УНИВЕРСИТЕТ ИТМО

Основы анализа данных.

Дисперсионный анализ

Лекция 8

КМАИ.

Секции

Понятие и назначение дисперсионного анализа

Постановка задачи дисперсионного анализа

Однофакторный дисперсионный анализ

Априорные контрасты и апостериорные критерии

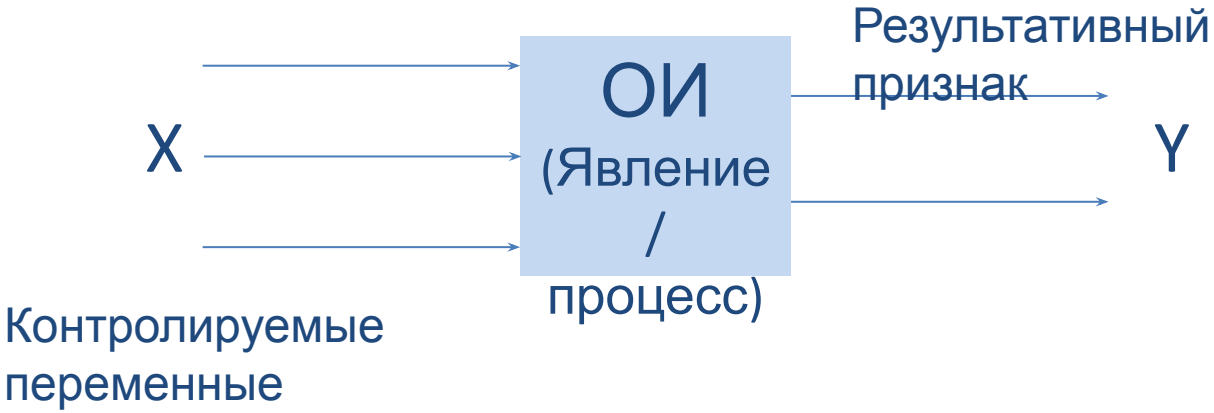
Многофакторный дисперсионный анализ

Условия применения дисперсионного анализа

- ✔ Количественный непрерывный тип данных, дискретные данные менее желательны.
- ✔ Независимые между собой выборки.
- ✔ Нормальное распределение признака в статистических совокупностях, из которых извлечены выборки.
- ✔ Равенство (гомогенность) дисперсий изучаемого признака в статистических совокупностях из которых извлечены выборки.
- ✔ Независимые наблюдения в каждой из выборок.

Дисперсионный анализ

анализ изменчивости результативного признака под влиянием каких-либо контролируемых переменных факторов. (В зарубежной литературе именуется ANOVA – «Analisis of Variance»)

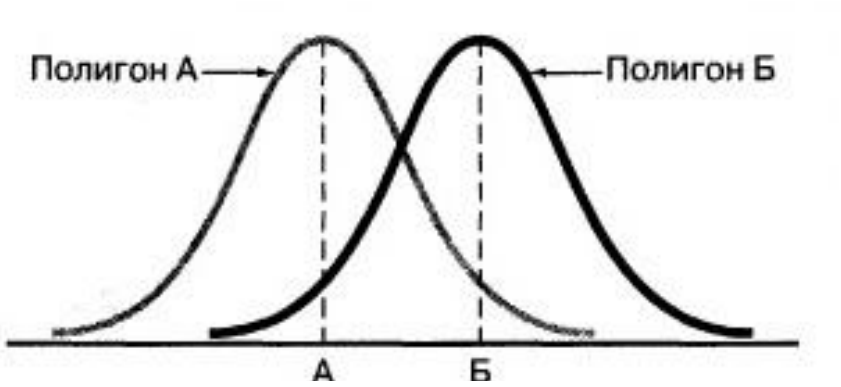


$$DY \sim DX$$

Формулировка гипотез в дисперсионном анализе

Нулевая гипотеза:
«Средние величины результативного признака во всех условиях действия фактора (или градациях фактора) одинаковы».

Альтернативная гипотеза:
«Средние величины результативного признака в разных условиях действия фактора различны».



Формулировка гипотез в дисперсионном анализе

Дисперсионный анализ можно подразделить на несколько категорий в зависимости:

1. от количества рассматриваемых независимых факторов;
2. от количества результативных переменных, подверженных действию факторов;
3. от характера, природы получения и наличия взаимосвязи сравниваемых выборок значений.

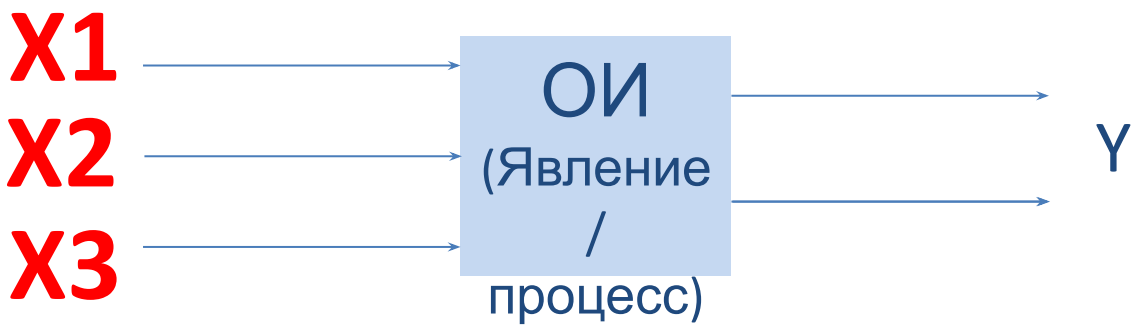
Формулировка гипотез в дисперсионном анализе

Дисперсионный анализ:

от количества рассматриваемых независимых факторов;

1. Однофакторный;

2. Многофакторный.



Формулировка гипотез в дисперсионном анализе

Дисперсионный анализ:

от количества рассматриваемых независимых факторов;

1. Однофакторный;

1.1. Анализ несвязанных (то есть – различных) выборок $corr(X_{\text{гр1}}, X_{\text{гр2}}) = 0$

1.2. Анализ связанных выборок $corr(X_{\text{гр1}}, X_{\text{гр2}}) \neq 0$

Формулировка гипотез в дисперсионном анализе

Дисперсионный анализ:

от количества результативных переменных, подверженных действию факторов;

1. Одномерный;

1. Многомерный.

Задача дисперсионного анализа

Определить:

1. вариативность, обусловленную действием каждой из исследуемых независимых переменных (факторов);

$$F_{X1} = \frac{var(Y|X1)}{var(Y)}$$

2. вариативность, обусловленную взаимодействием исследуемых независимых переменных;

$$F_{X2} = \frac{var(Y|X2)}{var(Y)}$$

3. вариативность случайную, обусловленную всеми неучтенными обстоятельствами

$$F_{X1,X2} = \frac{var(Y|var(X1|X2))}{var(Y)}$$

Пример медиального критерия

Из I: 21, 50, 6, 69, 42, 34, 26, 57, 14, 31
Из II: 10, 49, 22, 40, 24, 54, 12, 29, 25, 17, 32, 61
Из III: 3, 15, 9, 18, 1, 33, 11, 5, 16, 30, 41

Медиана по всем выборкам =

	Больше медианы	Меньше медианы	Всего
Выборка I	7	3	10
Выборка II	6	5	11
Выборка III	3	8	11
Всего	16	16	32

Пример медиального критерия

Ожидаемые наблюдения

	Больше меднаны	Меньше меднаны
Выборка I	5	5
Выборка II	5,5	5,5
Выборка III	5,5	5,5

Критерий согласия

$$X^2 = \sum_{\text{по всем клеткам}} \frac{(\text{число наблюдений})^2}{\text{ожидаемое число наблюдений}} - n$$

Пример медиального критерия

Значения критерия согласия

$$X^2 = \sum_{\text{по всем клеткам}} \frac{(\text{число наблюдений})^2}{\text{ожидаемое число наблюдений}} - n$$

$$X^2 = \frac{7^2}{5} + \frac{3^2}{5} + \frac{6^2}{5{,}5} + \frac{5^2}{5{,}5} + \frac{3^2}{5{,}5} + \frac{6^2}{5{,}5} + \frac{5^2}{5{,}5} + \frac{3^2}{5{,}5} + \frac{8^2}{5{,}5} - 32 = 3{,}96.$$

Табличное значение критерия согласия

$$A_2 = \{\text{наблюдения: } X^2 > \chi^2_{0{,}95}(2) = 5{,}99\},$$

Задание:

1. Разбиться на 3 команды и проверить зависимость выборок успеваемости по дисциплинам первого семестра:

   Урбанистика

   НИРС

   Социология города

   Экономика города

2. Разбиться на 2 команды (М/Ж) и проверить тоже самое для 98% доверительного интервала.

Секции

Понятие и назначение дисперсионного анализа

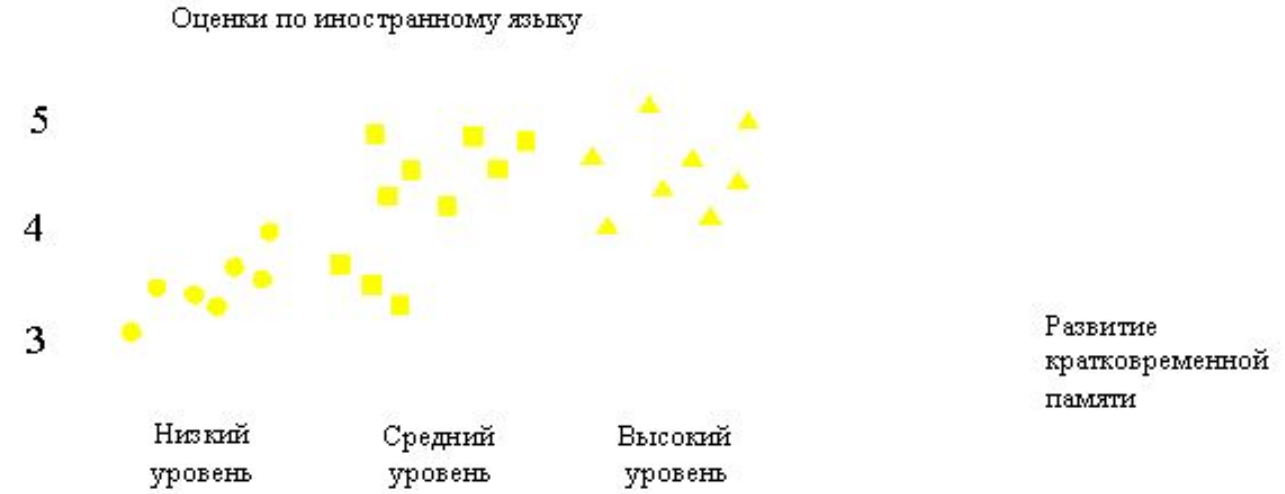
Постановка задачи дисперсионного анализа

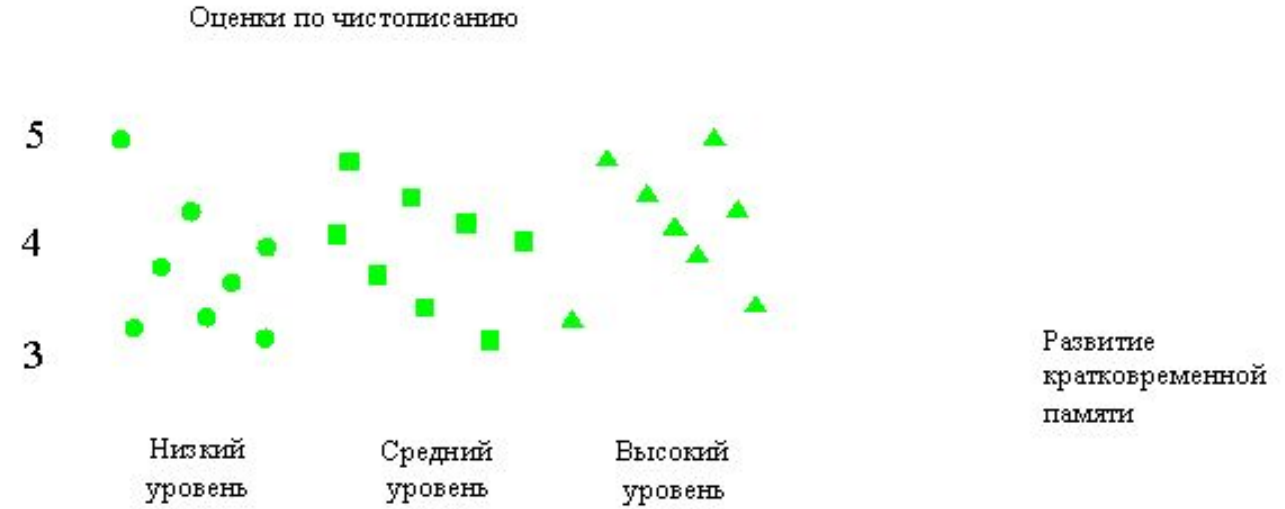
Однофакторный дисперсионный анализ

Априорные контрасты и апостериорные критерии

Многофакторный дисперсионный анализ

Постановка задачи



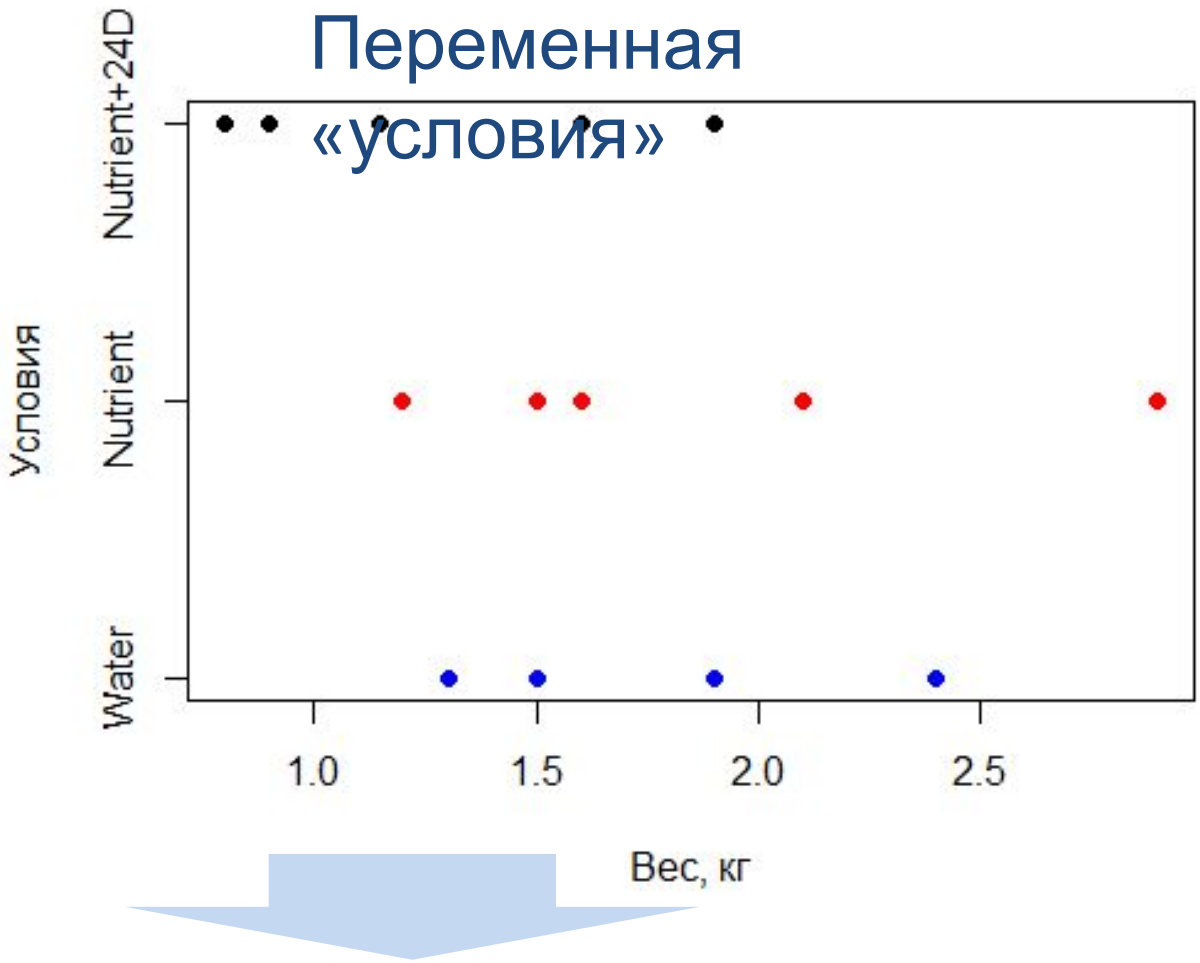


Влияние кратковременной памяти на успеваемость

Постановка задачи

№	вес	условия
1	1.50	Water
2	1.90	Water
3	1.30	Water
4	1.50	Water
5	2.40	Water
6	1.50	Water
7	1.50	Nutrient
8	1.20	Nutrient
9	1.20	Nutrient
10	2.10	Nutrient
11	2.90	Nutrient
12	1.60	Nutrient
13	1.90	Nutrient+24D
14	1.60	Nutrient+24D
15	0.80	Nutrient+24D
16	1.15	Nutrient+24D
17	0.90	Nutrient+24D
18	1.60	Nutrient+24D

Переменная «условия»



Условия	Water	Nutrient	Nutrient+24D
Средний вес, кг	1.683333	1.750000	1.325000

Постановка задачи

Условия	Water	Nutrient	Nutrient+24D
Средний вес, кг	1.683333	1.750000	1.325000

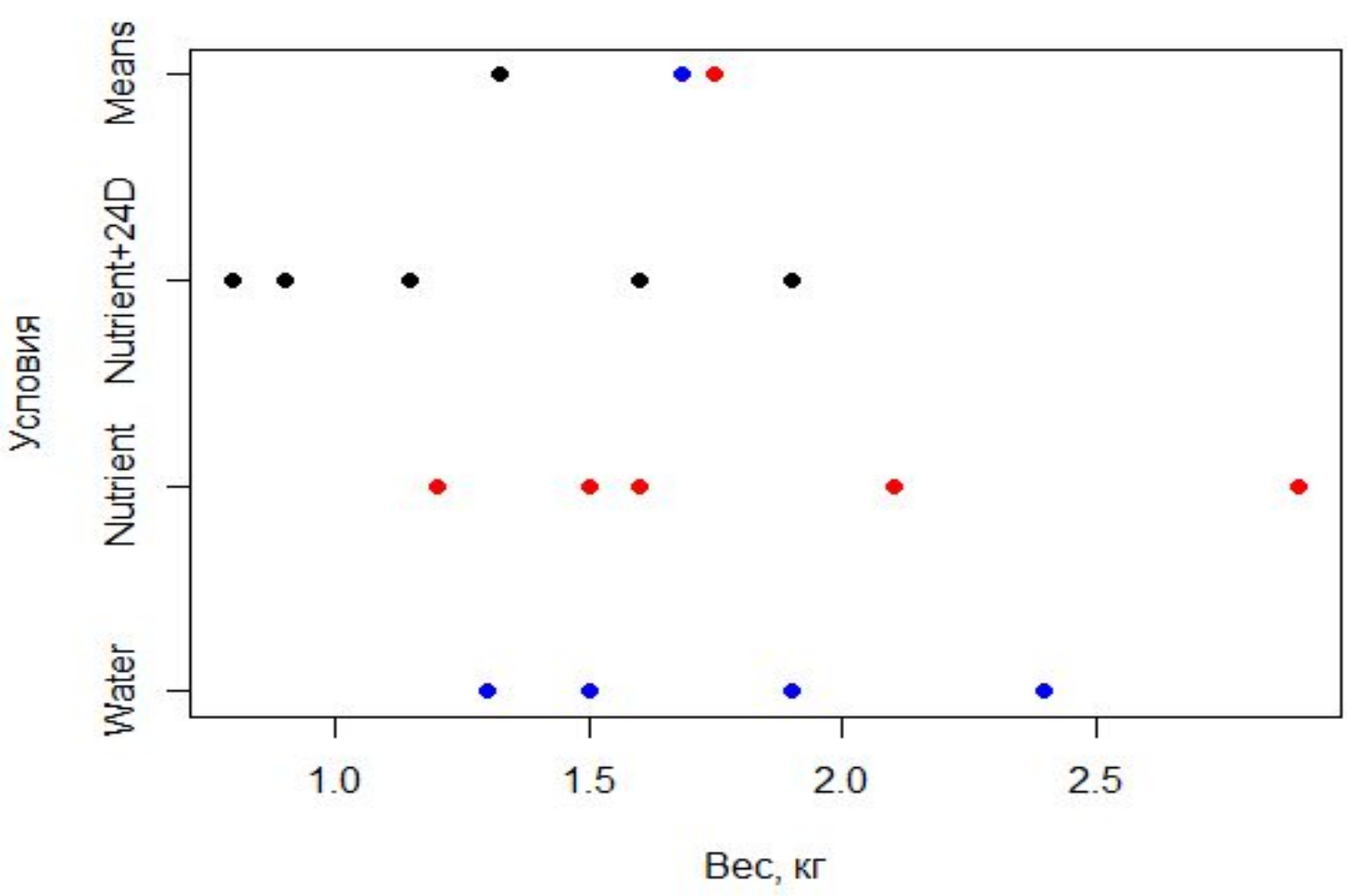
H0: исследованные условия выращивания растений не оказывают никакого влияния на вес последних.

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

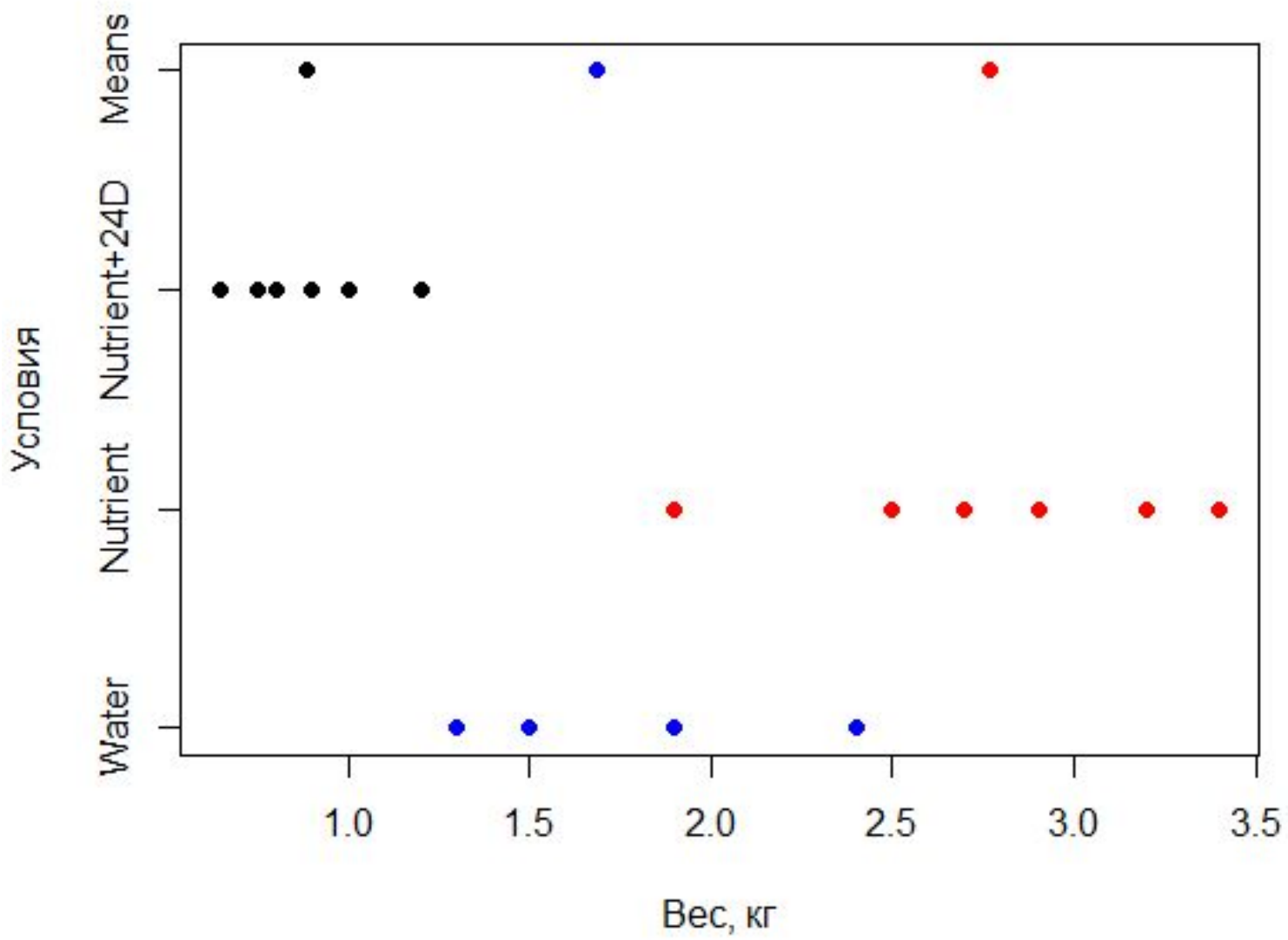
H1: исследованные условия выращивания растений оказывают влияние на вес последних.

Постановка задачи

Условия	Water	Nutrient	Nutrient+24D
Средний вес, кг	1.683333	1.750000	1.325000



Постановка задачи



Секции

Понятие и назначение дисперсионного анализа

Постановка задачи дисперсионного анализа

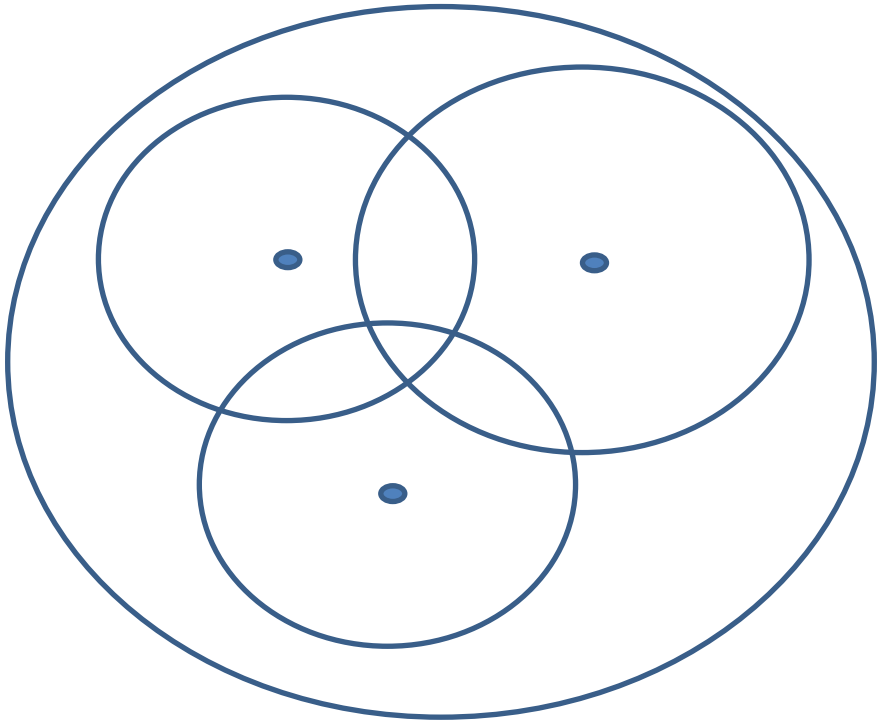
Однофакторный дисперсионный анализ

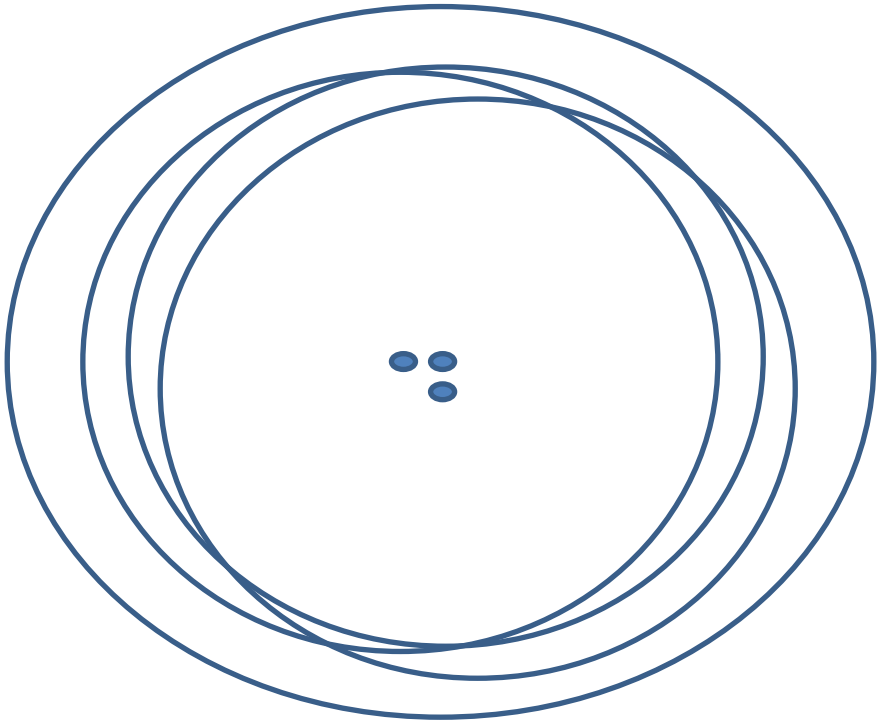
Априорные контрасты и апостериорные критерии

Многофакторный дисперсионный анализ

Однофакторный дисперсионный анализ

Дисперсионный анализ, который рассматривает только одну независимую переменную называется **однофакторным дисперсионным анализом** (One-Way ANOVA).





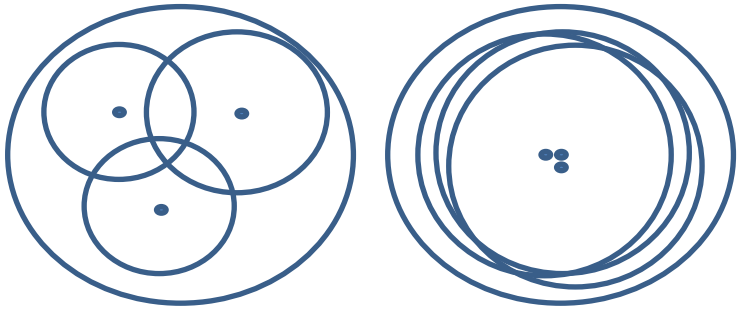
Однофакторный дисперсионный анализ

Процедура дисперсионного анализа состоит в определении соотношения систематической (межгрупповой) дисперсии к случайной (внутригрупповой) дисперсии в измеряемых данных.

Межгрупповая сумма квадратов SS_{BG}

Внутригрупповая сумма квадратов SS_{WG}

Общая сумма квадратов $SS_{Total} = SS_{BG} + SS_{WG}$



Однофакторный дисперсионный анализ

В случае если верна Но, то как внутригрупповая, так и межгрупповая дисперсии служат оценками одной и той же дисперсии и должны быть приблизительно равны.

$$F = \frac{MS_{BG}}{MS_{WG}}$$

$$MS_{BG} = \frac{SS_{BG}}{vBG}$$

$$MS_{wG} = \frac{SS_{wG}}{vWG}$$

Межгрупповое число степеней свободы:

$$vBG = m - 1$$

m – число групп

Внутригрупповое число степеней свободы:

$$vWG = n - m$$

n - число наблюдений в каждой из групп

Данные подготовленные для анализа.

Независимая переменная – фактор

(количество выборок)

	Уровень 1	Уровень 2	…	Уровень M
Измерения признака	X11	X21	---	XM1
	X12	X22	---	XM2
	X13	X23	---	XM3
	…	…	---	…
	X1N	X2N	---	XMN
Объем:	n1	n2	---	nm
Среднее:	MX1	MX2	---	MXM
Ст. отклонение:	SSX1	SSX2		SSXM

Однофакторный дисперсионный анализ

Межгрупповая вариация:

$$SS_{BG} = \sum_{i=1}^{M} n_i(\bar{x}_i - \bar{x})^2 \qquad MS_{BG} = \frac{SS_{BG}}{m-1}$$

Внутригрупповая вариация:

$$SS_{WG} = \sum_{i=1}^{M} \sum_{j} (x_{j,i} - \bar{x}_i)^2 \qquad MS_{WG} = \frac{SS_{WG}}{m-n_i}$$

Результаты анализа.

	Сумма квадратов	Степени свободы	Дисперсия
Между группами:		m-1	
Внутри групп:		n-m	
Общая:		n	

$$F \geq F_{\text{крит}}(\alpha, m-1, n-m)$$

Задание:

1. Разбиться на команды по базовому образованию и проверить зависимость выборок успеваемости по дисциплинам первого семестра для 95% интервала :

Урбанистика

НИРС

Социология города

Экономика города

Секции

Понятие и назначение дисперсионного анализа

Постановка задачи дисперсионного анализа

Однофакторный дисперсионный анализ

Априорные контрасты и апостериорные критерии

Многофакторный дисперсионный анализ

Априорные контрасты и апостериорные критерии

Критерии для сравнения средних значений

Априорные контрасты

- коэффициенты сравниваемых уровней (или комбинаций уровней) должны иметь разные знаки
- коэффициенты уровней, не представляющих интереса, приравниваются нулю

Апостериорные критерии

$$x_i = w_1 MS_{WG,1} + w_2 MS_{WG,2} + \cdots + w_m MS_{WG,m}$$

Однофакторный дисперсионный анализ для связанных выборок (ANOVA с повторными измерениями):

Проверяемые гипотезы:

1. H0(A): Различия независимой величины при разных градациях фактора являются не более выраженными, чем различия, обусловленные случайными причинами.

2. H1(A): Различия независимой величины при разных градациях фактора являются более выраженными, чем различия, обусловленные случайными причинами.

3. Но(Б): Индивидуальные различия между элементами выборки являются не более выраженными, чем различия, обусловленные случайными причинами.

4. Н1(Б): Индивидуальные различия между элементами выборки являются более выраженными, чем различия, обусловленные случайными причинами.

Результаты анализа:

	Сумма квадратов	Степени свободы	Дисперсия	F
Вариация, вызванная влиянием фактора		c-1		
Вариация между элементами выборки		n-1		
Вариация, вызванная случайными причинами		c*n-1		
Общая вариация		c*n-n-c+1		

Повторные измерения

Где:

$$SS_w = \frac{\sum_{i=1}^{c}\left(\sum_{j=1}^{n} x_{j,i}\right)^2}{n} - \frac{\left(\sum_{i=1}^{c}\sum_{j=1}^{n} x_{j,i}\right)^2}{n \cdot c}$$

$$SS_b = \frac{\sum_{j=1}^{n}\left(\sum_{i=1}^{c} x_{j,i}\right)^2}{n} - \frac{\left(\sum_{i=1}^{c}\sum_{j=1}^{n} x_{j,i}\right)^2}{n \cdot c}$$

$$SS = \left(\sum_{i=1}^{c}\sum_{j=1}^{n} \left(x_{j,i}\right)^2\right) - \frac{\left(\sum_{i=1}^{c}\sum_{j=1}^{n} x_{j,i}\right)^2}{n \cdot c}$$

Статистическая проверка гипотезы о наличии различий осуществляется на основании:

$$F_b = \frac{MS_b}{MS_e}$$

$$F_{крит}(\alpha, c-1, c \cdot n - c - n + 1)$$

$$F_w = \frac{MS_w}{MS_e}$$

$$F_{крит}(\alpha, n-1, c \cdot n - c - n + 1)$$

$SS_e = SS - SS_w - SS_b$ - сумма квадратов ошибки

Повторные измерения

Ограничения метода дисперсионного анализа для связанных выборок:

1. Дисперсионный анализ для связанных выборок требует не менее трех градаций фактора и не менее двух элементов выборки в каждой группе.

2. Должно соблюдаться правило равенства дисперсий в каждой группе. Это условие косвенно выполняется за счет одинакового количества наблюдений в каждой группе.

3. Результативный признак должен быть нормально распределен в исследуемой выборке. :

Повторные измерения

Способы реализации однофакторного дисперсионного анализа с повторными измерениями:

1) Одномерная модель основана на предположении, что каждому уровню внутригруппового фактора соответствует повторное измерение одной и той же зависимой переменной (следовательно, эти изменения положительно коррелируют).

1) Многомерная модель свободна от допущения о коррелированности измерений зависимой переменной (т.е. о сферичности).

Секции

Понятие и назначение дисперсионного анализа

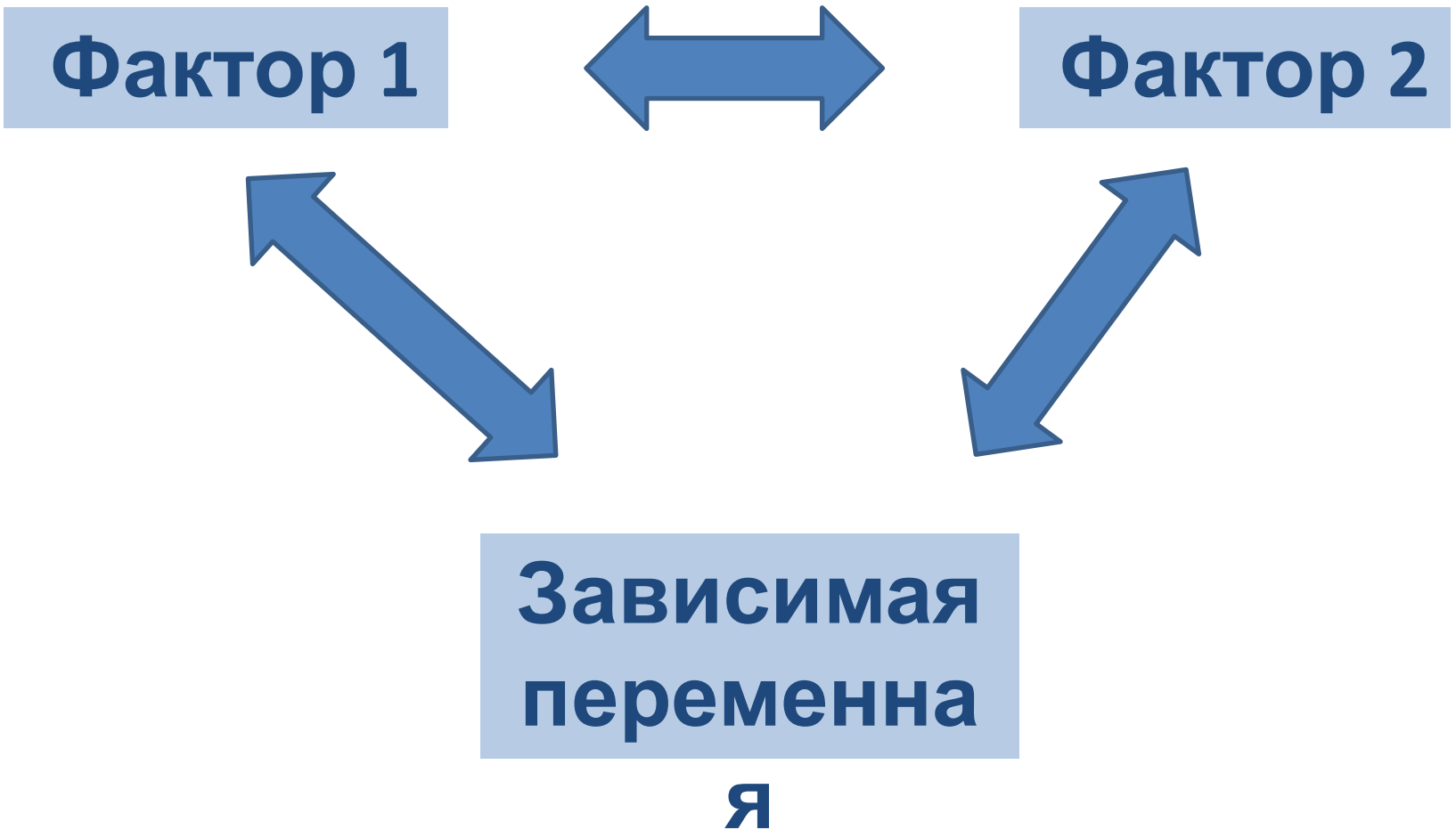
Постановка задачи дисперсионного анализа

Однофакторный дисперсионный анализ

Априорные контрасты и апостериорные критерии

Многофакторный дисперсионный анализ

Многофакторный дисперсионный анализ



Многофакторный дисперсионный анализ

Схема двухфакторного дисперсионного анализа имеет несколько нулевых гипотез:

H0: Фактор 1 и д Фактор 2 **не** имеют эффекта взаимодействия на Зависимую переменную.

H1: Фактор 1 и Фактор 2 имеют эффект взаимодействия на Зависимую переменную.

H0: Зависимая переменная **не** зависит от Фактора 1.

H1: Зависимая переменная зависит от Фактора 1.

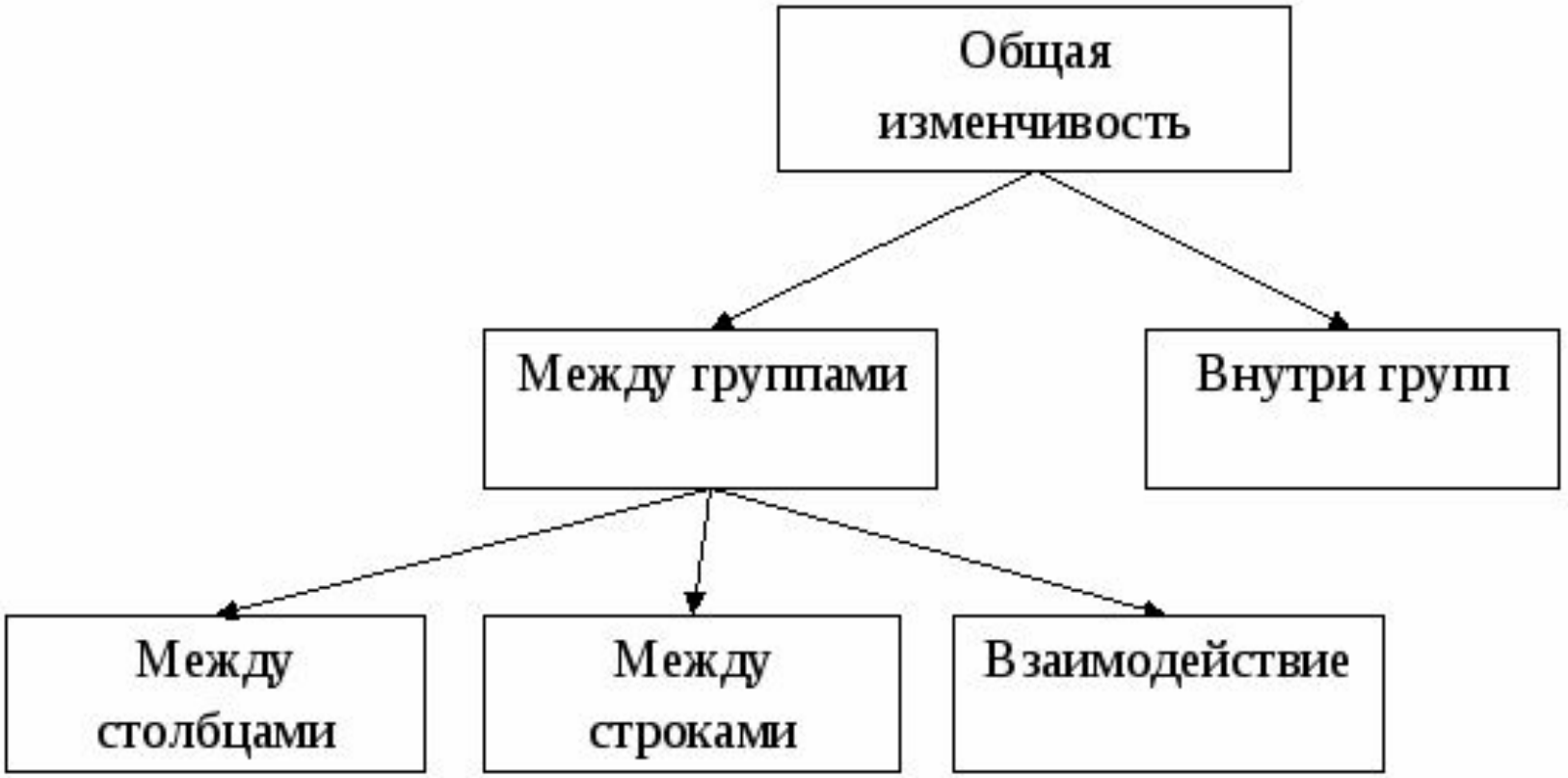
H0: Зависимая переменная **не** зависит от Фактора 2 .

H1: Зависимая переменная зависит от Фактора 2 .

Результаты анализа:

	Сумма квадратов	Степени свободы	Дисперсия	F
Фактор 1		a-1		
Фактор 2		b-1		
Взаимодействие Фактора 1 и Фактора 2		(a-1)*(b-1)		
Ошибка		a*b*(n-1)		
Общая вариация		n		

Общая изменчивость в двухфакторном дисперсионном анализе может быть разложена



Многофакторный дисперсионный анализ

Условия применения:

1. Генеральные совокупности, из которых извлечены выборки, должны быть нормально распределены.

2. Выборки должны быть независимыми.

3. Дисперсии генеральных совокупностей, из которых извлекались выборки, должны быть равными.

4. Группы должны иметь одинаковый объем выборки.

Пример применения:

Необходимо выяснить, оказывают ли влияние тип потребляемого бензина и тип автомобиля на расход топлива. Для этого будут использованы два типа бензина – обычный и высокооктановый, и для каждой группы будут использованы два типа автомобилей – с двумя ведущими колесами и с четырьмя. Для каждой группы будут использованы по два автомобиля, всего восемь.

Пробег автомобиля в милях на галлон:

Топливо	Тип автомобиля	
	два колеса	четыре колеса
Обычное	26,7	28,6
	25,2	29,3
Высокооктановое	32,3	26,1
	32,8	24,2

Алгоритм решения задачи:

1. Сформулировать гипотезы.

2. Найти критическое значение для каждого значения F-критерия при заданном α, например, α = 0,05.

3. Заполнить итоговую таблицу, чтобы получить фактические значения критерия.

4. **Принять решение.**

Формулировка гипотез.

1. для взаимодействия типа топлива и типа автомобиля:

H0: Тип топлива и тип автомобиля не оказывают эффекта взаимодействия на потребление бензина.

H1: Тип топлива и тип автомобиля оказывают эффект взаимодействия на потребление бензина.

2. для типов топлива:

H0: Для двух типов топлива нет разницы между средним потреблением бензина.

H1: Для двух типов топлива существует разница между средним потреблением бензина.

3. для типов автомобилей:

H0: Для автомобилей с двумя и четырьмя ведущими колесами нет разницы в среднем потреблении бензина.

H1: Для автомобилей с двумя и четырьмя ведущими колесами существует разница в среднем потреблении бензина.

Многофакторный дисперсионный анализ

Каждая независимая переменная имеет два уровня:

Фактор А - тип топлива: обычное и высокооктановое, a = 2.
Фактор В - тип автомобиля: также имеет два значения, b = 2.
Число объектов в каждой группе, n = 2.

Степени свободы для каждого фактора:

фактор А $df_A = a - 1 = 2 - 1 = 1$

фактор В $df_B = b - 1 = 2 - 1 = 1$

взаимодействие (А×В) $df_{A\times B} = (a-1)\cdot(b-1) = (2-1)\cdot(2-1) = 1$

ошибка внутри группы $df_{error} = a \cdot b \cdot (n-1) = 2 \cdot 2 \cdot (2-1) = 4$

Критические значения:

$$F_{критA}(0.05,1,4) = 7.71$$

$$F_{критB}(0.05,1,4) = 7.71$$

$$F_{критAB}(0.05,1,4) - 7.71$$

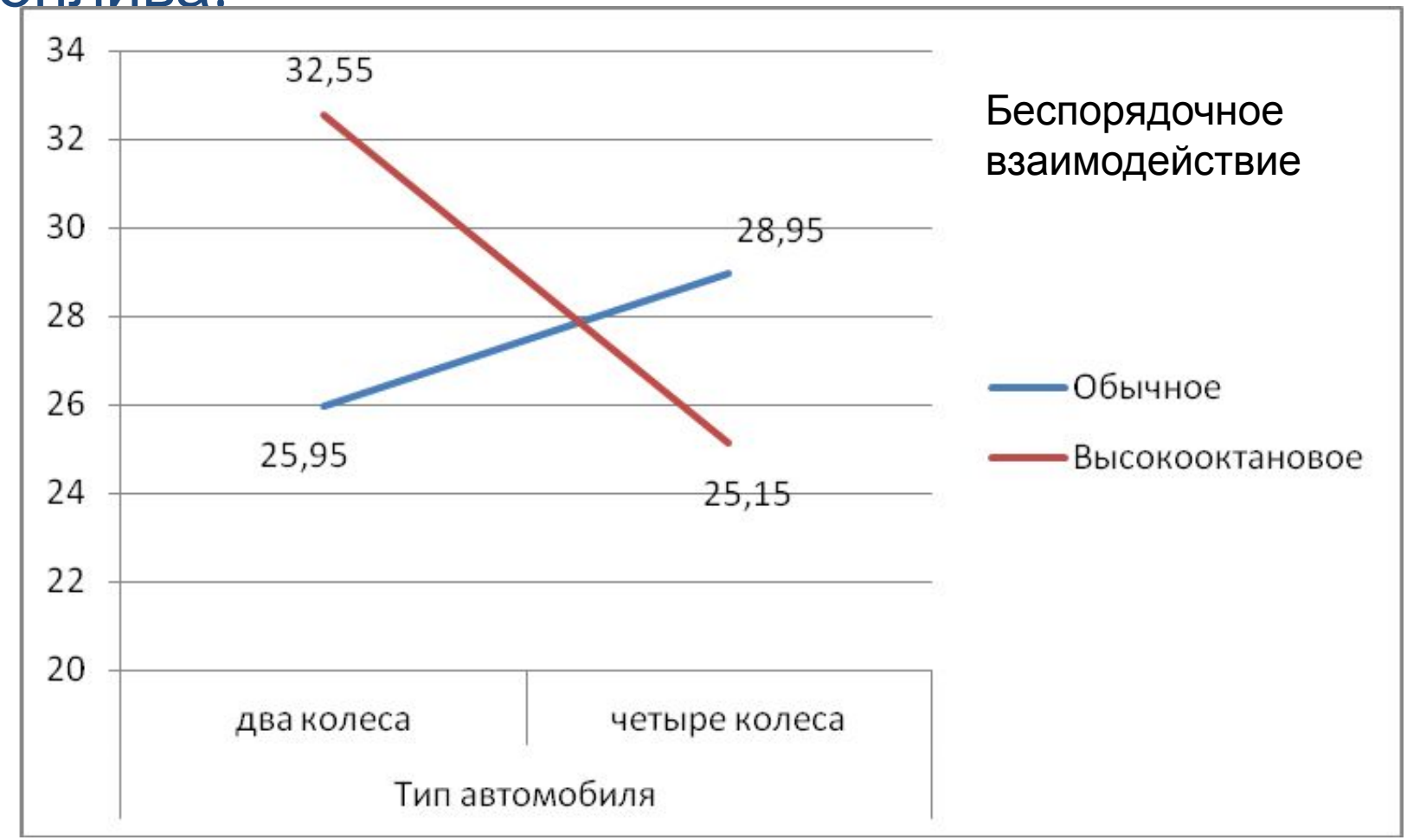
Результаты дисперсионного анализа:

	Сумма квадратов	Степени свободы	Дисперсия	F
Топливо, A	3,92	1	3,92	4,752
Автомобиль, B	9,68	1	9,68	11,733
Взаимодействие A и B	54,08	1	54,08	65,552
Ошибка (внутри группы)	3,3	4	0,825	
Общая	70,98	7		

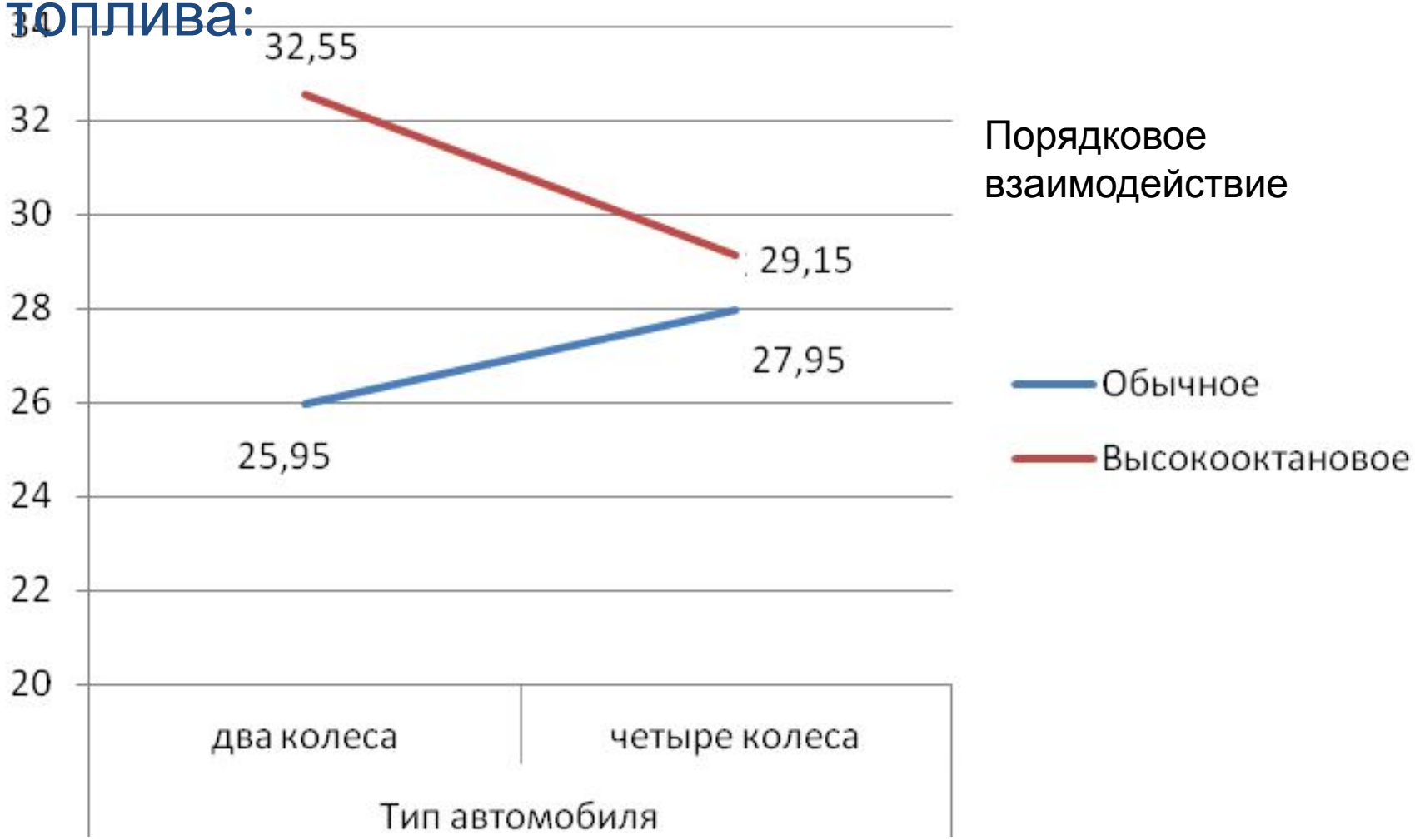
Средний пробег автомобиля в милях на галлон топлива:

Топливо	**Тип автомобиля**	
	два колеса	**четыре колеса**
Обычное	25.95	28,95
Высокооктановое	32.55	25.15

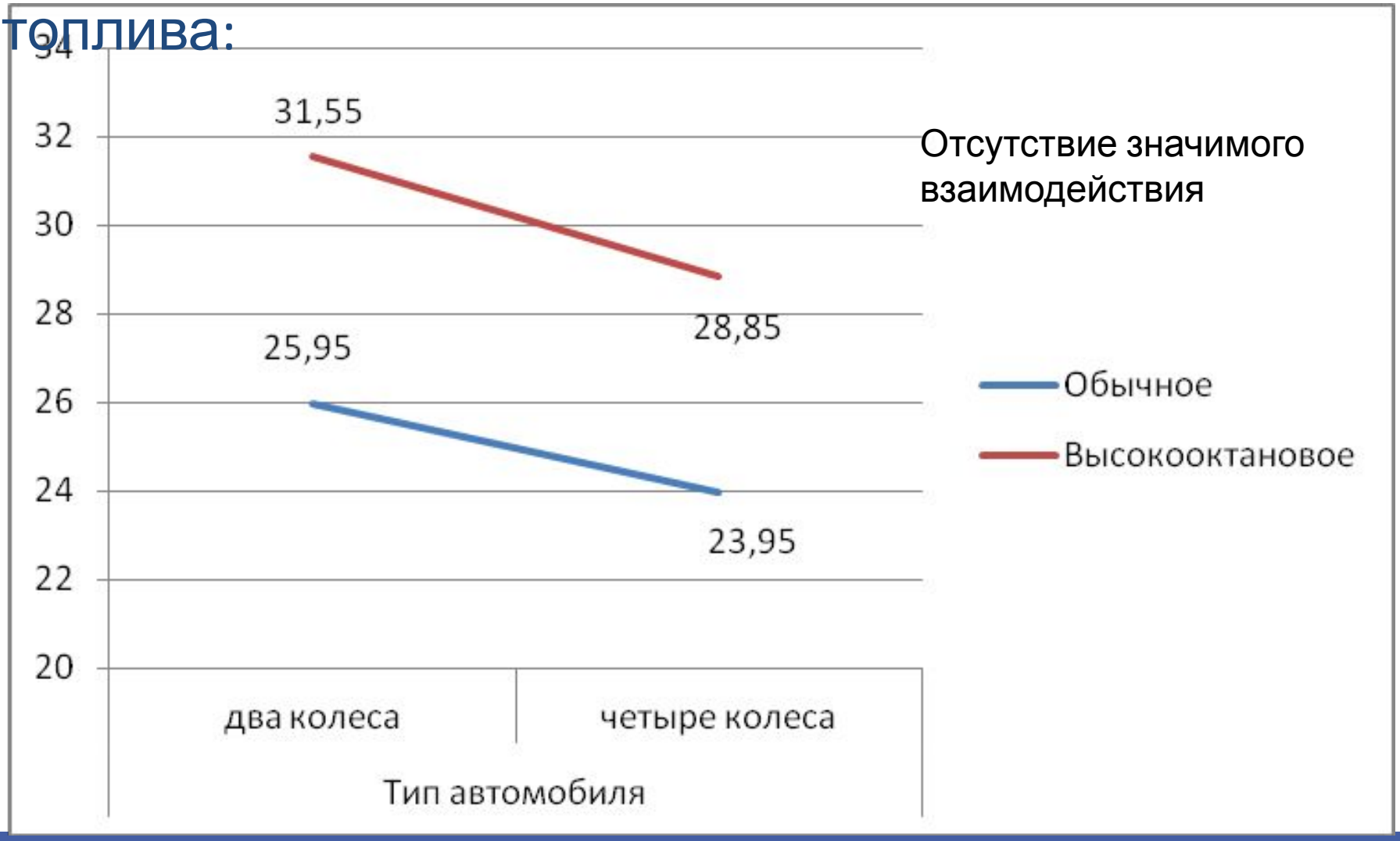
Средний пробег автомобиля в милях на галлон топлива:



Средний пробег автомобиля в милях на галлон топлива:



Средний пробег автомобиля в милях на галлон топлива:



Вопросы

1. Условия применения дисперсионного анализа.
2. Определение дисперсионного анализа. Формулировка гипотез.
3. Задача дисперсионного анализа.
4. Однофакторный дисперсионный анализ.
5. Априорные контрасты и апостериорные критерии
6. Однофакторный дисперсионный анализ для связанных выборок
7. Ограничения дисперсионного анализа для связанных выборок
8. Многофакторный дисперсионный анализ. Формулировка гипотез.