Дисперсионный анализ

Постановка проблемы

Дисперсионный анализ является статистическим методом анализа результатов наблюдений, зависящих от различных одновременно действующих факторов, с целью выбора наиболее значимых факторов и оценки их влияния на исследуемый процесс.

Методами дисперсионного анализа устанавливается наличие влияния заданного фактора на изучаемый процесс (на выходную переменную процесса) за счёт статистической обработки наблюдаемой совокупности выборочных данных.

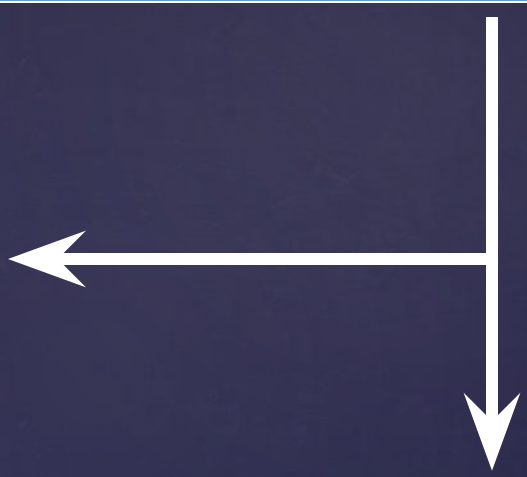
- Основной целью дисперсионного анализа является исследование значимости различия между средними.

- Установить различаются ли три группы или более по какому-либо одному количественному признаку

Например определить, зависит ли активность фермента от стадии заболевания

Классификация методов дисперсионного анализа

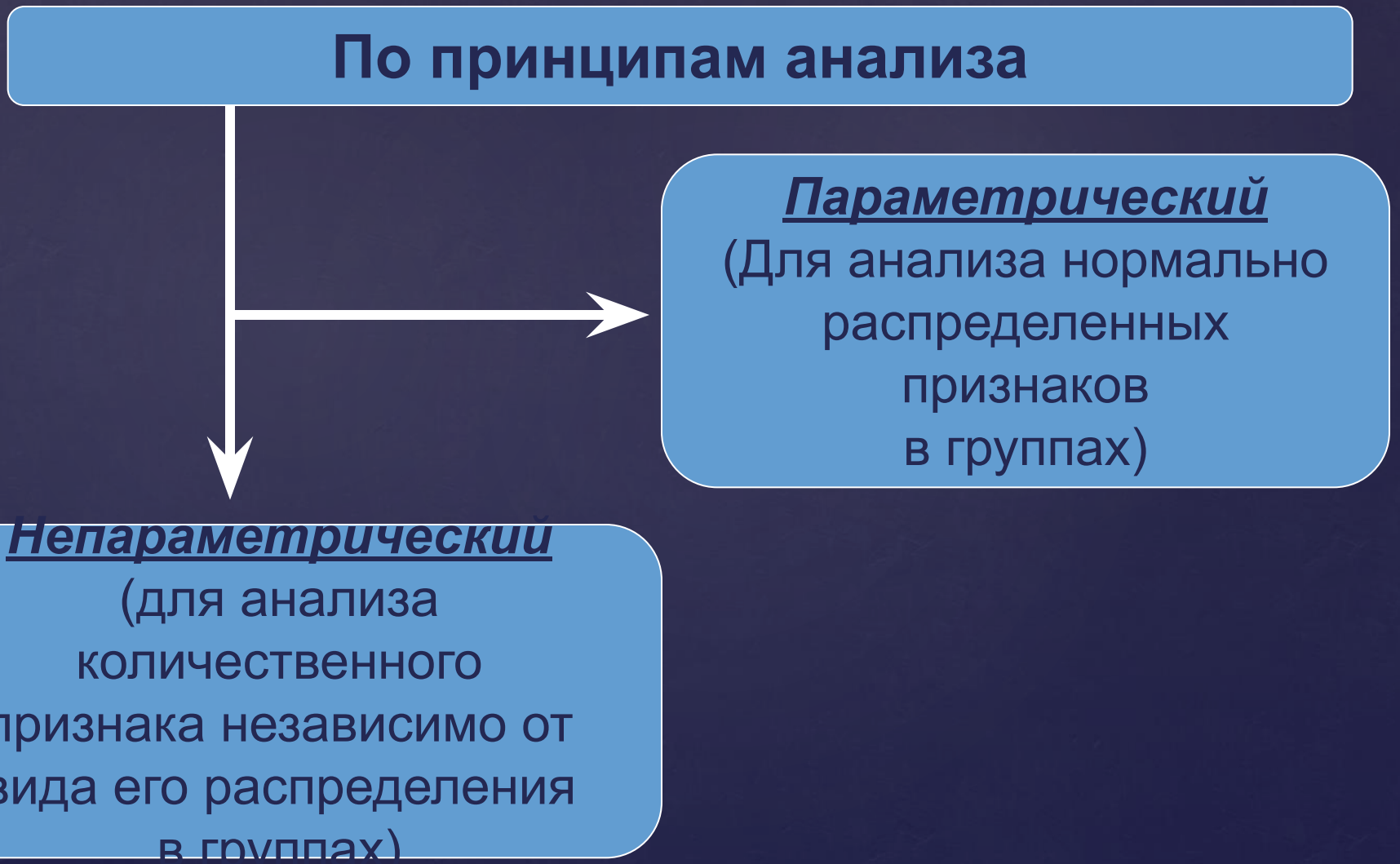
По количеству анализируемых признаков



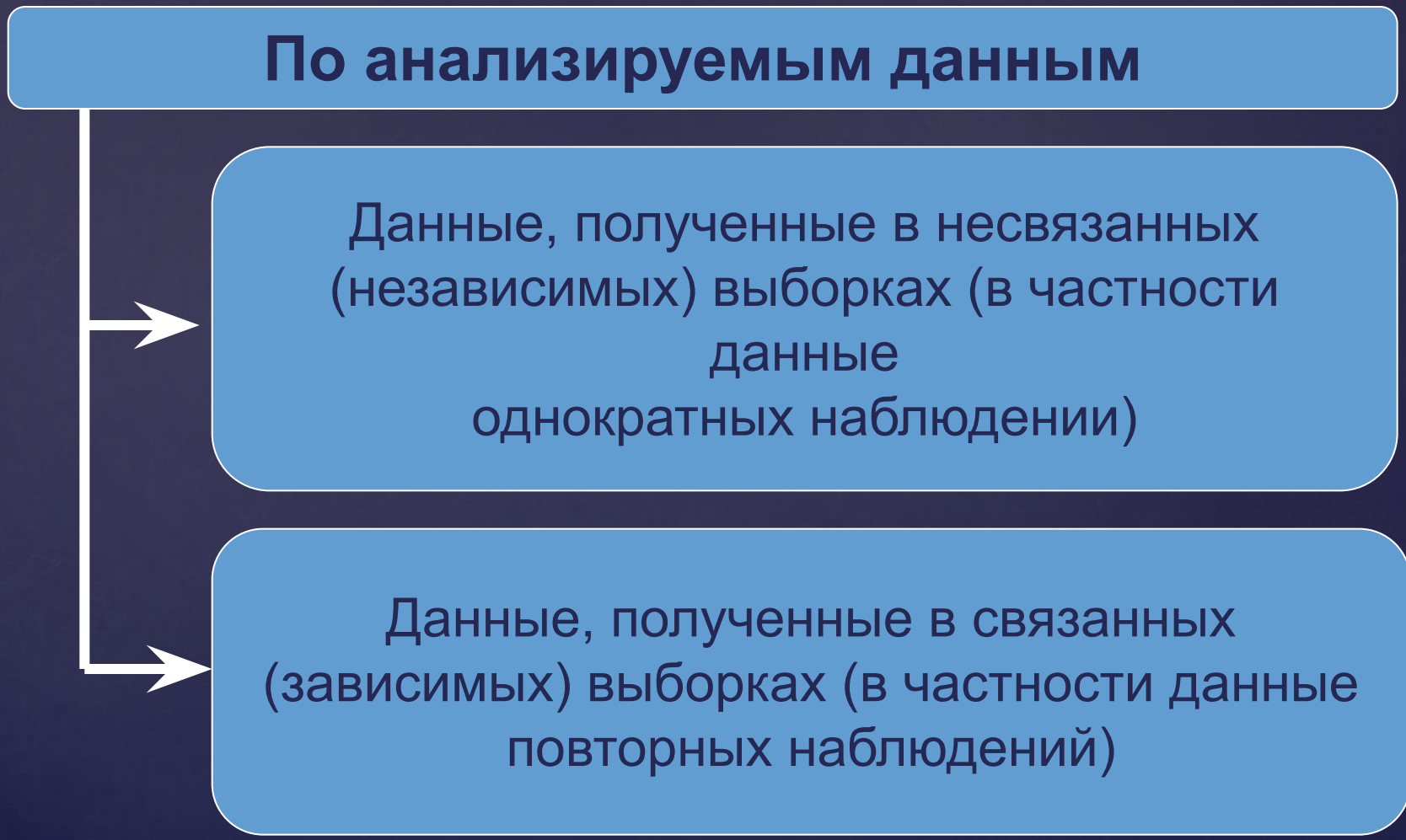
Однофакторный
(ANOVA)
(Анализ различий групп
по одному признаку)

Многофакторный
(MANOVA)
(Анализ различий групп
Одновременно по двум
признакам и более)

Классификация методов дисперсионного анализа



Классификация методов дисперсионного анализа



- Сравнить три или более группы по количественному нормально распределенному признаку

- В процедуре параметрического анализа вариаций общая вариация данных рассматривается как сумма двух видов вариаций:

Параметрический дисперсионный анализ

1. Межгрупповая вариация – вариация между средним каждой группы и общим средним значением всей выборки

2. Внутригрупповая вариация – вариация между каждым объектом исследования группы и средним значением соответствующей группы

Параметрический дисперсионный анализ

Этапы выполнения:

- Проверка гипотез о равенстве дисперсий
- Собственно анализ вариаций
- Апостериорное сравнение групп с помощью специализированных процедур, отличных от Т-критерия

Параметрический дисперсионный анализ

Происходит проверка нулевой гипотезы об отсутствии различий дисперсий в группах

- Если результат свидетельствует об отсутствии различия дисперсий ($p>0,05$), то применение параметрического дисперсионного анализа обосновано
- Если различие дисперсий имеется ($p<0,05$), то применять параметрический дисперсионный анализ не следует

Проверка гипотез о равенстве дисперсий (тест Левена)

Непараметрические методы исследования независимых групп (м-д Краскела-Уоллиса, медианный тест)

- Используется в случае необходимости сопоставить несколько групп по одному количественному или порядковому признаку независимо от вида его распределения в группах

Мощность

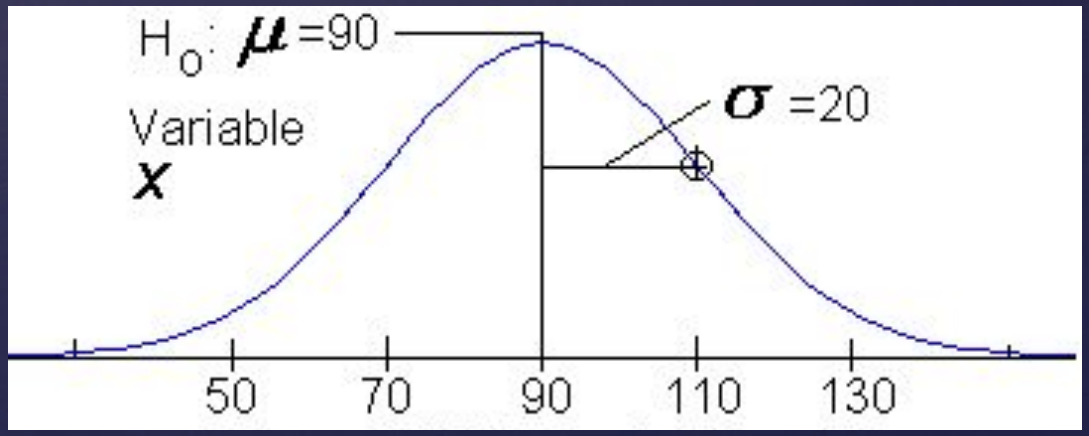
Мощность - вероятность отвергнуть H_0 в эксперименте, когда H_0 действительно неверна.

	Истинное (но неизвестное нам) положение дел	
	Верна H_0	Верна H_1
Мы «приняли» H_0	ПРАВИЛЬНО!	ОШИБКА 2-го рода = β
Мы отвергли H_0	ОШИБКА 1-го рода = α	ПРАВИЛЬНО! **мощность** критерия = 1-β

Мощность

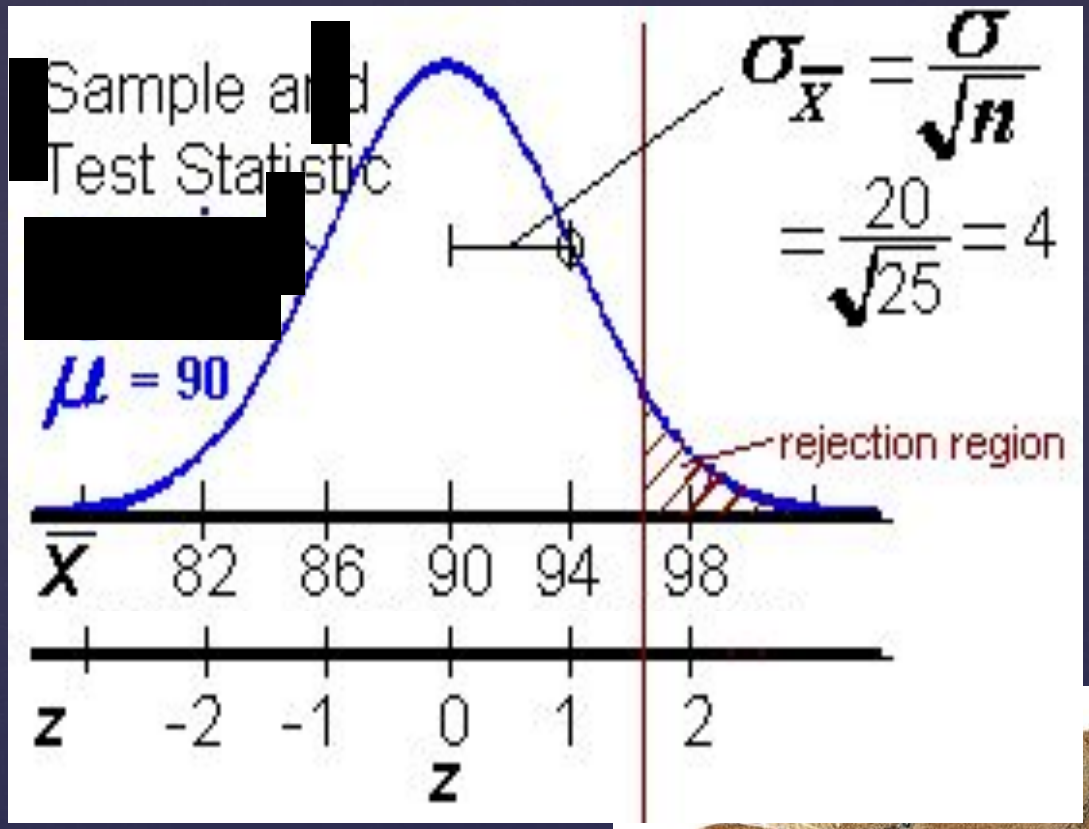
Мощность предполагаемого статистического теста - ключевой элемент планирования исследования

«Реальное значение» параметра:
Во всей мировой популяции землероек μ = 90 г.
Пусть «реальное значение» средней массы в заповеднике = 94 г.



Мощность

Нарисуем распределения выборочных средних для $\mu = 90$ и $\mu = 94$ (стандартное отклонение $\sigma = 20$).



Размер выборки n = 25 зверей

Как увеличить мощность?

Большей МОЩНОСТИ критерия способствуют:

1. Большой размер выборки;
2. Большие различия между популяциями (effect size);
3. Маленькое стандартное отклонение;
4. Большой уровень значимости (α=0.05 а не α=0.01);
5. Выбор одностороннего теста вместо двустороннего

Базовая модель

Математическая основа базовой модели:

$$SS_{общ} = SS_A + SS_B + SS_{ост}$$

Где SS – это сумма квадратов отклонений от среднего.

Рассмотрим случай, когда комбинация определенных значений A и B встречается у равного количества человек r, число возможных значений B равно b и число возможных значений A равно a. (сбалансированная модель).

Уровень фактора A \ Уровень фактора B	B_1	B_2	...	B_j	...	B_b
A_1	x_{111} x_{112} ... x_{11r}	x_{121} x_{122} ... x_{12r}	...	x_{1j1} x_{1j2} ... x_{1jr}	...	x_{1b1} x_{1b2} ... x_{1br}
A_2	x_{211} x_{212} ... x_{21r}	x_{221} x_{222} ... x_{22r}	...	x_{2j1} x_{2j2} ... x_{2jr}	...	x_{2b1} x_{2b2} ... x_{2br}
...	...	...	...	...	...	...
A_i	x_{i11} x_{i12} ... x_{i1r}	x_{i21} x_{i22} ... x_{i2r}	...	x_{ij1} x_{ij2} ... x_{ijr}	...	x_{ib1} x_{ib2} ... x_{ibr}
...	...	...	...	...	...	...
A_a	x_{a11} x_{a12} ... x_{a1r}	x_{a21} x_{a22} ... x_{a2r}	...	x_{aj1} x_{aj2} ... x_{ajr}	...	x_{ab1} x_{ab2} ... x_{abr}

Базовая модель

Тогда общее число человек в выборке

$$n = a \times b \times r$$

$$\begin{cases} SS_A = br\sum_{i=1}^{a}(x_{i\cdot} - x)^2 \\ SS_B = ar\sum_{j=1}^{b}(\bar{x}_{\cdot j} - \bar{x})^2 \end{cases}$$

$$SS_{ост} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}(x_{ijk} - \bar{x}_{ij})^2$$

Базовая модель

В основе лежит все та же основная модель дисперсионного анализа, что и в случае однофакторной статистики, только теперь мы изучаем действие двух или более факторов:

$$x = m + a + \beta + \ldots + \varepsilon$$

Базовая модель

Источник вариации	SS	df	MS	F
Общий	$SS_{общ}$	abr-1	$MS_{общ}$	
Фактор A	SS_A	a-1	MS_A	$MS_A / MS_{ост}$
Фактор B	SS_B	b-1	MS_B	$MS_B / MS_{ост}$
Главные эффекты	$SS_{мод} = SS_A + SS_B$	a + b -2	$MS_{мод}$	$MS_{мод} / MS_{ост}$
Случайные отклонения	$SS_{ост}$	ab(r-1)	$MS_{ост}$	

Модель с эффектом взаимодействия

Эффект взаимодействия предусматривает то, что **дисперсия общего влияния факторов** не равна простой **сумме их дисперсий**:

$$SS_{общ} = SS_A + SS_B + SS_{AB} + SS_{ост}$$

Вводится еще один компонент – взаимодействие A и B.

$$SS_{AB} = \sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{x}_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} - \bar{x})^2$$

Модель с эффектом взаимодействия

Источник вариации	SS	df	MS	F
Общий	$SS_{общ}$	abr-1	$MS_{общ}$	
Фактор A	SS_A	a-1	MS_A	$MS_A / MS_{ост}$
Фактор B	SS_B	b-1	MS_B	$MS_B / MS_{ост}$
Взаимодействие A и B	SS_{AB}	(a -1)(b -1)	MS_{AB}	$MS_{AB} / MS_{ост}$
Случайные отклонения	$SS_{ост}$	ab(r-1)	$MS_{ост}$	

Модель со случайными эффектами

Случайные факторы предусматривают другой подход к вычислению компонентов дисперсии. Если все факторы случайны, то в модели

$$\{ \quad x = m + a + b + e$$

при справедливости **нулевой гипотезы** a, b и e распределены **нормально** со **средним = 0** и разными дисперсиями.

Модель со случайными эффектами

Источник вариации	SS	df	MS	F
Общий	$SS_{общ}$	abr-1	$MS_{общ}$	
Между значениями фактора А	SS_A	a-1	MS_A	MS_A / MS_B
Между значениями фактора В при разных А	SS_B	a(b-1)	MS_B	$MS_B / MS_{ост}$
Случайные отклонения	$SS_{ост}$	ab(r-1)	$MS_{ост}$	

Модель со случайными эффектами

Поскольку подход к SS_B иной, рассчитывается он тоже по-другому:

$$SS_{AB} = r\sum_{i=1}^{a}\sum_{j=1}^{b}(x_{ij} - x_{i.})^2$$

Если в модели со случайными эффектами есть взаимодействия, их дисперсия считается так же, как и в модели постоянных эффектов.

Модель с несколькими эффектами

Чем больше факторов в модели, тем сложнее ее расчет и построение.

Так, например, если в модели три фактора, то оценка влияния одного фактора на модель в целом можно провести только после исключения его взаимодействия с другими факторами:

$MS_{ABC} / MS_{ост}$ – взаимодействие всех факторов

MS_{AB} / MS_{ABC} - взаимодействие двух факторов

MS_{AC} / MS_{ABC} - взаимодействие двух факторов

Немного терминологии

Уровень (level) – это одно из возможных значений фактора. В англоязычной литературе фактор принято обозначать в виде его номера и количества уровней: 2x2, 3x4 и т.п.

Ячейка/гнездо (cell) – это группа значений при заданной комбинации факторов (например, ячейка A=1, B=2, C=10)

Немного терминологии

Полный перекрестный дизайн (Completely crossed design) – каждый уровень каждого фактора встречается в комбинации со всеми уровнями остальных факторов.

Сбалансированный дизайн (balanced design) - в каждой ячейке равное количество значений.

Ортогональный дизайн (orthogonal design) – сбалансированный, полный перекрестный дизайн при условии случайной выборки.

Простой пример

Изучаются 2 фактора, влияющих на сдачу экзамена:

- Употребление кофе (да/нет)
- Наличие конспекта (да/нет)

Результат оценивается в количестве правильных ответов на вопросы единого междисциплинарного теста.

	Конспект (Фактор А)	
Кофеин (Фактор В)	Нет	Да
Да	Только кофеин	Оба
Нет	Контроль (ни одного)	Только конспект

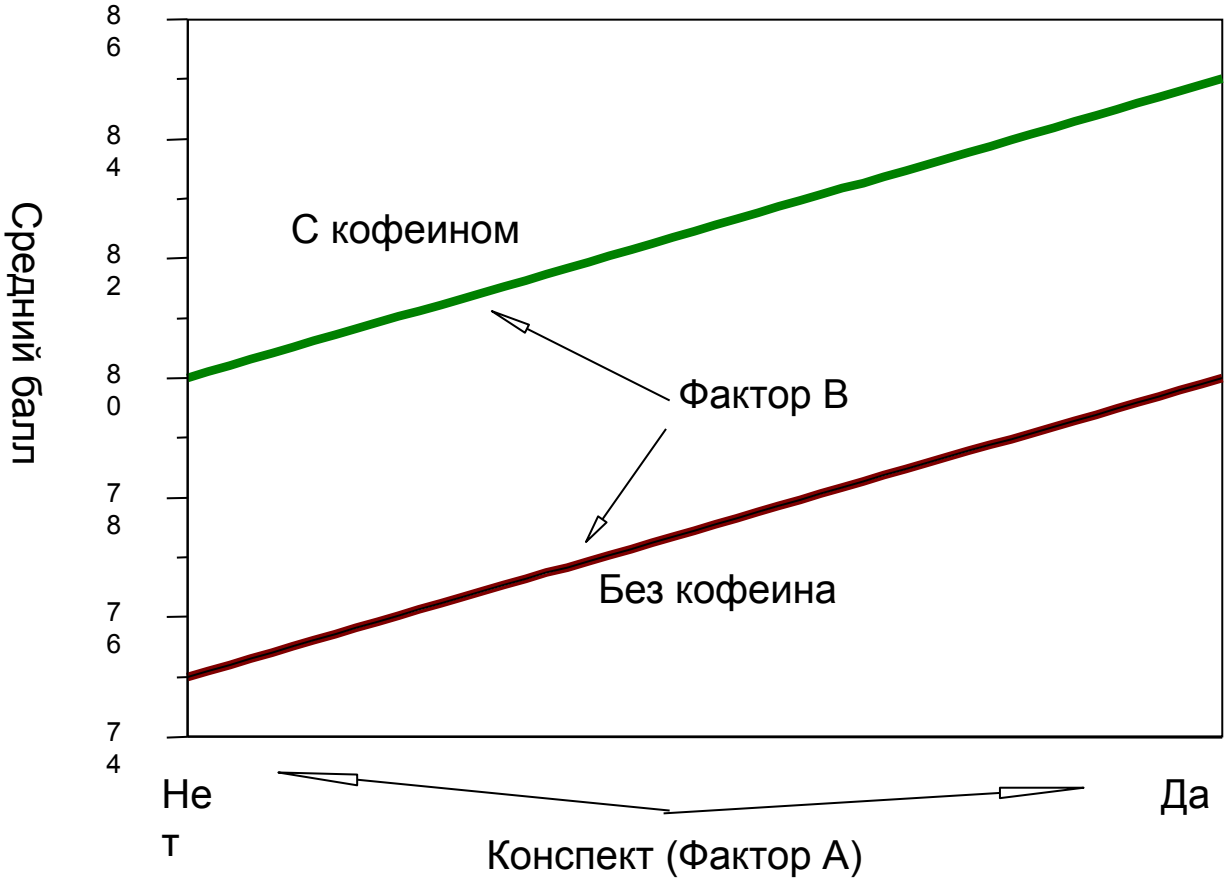
Простой пример

Основные эффекты:

N= по 30 в клетке	Конспект (Фактор А)		Средние по столбцам
Кофеин (Фактор В)	Нет	Да	
Да {	Кофеин Ср.балл = 80 СО = 5	Оба Ср.балл = 85 СО=5	82.5
Нет	Контроль Ср.балл = 75 СО = 5	Конспект Ср.балл = 80 СО = 5	77.5
Средние по строкам	77.5	82.5	80

Простой пример

Основные эффекты и их взаимодействие



Простой пример

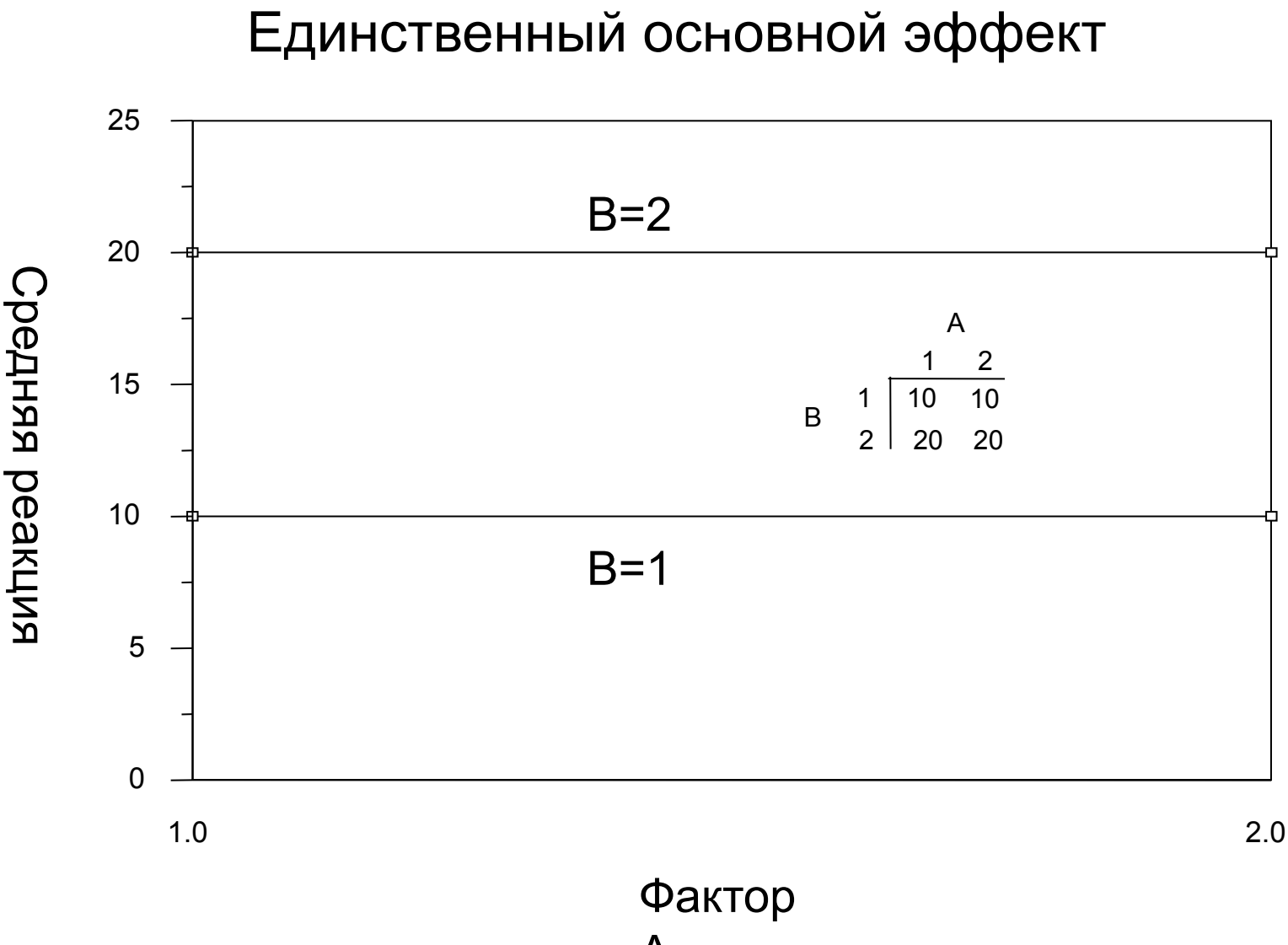
Основные эффекты и их взаимодействие

Эффекты факторов видны по наклону линий на графике (первый эффект) и точках пересечения линий с вертикальной осью (второй эффект)

Взаимодействие факторов проявляется в виде нарушения параллельности линий на графике.

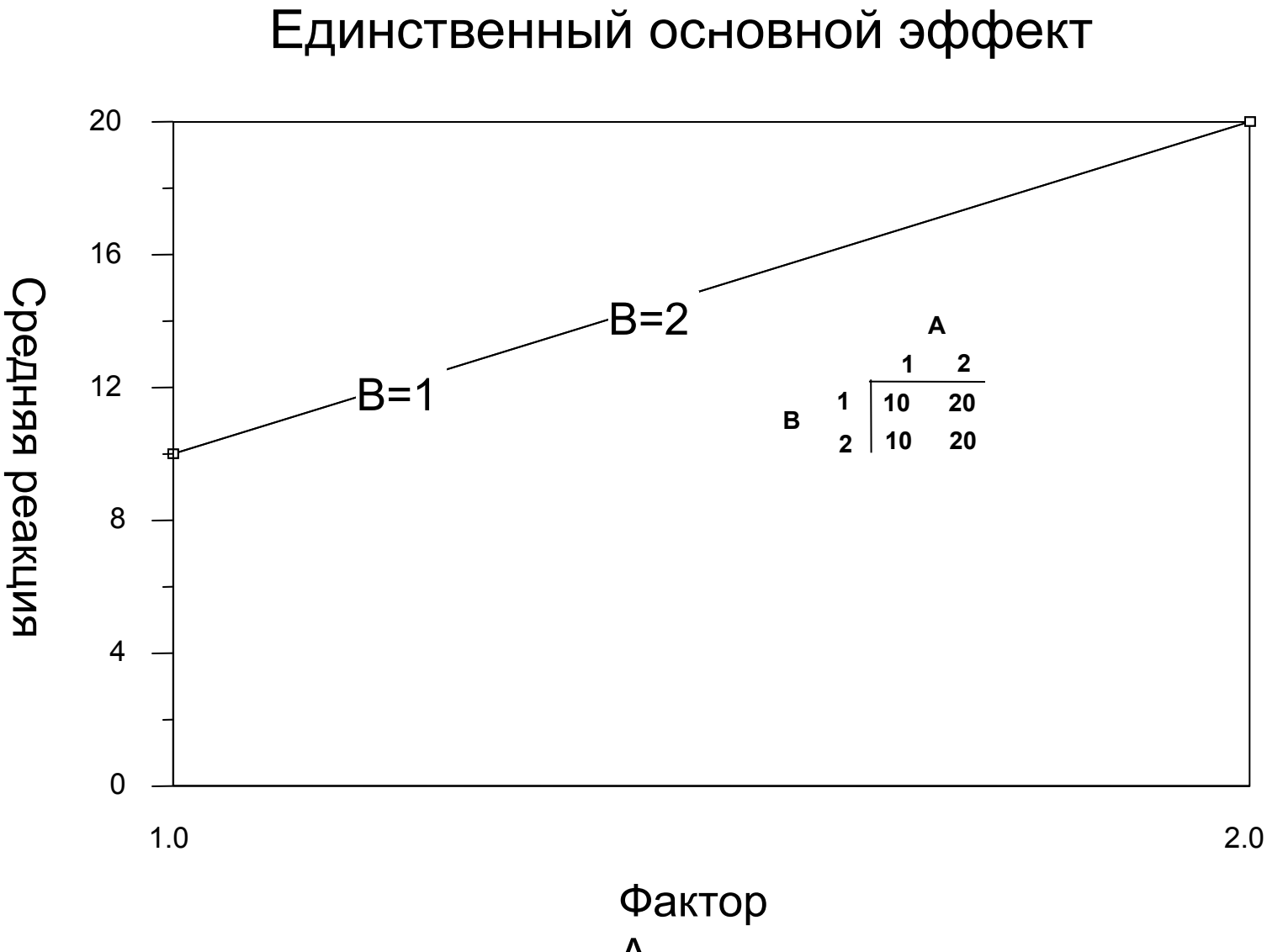
Простой пример

Единственный основной эффект за счет B (только кофе)



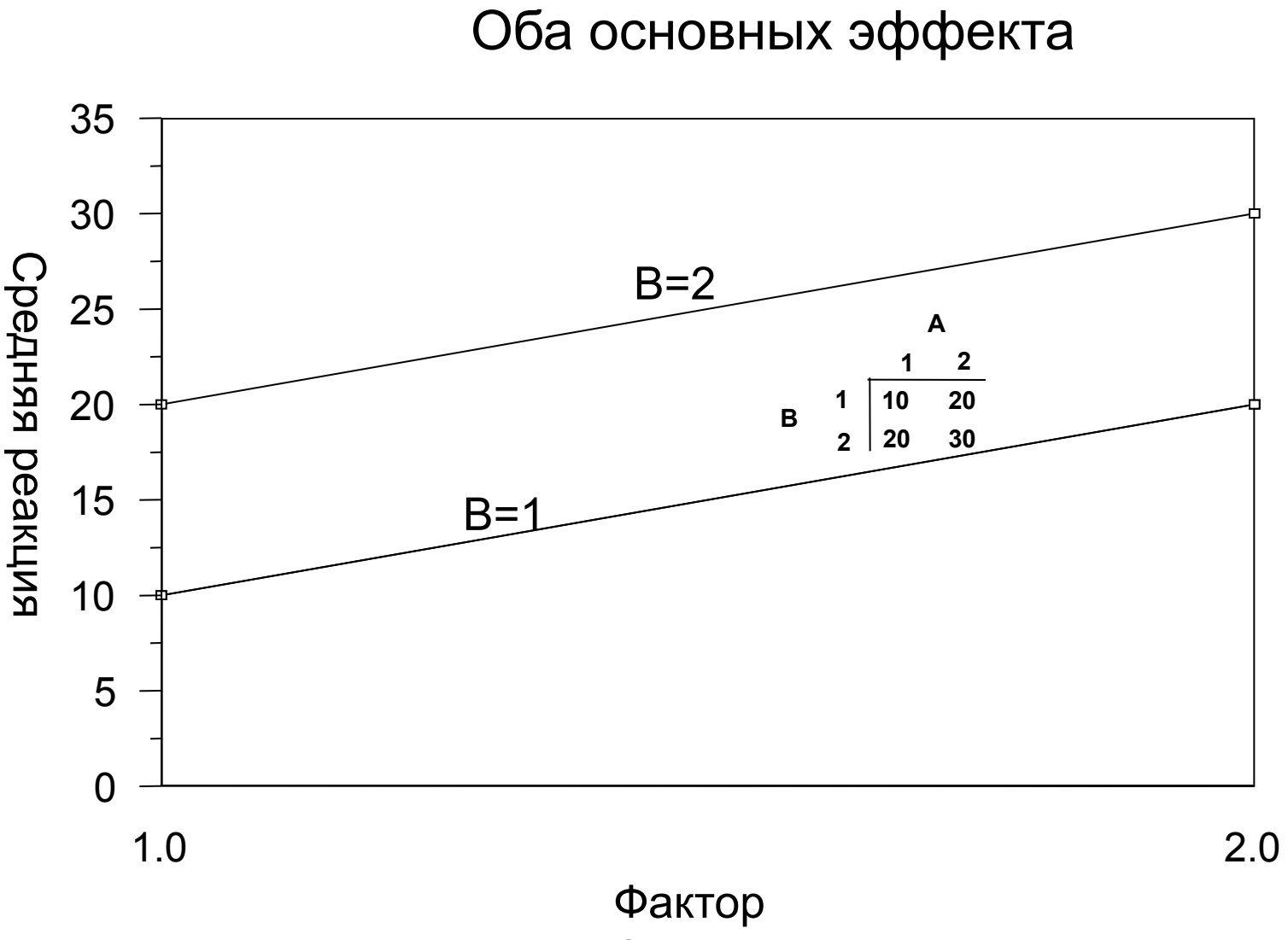
Простой пример

Единственный основной эффект за счет A (только конспект)



Простой пример

Оба основных эффекта A и B (кофе и конспект)



Однофакторный дисперсионный анализ

Рассмотрим оценки различных дисперсий, возникающие при анализе таблицы результатов наблюдений. Для оценки дисперсии, характеризующей изменение данных на уровне Ai (по строкам таблицы), имеем:

$$S_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_{ij} - \overline{x_i})^2 = \frac{1}{n-1}\left[\sum_{j=1}^{n} x_{ij}^2 - \frac{1}{n}\left(\sum_{j=1}^{n} x_{ij}\right)^2\right].$$

Из предпосылок дисперсионного анализа следует, что должно иметь место равенство всех дисперсий. При выполнении этого условия находим оценку дисперсии, характеризующей рассеяние значений x_{ij} вне влияния фактора A, по формуле:

$$S_0^2 = \frac{1}{k}\sum_{i=1}^{k} S_i^2 = \frac{1}{k(n-1)}\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij} - \overline{x_i})^2 = \frac{1}{k(n-1)}\left[\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}^2 - \frac{1}{n}\sum_{i=1}^{k}\left(\sum_{j=1}^{n} x_{ij}\right)^2\right]$$

Однофакторный дисперсионный анализ

Для упрощения вычислений приведем алгоритм их выполнения. Вычисляем последовательно суммы:

$$Q_1 = \sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}^2 \qquad Q_2 = \frac{1}{n}\sum_{i=1}^{k} X_i^2 \qquad Q_3 = \frac{1}{kn}\left(\sum_{i=1}^{k} X_i\right)^2$$

$$S_0^2 = \frac{Q_1 - Q_2}{k(n-1)} \qquad S_A^2 = \frac{Q_2 - Q_3}{k-1}$$

Сравниваем S_A^2 S_0^2 и устанавливаем наличие влияния фактора А.

Если $\frac{k(n-1)}{k-1}\frac{Q_2 - Q_3}{Q_1 - Q_2} > F_\alpha[k-1; k(n-1)]$, то влияние А – значимо.

Двухфакторный дисперсионный анализ

Рассмотренный ранее однофакторный дисперсионный анализ обладает информативностью, не большей, чем методы множественного сравнения средних. Информативность дисперсионного анализа возрастает при одновременном изучении влияния нескольких факторов.

Рассмотрим случай, когда анализируется влияние одновременно двух факторов A и B.

Двухфакторный дисперсионный анализ

Пусть результаты эксперимента представлены таблицей:

B	Уровни фактора A						Σ
	A_1	A_2	…	A_i	…	A_k	
B_1	x_{11}	x_{21}	…	x_{i1}	…	x_{k1}	X_1'
B_2	x_{12}	x_{22}	…	x_{i2}	…	x_{k2}	X_2'
….	…	…	…	…	…	…	…
B_j	x_{1j}	x_{2j}	…	x_{ij}	…	x_{kj}	X_i'
…	…	…	…	…	…	…	…
B_m	x_{1n}	X_{2n}	…	x_{in}	…	x_{kn}	X_m'
Σ	X_1	X_2	…	X_i	…	X_n	

Двухфакторный дисперсионный анализ

Дисперсионный анализ для двухфакторных таблиц проводится в следующей последовательности. Вычисляются суммы:

$$Q_1 = \sum_{i=1}^{k}\sum_{j=1}^{m} x_{ij}^2 \qquad Q_2 = \frac{1}{m}\sum_{i=1}^{k} X_i^2 \qquad Q_3 = \frac{1}{k}\sum_{j=1}^{m} X_j^2 \qquad Q_4 = \frac{1}{mk}\left(\sum_{i=1}^{k} X_i\right)^2 = \frac{1}{mk}\left(\sum_{j=1}^{k} X_{j'}\right)^2$$

Далее находятся оценки дисперсий:

$$S_0^2 = \frac{Q_1 + Q_4 - Q_2 - Q_3}{(k-1)(m-1)} \qquad S_A^2 = \frac{Q_2 - Q_4}{k-1} \qquad S_B^2 = \frac{Q_3 - Q_4}{m-1}$$

Если $\frac{S_A^2}{S_0^2} > F_\alpha(f_1, f_2)$, то влияние фактора А признается значимым.

Если $\frac{S_B^2}{S_0^2} > F_\alpha(f_1, f_2)$, то влияние фактора В признается значимым.

Двухфакторный дисперсионный анализ

Приведенный анализ предполагает независимость факторов A и B. Если они зависимы, то взаимодействие факторов C=AB также является фактором, которому соответствует своя дисперсия. Для того чтобы выделить такое взаимодействие, необходимы параллельные наблюдения в каждой клетке таблицы, т.е. при каждом сочетании факторов A и B на уровнях A_i и B_j соответственно необходимо не одно наблюдение, а серия наблюдений.

Для оценки влияния взаимодействия факторов AB вычисляем дополнительную сумму: $Q_5 = \sum_{i=1}^{k}\sum_{j=1}^{m}\sum_{v=1}^{n} x_{ijv}^2$

Далее анализ проводится, как и ранее, с той лишь разницей, что в клетках таблицы вместо отдельных значений используется их средние значения. Вычисляется оценка дисперсии и проверяется значимость взаимодействия факторов:

$$S_{AB}^2 = \frac{Q_5 - nQ_1}{mk(n-1)} \qquad \frac{nS_0^2}{S_{AB}^2} > F_\alpha(f_1, f_2) \qquad f_1 = (k-1)(m-1) \qquad f_2 = mk(n-1)$$

Планирование эксперимента при дисперсионном анализе

Дисперсионный анализ тесно связан с соответствующим планированием эксперимента. Удачно спланированный эксперимент, выявляя все необходимые эффекты, оказывается всегда либо более точным, либо менее трудоемким по сравнению с непродуманным экспериментом.

Если на результат эксперимента действуют одновременно несколько факторов, то наилучший эффект дает одновременный дисперсионный анализ всех этих факторов (многофакторный анализ).

Методы дисперсионного анализа позволяют исследовать и такой случай, когда некоторые сочетания уровней пропущены. Такой эксперимент называется дробным факторным экспериментом (ДФЭ). Планирование при ДФЭ приобретает особо важную роль, ибо пропущенные сочетания уровней не так-то просто нейтрализовать.

Планирование эксперимента при дисперсионном анализе

Такие способы планирования существуют и притом не единственные; согласно Фишеру их называют латинскими квадратами. Эти расположения приводятся в специальных справочниках; для примера приведен один вид такого квадрата:

	A_1	A_2	…	A_{k-1}	A_k
B_1	C_1	C_2	…	C_{k-1}	C_k
B_2	C_2	C_3	…	C_k	C_1
…	…	…	…	…	…
B_{k-1}	C_{k-1}	C_k	…	C_{k-3}	C_{k-2}
B_k	C_k	C_1	…	C_{k-2}	C_{k-1}

Планирование эксперимента при дисперсионном анализе

Схема расчетов для латинского квадрата очень похожа на обычный двухфакторный анализ:

$$Q_1 = \sum_{i=1}^{k}\sum_{j=1}^{k} x_{ij}^2$$

Находим сумму квадратов по столбцам, деленную на число наблюдений в столбце:

$$Q_2 = \frac{1}{k}\sum_{i=1}^{k} X_i^2$$

Находим сумму квадратов итогов по строкам, деленную на число наблюдений в строке:

$$Q_3 = \frac{1}{k}\sum_{j=1}^{k} X_j'^2$$

Находим квадрат общего итога, деленный на число всех наблюдений:

$$Q_4 = \frac{1}{k^2}\left(\sum_{i=1}^{k} X_i\right)^2 = \frac{1}{k^2}\left(\sum_{j=1}^{k} X_j'\right)^2$$

Находим сумму квадратов итогов по уровням фактора C, деленную на число уровней:

$$Q_5 = \frac{1}{k}\sum_{\nu=1}^{k} Y_\nu^2$$

Планирование эксперимента при дисперсионном анализе

Перейдем теперь к вычислению и оценке значимости дисперсий:

$$S_0^2 = \frac{Q_1 + 2Q_4 - Q_2 - Q_3 - Q_5}{(k-1)(k-2)}$$

$$S_A^2 = \frac{Q_2 - Q_4}{k-1},\ S_B^2 = \frac{Q_3 - Q_4}{k-1}$$

Если отличие будет значимым, то $\frac{S_A^2 - S_0^2}{k} \Rightarrow \sigma_A^2,\ \frac{S_B^2 - S_0^2}{k} \Rightarrow \sigma_B^2$

$$S_C^2 = \frac{Q_5 - Q_4}{k-1}$$

Если отличие будет значимым, то $\frac{S_C^2 - S_0^2}{k} \Rightarrow \sigma_C^2$