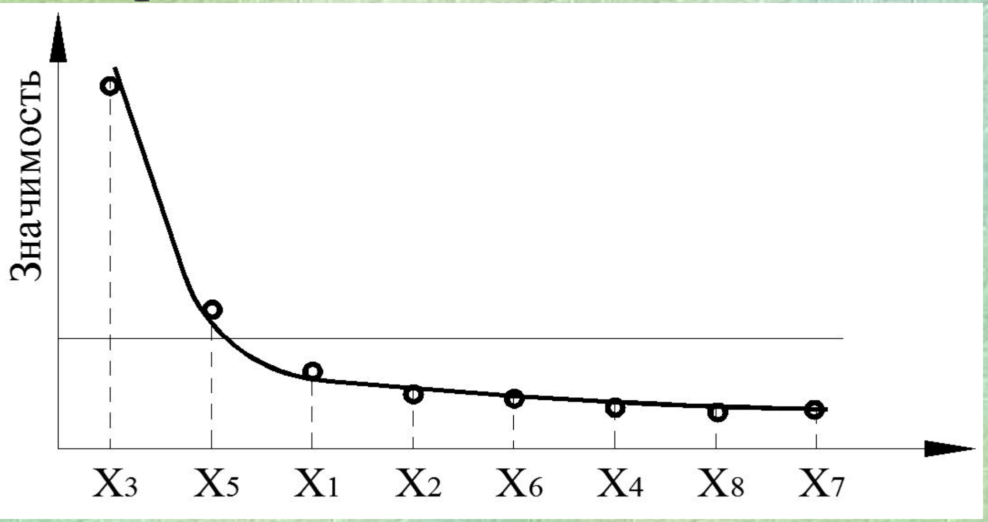
Дисперсионный анализ

Назначение дисперсионного анализа

Дисперсионный анализ (**ДА**) (от латинского Dispersio – рассеивание / на английском Analysis Of Variance - ANOVA) применяется для оценки влияния одного или нескольких входных параметров на выходной параметр (функцию).

ДА позволяет ранжировать входные параметры по степени их прямого и взаимного влияния на функцию.



Чем больше параметров требуется учитывать, тем дороже проведение эксперимента.

Согласно закону Парето (принцип 20/80), значимых факторов немного, т.е. примерно *20% параметров дают 80% результата, а остальные 80% параметров — лишь 20% результата.*

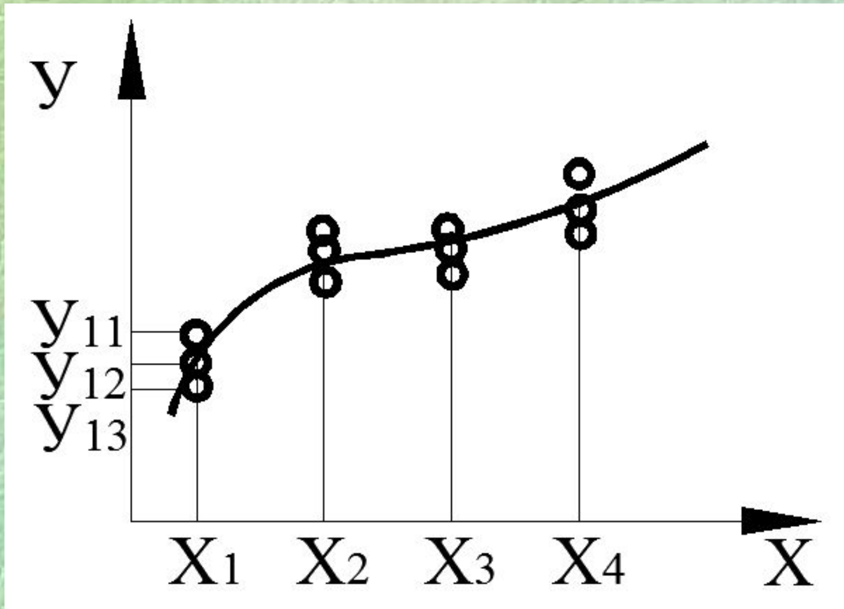
Особенности дисперсионного анализа, дисперсионные модели одно-, двух- и трех факторного эксперимента

Дисперсионный анализ предназначен **для качественного исследования модели процесса**: $y = f(x1, x2, \dots, xk)$ на предмет **оценки значимости каждого входного параметра на функцию У**.

Математический аппарат, который занимается **исследованием значимости входных параметров**, называется **дисперсионным анализом**. В его основе лежит **анализ вкладов каждого фактора в общую дисперсию эксперимента.**

Рассмотрим однофакторный эксперимент: $y = f(x_1)$. Дисперсионную модель этого эксперимента можно представить в виде: $y = A + \varepsilon$,

где **У** - общий вклад в общую дисперсию, который вносят все факторы; **А** - эффект фактора X_1, ε - эффект ошибки воспроизводимости.



ε рассчитывается в случае, если хотя бы в одной точке Xi проведено более одного эксперимента ($У_{i1}$, $У_{i2}$, $У_{i3}$). Если в каждой точке Xi проведен только один эксперимент, то ε **= 0.**

Дисперсионная модель двухфакторного эксперимента $y = f(x_1, x_2)$ строится с учетов эффекта совместного влияния факторов X_1 и X_2:

$$y = A + B + AB + \varepsilon ,$$

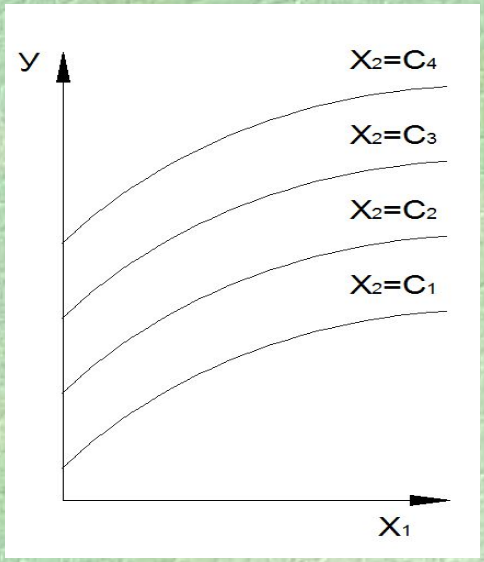
где А и В – эффекты факторов X_1 и X_2;
АВ – эффект совместного влияния (взаимодействия) факторов X_1 и X_2 (АВ=0, если функции сепарабельные);
ε – эффект ошибки воспроизводимости.

Дисперсионная модель трехфакторного эксперимента строится по аналогии и будет содержать не только **эффекты парных** (***AB, AC и BC)*** , но и **тройного взаимодействия** (***ABC)***:

$$y = A + B + C + AB + AC + BC + ABC + \varepsilon$$

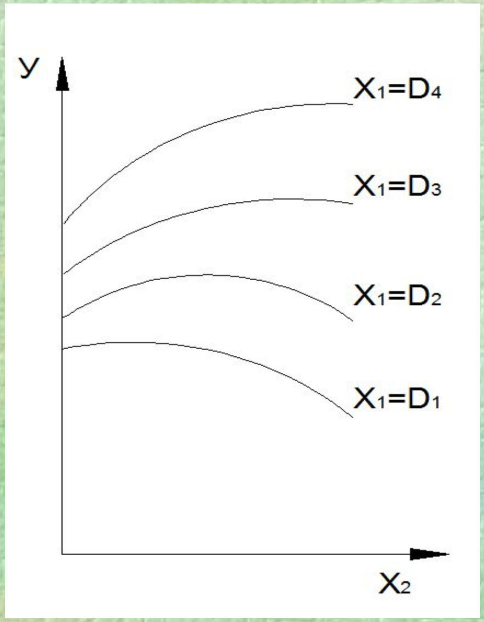
Вспомним о **сепарабельных функциях**:

Для первого случая:
$У = A + f(X_1) + f(X_2)$
Каждая функция $f(X_1)$ и $f(X_2)$ зависят только от одной переменной, а сами переменные (X_1 и X_2) независимы друг от друга.
Семейство функций $У = A + f(X_1) + f(X_2)$ называется **сепарабельными функциями**.



Для второго случая:
$У = A + f(X_1) + f(X_2) + f(X_1)*f(X_2)$

Член уравнения **$f(X_1)*f(X_2)$** показывает **степень взаимодействия** параметров X_1 и X_2 на функцию У.



В качестве количественного показателя, применяемого для сравнения эффектов факторов X_1, X_2 и др., используется **критерий Фишера:**

$$F_i = \frac{S_i^2}{S_0^2} = \frac{SS_i / f_i}{SS_0 / f_0} \leq F_T(p; f_i; f_0),$$

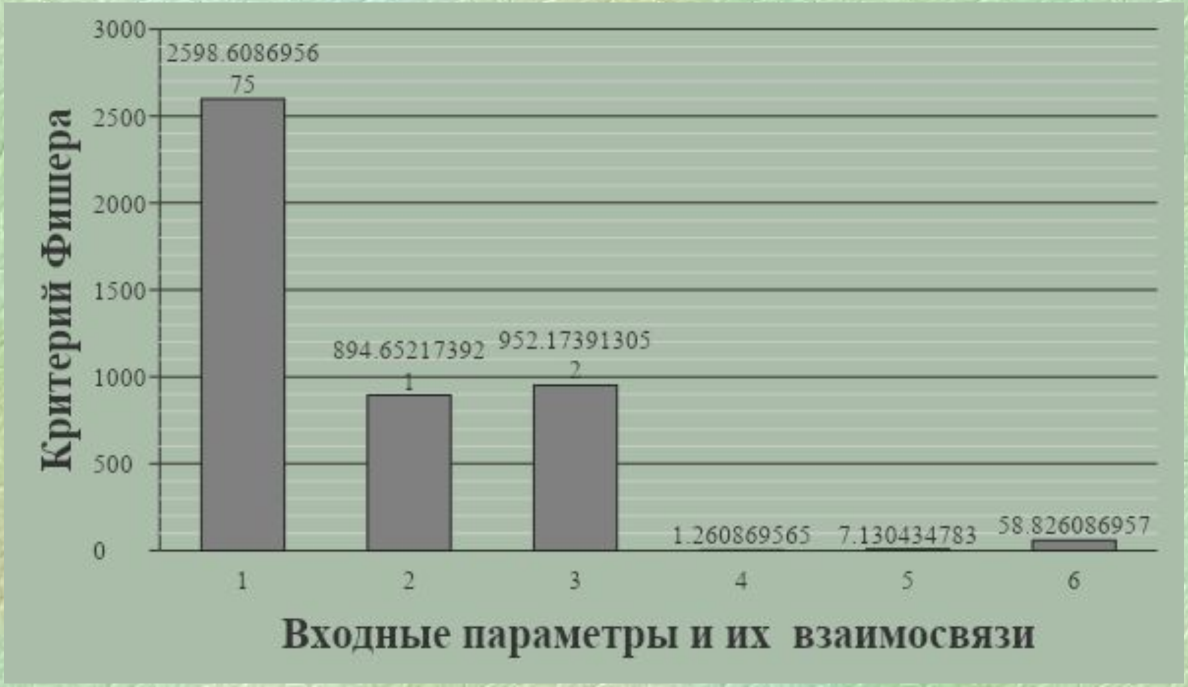
где S_i^2, S_0^2 – дисперсии соответственно i-того и наименее значимого фактора (обычно от эффекта ошибки воспроизводимости **ε**);

F_T – табличное (критическое) значение критерия Фишера;

f_i и f_0 – степени свободы i-того и 0-го факторов;

p – доверительная вероятность (обычно p=0,95).

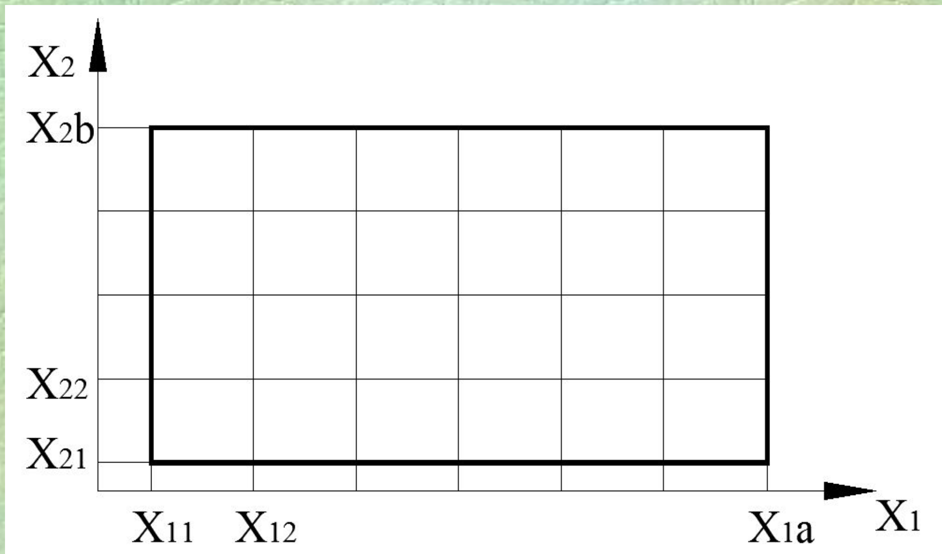
Дисперсионную модель наиболее удобно представлять в виде гистограммы:



Таким образом, для проведения ДА нужно уметь рассчитывать **критерии Фишера**, т.е. уметь определять значения дисперсий $\mathbf{S^2_i}$, среднеквадратических отклонений $\mathbf{SS_i}$ и степеней свободы $\mathbf{f_i}$.

Основные уравнения ДА

Рассмотрим **двухфакторный** эксперимент.



Уровни входных параметров (факторов) X_1 и X_2 откладываются по осям координат.

Фактор X_1 измеряется на **a** равностоящих уровнях.

Счетчик уровней для X_1:
i = 1, 2, …a.

Фактор X_2 измеряется на **b** равностоящих уровнях.
Счетчик уровней для X_2: **j = 1, 2, …b.**

В каждой узловой точке эксперимента проводится по **n** опытов.

n также является фактором, от которого зависит эффект ошибки воспроизводимости **ε**. Счетчик уровней по **n: k = 1, 2, …, n**

Таким образом, **полный факторный эксперимент (ПФЭ)** будет содержать **$N_n = a*b*n$** опытов. Если в каких-то точках опыты не проводятся, то эксперимент называется **дробным факторным (ДФЭ).**

Для ПФЭ **выходной параметр У** будет иметь **три индекса: i, j, k.** Т.е. обозначение **$У_{ijk}$ будет определять значение выходного параметра в ijk узловой точке, согласно ПФЭ.**

Определим **общее число степеней свободы $f_{общ}$** эксперимента:

$$f_{общ} = abn - 1.$$

Число степеней свободы каждого из факторов X_1 и X_2:

$$fx_1 = f_1 = a - 1 ; \qquad fx_2 = f_2 = b - 1 .$$

Число степеней свободы взаимодействия:

$$f_{12} = (a - 1)(b - 1) .$$

Число степеней свободы ошибки воспроизводимости по *ab точкам:*

$$f_0 = ab(n - 1).$$

Согласно **первому основному уравнению дисперсионного анализа:** $f_{общ} = f_1 + f_2 + f_{12} + f_o$.

Это уравнение легко получить, если преобразовать правую часть тождества:

$$abn - 1 = (a - 1) + (b - 1) + (a - 1)(b - 1) + ab(n - 1).$$

По аналогии можно получить первое основное уравнение для трехфакторного эксперимента:

$$f_{общ} = f_1 + f_2 + f_3 + f_{12} + f_{13} + f_{23} + f_{123} + f_o .$$

Число уровней фактора X_3 равно **c** (счетчик **s** = 1, 2, ... , **c**). Недостающие числа степеней свободы равны:

$f_3 = c - 1;$ $f_{123} = (a - 1)(b - 1)(c - 1)$; $f_o = abc(n - 1).$

Общее среднеквадратичное отклонение для двухфакторного эксперимента можно рассчитать по формуле:

$$SS_{\text{общ}} = \sum_{i}^{a}\sum_{j}^{b}\sum_{k}^{n}(y_{ijk} - y_{***})^2 \qquad (2)$$

В ДА для компактности записи расчетных формул знак суммирования заменяется звездочкой, т.е.:

$$y_{***} = \sum_{i}^{a}\sum_{j}^{b}\sum_{k}^{n} y_{ijk}$$

Раскрыв скобки и преобразовав уравнение (2), получим:

$$SS_{\text{общ}} = \sum_{i}^{a}\sum_{j}^{b}\sum_{k}^{n}\Big[(\bar{y}_{i**} - \bar{y}_{***})^2 + (\bar{y}_{*j*} - \bar{y}_{***})^2 + (\bar{y}_{ij*} - \bar{y}_{i**} - \bar{y}_{*j*} + \bar{y}_{***})^2 + (y_{ijk} - \bar{y}_{ij*})^2\Big] \qquad (3)$$

Уравнение (3) состоит из четырех слагаемых, каждое из которых соответственно равно SS_1, SS_2, SS_{12} и SS_0.
После преобразований уравнения (2) для $SS_{общ}$ и каждого из слагаемых в уравнении (3), получим:

$$SS_{общ} = \sum_i^a \sum_j^b \sum_k^n y_{ijk}^2 - \frac{y_{***}^2}{abn} \quad ;$$

$$SS_1 = \sum_1^a \frac{y_{i**}^2}{bn} - \frac{y_{***}^2}{abn} \quad ; \qquad SS_2 = \sum_1^b \frac{y_{*j*}^2}{an} - \frac{y_{***}^2}{abn} \quad ;$$

$$SS_{12} = \sum_i^a \sum_j^b \frac{y_{ij*}^2}{n} - \sum_1^a \frac{y_{i**}^2}{bn} - \sum_1^b \frac{y_{*j*}^2}{an} + \frac{y_{***}^2}{abn} \quad ;$$

$$SS_0 = \sum_i^a \sum_j^b \sum_k^n y_{ijk}^2 - \frac{y_{***}^2}{abn}$$

Соотношение между суммами квадратов отклонений подчиняется **второму основному уравнению ДА**:

$$SS_{общ} = SS_1 + SS_2 + SS_{12} + SS_o$$

По аналогии для трехфакторного эксперимента:

$$SS_{общ} = SS_1 + SS_2 + SS_3 + SS_{12} + SS_{13} + SS_{23} + SS_{123} + SS_o$$

Между выражениями для расчета числа степеней свободы и суммы квадратов отклонений существует аналогия.

Аналогии между выражениями для расчета SS_i и f_i

1) количество слагаемых и знаки перед ними в выражениях для числа степеней свободы и соответствующей суммы квадратов отклонений совпадают;
2) в каждом слагаемом для SS знаки Σ содержат индексы, аналогичные индексам при f ;
3) эти же индексы присутствуют в числителе при y^2, а недостающие индексы числителя заменены звездочками;
4) знаменатель можно записать по недостающим индексам числителя, которые в числителе обозначаются звездочками.

Например: $f_1 = a\text{-}1$

$$SS_1 = \sum_1^a \frac{y_{i**}^2}{bn} - \frac{y_{***}^2}{abn}$$

Эту аналогию используем в качестве правила для формального написания суммы квадратов отклонений.

Для этого сначала необходимо написать выражение для числа степеней свободы и раскрыть в нем скобки. Затем, придерживаясь п.1 – 4, написать соответствующие члены искомых сумм.

Вывод формул для расчета суммы квадратов отклонений SSi по формальным правилам

Эффект модели	Число степеней свободы *f*	Сумма квадратов отклонений SS
A (фактор X_1)	*a - 1*	$\sum_{i}^{a} \frac{y_{i**}^2}{bn} - \frac{y_{***}^2}{abn}$
B (фактор X_2)	*b -1*	$\sum_{j}^{b} \frac{y_{*j*}^2}{an} - \frac{y_{***}^2}{abn}$
AB (Взаимодействие X_1X_2)	*(a - 1)(b -1) = ab - a - b + 1*	$\sum_{i}^{a}\sum_{j}^{b} \frac{y_{ij*}^2}{n} - \sum_{i}^{a} \frac{y_{i**}^2}{bn} - \sum_{j}^{b} \frac{y_{*j*}^2}{an} + \frac{y_{***}^2}{abn}$
Ошибка воспроизводимости ε	*ab*(*n* - 1) = *abn - ab*	$\sum_{i}^{a}\sum_{j}^{b}\sum_{k}^{n} y_{ijk}^2 - \sum_{i}^{a}\sum_{j}^{b} \frac{y_{ij*}^2}{n}$
Общий эффект	*abn - 1*	$\sum_{i}^{a}\sum_{j}^{b}\sum_{k}^{n} y_{ijk}^2 - \frac{y_{***}^2}{abn}$

Для трехфакторного эксперимента имеем:

$$f_{123} = (a-1)(b-1)(c-1) = abc - ab - ac - bc + a + b + c - 1.$$

$$SS_{123} = \sum_{i}^{a}\sum_{j}^{b}\sum_{s}^{c}\frac{y_{ijs*}^{2}}{n} - \sum_{i}^{a}\sum_{j}^{b}\frac{y_{ij**}^{2}}{cn} - \sum_{i}^{a}\sum_{s}^{c}\frac{y_{i*s*}^{2}}{bn} - \sum_{j}^{b}\sum_{s}^{c}\frac{y_{*js*}^{2}}{an} + \sum_{i}^{a}\frac{y_{i***}^{2}}{bcn} + \sum_{j}^{b}\frac{y_{*j**}^{2}}{acn} + \sum_{s}^{c}\frac{y_{**s*}^{2}}{abn} - \frac{y_{****}^{2}}{abcn}$$

В ДА для компактности записи расчетных формул знак суммирования заменяется звездочкой:

$$y^2_{***} = (y_{111} + y_{112} + \ldots y_{11n} + y_{121} + \ldots y_{12n} + \ldots y_{abn})^2$$

$$\sum_i^a \sum_j^b \sum_s^n y_{ijs}^2 = y^2_{111} + y^2_{112} + \ldots y^2_{11n} + y^2_{121} + \ldots y^2_{12n} + \ldots y^2_{abn}$$

$$\sum_i^a y_{i**}^2 = (y_{111} + y_{112} + \ldots + y_{11n} + y_{121} + \ldots + y_{1bn})^2 + (y_{211} + y_{212} + \ldots + y_{21n} + y_{221} + \ldots + y_{2bn})^2 + \ldots + (y_{a11} + y_{a12} + \ldots + y_{a1n} + y_{a21} + \ldots + y_{abn})^2$$

$$\sum_i^a \sum_j^b y_{ij*}^2 = (y_{111} + y_{112} + \ldots + y_{11n})^2 + (y_{121} + y_{122} + \ldots + y_{1bn})^2 + \ldots + (y_{ab1} + y_{ab2} + \ldots + y_{abn})^2$$

Рассмотрим пример двухфакторного эксперимента

Пусть уровни варьирования параметров a и b меняются от 1 до 2. В каждой точке проводится по два эксперимента (n=2). Т.е. a = 2 (i=1,2); b = 2 (j=1,2); n = 2 (s=1,2).
В каждой ijs-эксперименте зафиксированы следующие значения выходного параметра У:

s	j	i	
		1	2
1	1	2	1
	2	4	3
2	1	6	5
	2	8	7

Рассчитать суммы квадратов для двухфакторного эксперимента.

Расчет сумм квадратов

s	j	i	
		1	2
1	1	2	1
	2	4	3
2	1	6	5
	2	8	7

Сумма	Порядок расчета	Значение	Число опытов	
y^2_{***}	$(2+4+6+8+1+3+5+7)^2$	1296	abc	8
$\sum_i^a \sum_j^b \sum_s^c y^2_{ijs}$	$2^2+4^2+6^2+8^2+1^2+3^2+5^2+7^2$	204		1
$\sum_j^b \sum_s^c y^2_{*js}$	$(2+1)^2+(4+3)^2+(6+5)^2+(8+7)^2$	404	a	2
$\sum_i^a \sum_s^c y^2_{i*s}$	$(2+4)^2+(6+8)^2+(1+3)^2+(5+7)^2$	392	b	2
$\sum_i^a \sum_j^b y^2_{ij*}$	$(2+6)^2+(4+8)^2+(1+5)^2+(3+7)^2$	344	c	2
$\sum_i^a y^2_{i**}$	$(2+4+6+8)^2+(1+3+5+7)^2$	656	bc	4
$\sum_j^b y^2_{*j*}$	$(2+6+1+5)^2+(4+8+3+7)^2$	680	ac	4
$\sum_s^c y^2_{**s}$	$(2+1+4+3)^2+(6+5+8+7)^2$	776	ab	4

Эффект модели	Степени свободы fi		Сумма квадратов отклонений SSi		S^2_i
	Формула	Значение	Формула	Значение	
A	f_1=a-1	1	$SS_1 = \sum_i^a \frac{y_{i**}^2}{bn} - \frac{y_{***}^2}{abn}$	2	2
B	f_2=b-1	1	$SS_2 = \sum_j^b \frac{y_{*j*}^2}{an} - \frac{y_{***}^2}{abn}$	8	8
AB	f_{12}=(a-1)(b-1)=ab-a -b+1	1	$SS_{12} = \sum_i^a \sum_j^b \frac{y_{ij*}^2}{n} - \sum_i^a \frac{y_{i**}^2}{bn} - \sum_j^b \frac{y_{*j*}^2}{an} + \frac{y_{***}^2}{abn}$	0	0
Ошибка ε	f_0=ab(n-1)=abn-ab	4	$SS_0 = \sum_i^a \sum_j^b \sum_s^n y_{ijs}^2 - \sum_i^a \sum_j^b \frac{y_{ij*}^2}{n}$	32	8
Общий эффект	$f_{общ}$=abn-1	7	$SS_{общ} = \sum_i^a \sum_j^b \sum_s^n y_{ijs}^2 - \frac{y_{***}^2}{abn}$	42	6

Проверка: $f_{общ.} = f_1+f_2+f_{12}+f_0 = 1+1+1+4=7$

$SS_{общ.}=SS_1+SS_2+SS_{12}+SS_o=2+8+0+32=42$

Формулы для расчета средних значений:

$$\bar{y}_{ij*} = y_{ij*}/n \qquad \bar{y}_{*j*} = y_{*j*}/an$$

$$\bar{y}_{i**} = y_{i**}/bn \qquad \bar{y}_{***} = y_{***}/abn$$

ДА проводится в несколько этапов:

1. Расчет сумм и средних значений внутригрупповых и межгрупповых выборок.
2. Расчет степеней свободы факторов.
3. Расчет суммы квадратов отклонений SSi.
4. Расчет дисперсий.
5. Оценка значимости факторов по критерию Фишера.

$$F_i = \frac{S_i^2}{S_0^2} = \frac{SS_i/f_i}{SS_0/f_0} \le F_T(p; f_i; f_0)$$