Лекция
Тема: «ЭЛЕМЕНТЫ ДИСПЕРСИОННОГО АНАЛИЗА»

План

1. **Основные понятия**
2. **Описание метода дисперсионного анализа**
3. **Решение типовой задачи (однофакторный дисперсионный анализ несвязанных выборок)**

Дисперсионный анализ (от латинского **DISPERSIO** – рассеивание / на английском ***Analysis Of Variance*** - **ANOVA**) буквально: **анализ факторных эффектов**

Рональд Эйлмер Фишер

(1890 - 1962)

Разработал:

- дисперсионный анализ
- теорию планирования эксперимента
- метод максимального правдоподобия оценки параметров.

- Фундаментальная концепция дисперсионного анализа предложена ***ФИШЕРОМ*** в 1920 году.
- Первоначально дисперсионный анализ был разработан для обработки данных, полученных в ходе специально поставленных экспериментов, и считался **единственным методом, корректно исследующим ПРИЧИННЫЕ связи.**
- Метод применялся для оценки экспериментов в **растениеводстве.**

- В дальнейшем выяснилась общенаучная значимость дисперсионного анализа для экспериментов **в психологии, педагогике, медицине и др.**
- Возможно, более естественным был бы термин анализ суммы квадратов или анализ вариации, но в силу традиции употребляется термин дисперсионный анализ.

- **Дисперсионный анализ** — метод в математической статистике, направленный на поиск зависимостей в экспериментальных данных путём исследования значимости различий в средних значениях.
- В отличие от t-критерия позволяет сравнивать средние значения трёх и более групп.

1. Основные понятия

- Сущность ДА заключается в изучении статистического влияния одного или нескольких факторов на результативный признак (результат)
- **Результативные признаки** — это те признаки, которые изменяются под влиянием факторных признаков.
- **Результативный признак** – это элементарное качество или свойство объектов, изучаемое **как результат влияния факторов:** организованных в исследовании и всех остальных, неорганизованных в данном исследовании

К ***результативным*** признакам можно отнести:

- точно измеряемые параметры объектов: рост, масса, АД, содержание гемоглобина в крови
- неточно измеряемые параметры: умственные способности, например
- комбинированные признаки
- качественные признаки

Фактор – это любое влияние, воздействие или состояние, разнообразие которых может так или иначе отражаться в разнообразии результативного признака

Факторами могут быть

- Физические воздействия (температура, влажность, радиация)
- Химические воздействия: питание, стимуляторы, мутагены, алкоголь
- Биологические: здоровье, болезни, наследственность, талантливость, идиотизм
- Окружающая среда: ареал обитания, условия жизни
- Возраст, пол и др.

- Факторы могут иметь различные **ГРАДАЦИИ** или различные условия действия

  Градация (с лат. ***GRADATIO*** **–** постепенное возвышение, усиление**)** фактора – это изменение его величины при переходе от одной группы к другой
- Пример (шутка),

  если отыщется исследователь, желающий **определить зависимость яйценоскости от цвета курицы**, то ничто не помешает ему применить дисперсионный анализ, и в качестве **условий действия фактора «цвет»** избрать, скажем, **ЧЕРНЫХ, БЕЛЫХ И ПЕСТРЫХ** кур.

- Фактор
  - регулируемый
    - Уровень 1
    - Уровень 2
  - неконтролируемый
    - Случайный

Виды дисперсионного анализа

По количеству выявляемых регулируемых факторов дисперсионный анализ может быть

однофакторным (при этом изучается влияние одного фактора на результаты эксперимента),

двухфакторным (при изучении влияния двух факторов)

многофакторным (позволяет оценить не только влияние каждого из факторов в отдельности, но и их взаимодействие).

- **ДА несвязанных (различных, независимых) выборок**.

В зависимости от поставленной цели и задач выборочные ***группы формируются случайным*** образом независимо друг от друга (контрольная и экспериментальная группы для изучения некоторого показателя, ***например, влияние высокого артериального давления на развитие инсульта).***

- **ДА связанных выборок (зависимых).**

  Результаты воздействия факторов исследуются у одной и той же выборочной группы (например, у одних и тех же пациентов) до и после воздействия (лечение, профилактика, реабилитационные мероприятия)

- **дисперсионный анализ одномерный и многомерный** (одна или несколько зависимых переменных)

Условия применения дисперсионного анализа

- выборочные данные должны быть взяты из НОРМАЛЬНЫХ совокупностей
- исправленные выборочные дисперсии каждого уровня контролируемого фактора должны быть равны (оценки выборочных дисперсий)
- результаты наблюдений должны быть независимыми

2. Принцип применения метода дисперсионного анализа

- Формулируется **НУЛЕВАЯ ГИПОТЕЗА**, то есть предполагается, что исследуемые факторы не оказывают никакого влияния на значения результативного признака и полученные различия случайны.

- Очевидно, что если регулируемый фактор **ОКАЗЫВАЕТ** влияние на признак, то при различных уровнях этого фактора будут **наблюдаться существенные изменения средних значений признака**.

- Следовательно, **ИЗМЕНЕНИЯ**, вызванные влиянием контролируемого фактора будут **БОЛЕЕ ЗНАЧИМЫ**, чем влияние неконтролируемых (случайных) факторов.
- Оценить изменения можно с помощью дисперсий.

- **ОСНОВНАЯ ЗАДАЧА ДИСПЕРСИОННОГО АНАЛИЗА**

заключается в разложении общей дисперсии признака на дисперсию, вызванную действием контролируемого фактора (факторную дисперсию $D_{факт}$) и дисперсию остаточную (остаточную дисперсию $D_{ост}$), т.е. вызванную неконтролируемыми факторами:

$$D_{общ.} = D_{факт} + D_{ост}$$

- $D_{общ.}$ - общая дисперсия наблюдаемых значений (вариант), характеризуется разбросом вариант от ***общего среднего***. Измеряет вариацию признака во всей совокупности под влиянием всех факторов, обусловивших эту вариацию.

ОБЩЕЕ РАЗНООБРАЗИЕ СКЛАДЫВАЕТСЯ ИЗ МЕЖГРУППОВОГО И ВНУТРИГРУППОВОГО

- $D_{факт}$ - факторная (межгрупповая) дисперсия, характеризуется ***различием средних в каждой группе*** и зависит от влияния исследуемого фактора, по которому дифференцируется каждая группа.

  Например, в группах различных по этиологическому фактору клинического течения пневмонии средний уровень проведенного койко-дня неодинаков — наблюдается межгрупповое разнообразие.

- $D_{ост.}$ - остаточная (внутригрупповая) дисперсия, которая характеризует ***рассеяние вариант внутри групп.*** Отражает случайную вариацию, т.е. часть вариации, происходящую под влиянием неучтенных факторов и не зависящую от признака — фактора, положенного в основание группировки.
- Вариация изучаемого признака зависит от силы влияния каких-то неучтенных случайных факторов, как от организованных (заданных исследователем), так и от случайных (неизвестных) факторов.

Этапы дисперсионного анализа

1. **Построение дисперсионного комплекса.**
2. **Вычисление квадратов отклонений.**
3. **Вычисление дисперсий.**
4. **Сравнение факторной и остаточной дисперсий.**
5. **Статистическая проверка нулевой гипотезы о несущественности различий факторной и остаточной дисперсий**

Замечание

- Для проверки нулевой гипотезы используется F-статистика
- С помощью *критерия Фишера-Снедекора* можно определить значимость отличия факторной и остаточной дисперсий и тем самым подтвердить или опровергнуть гипотезу о значимости влияния изучаемого фактора на контролируемый признак.

Например, пусть число наблюдений при действии каждого из уровней фактора одинаково (q) и результаты представлены в таблице.

Номер испытания	**Уровень фактора** A_j				
1	x_{11}	x_{12}	x_{13}	…	x_{1k}
2	x_{21}	x_{22}	x_{23}	…	x_{2k}
3	x_{31}	x_{32}	x_{33}	…	x_{3k}
…	…	…	…	…	…
q	x_{q1}	x_{q2}	x_{q3}	…	x_{qk}
Групповая средняя $\bar{x}_i$	$\bar{x}_1$	$\bar{x}_2$	$\bar{x}_3$	…	$\bar{x}_k$

- Все значения величины *x*, наблюдаемые при каждом фиксированном уровне фактора , составляют группу, и в последней строке таблицы A_j представлены соответствующие выборочные групповые средние, вычисленные по формуле:

$$\overline{x_j} = \frac{\sum_{i=1}^{q} x_{ij}}{q}$$

- Скорее всего выборочные средние по каждому уровню будут отличаться друг от друга. Но является ли это отличие значимым и вызвано ли это отличие действием фактора?

Выдвигаются две гипотезы:

- **H_0 – фактор не влияет на признак и, следовательно, средние значения величины признака на разных уровнях равны, т.е.** $\overline{x}_1 = \overline{x}_2 = \ldots = \overline{x}_j$

- **H_1 – фактор влияет на признак, и следовательно, хотя бы одно выборочное среднее <u>значимо</u> отличается от других.**

- *Пример.* Методом дисперсионного анализа на уровне $\alpha = 0{,}05$ значимости установить существенность влияния реагента ***A*** (фактора ***F***) на синтез лекарственного препарата (выход ***X*** в условных единицах – результативный признак). Установить силу влияния фактора на признак.

№ испытания	***Уровни фактора F***		
	A_1	A_2	A_3
1	***59***	***58***	***56***
2	***60***	***57***	***56***
3	***58***	***58***	***55***
4	***60***	***56***	
5	***59***		

- Найдем групповые среднии:

$$\overline{x_j} = \sum_{i=1}^{n} \frac{x_{ij}}{n_i};$$

$$\overline{x_1} = \frac{59 + 60 + 58 + 60 + 59}{5} = 59{,}2;$$

$$\overline{x_2} = \frac{58 + 57 + 58 + 56}{4} = 57{,}3;$$

$$\overline{x_3} = \frac{56 + 56 + 55}{3} = 55{,}7.$$

- Выборочные средние по каждому уровню отличаются друг от друга. Но является ли это отличие значимым и вызвано ли это отличие действием фактора?
- Выдвигаем нулевую гипотезу:

  фактор не влияет на признак и, следовательно, средние значения величины признака на разных уровнях равны, т.е. H_0:

$$\bar{x}_1 = \bar{x}_2 = \bar{x}_3$$

Для проверки предположения о случайном различи средних воспользуемся методом

дисперсионного анализа

ФОРМУЛЫ для вычислия сумм квадратов отклонений

$$TSS = z_2 - \frac{z_1^2}{N} \qquad ESS = z_3 - \frac{z_1^2}{N}$$

$$USS = Z_2 - Z_3$$

ФОРМУЛЫ ДЛЯ ВЫЧИСЛЕНИЯ ДИСПЕРСИЙ

$$S^2_{факт} = \frac{ESS}{a-1}$$

$$S^2_{ост} = \frac{USS}{N-a}$$

Нужные суммы вычислим в таблице

№ испытания	Уровни фактора F (a – количество уровней или градаций)			
	A_1	A_2	A_3	a=3
1	59	58	56	
2	60	57	56	
3	58	58	55	
4	60	56		
5	59			
n_i	5	4	3	$N = \sum n_i = 12$
$\sum x_i$	296	229	167	$z_1 = 692$
Групповые средние	59,2	57,3	55,7	

№ испытания	Уровни фактора F (*a* – количество уровней)			
	A_1	A_2	A_3	a=3
1	59	58	56	
2	60	57	56	
3	58	58	55	
4	60	56		
5	59			
n_i	5	4	3	$N = \sum n_i = 12$
$\sum x_i$	296	229	167	$z_1 = 692$
$\sum x_i^2$	17526	13113	9297	$z_2 = 39936$

№ испытания	Уровни фактора F (a – количество уровней, градаций)			
	F_1	F_2	F_3	a=3
1	59	58	56	
2	60	57	56	
3	58	58	55	
4	60	56		
5	59			
n_i	5	4	3	$N = \sum n_i = 12$
$\sum x_i$	296	229	167	$z_1 = 692$
Групповые средние	59,2	57,3	55,7	
$\sum x_i^2$	17526	13113	9297	$z_2 = 39936$
$\frac{\left(\sum x_i\right)^2}{n_i}$	17523,2	13110,25	9296,3	$z_3 = 39929,75$

Вычислим суммы квадратов отклонений

$$TSS = 39936 - \frac{692^2}{12} = 30,7$$

$$ESS = 39929,75 - \frac{692^2}{12} = 24,45$$

$$USS = 30,7 - 24,45 = 6,25$$

- **Вычислим дисперсии**

$$S^2_{факт} = \frac{ESS}{a-1}$$

$$S^2_{факт} = \frac{24,45}{3-1} = 12,2$$

$$S^2_{ост} = \frac{USS}{N-a}$$

$$S^2_{ост} = \frac{6,25}{12-3} = 0,7$$

- Сравнение факторной и остаточной дисперсий показывает, что

$$S^2_{факт.} > S^2_{ост..}$$

- Прежде, чем делать окончательный вывод о влиянии фактора на признак, необходимо проверить статистическую значимость различий дисперсий

Проверка гипотез для дисперсий.

1. Нулевая гипотеза $H_0: S^2_{факт} = S^2_{остат}$
2. Конкурирующая гипотеза $H_1: S^2_{факт} \neq S^2_{остат}$
3. Для проверки нулевой гипотезы используем *F*-критерий Фишера

$$F_{набл} = \frac{S^2_{больш}}{S^2_{меньш}}$$

$$F_{табл.}\left(\gamma, \nu_1 = a - 1, \nu_2 = N - a\right);$$

- Проверим значимость различия дисперсий:
- найдем наб
критерия р

$$F_{набл.} = \frac{S^2_{факт.}}{S^2_{ост.}} = \frac{12,2}{0,7} = 17,4;$$

- найдем табличное значение критерия достоверности (используя таблицу Фишера – Снедекора:

$$F_{табл.}(\gamma; v_1 = a - 1; v_2 = N - a);$$

$$F_{табл.}(0,05;2;9) = 4,26.$$

- Сравним $F_{набл.}$ и $F_{табл.}$

- *Вывод*: дисперсии различаются значимо на уровне значимости 0,05 . Следовательно, фактор (указать какой) оказывает существенное влияние на признак (указать признак) .

- **ОЦЕНИМ СИЛУ ВЛИЯНИЯ ФАКТОРА НА ПРИЗНАК**

$$\eta^2 = 1 - \frac{D_{ост.}}{D_{общ.}} \cdot \frac{N-1}{N-a};$$

$$\eta^2 = 1 - \frac{6,25}{30,7} \cdot \frac{11}{9} = 1 - 0,2 \cdot 1,22 = 0,76.$$

- *Вывод:* Изменения признака (выхода лекарственного препарата при его синтезе) на 76% обусловлены влиянием регулируемого фактора (реагента *A*) и на 24% влиянием всех других нерегулируемых факторов.

Математики шутят

ТЕОРВЕР БОЛЬШОЙ...

Во время сессии в коридоре института встречаются преподаватели В. и К., только что закончившие принимать экзамены в своих группах.

— Ну, как студенты? — спрашивает В. — Нормально сдают?

— Да как сказать, — мнется К. — Вот сейчас мне сдавал один студент. По билету ничего не сказал, на дополнительные вопросы не ответил. Но я ему все-таки поставил «четыре».

— Как?! За что? — поражается собеседник. — Он же ничего не знает!

— Теорвер большой, — задумчиво отвечает К. — что-нибудь да знает...

Потом спрашивает В.

— А у тебя как студенты?

— Да тоже не очень, — отвечает тот. — Только что принимал экзамен у студента. По билету все рассказал без запинки, на все дополнительные вопросы ответил, однако я ему поставил-таки «три».

— Но почему?! — теперь уже поражается К.

— Теорвер большой, — невозмутимо говорит В., — что-нибудь да не знает.

Критическое значение распределения Фишера-Снедекора

f2 \f1	1	2	3	4	5	6	7	8	9	10	11	12
					При α=0,05							
1	161	200	216	225	230	234	237	239	241	242	243	244
2	18,51	19	19,16	19,25	19,3	19,33	19,36	19,37	19,38	19,39	19,4	19,41
3	10,13	9,55	9,28	9,12	9,01	8,94	8,88	8,84	8,81	8,78	8,76	8,74
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6	5,96	5,93	5,91
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,78	4,74	4,7	4,68
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,1	4,06	4,03	4
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,63	3,6	3,57
8	5,32	4,46	4,07	3,84	3,69	3,58	3,5	3,44	3,39	3,34	3,31	3,28
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,13	3,1	3,07
10	4,96	4,1	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,97	2,94	2,91
11	4,84	3,98	3,59	3,36	3,2	3,09	3,01	2,95	2,9	2,86	2,82	2,79
12	4,75	3,88	3,49	3,26	3,11	3	2,92	2,85	2,8	2,76	2,72	2,69
13	4,67	3,8	3,41	3,18	3,02	2,92	2,84	2,77	2,72	2,67	2,63	2,6
14	4,6	3,74	3,34	3,11	2,96	2,85	2,77	2,7	2,65	2,6	2,56	2,53
15	4,54	3,68	3,29	3,06	2,9	2,79	2,7	2,64	2,59	2,55	2,51	2,48
					При ***α=0,025***							
1	648	800	864	900	922	937	948	957	963	968	985	993
2	38,51	39	39,17	39,25	39,3	39,33	39,36	39,37	39,39	39,4	39,43	39,45
3	17,44	16,04	15,44	15,1	14,89	14,74	14,62	14,54	14,47	14,42	14,25	14,17
4	12,22	10,65	9,98	9,6	9,36	9,2	9,07	8,98	8,9	8,84	8,66	8,56
5	10	8,43	7,76	7,39	7,15	6,98	6,85	6,76	6,68	6,62	6,43	6,33
6	8,81	7,26	6,6	6,23	5,99	5,82	5,7	5,6	5,52	5,46	5,27	5,17
7	8,07	6,54	5,89	5,52	5,29	5,12	5	4,9	4,82	4,76	4,57	4,47
8	7,57	6,06	5,42	5,05	4,82	4,65	4,53	4,43	4,36	4,3	4,1	4
9	7,21	5,71	5,08	4,72	4,48	4,32	4,2	4,1	4,03	3,96	3,77	3,67
10	6,94	5,46	4,83	4,47	4,24	4,07	3,95	3,85	3,78	3,72	3,52	3,42
11	6,72	5,26	4,63	4,28	4,04	3,88	3,76	3,66	3,59	3,53	3,33	3,23
12	6,55	5,1	4,47	4,12	3,89	3,72	3,61	3,51	3,44	3,37	3,18	3,07
13	6,41	4,97	4,35	4	3,77	3,6	3,48	3,39	3,31	3,25	3,05	2,95
14	6,3	4,86	4,24	3,89	3,66	3,5	3,38	3,89	3,21	3,15	2,95	2,84
15	6,2	4,77	4,15	3,8	3,58	3,41	3,29	3,2	3,12	3,06	2,86	2,76