Лекция 10. Однофакторный дисперсионный анализ

10-1. Задача дисперсионного анализа

10-2. Однофакторный дисперсионный анализ

10-3. Решение в SPSS

Пример данных

Имеется ли разница в среднем возрасте учителей, администрации и обслуживающего персонала школы? Взяты выборки из трех генеральных совокупностей.

Учителя	Администрация	Обслуживающий персонал
24	59	34
27	35	29
26	29	35
50	40	31
48	39	40
40	54	45
	56	

© Иванов О.В., 2005

10-1.
Задача дисперсионного анализа

Однофакторный и двухфакторный дисперсионный анализ

Дисперсионный анализ (Analysis of Variance)

F-критерий, который мы использовали при сравнении дисперсий, может применяться для сравнения трех и более средних.

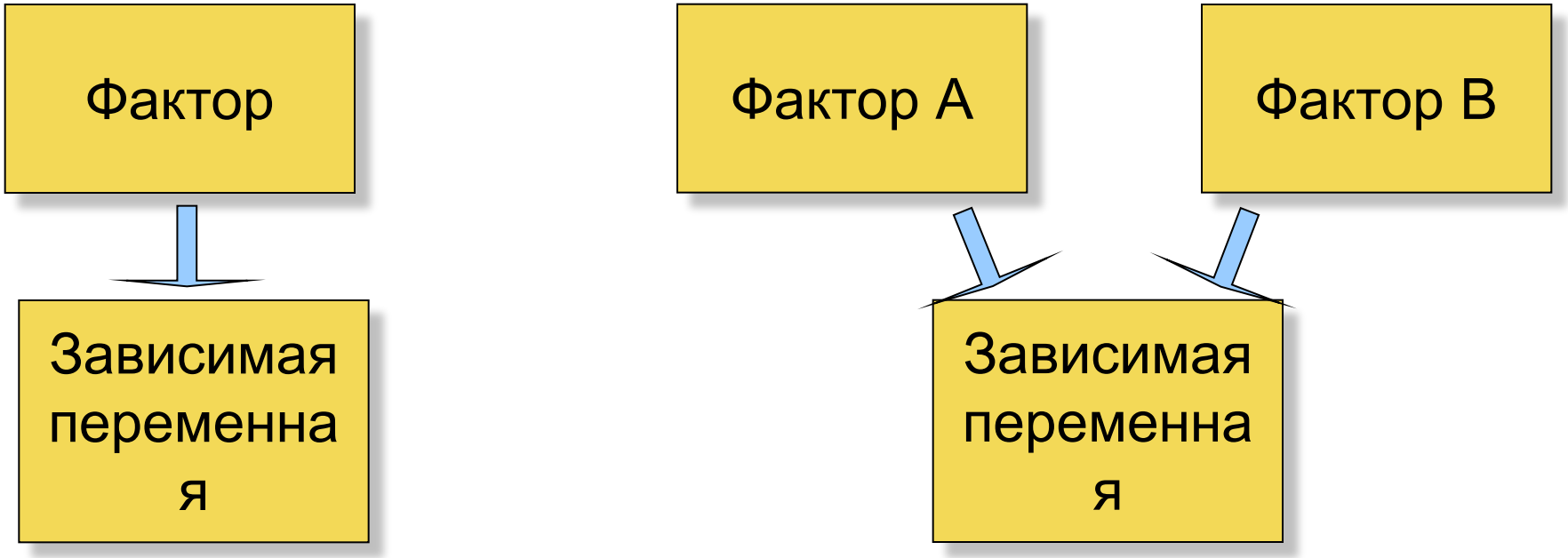
Этот метод называется **дисперсионным анализом** или в англоязычной аббревиатуре *ANOVA* (Analysis of Variance).

F-критерий можно использовать при сравнении двух средних. Но в этом случае он становится идентичным t-критерию.

© Иванов О.В., 2005

Однофакторный и двухфакторный анализ

Дисперсионный анализ, который рассматривает только одну переменную называется **однофакторным дисперсионным анализом** (One-Way ANOVA). Дисперсионный анализ может также применяться в случае двух переменных - это **двухфакторный дисперсионный анализ** (Two-Way ANOVA).



© Иванов О.В., 2005

10-2. Однофакторный дисперсионный анализ

Постановка задачи

Описание метода

Пример

Признак, фактор и уровни фактора

Исследуется **только один признак** или переменная: возраст сотрудников.

Рассматривается **только один фактор**: категория персонала.

Три уровня фактора: учителя, администрация, обслуживающий персонал.

© Иванов О.В., 2005

Представление данных

Данные удобно представлять в виде таблицы. Выборки не обязаны иметь иметь одинаковый объем.

	Уровень 1	Уровень 2	…	Уровень *k*
Уровни фактора				
Измерения признака	x_{11}	x_{21}	…	x_{k1}
	x_{12}	x_{22}	…	x_{k2}
		x_{23}	…	
			…	
Объемы выборок	n_1	n_2		n_k

Имеется k уровней.
Всего проведено N измерений.

© Иванов О.В., 2005

Условия применения

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

2. Выборки должны быть независимы.

3. Дисперсии генеральных совокупностей должны быть равны.

© Иванов О.В., 2005

Гипотезы

Для выявления различия между тремя и более средними, выдвигаются следующие гипотезы:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_m$$

H_1: не все средние равны

© Иванов О.В., 2005

Метод

Вычисляются две оценки: **межгрупповая дисперсия** и **внутригрупповая дисперсия**.

Если нет разницы в средних, то оценки межгрупповой и внутригрупповой дисперсий приблизительно равны и значение *F*-критерия близко к 1, поэтому нулевая гипотеза принимается.

Если различие в средних значительно, межгрупповая дисперсия будет гораздо больше, чем внутригрупповая. Значение *F*-критерия будет значительно больше 1 и нулевая гипотеза будет отвергнута.

Поскольку при проверке гипотезы мы сравниваем дисперсии, метод и получил название *дисперсионный анализ*.

© Иванов О.В., 2005

Степени свободы и критическая область

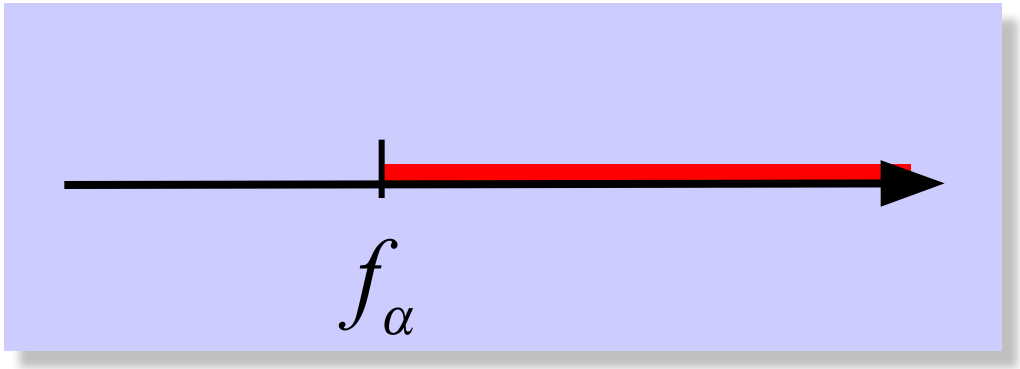
Степени свободы F-распределения задаются двумя значениями:

Числителя: df = k – 1

Знаменателя: df = N – k

Уравнение критической области (правосторонняя):

$$P(F > f_\alpha) = \alpha$$



© Иванов О.В., 2005

Суммы квадратов отклонений

Межгрупповая сумма квадратов отклонений:

$$SS_b = \sum n_i (\bar{x}_i - \bar{\bar{x}})^2$$

Sum **S**quare **B**etween Groups

Внутригрупповая сумма квадратов отклонений:

$$SS_w = \sum (x - \bar{x}_i)^2$$

Sum **S**quare **W**ithin Groups

Общая сумма квадратов отклонений:

$$SS = \sum (x - \bar{x})^2 = SS_b + SS_w$$

Sum **S**quare

© Иванов О.В., 2005

Факторная и остаточная дисперсия. Критерий

Межгрупповая (факторная) дисперсия:

$$MS_B = \frac{SS_B}{k-1}$$

Mean **S**quare
Between Groups

Внутригрупповая (остаточная) дисперсия:

$$MS_W = \frac{SS_W}{N-k}$$

Mean **S**quare
Within Groups

F-критерий:

$$F = \frac{MS_B}{MS_W}$$

© Иванов О.В., 2005

Таблица результатов

Результаты вычислений принято представлять в виде следующей таблицы:

	Сумма квадратов	**df**	**Среднее квадратичное**	***F***
Между группами	SS_B	$k - 1$	MS_B	F-значение
Внутри групп	SS_W	$N - k$	MS_W	
Итого	$SS_B + SS_W$	$N - 1$	$MS_B + MS_W$	

© Иванов О.В., 2005

Пример

Учителя	Администрация	Обслуживающий персонал
24	59	34
27	35	29
26	29	35
50	40	31
48	39	40
40	54	45
	56	

Шаг 1. Гипотезы: $H_0 : \mu_1 = \mu_2 = ... = \mu_m$

$H_1 :$ не все средние равны

© Иванов О.В., 2005

Шаг 2. Критическая область

Найдем критическое значение по таблице критических точек распределения Фишера.

Уровень значимости $\alpha = 0{,}05$.

Так как $k = 3$ и $N = 19$, то

числитель $df = k - 1 = 3 - 1 = 2$

знаменатель $df = N - k = 19 - 3 = 16$

Критическое значение равно 3,633.

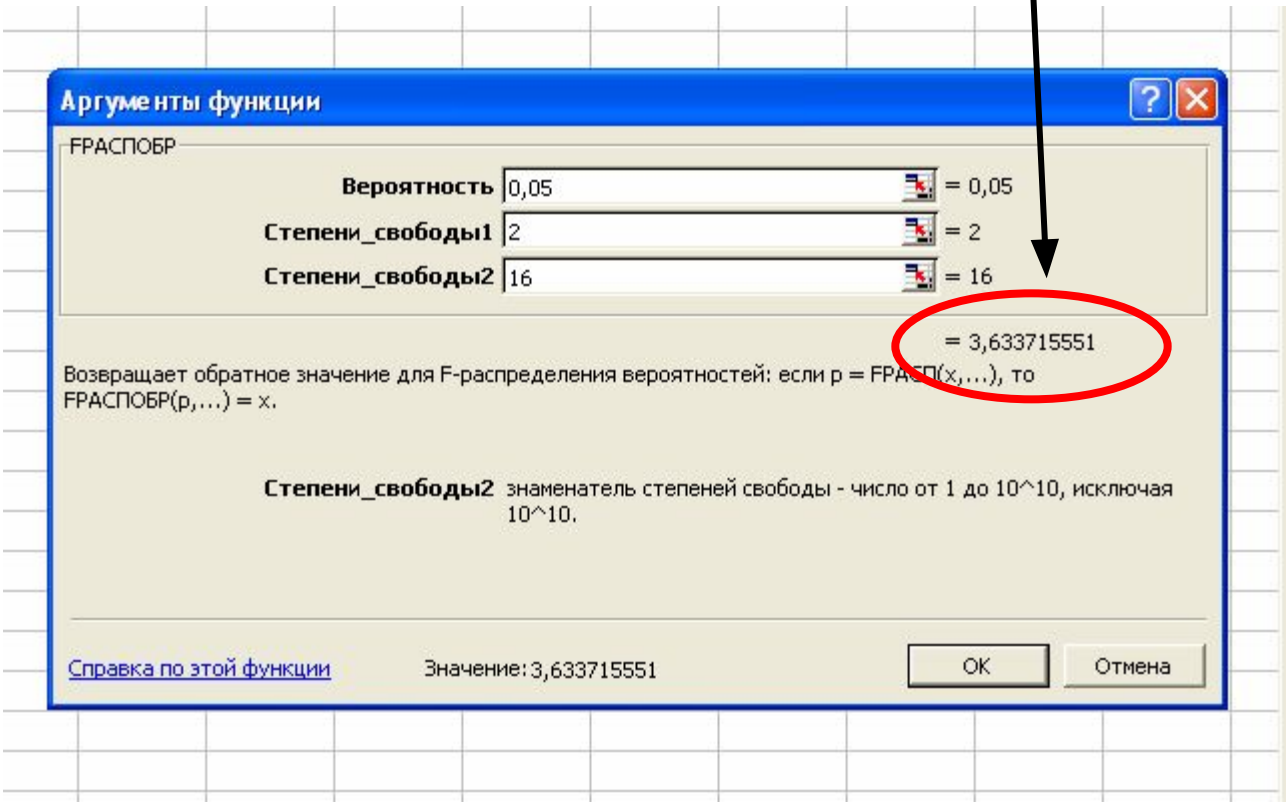
Критическая область $F > 3{,}633$

© Иванов О.В., 2005

Нахождение F-значения в Excel

Критическое значение можно найти, используя функцию в Excel:

FРАСПОБР (0,05; 2; 16) = 3,633…



© Иванов О.В., 2005

Шаг 3. Вычисление статистики F

Учителя	Администрация	Обслуживающий персонал
24	59	34
27	35	29
26	29	35
50	40	31
48	39	40
40	54	45
	56	

Шаг 3а. Подсчет средних

$$\bar{x}_1 = 35{,}8 \qquad n_1 = 6$$

$$\bar{x}_2 = 44{,}6 \qquad n_2 = 7$$

$$\bar{x}_3 = 35{,}7 \qquad n_3 = 6$$

$$\bar{\bar{x}} = 39$$

$$N = n_1 + n_2 + n_3 = 19$$

© Иванов О.В., 2005

Шаг 3b. Расчет отклонений

$$SS_b = \sum n_i (\bar{x}_i - \bar{\bar{x}})^2 =$$

$$= 6 \cdot (35{,}8 - 39)^2 + 7 \cdot (44{,}6 - 39)^2 + 6 \cdot (35{,}7 - 39)^2 =$$

$$= 344{,}1$$

$$SS_w = \sum (x - \bar{x}_i)^2 =$$

$$= (24 - 35{,}8)^2 + (27 - 35{,}8)^2 + \ldots + (48 - 35{,}8)^2 + (40 - 35{,}8)^2 +$$

$$+ (59 - 44{,}6)^2 + (35 - 44{,}6)^2 + \ldots + (54 - 44{,}6)^2 + (56 - 44{,}6)^2 +$$

$$+ (34 - 35{,}7)^2 + (29 - 35{,}7)^2 + \ldots + (40 - 35{,}7)^2 + (45 - 35{,}7)^2 =$$

$$= 1669{,}9$$

© Иванов О.В., 2005

Шаг 3с. Расчет дисперсий

$$MS_B = \frac{SS_B}{k-1} = \frac{344{,}1}{2} = 172{,}06$$

$$MS_W = \frac{SS_W}{N-k} = \frac{1669{,}9}{16} = 104{,}37$$

© Иванов О.В., 2005

Шаг 3d. Расчет статистики

$$F = \frac{MS_B}{MS_W} = \frac{172{,}06}{104{,}37} = 1{,}649$$

© Иванов О.В., 2005

Шаг 4-5. Получение выводов, ответ

$$1{,}649 < 3{,}633$$

Полученное значение статистики не попало в критическую область. У нас нет оснований думать, что средние значения отличаются.

Ответ.

Средний возраст рассматриваемых категорий персонала не различается значимо.

© Иванов О.В., 2005

10-3.
Решение задачи в SPSS

Ввод данных

Анализ

Отчет

Отчет в SPSS

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	3,444E+06	2	1,722E+06	1,646	,223
Within Groups	2,098E+08	196	1,046E+06		
Total	2,101E+08	198			