Дисперсионный анализ

План.

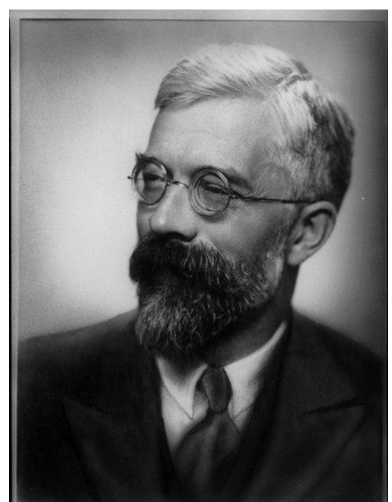
1. Сущность дисперсионного анализа. Градации факторов и их анализ.
2. Простейшая схема варьирования при различии по одному фактору.
3. Рабочие формулы для вычисления дисперсий.
4. Вычисление F- критерия для определения влияния изучаемого фактора в общей изменчивости изучаемого признака.
5. Количественная оценка влияния отдельных факторов.

Сущность дисперсионного анализа.

Установление роли отдельных факторов в изменчивости того или иного признака.

Дисперсионный анализ позволяет оценивать значимость влияния отдельных факторов, а так же их относительную роль в общей изменчивости признака.

Основатель дисперсионного анализа – Рональд Фишер.



В фактическом отклонении варианты от средней генеральной совокупности фигурируют два компонента:

- та часть отклонения, которая зависит именно от данного фактора – **А**;
- остаточная часть, независящая от данного фактора – **е**.

В таком случае можно сравнивать **А** и **е**.

При достоверном влиянии изучаемого фактора значение **А** будет превышать значение **е**.

По степени превышения **А** над **е** можно судить о том, насколько достоверно влияние данного фактора.

Разберём простейшую схему, когда анализируется влияние одного фактора, могущего принимать разные градации, или количественные уровни: **1,2,3… i …a.**

Отдельные наблюдения (варианты) разбиваются на группы согласно этим градациям фактора.

Количество наблюдений в одном уровне: **1,2,3 …j …n.**

Распределение вариант при различии по одному фактору

	Отдельные варианты (наблюдения)						$\sum X_i = T_i$	T_i^2
	1	**2**	**…**	**j**	**…**	**n**		
1	X_{11}	X_{12}		X_{1j}		X_{1n}	$\sum X_1 = T_1$	T_1^2
2	X_{21}	X_{22}		X_{2j}		X_{2n}	$\sum X_2 = T_2$	T_2^2
:	**:**	**:**	**:**	**:**	**:**	**:**	**:**	**:**
i	X_{i1}	X_{i2}		X_{ij}		X_{in}	$\sum X_i = T_i$	T_i^2
:	**:**	**:**	**:**	**:**	**:**	**:**	**:**	**:**
a	X_{a1}	X_{a2}		X_{aj}		X_{an}	$\sum X_a = T_a$	T_a^2
							$\sum T_i = T$	$\sum T_i^2 =$
							$T^2 =$	

Найдём сумму квадратов, составив дополнительную таблицу ($\sum X_{ij}^2$).

						$\sum X_i^2$
X_{11}^2	X_{12}^2	…	X_{1j}^2	…	X_{1n}^2	
X_{21}^2	X_{22}^2	…	X_{2j}^2	…	X_{2n}^2	
…	….	…	…	…		
X_{i1}^2	X_{i2}^2	…	X_{ij}^2	…	X_{in}^2	
…	…	…		…		
X_{a1}^2	X_{a2}^2	…	X_{aj}^2	…	X_{an}^2	
						$\sum X_{ij}^2$

Степени свободы.

Для общей дисперсии: k_0=N-1

Для факториальной дисперсии: $к_A$=a-1

Для остаточной дисперсии: $к_e$=N-a

N=a·n

Формулы для вычисления дисперсий.

1. **Общая дисперсия** $\sigma_0^2 = \frac{1}{N-1}\left(\sum X_{ij}^2 - \frac{T^2}{N}\right)$

2. **Факториальная дисперсия** $\sigma_A^2 = \frac{1}{a-1}\left(\sum \frac{T_i^2}{n} - \frac{T^2}{N}\right)$

3. **Остаточная дисперсия** $\sigma_e^2 = \frac{1}{N-a}\left(\sum X_{ij}^2 - \sum \frac{T_i^2}{n}\right)$

Нулевая гипотеза: данный фактор **А** не влияет на изменчивость данного признака.

Для того, чтобы отбросить нулевую гипотезу, нужно доказать, что σ_A^2 – достоверно (т. е. с вероятностью, не меньше, чем 0.95, или с α=0.05) отличается от нуля.

Достоверность значения σ_A^2 может быть установлена путём деления на ошибку, т. е. на σ_e^2:

$$F = \frac{\sigma_A^2}{\sigma_e^2}$$

Так как фактически полученное дисперсионное отношение $F_ф$ является величиной случайной, его необходимо сравнить с табличным (стандартным) значением критерия Фишера $F_{st.}$ для принятого уровня значимости α и чисел степеней свободы $к_A$ и $к_e$

При этом число степеней свободы для большей дисперсии находят в верхней строке, а для меньшей – в первом столбце таблицы.

Нулевую гипотезу H_0 отвергают и эффективность действия фактора **А** на результативный признак признают статистически достоверной, если $F_ф \geq F_{st.}$

Если $F_ф < F_{st}$, то нулевая гипотеза H_0 сохраняется, и фактор **А** не влияет на результативный признак.

Количественная оценка влияния отдельных факторов.

Наряду с доказательством влияния того или иного фактора на результативный признак, часто возникает необходимость установления меры этого влияния и его доли в сумме влияния всех факторов. Доля влияния фактора **А** равна:

$$P_A = \frac{\sigma_A^2 - \sigma_e^2}{\sigma_A^2 + \sigma_e^2 (n-1)}$$

Задача. Получены следующие данные о плодовитости самок мышей при облучении их рентгеновскими лучами:

Группы	**Число мышат от отдельных самок**			
Доза 0 р	10	12	11	10
Доза 100 р	8	10	7	9
Доза 200р	7	9	6	4

Влияет ли облучение на плодовитость мышей?

Схема решения задачи на дисперсионный анализ:

1.Суммируют данные задачи по каждому уровню фактора А **(T_i).**

2.Находят общую сумму **($\sum T_i$)** по всем уровням и получают значение **(T).**

3. Возводят в квадрат общую сумму **($\sum T_i$)** и получают значение **T^2**

4. Возводят полученные суммы **(T_i)** в квадрат и получают значения **(T_i^2)** по каждому уровню.

5. Находят общую сумму **($\sum T_i^2$)**

a	Отдельные наблюдения (n)				$T_i = \sum X_i$	T_i^2
	1	2	3	4		
1	10	12	11	10	43	1849
2	8	10	7	9	34	1156
3	7	9	6	4	26	676
					T=103	$\sum$ **T_i^2=**3681
					T^2=10609	

6.Находят сумму квадратов данных задачи, составив дополнительную таблицу.

1	2	3	4	$\sum X_i^2$
100	144	121	100	465
64	100	49	81	294
49	81	36	16	182
				$\sum X_{ij}^2=941$

7. Определяют число степеней свободы:

- Для факториальной дисперсии

$$K_A = a - 1 = 3 - 1 = 2$$

- Для остаточной дисперсии

$$K_e = N - a = 12 - 3 = 9$$

8. Рассчитывают дисперсии:

- Факториальная дисперсия

$$\sigma_A^{\,2} = \frac{1}{a-1}\left(\sum \frac{T_i^{\,2}}{n} - \frac{T^2}{N}\right)$$

$$\sigma_A^2 = \frac{1}{2} \cdot \left(\frac{3681}{4} - \frac{10609}{12}\right) = 18{,}08$$

➢ Остаточная дисперсия

$$\sigma_e^2 = \frac{1}{N-a}\left(\sum X_{ij}^2 - \sum \frac{T_i^2}{n}\right)$$

$$\sigma_e^2 = \frac{1}{9} \cdot \left(941 - \frac{3681}{4}\right) = 2{,}3$$

9. Находят фактическое значение F-критерия Фишера:

$F_ф$ =18,08/2,3=**7,86**

10. Находят стандартное значение F-критерия Фишера (по таблице №4)

F_{st} =4,26 при уровне значимости 0,05 и числа степеней свободы для большей дисперсии **(K_A=3-1=2)** и меньшей дисперсии **(K_e=12-3=9).**

11. Сравнивают фактически найденное значение F-критерия Фишера со стандартным значением и делают вывод.

$F_{ф} \geq F_{st}$, нулевую гипотезу отвергают на 5% уровне значимости.

Вывод: С вероятностью 0,95 можно заключить, что облучение рентгеновскими лучами влияет на плодовитость мышей.

12. Определяют степень влияния фактора А по формуле:

$$P = \frac{\sigma_A^2 - \sigma_e^2}{\sigma_A^2 + \sigma_e^2(n-1)}$$

$$P = \frac{18{,}08 - 2{,}3}{18{,}08 + 2{,}3(4-1)} = 0{,}63$$

Это означает, что примерно около 63% от общего варьирования плодовитости мышей обусловлено облучением.

Ответ:

1. С вероятностью 0,95 можно заключить, что облучение рентгеновскими лучами влияет на плодовитость мышей.

2. Сила влияния фактора А на результативный признак равна Р=0,63.