Двухфакторный дисперсионный анализ и случайные эффекты

Н.П. Алексеева

СПбГУ, математико-механический факультет

2020 г.

Двухфакторный дисперсионный анализ

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:

$$x_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk}, \qquad (1)$$

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

- μ - генеральное среднее,
- α_i - дифференциальный эффект фактора A,
- β_j - дифференциальный эффект фактора B.
- Величина $(\alpha\beta)_{ij}$ называется взаимодействием факторов. Эта величина 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 $\mathcal{N}(0, \sigma)$.

Оценки параметров модели

При ограничениях на параметры

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

$$\forall j \quad \sum_{i=1}^{I}(\alpha\beta)_{ij} = 0, \quad \forall i \quad \sum_{j=1}^{J}(\alpha\beta)_{ij} = 0,$$

оценками параметров модели (1) 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;

$\widehat{(\alpha\beta)}_{ij} = \bar{x}_{ij\cdot} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{x}$, где $\bar{x}_{ij\cdot}$ – среднее в ячейке ij.

Двухфакторный дисперсионный анализ

Уровень адреналина при разном вскармливании и содержании

вскармл. с матерью 3.35 ± 0.15			вскармл. без матери 4.48 ± 0.15		
отд.кл.ξ_1	сооб.кл.ξ_2	общ.кл.ξ_3	отд.кл.ξ_1	сооб.кл.ξ_2	общ.кл.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
$\bar{x}_{11} = 2.38$	$\bar{x}_{12} = 4.85$	$\bar{x}_{13} = 2.83$	$\bar{x}_{21} = 3.44$	$\bar{x}_{22} = 5.51$	$\bar{x}_{23} = 4.48$
$\bar{x}_{\cdot 1} = 2.91$	$\bar{x}_{\cdot 2} = 5.18$	$\bar{x}_{\cdot 3} = 3.65$			

Матрица плана

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	$(\alpha\beta)_{11}$	$(\alpha\beta)_{12}$	$(\alpha\beta)_{13}$	$(\alpha\beta)_{21}$	$(\alpha\beta)_{22}$
y_{11*}	1	1	0	1	0	0	1	0	0	0	0
y_{12*}	1	1	0	0	1	0	0	1	0	0	0
y_{13*}	1	1	0	0	0	1	0	0	1	0	0
y_{14*}	1	1	0	−1	−1	−1	−1	−1	−1	0	0
y_{21*}	1	0	1	1	0	0	0	0	0	1	0
y_{22*}	1	0	1	0	1	0	0	0	0	0	1
y_{23*}	1	0	1	0	0	1	0	0	0	0	0
y_{24*}	1	0	1	−1	−1	−1	0	0	0	−1	−1
y_{31*}	1	−1	−1	1	0	0	−1	0	0	−1	0
y_{32*}	1	−1	−1	0	1	0	0	−1	0	0	−1
y_{33*}	1	−1	−1	0	0	1	0	0	−1	0	0
y_{34*}	1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1

Двухфакторный дисперсионный анализ

Вектор параметров
$\beta = (\mu, \alpha_1, \alpha_{I-1}, \beta_1, \ldots, \beta_{J-1}, (\alpha\beta)_{11}, \ldots, (\alpha\beta)_{I-1,J-1})'$. Общее число наблюдений равно IJK. Остаточная сумма квадратов основной модели равна

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

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

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

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

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

Аналогично строятся остальные статистики.

Двухфакторный дисперсионный анализ

Таблица двухфакторного дисперсионного анализа

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	$Q_B = IK\sum_{j=1}^{J}(\bar{x}_{\cdot j} - \bar{x})^2$	$\nu_B = J - 1$	$MQ_B = \frac{Q_B}{\nu_B}$
взаимо-действие AB	$Q_{AB} = K\sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{x}_{ij\cdot} - -\bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{x})^2$	$\nu_{AB} = = (I-1)(J-1)$	$MQ_{AB} = \frac{Q_{AB}}{\nu_{AB}}$
остаток (ошибка)	$Q_R = = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(x_{ijk} - \bar{x}_{ij\cdot})^2$	$\nu_R = IJ(K-1)$	$MQ_R = \frac{Q_R}{\nu_R}$
полная	$Q_T = \sum_{i=1}^{I}\sum_{j=1}^{J}s\sum_{k=1}^{K}(x_{ijk} - \bar{x})^2$	$\nu_T = IJK - 1$	

Проверка гипотез

- средний уровень адреналина 3.35 ± 0.15 в группе с матерью, ниже среднего уровня адреналина 4.48 ± 0.15 в группе без матери.
- средние 2.91 ± 0.18, 5.18 ± 0.18 и 3.65 ± 0.18 в группах, отличающихся условиями содержания

Нужно выяснить, насколько это значимо.

$$\mathrm{H_0}: \text{все } (\alpha\beta)_{ij} = 0 \text{ эффект взаимодействия } \mathrm{F} = \frac{\mathrm{MQ_{AB}}}{\mathrm{MQ_R}} \sim \mathrm{F}(\nu_{\mathrm{AB}}, \nu_{\mathrm{R}})$$

$$\mathrm{H_0}: \text{все } \alpha_i = 0 \text{ эффект A-фактора } \mathrm{F} = \frac{\mathrm{MQ_A}}{\mathrm{MQ_R}} \sim \mathrm{F}(\nu_{\mathrm{A}}, \nu_{\mathrm{R}})$$

$$\mathrm{H_0}: \text{все } \beta_j = 0 \text{ эффект B-фактора } \mathrm{F} = \frac{\mathrm{MQ_B}}{\mathrm{MQ_R}} \sim \mathrm{F}(\nu_{\mathrm{B}}, \nu_{\mathrm{R}})$$

Двухфакторный дисперсионный анализ

ПРИМЕР 1 (продолжение) Проверяем гипотезы.
H_0 : все $\alpha_i = 0$. $p = 0.000003$ — влияние фактора вскармливания значимо для уровня адреналина,
H_0 : все $\beta_j = 0$. $p < 0.000001$ – влияние фактора условия содержания значимо для уровня адреналина:
H_0 : все $(\alpha\beta)_{ij} = 0$. $p = 0.16$ – фактор взаимодействия незначим.

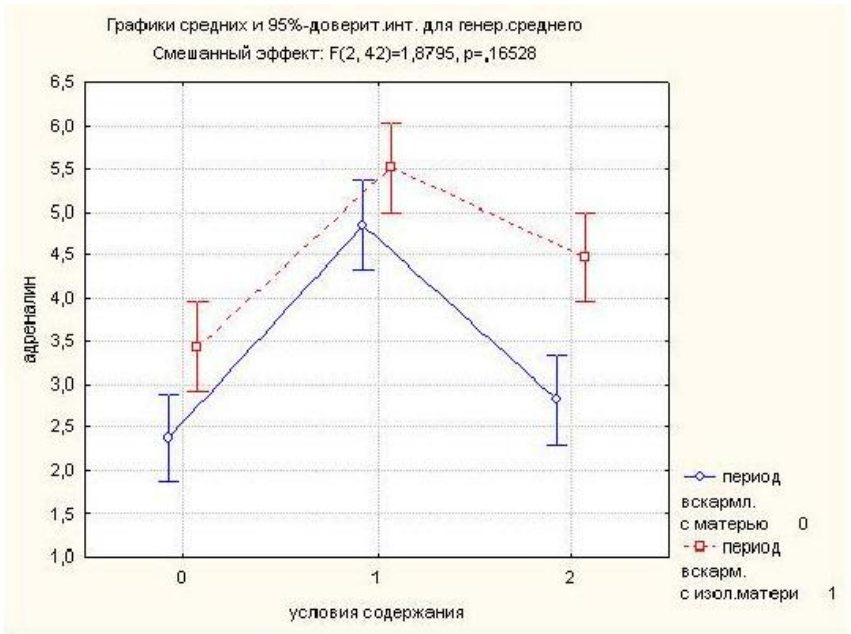

Figure: Иллюстрация отсутствия эффекта взаимодействия факторов вскармливания и содержания.

Модель со случайными эффектами

Подпопуляции выбираются случайно из большого (бесконечного) числа подпопуляций

$$x_{ij} = \mu + a_i + e_{ij}, \quad i = 1, \ldots, r, \ j = 1, \ldots, n_i, \tag{2}$$

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

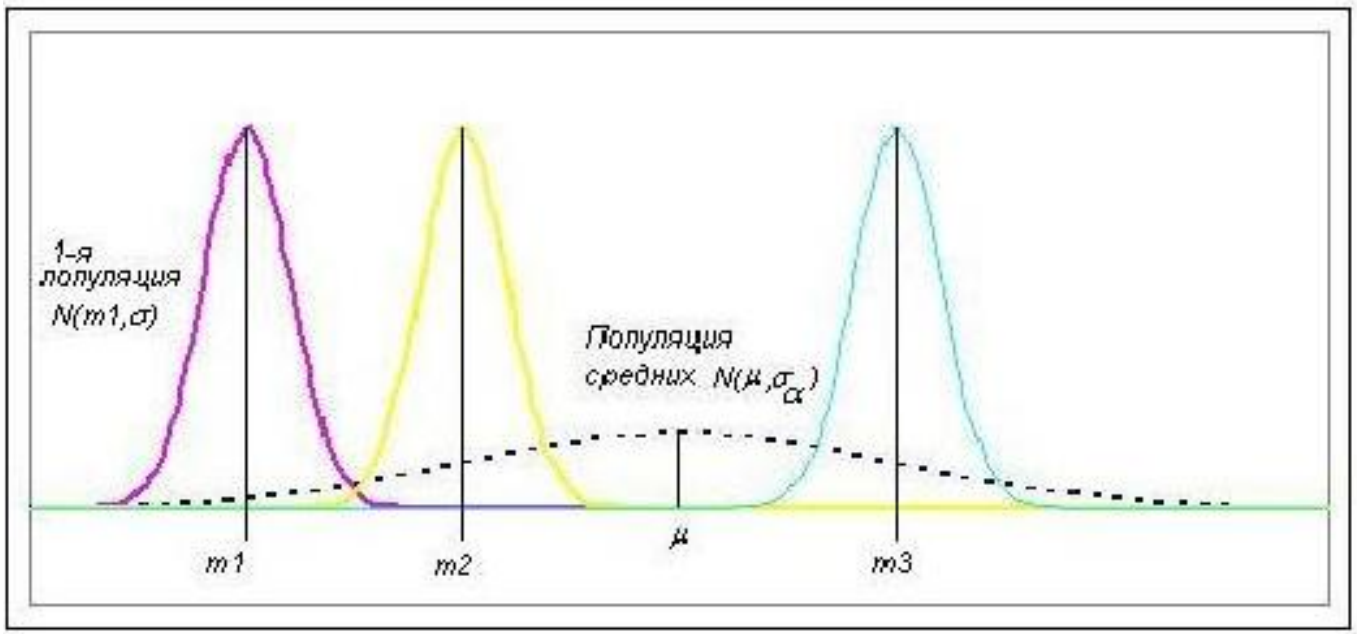

Figure: $m_1, \ldots, m_r$ выбираются случайно.

Проверка значимости случайных эффектов

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 $\frac{Q_2}{n-r}$.

$$\mathbb{E}\left(\frac{Q_2}{n-r}\right) = \sigma^2,$$

$$\mathbb{E}\left(\frac{Q_1}{r-1}\right) = \sigma^2 + k\sigma_a^2,$$

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

Используем выражения $x_{ij} = \mu + a_i + \epsilon_{ij}$ и $\bar{x}_i = \mu + a_i + \bar{\epsilon}_i$ для вычисления $\mathbb{E}Q_2$.

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

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

$$\mathbb{E}Q_2 = \sum_{i=1}^{r}\sum_{j=1}^{n_i}\mathbb{E}\epsilon_{ij}^2 - \sum_{i=1}^{r} n_i\mathbb{E}\bar{\epsilon}_i^2 = n\sigma^2 - \sum_{i=1}^{r} n_i\frac{\sigma^2}{n_i} = (n-r)\sigma^2\,.$$

Для вычисления $\mathbb{E}Q_1$ используем выражения $\bar{x}_i = \mu + a_i + \bar{\epsilon}_i$ и

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

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

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

Считаем математическое ожидание.

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

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

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

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

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

Если $n_1 = \ldots = n_r = m$, то $n = rm$,

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

Статистика критерия и оценка дисперсии

Таким образом, $\mathbb{E}Q_1 = (r-1)(\sigma^2 + k\sigma_a^2)$, $\mathbb{E}Q_2 = (n-r)\sigma^2$.

При $H_0 : \sigma_a^2 = 0$ статистики $\frac{Q_2}{n-r}$ и $\frac{Q_1}{r-1}$ имеют одинаковые математические ожидания σ^2, поэтому

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

Заметим, что в случае однофакторного дисперсионного анализа значимости моделей со случайными и фиксированными эффектами совпадают.

Оценка дисперсии σ_a^2

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

Двухфакторный дисперсионный анализ со случайными эффектами

Модель со случайными эффектами имеет вид:

$$x_{ijk} = \mu + a_i + b_j + (ab)_{ij} + e_{ijk}, \quad \text{где} \tag{3}$$

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

Математические ожидания статистик и оценки дисперсий

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:

$$\sigma^2_{\mathrm{ab}} = \frac{\mathrm{MQ_{AB}} - \mathrm{MQ_R}}{\mathrm{K}}$$

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

Критерии значимости случайных эффектов

Для проверки гипотез

- $\mathrm{H}_0 : \sigma^2_{\mathrm{ab}} = 0$ используется отношение $\mathrm{F} = \frac{\mathrm{MQ_{AB}}}{\mathrm{MQ_R}}$, $\nu_{\mathrm{AB}} = (\mathrm{I}-1)(\mathrm{J}-1)$,
- $\mathrm{H}_0 : \sigma^2_{\mathrm{a}} = 0$ используется отношение $\mathrm{F} = \frac{\mathrm{MQ_A}}{\mathrm{MQ_{AB}}}$, $\nu_{\mathrm{a}} = \mathrm{I}-1$,
- $\mathrm{H}_0 : \sigma^2_{\mathrm{b}} = 0$ отношение $\mathrm{F} = \frac{\mathrm{MQ_B}}{\mathrm{MQ_{AB}}}$, $\nu_{\mathrm{b}} = \mathrm{J}-1$.

ПРИМЕР (продолжение)

- 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

Модель со смешанными эффектами имеет вид:

$$x_{ijk} = \mu + \alpha_i + b_j + (\alpha b)_{ij} + e_{ijk}, \text{ где} \tag{4}$$

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

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 если в двуфакторном плане имеется один случайный фактор, то результаты дисперсионного анализа будут такие же, как и в случае двух случайных факторов.

Способ вычислений в R

```
data <- read.csv("data_big.csv")
df<-data.frame(x=data$HR.1, f1=as.factor(data$craving.to.alcohol.1),
f2=as.factor(data$depressed.mood.1))
ao<-aov(x~ f1*f2,df); summary(ao)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
f1	2	2589	1294.3	4.754	0.0167 *
f2	1	59	59.2	0.218	0.6445
f1:f2	2	591	295.5	1.085	0.3516
Residuals	28	7623	272.2		

Для вывода средних значений можно использовать функцию model.tables

Построение графика взаимодействия факторов

```
K1<-length(table(df$f1)); K2<-length(table(df$f2))
interaction.plot(x.factor=df$f1, trace.factor=df$f2, response=df$x,
fun = mean, type = "b", legend = FALSE, trace.label ="group",
xlab = "craving", ylab = "HR", lty = seq(K2), col = seq(K2), pch = 20)
legend('bottomright',paste("depress",names(table(df$f2))), lty = seq(K2),
col=seq(K2), cex=0.7, pch=20)
```

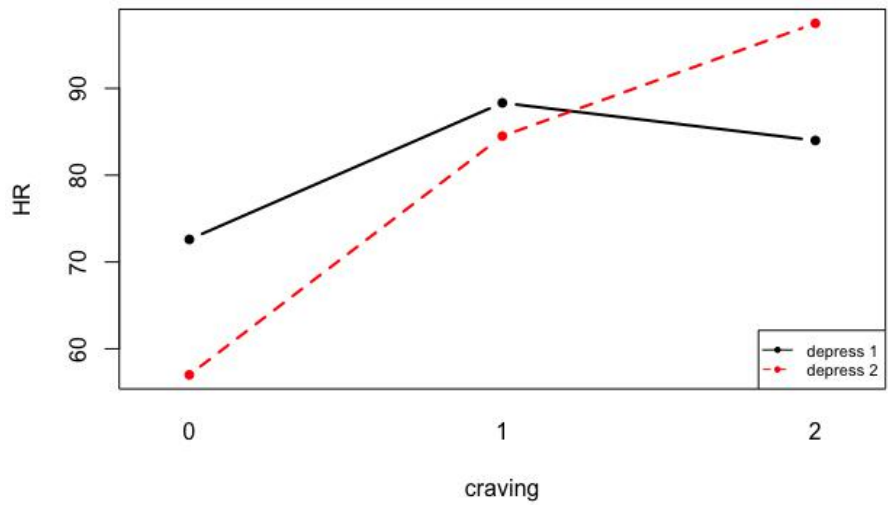

Figure: Эффекты взаимодействия факторов: тяга к алкоголю и депрессия.