ТЕМА №10
ОСНОВЫ ДИСПЕРСИОННОГО АНАЛИЗА

ПРОГРАММНЫЕ ВОПРОСЫ:

10.1.Понятие о дисперсионном анализе.
10.2. Однофакторный дисперсионный анализ при одинаковом числе испытаний на уровнях.
10.3. Однофакторный дисперсионный анализ при неодинаковом числе испытаний на уровнях.
10.4. Понятие о двухфакторном и многофакторном анализах.

10.1. Понятие о дисперсионном анализе

В практической деятельности врачей при проведении медико-биологических, социологических и экспериментальных исследований возникает необходимость установить влияние факторов на результаты изучения состояния здоровья населения, при оценке профессиональной деятельности, эффективности нововведений.

Существует ряд статистических методов, позволяющих определить силу, направление, закономерности влияния факторов на результат в генеральной или выборочной совокупностях (корреляционный анализ, регрессия и др.). Дисперсионный анализ был разработан и предложен английским ученым, математиком и генетиком Рональдом Фишером в 20-х годах XX века.

Дисперсионный анализ чаще используют в научно-практических исследованиях общественного здоровья и здравоохранения для изучения влияния одного или нескольких факторов на результативный признак.

Сущность метода дисперсионного анализа заключается в измерении отдельных дисперсий (общая, факторная, остаточная), и дальнейшем определении силы влияния изучаемых факторов (оценки роли каждого из факторов, либо их совместного влияния) на результативный признак.

Дисперсионный анализ – это статистический метод анализа результатов наблюдений, зависящих от различных одновременно действующих факторов, основанный на сравнении оценок дисперсий соответствующих групп выборочных данных.

Под ***фактором*** понимают различные, независимые, качественные показатели, влияющие на изучаемей признаки. Факторы обозначаются прописными начальными буквами латинского алфавита. Например: *A*, *B*, *C*…. Факторы контролируемые и измеряемые в процессе исследования, называются ***регулируемыми***.

Важным методическим значением для применения дисперсионного анализа является правильное формирование выборки. В зависимости от поставленной цели и задач выборочные группы могут формироваться случайным образом независимо друг от друга (контрольная и экспериментальная группы для изучения некоторого показателя, например, влияние высокого артериального давления на развитие инсульта). Такие выборки называются ***независимыми***.

Нередко 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), такие выборки называются ***зависимыми***.

Дисперсионный анализ, в котором проверяется влияние одного фактора, называется ***однофакторным*** (одномерный анализ). При изучении влияния более чем одного фактора используют ***многофакторный*** дисперсионный анализ (многомерный анализ).

Признаки, изменяющиеся под воздействием тех или иных факторов, называют ***результативными***. Для их обозначения используются конечные буквы латинского алфавита: X, Y, Z.

Для проведения дисперсионного анализа могут использоваться как качественные (пол, профессия), так и количественные признаки (число инъекций, больных в палате, число койко-дней).

Методы дисперсионного анализа:

1. Метод по Фишеру (Fisher) – критерий F. Метод применяется в однофакторном дисперсионном анализе, когда совокупная дисперсия всех наблюдаемых значений раскладывается на дисперсию внутри отдельных групп и дисперсию между группами.

2. Метод "общей линейной модели". В его основе лежит корреляционный или регрессионный анализ, применяемый в многофакторном анализе.

Обычно в медико-биологических исследованиях используются только однофакторные, максимум двухфакторные дисперсионные комплексы. Многофакторные комплексы можно исследовать, последовательно анализируя одно- или двухфакторные комплексы, выделяемые из всей наблюдаемой совокупности.

Условия применения дисперсионного анализа:

1. Задачей исследования является определение силы влияния одного (до 3) факторов на результат или определение силы совместного влияния различных факторов (пол и возраст, физическая активность и питание и т.д.).

2. Изучаемые факторы должны быть независимые (несвязанные) между собой. Например, нельзя изучать совместное влияние стажа работы и возраста, роста и веса детей и т.д. на заболеваемость населения.

3. Подбор групп для исследования проводится рандомизированно (случайный отбор). Организация дисперсионного комплекса с выполнением принципа случайности отбора вариантов называется рандомизацией (перев. с англ.– random), т.е. выбранные наугад.

4. Можно применять как количественные, так и качественные (атрибутивные) признаки.

10.2. Однофакторный дисперсионный анализ при одинаковом числе испытаний на уровнях

Рассмотрим нормально распределенную величину *X,* на которую действует некоторый фактор *A,* имеющий *l* постоянных уровней, причем на всех уровнях распределение значений *X* является нормальным, а дисперсии одинаковы, хотя и неизвестны.

Пусть число проведенных наблюдений при действии каждого из уровней фактора одинаково (*q*) и результаты представлены в таблице 10.1

Таблица 10.1

Номер испытания	Уровень фактора A_j				
	A_1	A_2	A_3	...	A_l
1	x_{11}	x_{12}	x_{13}	...	x_{1l}
2	x_{21}	x_{22}	x_{23}	...	x_{2l}
3	x_{31}	x_{32}	x_{33}	...	x_{3l}
...	...	...	...	...	...
q	x_{q1}	x_{q2}	x_{q3}	...	x_{ql}
Групповая средняя $\overline{x_j}$	$\overline{x_1}$	$\overline{x_2}$	$\overline{x_3}$	...	$\overline{x_l}$

Все значения величины *X,* наблюдаемые при каждом фиксированном уровне фактора A_j, составляют группу, и в последней строке табл. 10.1 представлены соответствующие *выборочные групповые средние,* вычисленные по формуле

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

В основе однофакторного дисперсионного анализа лежит тесная связь между различием в групповых средних $\overline{x_j}$ и соотношением между двумя видами дисперсий – остаточной и факторной, которые рассчитываются по формулам:

$$S^2_{факт} = \frac{\sum_{j=1}^{l} \frac{R_j^2}{q} - \frac{1}{lq}\left(\sum_{j=1}^{l} R_j\right)^2}{l-1}, \qquad S^2_{ост} = \frac{\sum_{j=1}^{l} P_j - \frac{1}{q}\sum_{j=1}^{l} R_j^2}{l(q-1)}.$$

Здесь $R_j = \sum_{i=1}^{q} x_{ij}$ – сумма значений величины X на уровне A_j; $P_j = \sum_{i=1}^{q} {x_{ij}}^2$ – сумма квадратов значений величины X на уровне A_j.

В методе однофакторного дисперсионного анализа факторная дисперсия характеризует влияние фактора на исследуемую величину, а остаточная – влияние случайных причин, в связи с чем справедливо следующее правило.

Правило 10.1: Если $S^2_{факт} < S^2_{ост}$, то следует сделать вывод об отсутствии существенного влияния фактора A на величину X. Если же $S^2_{факт} > S^2_{ост}$, то необходимо проверить значимость различия этих дисперсий, т. е. при заданном (или выбранном) уровне значимости p проверить нулевую гипотезу о равенстве соответствующих генеральных дисперсий $\sigma^2_{факт} = \sigma^2_{ост}$ при конкурирующей гипотезе вида $\sigma^2_{факт} > \sigma^2_{ост}$. При этом возможны следующие варианты.

1. Если проверка покажет значимость различия между $S^2_{факт}$, для которой число степеней свободы $f_1 = l - 1$, и $S^2_{ост}$ для которой $f_2 = l(q-1)$, то отсюда также следует значимость различия между найденными по результатам наблюдений выборочными групповыми средними, что соответствует выводу о существенном влиянии фактора A на величину X.

2. Если различие между факторной и остаточной дисперсиями окажется незначимым, то нельзя сделать вывод о существенном влиянии фактора A на величину X.

Обычно для упрощения расчетов факторную и остаточную дисперсию рассчитывают не по экспериментальным значениям x_{ij} величины X, а по значениям $y_{ij} = x_{ij} - C$, где C представляет собой произвольное число, близкое к среднему значению $\bar{x}$ всех результатов наблюдений x_{ij}.

Пример 10.1: При уровне значимости α=0,05 методом дисперсионного анализа проверить эффективность внешнего воздействия (факторы A_1, A_2, A_3) на темп размножения определенного вида бактерий по данным, приведенным ниже в таблице:

Номер испытания	Уровень фактора A_j		
	A_1	A_2	A_3
1	426	429	430
2	425	433	431
3	427	427	436
4	423	426	429
$\overline{x_j}$	425,3	428,8	431,5

Решение:

Пусть число проведенных наблюдений при действии каждого из уровней фактора A одинаково и равно q =4. Все значения величины X, наблюдаемые при каждом фиксированном уровне фактора A_j, составляют группу, и в последней строке таблицы представим соответствующие выборочные групповые средние, вычисленные по формуле:

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

Для упрощения расчетов будем использовать не экспериментальные значения x_{ij} величины X, а значения $y_{ij} = x_{ij} - C$, где постоянная C представляет собой произвольное число, близкое к среднему значению $\overline{x}$ всех результатов наблюдений x_{ij}, таким образом:

$$\overline{x} = \sum_{j=1}^{l} \frac{\overline{x_j}}{l} = \frac{425{,}3 + 428{,}8 + 431{,}5}{3} = 428{,}5,$$

где l - количество уровней фактора A (l=3).

Введем новые переменные:

$$y_{ij} = x_{ij} - C = x_{ij} - \overline{x} = x_{ij} - 428{,}5$$

Преобразуем таблицу с дальнейшим использованием ее в качестве расчетной.

Номер испытания	Уровень фактора A_j		
	A_1	A_2	A_3
	y_{i1}	y_{i2}	y_{i3}
1	-2,5	0,5	1,5
2	-3,5	4,5	2,5
3	-1,5	-1,5	7,5
4	-5,5	-2,5	0,5

С помощью таблицы определим вспомогательные величины:

$$P_j = \sum_{i=1}^{q} {y_{ij}}^2; \qquad R_j = \sum_{i=1}^{q} y_{ij}$$

$$P_1 = \sum_{i=1}^{4} {y_{i1}}^2 = (-2{,}5)^2 + (-3{,}5)^2 + (-1{,}5)^2 + (-5{,}5)^2 = 51;$$

$$P_2 = \sum_{i=1}^{4} {y_{i2}}^2 = (0{,}5)^2 + (4{,}5)^2 + (-1{,}5)^2 + (-2{,}5)^2 = 29;$$

$$P_3 = \sum_{i=1}^{4} {y_{i3}}^2 = (1{,}5)^2 + (2{,}5)^2 + (7{,}5)^2 + (0{,}5)^2 = 65;$$

$$R_1 = \sum_{i=1}^{4} y_{i1} = -2{,}5 - 3{,}5 - 1{,}5 - 5{,}5 = -13;$$

$$R_2 = \sum_{i=1}^{4} y_{i2} = 0{,}5 + 4{,}5 - 1{,}5 - 2{,}5 = 1;$$

$$R_3 = \sum_{i=1}^{4} y_{i3} = 1{,}5 + 2{,}5 + 7{,}5 + 0{,}5 = 12\,.$$

Рассчитаем значения факторной и остаточной дисперсии:

$$S^2_{факт} = \frac{\sum_{j=1}^{l} \frac{R_j^2}{q} - \frac{1}{lq}\left(\sum_{j=1}^{l} R_j\right)^2}{l-1}, \qquad S^2_{ост} = \frac{\sum_{j=1}^{l} P_j - \frac{1}{q}\sum_{j=1}^{l} R_j^2}{l(q-1)},$$

таким образом, получим:

$$S^2_{факт} = \frac{\frac{1}{4}\cdot\left[(-13)^2 + 1^2 + 12^2\right] - \frac{1}{4\cdot 3}\left[-13+1+12\right]^2}{3-1} = 39{,}25\,,$$

$$S^2_{ост} = \frac{(51+29+65) - \frac{1}{4}\cdot\left[(-13)^2 + 1^2 + 12^2\right]}{3(4-1)} = 7{,}4\,.$$

Так как $S^2_{факт} > S^2_{ост}$, следует проверить значимость их различия. Вычислим экспериментальное значение критерия:

$$F_{эксп} = \frac{S^2_{ф}}{S^2_{ост}} = \frac{39{,}25}{7{,}4} = 5{,}3\,.$$

По таблице критических значений распределений Фишера – Снедекора при уровне значимости α = 0,05, определим критическое значение $F_{кр}(\alpha, f_1, f_2)$. Здесь $f_1 = l - 1 = 3 - 1 = 2$ – число степеней свободы факторное, $f_2 = l(q-1) = 3(4-1) = 9$ – число степеней свободы остаточное. Таким образом:

$$F_{кр}(\alpha, f_1, f_2) = F_{кр}(0{,}05;2;9) = 4{,}26$$

Так как $F_{эксп} > F_{кр}$, то нулевая гипотеза отвергается в пользу конкурирующей, т.е. групповые средние различаются значимо. Внешнее воздействие надо признать эффективным.

9.3. Однофакторный дисперсионный анализ при неодинаковом числе испытаний на уровнях

Если число испытаний, проведенных на различных уровнях действия фактора, различно, а именно: на уровне A_1 проведено q_1 испытаний, на уровне A_2 – q_2 испытаний и т.д., на уровне A_l – q_l испытаний, то факторную и остаточную дисперсии находят по следующим формулам:

$$S^2_{факт} = \frac{\sum_{j=1}^{l} \frac{R_j^2}{q_j} - \left(\sum_{j=1}^{l} R_j\right)^2 / N}{l-1}, \qquad S^2_{ост} = \frac{\sum_{j=1}^{l} P_j - \sum_{j=1}^{l} \frac{R_j^2}{q_j}}{N-l}.$$

Здесь $R_j = \sum_{i=1}^{q_j} x_{ij}$ – сумма значений величины X на уровне A_j; $N = \sum_{j=1}^{l} q_j$ – общее количество результатов испытаний; $P_j = \sum_{i=1}^{q_j} x_{ij}^{\ 2}$ – сумма квадратов значений величины X на уровне A_j.

Далее следует действовать так же, как и в случае одинакового числа наблюдений на различных уровнях, но при нахождении критического значения распределения Фишера – Снедекора число степеней свободы факторной дисперсии принимают равным $f_1 = l - 1$, остаточной – $f_2 = N - l$.

Обычно для упрощения расчетов от значений x_{ij} переходят к рассмотрению значений $y_{ij} = x_{ij} - C$, где постоянная C близка по величине к общей средней $\bar{x}$ всех измерений величины X.

Пример 10.2: По результатам, представленным в таблице, при уровне значимости $\alpha=0,05$ проверить, имеется ли различие между способами A_1, A_2, A_3 и A_4 обработки фармацевтического сырья с точки зрения большего выхода продукта.

Номер испытания	Способ обработки A_j			
	A_1	A_2	A_3	A_4
1	70	71	78	73
2	75	70	74	72
3	73	72	70	71
4	72	73	76	74
5	71		72	72
6	74			
$\overline{x_j}$	72,5	71,5	74,0	72,4

Решение:

Предполагая, что распределения значений, характеризующих выход сырья, при каждом способе обработки являются нормальными, а соответствующие генеральные дисперсии равны, применим метод однофакторного дисперсионного анализа. Обработку сырья будем считать фактором, а способы обработки – уровнями.

Общая средняя всех измерений

$$\bar{x} = \sum_{j=1}^{l} \frac{\overline{x_j}}{l} = \frac{72{,}5 + 71{,}5 + 74{,}0 + 72{,}4}{4} = 72{,}65,$$

где l - количество уровней фактора A (l=4).

Введем новые переменные:

$$y_{ij} = x_{ij} - C = x_{ij} - \bar{x} = x_{ij} - 73$$

Преобразуем таблицу с дальнейшим использованием ее в качестве расчетной.

Номер испытания	Уровень фактора A_j			
	A_1	A_2	A_3	A_4
	y_{i1}	y_{i2}	y_{i3}	y_{i4}
1	-3	-2	5	0
2	2	-3	1	-1
3	0	-1	-3	-2
4	-1	0	3	1
5	-2		-1	-1
6	1			

С помощью таблицы определим вспомогательные величины:

$$P_j = \sum_{i=1}^{q} y_{ij}^{\ 2}; \qquad R_j = \sum_{i=1}^{q} y_{ij}$$

$$P_1 = \sum_{i=1}^{6} y_{i1}^{\ 2} = (-3)^2 + 2^2 + 0^2 + (-1)^2 + (-2)^2 + 1^2 = 19;$$

$$P_2 = 14; \qquad P_3 = 45; \qquad P_4 = 7.$$

$$R_1 = \sum_{i=1}^{6} y_{i1} = (-3) + 2 + 0 + (-1) + (-2) + 1 = -3;$$

$$R_2 = -6; \qquad R_3 = 5; \qquad R_4 = -3.$$

Рассчитаем значения факторной и остаточной дисперсии:

$$S_{факт}^2 = \frac{\sum_{j=1}^{l} \frac{R_j^2}{q} - \frac{1}{lq}\left(\sum_{j=1}^{l} R_j\right)^2}{l-1}, \qquad S_{ост}^2 = \frac{\sum_{j=1}^{l} P_j - \frac{1}{q}\sum_{j=1}^{l} R_j^2}{l(q-1)}.$$

$$S_{факт}^2 = 4{,}95, \qquad S_{ост}^2 = 4{,}23,$$

Так как $S_{факт}^2 > S_{ост}^2$, следует проверить значимость различия между этими дисперсиями. Вычислим экспериментальное значение критерия:

$$F_{эксп} = \frac{S_ф^2}{S_{ост}^2} = \frac{4{,}95}{4{,}23} \approx 1{,}17.$$

По таблице критических значений распределений Фишера – Снедекора при уровне значимости α = 0,05, определим критическое значение $F_{кр}(\alpha, f_1, f_2)$. Здесь $f_1 = l - 1 = 4 - 1 = 3$ – число степеней свободы факторное, $f_2 = N - l = 20 - 4 = 16$ – число степеней свободы остаточное. Таким образом:

$$F_{кр}(\alpha, f_1, f_2) = F_{кр}(0{,}05;3;16) \approx 3{,}24$$

Так как $F_{эксп} < F_{кр}$, то различие между факторной и остаточной дисперсиями не является значимым. В соответствии с методом дисперсионного анализа отсюда следует вывод о незначимости различия между групповыми средними, что соответствует отсутствию существенного различия между рассмотренными способами обработки сырья.

9.4. Понятие о двухфакторном и многофакторном анализах

Оценим влияние двух одновременно действующих факторов A и B на формирование значений нормально распределенной случайной величины X, причем фактор A имеет m уровней действия (A_1, A_2, …, A_m), фактор B – n уровней (B_1, B_2, …, B_n). Для простоты ограничимся рассмотрением результатов таких экспериментов, когда при действии каждой пары уровней фактора (A_i, B_i) (i=1,2,…,m; j=1,2,…, n) производится лишь одно наблюдение величины X. В этом случае результаты эксперимента могут быть представлены в виде таблицы 10.2.

Таблица 10.2

Уровни A_i	Уровни B_j				$\overline{x}_i^A$
	B_1	B_2	…	B_n	
A_1	x_{11}	x_{12}	…	x_{1n}	$\overline{x}_1^A$
A_2	x_{21}	x_{22}	…	x_{2n}	$\overline{x}_2^A$
…	…	…	…	…	…
A_m	x_{m1}	x_{m2}	…	x_{mn}	$\overline{x}_m^A$
$\overline{x}_j^B$	$\overline{x}_1^B$	$\overline{x}_2^B$	…	$\overline{x}_n^B$	

В последнем столбце таблицы 10.2 представлены средние групповые (по строкам) значения $\overline{x}_i^A = \sum_{j=1}^{n} \frac{x_{ij}}{n}$, в последней строке – средние групповые (по столбцам) $\overline{x}_j^B = \sum_{i=1}^{m} \frac{x_{ij}}{m}$.

В соответствии с методом дисперсионного анализа для проверки значимости различий между отдельными значениями $\overline{x}_i^A$, а также между отдельными значениями $\overline{x}_j^B$ рассчитывают оценки двух факторных S_A^2, S_B^2 и остаточной $S_{ост}^2$ дисперсий по формулам:

$$S_A^2 = \frac{\sum_{i=1}^{m} R_i^2 - \left(\sum_{i=1}^{m} R_i\right)^2 / m}{n(m-1)}; \qquad S_B^2 = \frac{\sum_{j=1}^{n} V_j^2 - \left(\sum_{i=1}^{m} R_i\right)^2 / n}{m(n-1)};$$

$$S_{ост}^2 = \frac{\sum_{i=1}^{m} P_i - \sum_{i=1}^{m} R_i^2 / n - \sum_{j=1}^{n} V_j^2 / m + \left(\sum_{i=1}^{m} R_i\right)^2 / (mn)}{(m-1)(n-1)},$$

где $\quad R_i = \sum_{j=1}^{n} x_{ij}$; $\quad V_j = \sum_{i=1}^{m} x_{ij}$; $\quad P_i = \sum_{j=1}^{n} x_{ij}^2$.

Поскольку оценки S_A^2, S_B^2 и $S_{ост}^2$ характеризуют роль соответственно фактора *A*, фактора *B* и случайных причин, то значимость влияния каждого из факторов *A* и *B* на величину *X* определяют, сравнивая S_A^2 и S_B^2 (в отдельности) с $S_{ост}^2$ так же, как и при однофакторном анализе. При этом числа степеней свободы S_A^2, S_B^2 и $S_{ост}^2$ равны соответственно *m–1*, *n–1*, (*m–1*)(*n–1*).

Пример 10.3: При выяснении влияния реагентов *A* и *B* на синтез лекарственного препарата получены результаты (выход *X* в условных единицах), представленные в таблице.

Уровни A_i	Уровни B_j				$\bar{x}_i^A$
	B_1	B_2	B_3	B_4	
A_1	4,5	3,0	4,0	3,5	3,75
A_2	3,5	2,5	3,5	2,0	2,88
A_3	6,5	5,5	4,5	6,0	5,62
A_4	7,5	7,0	8,5	7,0	7,50
$\bar{x}_j^B$	5,5	4,5	5,12	4,62	

В предположении нормальности распределения величины *X* методом дисперсионного анализа при уровне значимости 0,05 проверить значимость влияния каждого из реагентов *A* и *B* на выход препарата.

Решение:

Найдем значения P_i, R_i и V_j:

$P_1 = 57{,}5$; $\quad P_2 = 34{,}75$; $\quad P_3 = 128{,}75$; $\quad P_4 = 226{,}5$;

$R_1 = 15$; $\quad R_2 = 11{,}5$; $\quad R_3 = 22{,}5$; $\quad R_4 = 30$;

$V_1 = 22$; $\quad V_2 = 18$; $\quad V_3 = 20{,}5$; $\quad V_4 = 18{,}5$.

Вычислим S_A^2, S_B^2 и $S_{ост}^2$:

$S_A^2 = 16{,}9$, $\quad S_B^2 = 0{,}854$, $\quad S_{ост}^2 = 0{,}451$.

Поскольку S_A^2 и S_B^2 превышают $S_{ост}^2$, рассчитаем экспериментальные значения критерия Фишера – Снедекора:

$$F_{эксп}^A = \frac{S_A^2}{S_{ост}^2} \approx 37{,}5; \qquad F_{эксп}^B = \frac{S_B^2}{S_{ост}^2} \approx 1{,}89$$

и сравним их с критическими :

$$F_{кр}^A = F_{кр}(\alpha, m-1, (m-1)(n-1)) = F_{кр}(0{,}05;3;9) \approx 3{,}86;$$

$$F_{кр}^B = F_{кр}(\alpha, n-1, (m-1)(n-1)) = F_{кр}(0{,}05;3;9) \approx 3{,}86.$$

Так как $F^{A}_{эксп} > F^{A}_{кр}$, при уровне значимости $\alpha = 0{,}05$ следует сделать вывод о существенном влиянии реагента *A* на выход препарата. В отношении реагента *B* такого вывода сделать нельзя, поскольку $F^{B}_{эксп} < F^{B}_{кр}$.