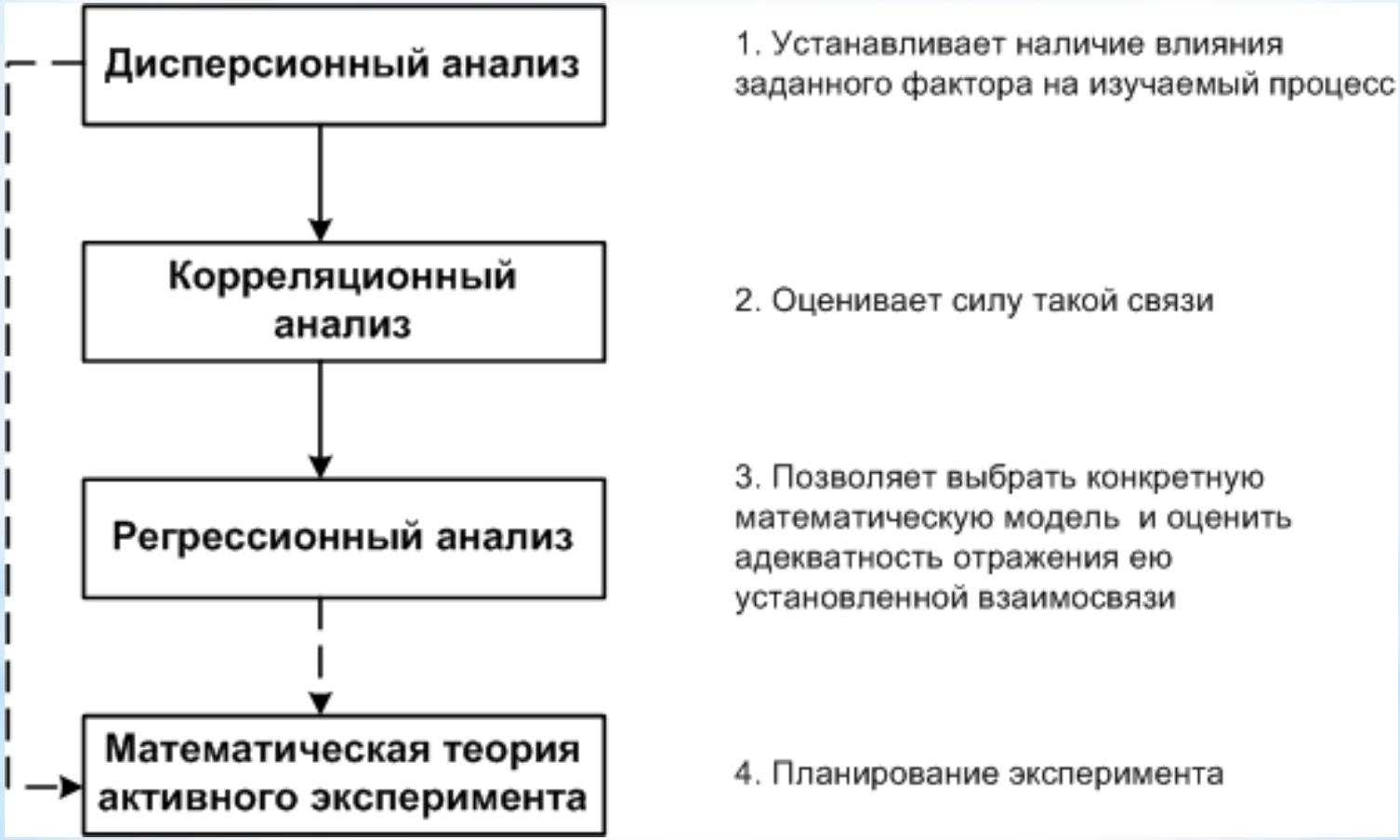
* Методы исследования связей между случайными величинами



Основы дисперсионного анализа

Это статистический метод анализа результатов наблюдений, зависящих от различных одновременно действующих факторов, с целью выбора наиболее значимых факторов и оценки их влияния на исследуемый процесс.

Предположения:

- *распределения с.в. нормально;*
- *дисперсии экспериментальных данных одинаковы для всех условий эксперимента.*

Основная идея: сравнение «факторной дисперсии», порождаемой воздействием фактора, и «остаточной дисперсии», обусловленной случайными признаками.

Если дисперсии отличаются значимо, то следует вывод о значимом влиянии фактора на среднее значение наблюдаемой случайной величины.

Основы дисперсионного анализа

Виды дисперсионного анализа:

◊ одномерный (одна зависимая переменная) и многомерный (несколько зависимых переменных);

◊ однофакторный (одна группирующая переменная) и многофакторный (несколько группирующих переменных) с возможным взаимодействием между факторами;

◊ с простыми измерениями (зависимая переменная измеряется лишь один раз) и с повторными (зависимая переменная измеряется несколько раз).

Гипотезы:

Основная H_0: $\mu_1 = \mu_2 = \ldots = \mu_c$ все средние одинаковы;

Альтернативная H_1: не все μ_j одинаковы $j = 1, 2, \ldots, c$.

Основы дисперсионного анализа

Результативные признаки (показатели, зависимые переменные) – их значения определяется с помощью измерений в ходе эксперимента - количественная шкала (например, цена, объем производства, урожайность).

Факторы - переменные, вызывающие изменчивость средних значений результативных признаков – номинальная шкала (например, тип производства, сорт сырья, изменения в законодательстве, климатические условия).

Уровень фактора – значения, которые может принимать фактор.

Отклик – значение измеряемого признака.

Основы дисперсионного анализа

Однофакторный дисперсионный анализ

Анализируется влияние фактора *A*, изучаемого на *k* уровнях (A_1, A_2,…, A_k). На каждом уровне A_i проведены *n* наблюдений (x_{i1}, x_{i2},…, x_{in}), т.е. на всех *k* уровнях фактора *A* произведены *kn* наблюдений.

Номер наблюдения	Уровни фактора A					
	A_1	A_2	…	A_i	…	A_k
1	x_{11}	x_{21}	…	x_{i1}	…	x_{k1}
2	x_{12}	x_{22}	…	x_{i2}	…	x_{k2}
.	.	.	…	.	…	.
.	.	.	…	.	…	.
.	.	.	…	.	…	.
j	x_{1j}	x_{2j}	…	x_{ij}	…	x_{kj}
.	.	.	…	.	…	.
.	.	.	…	.	…	.
.	.	.	…	.	…	.
n	x_{1n}	x_{2n}	…	x_{in}	…	x_{kn}
$\sum$	X_1	X_2	…	X_i	…	X_k

Однофакторный дисперсионный анализ

Алгоритм:

1. Вычислить суммы

$$Q_1 = \sum_{i=1}^{k} \sum_{j=1}^{n} x_{ij}^2; \quad Q_2 = \frac{1}{n} \sum_{i=1}^{k} X_i^2; \quad Q_3 = \frac{1}{kn} \left(\sum_{i=1}^{k} X_i \right)^2$$

2. Получить оценки $S_0^2 = \dfrac{Q_1 - Q_2}{k\,(n-1)}; \quad S_A^2 = \dfrac{Q_2 - Q_3}{k-1}$

3. Если $\dfrac{k\,(n-1)}{k-1} \dfrac{Q_2 - Q_3}{Q_1 - Q_2} > F_\alpha [k-1; k\,(n-1)],$

где $F\alpha(f_1, f_2)$— α-квантиль F-распределения с f_1 и f_2 степенями свободы,

то влияние фактора A признается значимым, в противном случае всю выборку наблюдений считают однородной с общей дисперсией $S^2 = \dfrac{Q_1 - Q_3}{kn-1}.$

Однофакторный дисперсионный анализ

Пример. Провести однофакторный дисперсионный анализ данных при доверительной вероятности α=0,95.

i	Уровни фактора A_i				
	A_1	A_2	A_3	A_4	A_5
1	3,2	2,6	2,9	3,6	3,0
2	3,1	3,1	2,6	3,4	3,4
3	3,1	2,7	3,0	3,2	3,2
4	2,8	2,9	3,1	3,3	3,5
5	3,3	2,7	3,0	3,5	2,9
6	3,0	2,8	2,8	3,3	3,1
$\sum$	18,5	16,8	17,4	20,3	19,1

$$Q_1 = \sum_{i=1}^{5} \sum_{j=1}^{6} x_{ij}^2 = 284{,}8;$$

$$Q_2 = \frac{1}{6} \cdot \sum_{i=1}^{5} X_i^2 = \frac{1}{6} \cdot \left(18{,}5^2 + 16{,}8^2 + \ldots + 19{,}1^2\right) = 284{,}025;$$

$$Q_3 = \frac{1}{5 \cdot 6} \cdot \left(\sum_{i=1}^{5} X_i \right)^2 = \frac{1}{30} \cdot (18{,}5 + 16{,}8 + 17{,}4 + 20{,}3 + 19{,}1)^2 = 282{,}747.$$

$$S_0^2 = \frac{284{,}87 - 284{,}025}{5 \cdot (6-1)} = 0{,}0338; \quad S_A^2 = \frac{284{,}025 - 282{,}747}{5-1} = 0{,}319;$$

$$\frac{S_A^2}{S_0^2} = \frac{0{,}319}{0{,}0338} = 9{,}45.$$

Для $f_1 = k - 1 = 4$ и $f_2 = k(n - 1) = 25$ находим $F_{0,95}(4; 25) = 2{,}8$

Так как

$$\frac{S_A^2}{S_0^2} = 9{,}45 > F_{0,95}(4; 25) = 2{,}8,$$

то влияние фактора A на с.в. следует признать значимым.

Двухфакторный дисперсионный анализ

Предположения:

- при различных сочетаниях уровней факторов *A* и *B* наблюдения независимы;

- при каждом сочетании уровней факторов *A* и *B* результативный признак ***Y*** имеет нормальный закон распределения с постоянной для различных сочетаний генеральной дисперсией σ^2.

Разновидности:

- без повторений - каждому уровню факторов соответствует только одна выборка данных,

- с повторениями - определенным уровням факторов может соответствовать более одной выборки данных.

Двухфакторный дисперсионный анализ

Факторы A и B независимы и фиксируются на уровнях $A_1, A_2, \ldots, A_k$ и $B_1, B_2, \ldots, B_m$ соответственно.

Результаты эксперимента представляют в виде таблицы:

B	A						$\sum$
	A_1	A_2	$\ldots$	A_i	$\ldots$	A_k	
B_1	x_{11}	x_{21}	$\ldots$	x_{i1}	$\ldots$	x_{k1}	$X_{1'}$
B_2	x_{12}	x_{22}	$\ldots$	x_{i2}	$\ldots$	x_{k2}	$X_{2'}$
.	.	.	$\ldots$	.	$\ldots$	$\ldots$	.
.	.	.	$\ldots$	.	$\ldots$	$\ldots$	.
.	.	.	$\ldots$	.	$\ldots$	$\ldots$	.
B_j	x_{ij}	x_{2j}	$\ldots$	x_{ij}	x_{kj}	x_{kj}	$X_{j'}$
.	.	.	$\ldots$	.	$\ldots$	$\ldots$	.
.	.	.	$\ldots$	.	$\ldots$	$\ldots$	.
.	.	.	$\ldots$	.	$\ldots$	$\ldots$	.
B_m	x_{1m}	x_{2m}	$\ldots$	x_{im}	$\ldots$	x_{km}	$X_{m'}$
$\sum$	X_1	X_2	$\ldots$	X_i	$\ldots$	X_k	

Двухфакторный дисперсионный анализ

Алгоритм:

1. Вычислить суммы

$$Q_1 = \sum_{i=1}^{k}\sum_{j=1}^{m} x_{ij}^2; \quad Q_2 = \frac{1}{m}\sum_{i=1}^{k} X_i^2; \quad Q_3 = \frac{1}{k}\sum_{j=1}^{m} X_{j'}^2;$$

$$Q_4 = \frac{1}{mk}\left(\sum_{i=1}^{k} X_i\right)^2 = \frac{1}{mk}\left(\sum_{j=1}^{m} X_{j'}\right)^2.$$

2. Оценить дисперсии

$$S_0^2 = \frac{Q_1 + Q_4 - Q_2 - Q_3}{(k-1)(m-1)}; \quad S_A^2 = \frac{Q_2 - Q_4}{k-1}; \quad S_B^2 = \frac{Q_3 - Q_4}{m-1}.$$

3. Принятие решения: если $\frac{S_A^2}{S_0^2} > F_\alpha(f_1, f_2)$,

где $f_1 = k - 1$ и $f_2 = (k - 1)(m - 1)$, то влияние фактора A с достоверностью α признается значимым.

Двухфакторный дисперсионный анализ

Если *A* и *B* зависимы, то при каждом сочетании факторов *A* и *B* на уровнях A_i, B_j необходима серия наблюдений x_{ij1}, x_{ij2},…, x_{ijn} со средним

$$x_{ij} = \frac{1}{n}\sum_{\nu=1}^{n} x_{ij\nu}$$

Далее вычисления аналогичны:

1. Вычислить дополнительную сумму $$Q_5 = \sum_{i=1}^{k}\sum_{j=1}^{m}\sum_{\nu=1}^{n} x_{ij\nu}^2.$$

2. Вычислить дисперсию $$S_{AB}^2 = \frac{Q_5 - nQ_1}{mk\,(n-1)}$$

3. Проверить значимость взаимодействия: если

$$\frac{nS_0^2}{S_{AB}^2} > F_\alpha(f_1, f_2)$$

где f_1= (*k* – 1)(*n* - 1) и f_2 = *mk*(*n* – 1), то влияние признается значимым

Пример. Провести двухфакторный дисперсионный анализ данных, представленных следующей таблицей при доверительной вероятности α= 0,95:

B	A								
	A_1			A_2			A_3		
B_1	3,6	3,8	4,1	2,9	3,1	3,0	2,6	2,5	2,9
B_2	4,2	4,0	4,1	3,3	2,9	3,2	3,7	3,5	3,6
B_3	3,8	3,5	3,6	3,6	3,7	3,5	3,2	3,0	3,4
B_4	3,4	3,2	3,2	3,4	3,6	3,5	3,6	3,8	3,7

Для серий значений вычисляют средние:

B	A			$\sum$
	A_1	A_2	A_3	
B_1	3,83	3,00	2,67	9,50
B_2	4,10	3,13	3,60	10,83
B_3	3,63	3,60	3,20	10,43
B_4	3,27	3,50	3,70	10,47
$\sum$	14,83	13,23	13,17	41,23

По данным таблицы вычисляем

$$Q_1 = \sum_{i=1}^{3}\sum_{j=1}^{4} x_{ij}^2 = 143{,}34; \quad Q_2 = \frac{1}{4}\cdot\sum_{i=1}^{3} X_i^2 = 142{,}102675;$$

$$Q_3 = \frac{1}{3}\cdot\sum_{j=1}^{4} X_j^2 = 141{,}98157;$$

$$Q_4 = \frac{1}{4\cdot 3}\cdot\left(\sum_{i=1}^{3} X_i\right)^2 = 141{,}6594; \quad Q_5 = \sum_{i=1}^{3}\sum_{j=1}^{4}\sum_{\nu=1}^{3} x_{ij\nu}^2 = 430{,}79.$$

$$S_0^2 = \frac{Q_1 + Q_4 - Q_2 - Q_3}{(k-1)\cdot(m-1)} =$$

$$= \frac{143{,}3745 + 141{,}6594 - 142{,}102675 - 141{,}98157}{2\cdot 3} = 0{,}1582;$$

$$S_A^2 = \frac{Q_2 - Q_4}{k-1} = \frac{142{,}3745 - 141{,}6594}{2} = 0{,}223675;$$

$$S_B^2 = \frac{Q_3 - Q_4}{m-1} = \frac{141{,}98157 - 141{,}6594}{3} = 0{,}10739;$$

$$S_{AB}^2 = \frac{Q_5 - n \cdot Q_1}{mk \cdot (n-1)} = \frac{430{,}79 - 3 \cdot 143{,}3745}{4 \cdot 3 \cdot 2} = 0{,}02777;$$

$$\frac{S_A^2}{S_0^2} = \frac{0{,}223675}{0{,}1582} = 1{,}41; \quad \frac{S_B^2}{S_0^2} = \frac{0{,}10739}{0{,}1582} = 0{,}679;$$

$$\frac{n \cdot S_0^2}{S_{AB}^2} = \frac{3 \cdot 0{,}1582}{0{,}02777} = 19{,}98$$

Проверка значимости

$$F_{0,95}[k-1; (k-1)\cdot(m-1)] = F_{0,95}(2; 6) = 5{,}1;$$

$$F_{0,95}[m-1; (k-1)\cdot(m-1)] = 4{,}8;$$

$$F_{0,95}[(k-1)\cdot(m-1); mk\cdot(n-1)] = F_{0,95}(6; 24) = 2{,}5.$$

$$\frac{S_A^2}{S_0^2} = 1{,}41 < F_{0{,}95}(2;6) = 5{,}1;$$

$$\frac{S_B^2}{S_0^2} = 0{,}679 < F_{0{,}95}(3;6) = 4{,}8;$$

$$\frac{n \cdot S_0^2}{S_{AB}^2} = 17{,}09 > F_{0{,}95}(6;24) = 2{,}5.$$

Следовательно, влияние факторов *A* и *B* должно быть признано незначимым. Однако существенно значимым является взаимодействие факторов A и B.