Тема 8.
Дисперсионный анализ

Дисперсионный анализ (Analysis of Variance)

Для проверки равенства средних **двух** генеральных совокупностей использовался t-критерий Стьюдента.

Для проверки равенства средних в **3-х и более** генеральных совокупностей используется F-критерий Фишера.

F-критерий можно использовать и при сравнении двух средних. Он даст те же результаты, что и t-критерий.

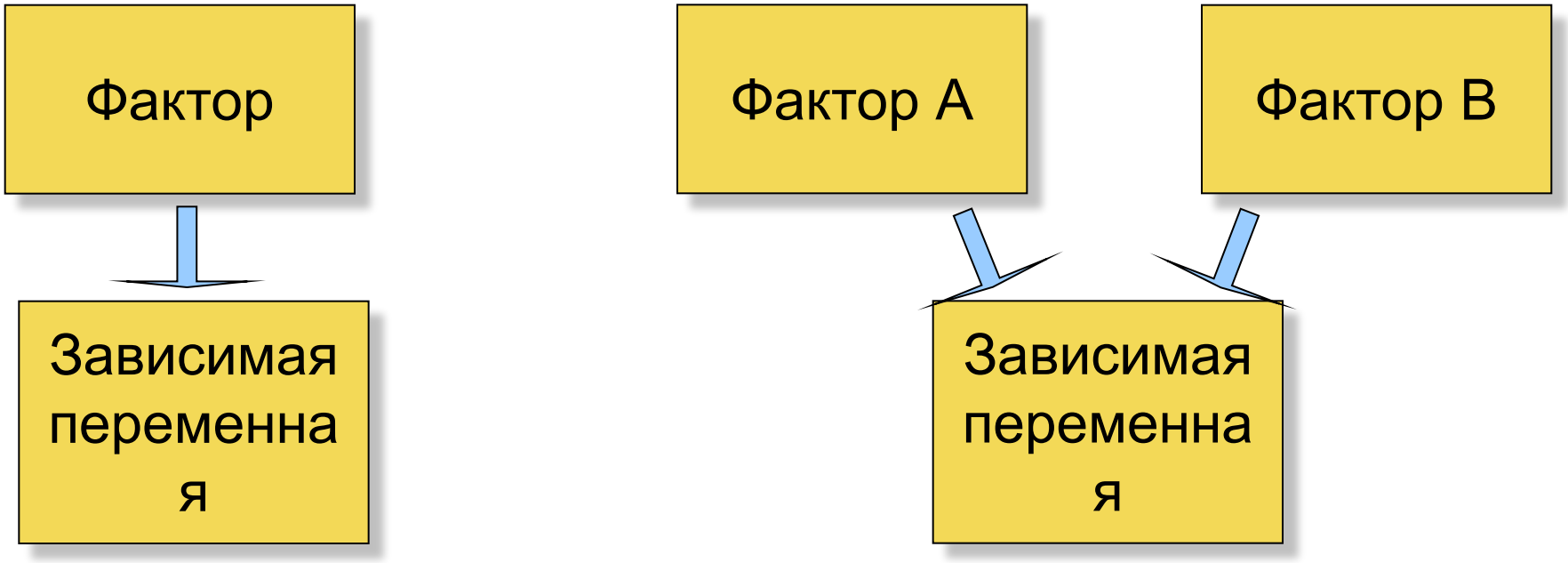
Этот метод называется **дисперсионным анализом** или в англоязычной аббревиатуре *ANOVA* (Analysis of Variance).

Дисперсионный анализ предназначен для выявления влияния на изучаемую количественную переменную одного или нескольких качественных факторов.

Одномерный и двумерный дисперсионный анализ

Дисперсионный анализ, который рассматривает только один качественный фактор называется **однофакторным дисперсионным анализом** (One-Way ANOVA).

Дисперсионный анализ может также применяться в случае двух факторов - это **двуфакторный дисперсионный анализ** (Two-Way ANOVA).



Пример задачи однофакторного анализа.
Зависимая переменная X – цена 1 кв.м на рынке жилья.

Фактор– район города

Задача дисперсионного анализа – выяснить влияют ли на переменную X фактор A.

Пример. Зависимая переменная X – цена 1 кв.м на рынке жилья.

Фактор А – район города
фактор В «тип жилья» (первичное или вторичное).

Задача дисперсионного анализа – выяснить влияют ли на переменную X фактор А, фактор В, а также взаимодействие этих факторов.

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	

Признак, фактор и уровни фактора

Исследуется **только одна количественная переменная**: возраст сотрудников.

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

Три уровня фактора: учителя, администрация, обслуживающий персонал.

Представление данных

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

Уровни фактора

Уровень 1	**Уровень 2**	…	**Уровень *k***
x_{11}	x_{21}	…	x_{k1}
x_{12}	x_{22}	…	x_{k2}
	x_{23}	…	
		…	
n_1	n_2		n_k

Измерения признака

Объемы выборок

Имеется k уровней.
Всего проведено N измерений.

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.

Гипотезы

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

$$H_0 : a_1 = a_2 = \ldots = a_m$$

H_1: не все средние равны

Метод

Берутся две различные оценки дисперсии генеральной совокупности: **межгрупповая дисперсия** и **внутригрупповая дисперсия**.

Если нет разницы в средних, то оценки межгрупповой и внутригрупповой дисперсий приблизительно равны.

Если различие в средних значительно, межгрупповая дисперсия будет гораздо больше, чем внутригрупповая.

Тем самым, при проверке гипотезы о равенстве средних, мы используем сравнение дисперсий. Собственно поэтому метод получил такое название – *дисперсионный анализ*.

Межгрупповые и внутригрупповые отклонения

Межгрупповая сумма квадратов отклонений:

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

Sum **S**quare
Between Groups

Внутригрупповая сумма квадратов отклонений:

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

Sum **S**quare
Within Groups

Общая сумма квадратов отклонений:

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

Sum **S**quare

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-статистика:

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

Если выполнена гипотеза равенства средних, F близко к 1.

Если гипотеза равенства средних неверна, то F существенно больше 1.

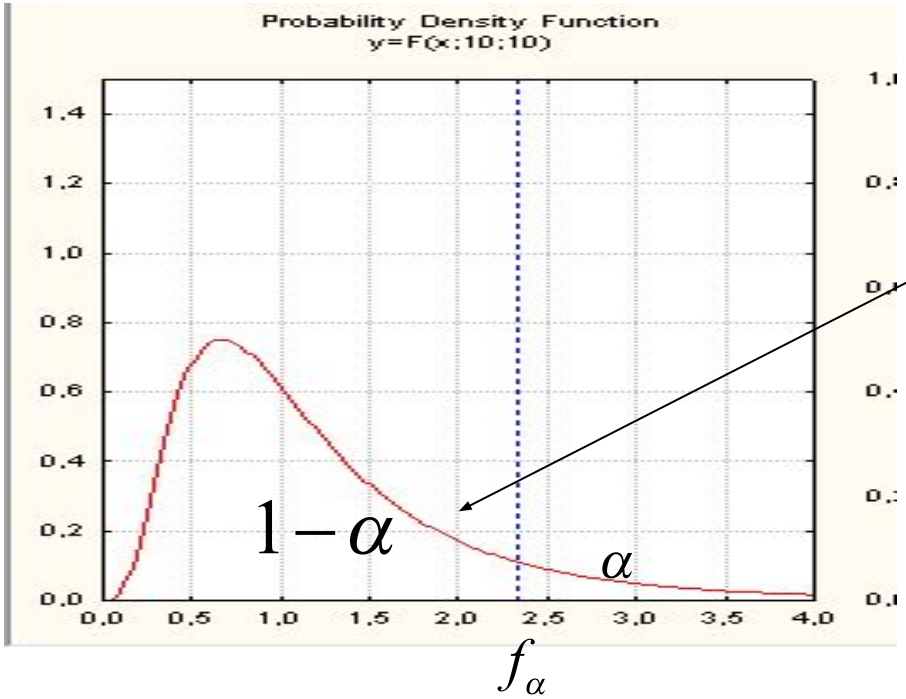
Распределение статистики F

В условиях нулевой гипотезы статистика F имеет распределение Фишера.

Это распределение имеет два параметра:

Степени свободы числителя: df = k – 1

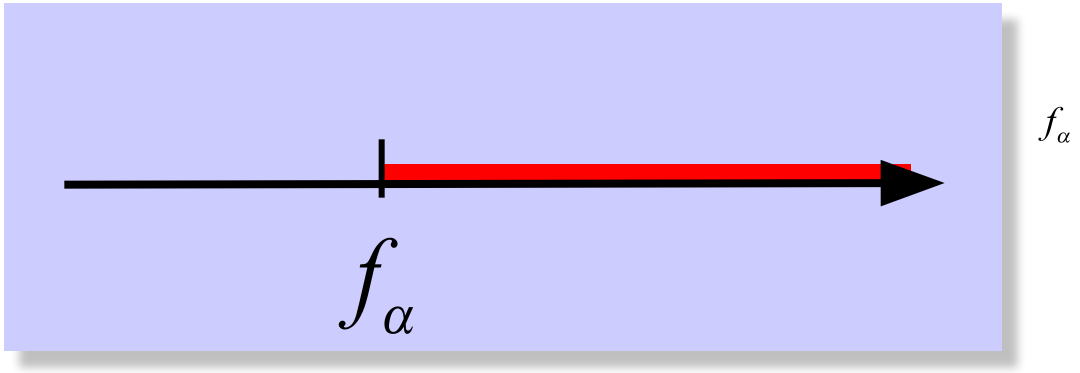
Степени свободы знаменателя: df = N – k



Плотность распределения Фишера F(k-1,N-k)

Степени свободы и критическая область

Критическая область (правосторонняя):



можно найти по таблице или с помощью функции Excel

=FРАСПОБР(α ; $k-1$; $N-k$)

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	$SS_B + SS_W$	$N - 1$	$MS_B + MS_W$	

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: $H_0 : a_1 = a_2 = \ldots = a_k$

$H_1 :$ í å âñå ñðåäí èå ðàâí û

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

=FРАСПОБР(0,05;2;16)

Критическое значение равно 3,633.

Критическая область $F > 3{,}633$

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

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

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

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

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

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

Шаг 3b. Расчет отклонений

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

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

$$= 344{,}1$$

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

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

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

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

$$= 1669{,}9$$

Шаг 3c. Расчет дисперсий

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

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

Шаг 3d. Расчет статистики

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

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.

Отчет в EXCEL

www.zeallsoft.com	B	C	D	E	F	G
Однофакторный дисперсионный анализ						
ИТОГИ						
Группы	*Счет*	*Сумма*	*Среднее*	*Дисперсия*		
Столбец 1	6	215	35,83	136,17		
Столбец 2	7	312	44,57	135,62		
Столбец 3	6	214	35,67	35,07		
Дисперсионный анализ						
очник вари	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
Между гру	344,119	2,000	172,060	1,649	0,223	3,634
Внутри гру	1669,881	16,000	104,368			
Итого	2014	18				