}(\bar{x}_{i\cdot} - \bar{x})^2$	$\nu_A = I - 1$	$MQ_A = \frac{Q_A}{\nu_A}$
фактор B (внутриA)	$Q_{B(A)} = K\sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{x}_{ij\cdot} - \bar{x}_{i\cdot})^2$	$\nu_{B(A)} = I(J-1)$	$MQ_{B(A)} = \frac{Q_{B(A)}}{\nu_{B(A)}}$
остаток (ошибка)	$Q_R = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(x_{ijk} - \bar{x}_{ij\cdot})^2$	$\nu_R = IJ(K-1)$	$MQ_R = \frac{Q_R}{\nu_R}$
полная	$Q_T = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(x_{ijk} - \bar{x})^2$	$\nu_T = IJK - 1$	

Смешанный двуфакторный план с группировкой можно записать в виде:

$$x_{ijk} = \mu + \alpha_i + b_{j(i)} + e_{ijk},$$

α_i – дифференциальный эффект, определяемый i-м уровнем фактора A. Приведем ожидания средних квадратов для двуфакторного плана с группировкой в случае модели со случайными эффектами:

$$E\left(MQ_A\right) = \sigma^2 + K\sigma^2_{b(a)} + KJ\sigma^2_a,$$
$$E\left(MQ_{B(A)}\right) = \sigma^2 + K\sigma^2_{b(a)}, \quad E\left(MQ_R\right) = \sigma^2,$$

и в случае модели со смешанными эффектами:

$$E\left(MQ_A\right) = \sigma^2 + K\sigma^2_{b(a)} + \frac{JK\sum_{i=1}^{I}\alpha_i^2}{I-1},$$
$$E\left(MQ_{B(A)}\right) = \sigma^2 + K\sigma^2_{b(a)}, \quad E\left(MQ_R\right) = \sigma^2.$$

Для проверки гипотез используем отношения

$$H_0 : \sigma^2_{b(a)} = 0 \qquad F = \frac{MQ_{B(A)}}{MQ_R},$$

$$H_0 : \sigma^2_a = 0 \text{ илиH}_0 : \text{все } \alpha_i = 0 \quad F = \frac{MQ_A}{MQ_{B(A)}}.$$

Результаты дисперсионного анализа

источник дисперсии	сумма квадратов	число ст.св.	средний квадрат	F	p
Диета	3.711	3	1.237	8.14	0.0032
Объект(диета)	1.828	12	0.152	0.20	0.9960
Остаток	12.020	16	0.751		
Полная	17.559	31			

Дисперсионный анализ для зависимых выборок

Рассмотрим данные вида $x_{(i)jk}$, где $i = 1, \ldots, r$ уровни фактора A (пол) с фиксированными эффектами, $j = 1, \ldots, n$ – индивиды (фактор B), $k = 1, \ldots, t$ уровни временного фактора C со случайными эффектами. Поскольку фактор A является следствием фактора индивида, то его индекс при наличии индекса j заключаем в скобки, $n_1 + n_2 + \ldots + n_r = n$. Требуется выяснить влияние факторов A и C на переменную X (количество пройденных метров по ТШХ в три момента времени: при поступлении в стационар, при выписке и через три месяца). Построим модель:

$$x_{(i)jk} = \mu + \alpha_i + c_k + e^1_{(i)j} + (\alpha c)_{ik} + e_{(i)jk},$$

где $e^1_{(i)j}$ – ошибка, обусловленная влиянием индивида, $e_{(i)jk}$ – ошибка наблюдения.

Оценкой влияния фактора A является разность

$$\widehat{\alpha} = \bar{x}_{i..} - \bar{x}_{...},$$

где $x_{i..}$ есть среднее по каждому полу, $x_{...}$ – общее среднее. Общий источник вариации Q с числом степеней свободы, равным $\nu = nt - 1$, имеет вид:

$$Q = \sum_{j=1}^{n}\sum_{k=1}^{t}(x_{(i)jk} - \bar{x}_{...})^2 = Q_1 + Q_2\,,$$

$$Q_1 = t\sum_{j=1}^{n}(\bar{x}_{(i)j.} - \bar{x}_{...})^2, \quad Q_2 = \sum_{j=1}^{n}\sum_{k=1}^{t}(x_{(i)jk} - \bar{x}_{(i)j.})^2$$

Q_1 – источник вариации, обуславливающий различие индивидов, с числом степеней свободы, равным $\nu_1 = n - 1$.

Источник вариации Q_A, обусловленный влиянием фактора A, с числом степеней свободы, равным $\nu_A = r - 1$, имеет вид:

$$Q_A = t \sum_{i=1}^{r} n_i (\bar{x}_{i..} - \bar{x}_{...})^2.$$

Вариация Q_A является частью вариации Q_1, обуславливающей различие индивидов, с числом степеней свободы, равным $\nu_1 = n - 1$:

$$Q_1 = t \sum_{j=1}^{n} (\bar{x}_{(i)j.} - \bar{x}_{...})^2 =$$

$$= t \sum_{i=1}^{r} n_i (\bar{x}_{i..} - \bar{x}_{...})^2 + t \sum_{j=1}^{n} (\bar{x}_{(i)j.} - \bar{x}_{i..})^2 = Q_A + Q_{1e}\,,$$

где усредненный показатель по каждому индивиду вычисляется как

$$\bar{x}_{(i)j.} = \frac{1}{t} \sum_{k=1}^{t} x_{(i)jk}\,.$$

Ошибка Q_{1e} имеет число степеней свободы, равное $\nu_{1e} = n - r$.

Из оставшегося после исключения влияния индивида источник вариации

$$Q_2 = Q - Q_1 = \sum_{j=1}^{n}\sum_{k=1}^{t}(x_{(i)jk} - \bar{x}_{(i)j\cdot})^2$$

выделим вариацию, обуславливающую различие по времени. Источник вариации, обусловленный влиянием временного фактора C, с числом степеней свободы, равным $\nu_C = t - 1$, имеет вид:

$$Q_C = n\sum_{k=1}^{t}(x_{\cdot\cdot k} - \bar{x}_{\cdot\cdot\cdot})^2,$$

а источник вариации, обусловленный влиянием взаимодействия фактора A и фактора времени C, с числом степеней свободы, равным $\nu_{AC} = (t-1)(r-1)$, имеет вид:

$$Q_{AC} = \sum_{i=1}^{r} n_i \sum_{k=1}^{t}(\bar{x}_{i\cdot k} - \bar{x}_{i\cdot\cdot} - \bar{x}_{\cdot\cdot k} + \bar{x}_{\cdot\cdot\cdot})^2.$$

Ошибка имеет вид:

$$Q_{err} = Q_2 - Q_C - Q_{AC} = \sum_{j=1}^{n}\sum_{k=1}^{t}(x_{(i)jk} - \bar{x}_{(i)j\cdot} - \bar{x}_{i\cdot k} + \bar{x}_{i\cdot\cdot})^2.$$

Ее число степеней свободы равно $\nu_{err} = (t-1)(n-r)$.

Для проверки гипотезы о том, что все дифференциальные эффекты фактора A равны нулю, вычисляется статистика

$$\mathrm{F} = \frac{\mathrm{MQ_A}}{\mathrm{MQ_{1e}}} = \frac{\mathrm{Q_A}/\nu_\mathrm{A}}{\mathrm{Q_{1e}}/\nu_{1\mathrm{e}}},$$

которая в случае справедливости нулевой гипотезы имеет распределение Фишера с числом степеней свободы $\nu_\mathrm{A} = \mathrm{r} - 1$ и $\nu_{1\mathrm{e}} = \mathrm{n} - \mathrm{r}$.

Для проверки гипотезы о том, что случайные эффекты временного фактора имеют нулевую дисперсию, вычисляется статистика

$$\mathrm{F} = \frac{\mathrm{MQ_C}}{\mathrm{MQ_{AC}}} = \frac{\mathrm{Q_C}/\nu_{\mathrm{C}}}{\mathrm{Q_{AC}}/\nu_{\mathrm{AC}}},$$

которая в случае справедливости нулевой гипотезы имеет распределение Фишера с числом степеней свободы $\nu_{\mathrm{C}} = \mathrm{t} - 1$ и $\nu_{\mathrm{AC}} = (\mathrm{t} - 1)(\mathrm{r} - 1)$.

В случае предположения фиксированных эффектов временного фактора C вычисляется статистика

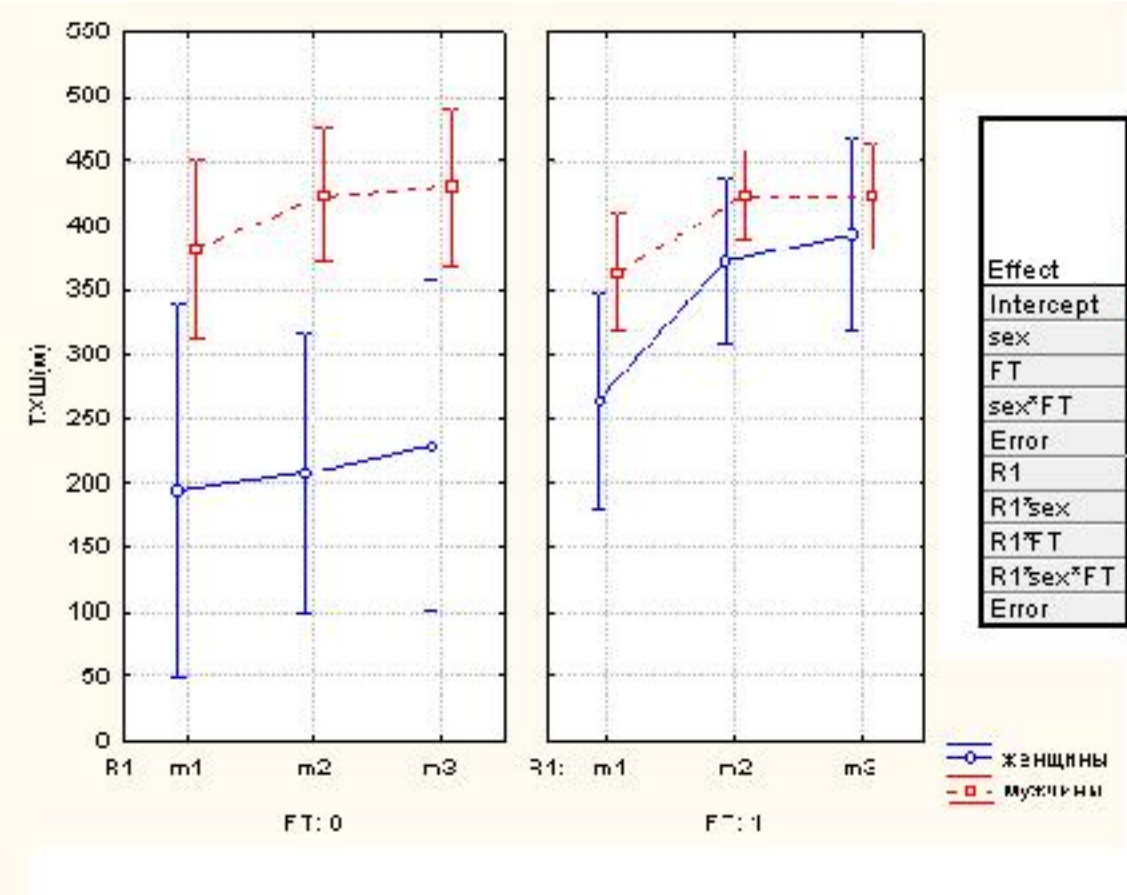
$$\mathrm{F} = \frac{\mathrm{MQ_C}}{\mathrm{MQ_{err}}} = \frac{\mathrm{Q_C}/\nu_{\mathrm{C}}}{\mathrm{Q_{err}}/\nu_{\mathrm{err}}},$$

которая в случае справедливости нулевой гипотезы имеет распределение Фишера с числом степеней свободы $\nu_{\mathrm{C}} = \mathrm{t} - 1$ и $\nu_{\mathrm{err}} = (\mathrm{t} - 1)(\mathrm{n} - \mathrm{r})$.

Для проверки эффекта взаимодействия вычисляется статистика

$$\mathrm{F} = \frac{\mathrm{MQ_{AC}}}{\mathrm{MQ_{err}}} = \frac{\mathrm{Q_{AC}}/\nu_{\mathrm{AC}}}{\mathrm{Q_{err}}/\nu_{\mathrm{err}}},$$

которая в случае справедливости нулевой гипотезы имеет распределение Фишера с числом степеней свободы $\nu_{\mathrm{AC}} = (\mathrm{t}-1)(\mathrm{r}-1)$ и $\nu_{\mathrm{err}} = (\mathrm{t}-1)(\mathrm{n}-\mathrm{r})$.



Effect	SS	Degr. of Freedom	MS	F	p
Repeated Measures Analysis of Variance (kutuso Sigma-restricted parameterization Effective hypothesis decomposition					
Intercept	678713	1	678713	349,646	0,00000
sex	24821	1	24821	12,787	0,00110
FT	5596(	1	5596(	2,882	0,09894
sex*FT	7223	1	7223	3,721	0,06235
Error	64057	33	1941		
R1	5050(	2	2525(	11,061	0,00007
R1*sex	191	2	95	0,419	0,65900
R1*FT	950	2	475	2,081	0,13286
R1*sex*FT	494	2	247	1,083	0,34432
Error	15065	66	228		

Figure: Динамика ТХШ в зависимости от пола и лечебной физкультуры.