



Непараметрический дисперсионный анализ

Лекция №10
для студентов 2 курса,
обучающихся по специальности 060609 –
Медицинская кибернетика
доц. Шапиро Л.А.
Красноярск, 2015 г.

План лекции:

- Актуальность темы.
- Непараметрический дисперсионный анализ для зависимых выборок.
- Непараметрический дисперсионный анализ для независимых выборок.
- Критерий Колмогорова-Смирнова.
- Заключение.

Сравнение более двух зависимых выборок.

Критерий Фридмана (χ^2) - это непараметрический аналог дисперсионного анализа повторных измерений (ANOVA).

Проверяется гипотеза о различии более двух зависимых выборок по уровню выраженности изучаемого признака.

1. Результаты наблюдения у каждого объекта упорядочиваются (по строке). Причем отдельно упорядочиваем значения у каждого объекта независимо от всех остальных. Таким образом получается столько упорядоченных рядов, сколько объектов участвует в исследовании.
2. Вычисляется сумма рангов для каждого уровня фактора (по столбцам).
3. Вычисляется эмпирическое значение критерия χ^2 - Фридмана

$$\chi^2 = \left[\frac{12}{Nk(k+1)} \sum_{i=1}^k R_i^2 \right] - 3N(k+1), \quad df = k - 1$$

Чем больше различаются зависимые выборки по изучаемому признаку, тем больше эмпирическое значение критерия χ^2 –Фридмана.

где N -число объектов, k -число уровней фактора (повторных измерений), R_i -сумма рангов для соответствующего уровня i .

4. Находится $\chi^2_{\text{крит}}$ для $df=k-1$ и $\alpha=0,05$.

При $k=3$, $N>9$ или $k>3$, $N>4$ пользуются обычной таблицей распределения χ^2 .

При $k=3$, $N<10$ или $k=4$, $N<5$ пользуются дополнительной таблицей критических значений χ^2 - Фридмана.

5. Определяется уровень значимости.

Если $\chi^2_{\text{эмп}} \geq \chi^2_{\text{крит}}$ нулевая гипотеза отвергается.
Различия статистически значимы.

Если $\chi^2_{\text{эмп}} < \chi^2_{\text{крит}}$ нулевая гипотеза не отвергается.
Различия статистически не значимы.

Если разброс сумм велик и различия статистически значимы, переходим к межгрупповым сравнениям по критерию Вилкоксона с поправкой Бонферрони.

Пример:

Результаты тестирования студентов по семестрам

№	1 семестр	2 семестр	3 семестр	4 семестр
1	6	14	5	14
2	11	5	4	12
3	12	8	7	10
4	8	10	11	12
5	5	14	10	14
6	10	7	6	12

H_0 - результаты тестирования по семестрам статистически значимо не различаются

Ранжируем по строкам

№				
1	5	6	14	14
Ранг	1	2	3,5	3,5
2	4	5	11	12
Ранг	1	2	3	4
3	7	8	10	12
Ранг	1	2	3	4
4	8	10	11	12
Ранг	1	2	3	4
5	5	10	14	14
Ранг	1	2	3,5	3,5
6	6	7	10	12
Ранг	1	2	3	4

Вычислим сумму рангов для каждого семестра R_i

№	1 сем	ранг	2 сем	ранг	3 сем	ранг	4 сем	ранг
1	6	2	14	3,5	5	1	14	3,5
2	11	3	5	2	4	1	12	4
3	12	4	8	2	7	1	10	3
4	8	1	10	2	11	3	12	4
5	5	1	14	3,5	10	2	14	3,5
6	10	3	7	2	6	1	12	4
Сумма рангов		14		15		9		22

Вычислим эмпирическое значение критерия χ^2 -Фридмана

$$\chi_{\text{эмп}}^2 = \left[\frac{12}{6 \cdot 4(4+1)} (14^2 + 15^2 + 9^2 + 22^2) \right] - 3 \cdot 6(4+1) = 8,6$$
$$df = 4 - 1$$

Найдем $\chi^2_{\text{крит}}$ для $df=3$ и $\alpha=0,05$. $\chi^2_{\text{крит}}=7,815$
Так как $8,6 > 7,815$ нулевая гипотеза отвергается.

Различия результатов тестирования по семестрам статистически значимы на уровне $\alpha < 0,05$.

По каким семестрам результаты различаются, проверяем по критерию Вилкоксона с поправкой Бонферрони:

T_{12} T_{13} T_{14} T_{23} T_{24} T_{34}

Сравнение более двух независимых выборок. Критерий Краскэла-Уоллиса.

Критерий Краскэла-Уоллиса (H) - это непараметрический аналог однофакторного дисперсионного анализа для независимых выборок.

Так же как критерий Манна-Уитни U показывает насколько совпадают (пересекаются) несколько рядов значений измеренного признака. Чем меньше совпадений, тем больше различаются ряды, соответствующие сравниваемым выборкам.

1. Значения выборок объединяются в один упорядоченный ряд.
2. Значения объединенного ряда ранжируются.
3. Записываются ранги отдельно для каждой выборки.
4. Вычисляются суммы рангов для каждой выборки.
5. Вычисляется эмпирическое значение критерия $H_{\text{эмп}}$ по формуле:

$$H_{\text{эмп}} = \left[\frac{12}{N(N+1)} \sum_{i=1}^k \frac{R_i^2}{n_i} \right] - 3(N+1)$$

N -суммарная численность всех выборок, k - количество сравниваемых выборок, R_i -сумма рангов для выборки i , n_i -численность выборки i .

Чем сильнее различаются выборки, тем больше критерий H и тем меньше уровень значимости.

6. Находится критическое значение критерия $H_{\text{крит}}$ ($\alpha=0,05$, $df=k-1$)

Если сравниваются 3 выборки и объем каждой выборки меньше 5, пользуются таблицами критических значений H -Краскэла-Уоллиса.

Если объем каждой выборки больше 5 и число выборок больше трех, пользуются таблицами распределения χ^2 .

7. Определяем уровень значимости.

Если $\chi^2_{\text{эмп}} \geq \chi^2_{\text{крит}}$ нулевая гипотеза отвергается.

Различия статистически значимы.

Если $\chi^2_{\text{эмп}} < \chi^2_{\text{крит}}$ нулевая гипотеза не отвергается.

Различия статистически не значимы.

Пример:

№	1 группа		2 группа		3 группа	
1	3	1	5	3	14	12
2	4	2	9	7	16	14
3	6	4	12	10	17	15
4	7	5	15	13		
5	8	6	19	16		
6	10	8				
7	11	9				
8	13	11				
		$R_1=46$		$R_2=49$		$R_3=41$

группа	Балл	Ранг
1	3	1
1	4	2
2	5	3
1	6	4
1	7	5
1	8	6
2	9	7
1	10	8
1	11	9
2	12	10
1	13	11
3	14	12
2	15	13
3	16	14
3	17	15
2	19	16

Проверяем правильность расчетов.

Общая сумма рангов должна равняться: $N(N+1)/2=16 \cdot 17/2=136$

$$R_1+R_2+R_3=46+49+41=136$$

Вычисляем H:

$$H_{\text{эмп}} = \left[\frac{12}{16(16+1)} \left(\frac{46^2}{8} + \frac{49^2}{5} + \frac{41^2}{3} \right) \right] - 3(16+1) = 6,575$$

По таблице критических значений находим χ^2 для $\alpha=0,05$ и $df=3-1=2$ $\chi^2_{\text{крит}}=5,992$

Так как $6,575 > 5,992$ нулевая гипотеза отвергается.

Различия в группах статистически значимы.

По каким группам результаты различаются, проверяем по критерию Манна-Уитни с поправкой Бонферрони:

$$U_{12} \quad U_{13} \quad U_{23}$$

Критерий Колмогорова-Смирнова

Критерий Колмогорова-Смирнова используется для сравнения эмпирического распределения с теоретическим или двух эмпирических распределений друг с другом.

При применении этого критерия сравниваются теоретическая $F(x)$ и эмпирическая $F_n(x)$ функции распределения случайной величины (накопленные частоты).

Если разность накопленных частот в двух распределениях оказывается большой, то различия между двумя распределениями являются существенными.

В качестве меры расхождения между теоретической $F(x)$ и эмпирической $F_n(x)$ функциями распределения непрерывной случайной величины X используется модуль максимальной разности

$$D_n = \max |F(x) - F_n(x)|.$$

Процедура расчетов

1. Данные в выборке ранжируются по возрастанию.
2. Вычисляются кумулятивные разности:

$$D_i = F(x_i) - F_n(x_i)$$

3. Находится абсолютное наибольшее значение кумулятивных разностей $|D_i|_{max}$

4. Вычисляется значение D критерия Колмогорова-Смирнова и сравнивается с соответствующим табличным значением.

$$D = |D_i|_{max} / n$$

Упорядочим эмпирические частоты по возрастанию:

8 8 9 10 13 15 24 25

Найдем функции распределения вероятностей (накопленные частоты):

Градации цвета	1	2	3	4	5	6	7	8
$F_{\text{теор}}$	14	28	42	56	70	84	98	112
$F_{\text{эмп}}$	8	16	25	35	48	63	87	112
$ F_{\text{теор}} - F_{\text{эмп}} $	6	12	17	21	22	21	11	

Эмпирическое значение критерия равно:

$$D_{\text{эмп}} = |D_i|_{\text{max}} / n = 22 / 112 = 0,196$$

Критическое значение критерия находим по таблице.

Если число элементов выборки больше 100, критические значения критерия Колмогорова-Смирнова вычисляются по формулам:

$$\text{для } \alpha=0,05 \quad D_{\text{кр}} = 1,36/\sqrt{n}$$

$$\text{для } \alpha=0,01 \quad D_{\text{кр}} = 1,63/\sqrt{n}$$

$$\text{Так как } D_{\text{кр}} = 1,36/\sqrt{112} = 0,128; \quad D_{\text{кр}} = 1,63/\sqrt{112} = 0,154$$

$D_{\text{эмп}} > D_{\text{кр}} \quad 0,196 > 0,154$. Нулевая гипотеза отвергается, распределение желтого цвета по 8 позициям отличается от равномерного.

Для применения критерия необходимо выполнение следующих условий:

1. Измерения должны быть проведены в шкале интервалов и отношений
2. Выборки должны быть случайными и независимыми
3. Эмпирические данные должны допускать упорядочение по возрастанию или убыванию
4. Суммарный объем двух выборок ≥ 50 . С увеличением объема выборки точность критерия повышается.

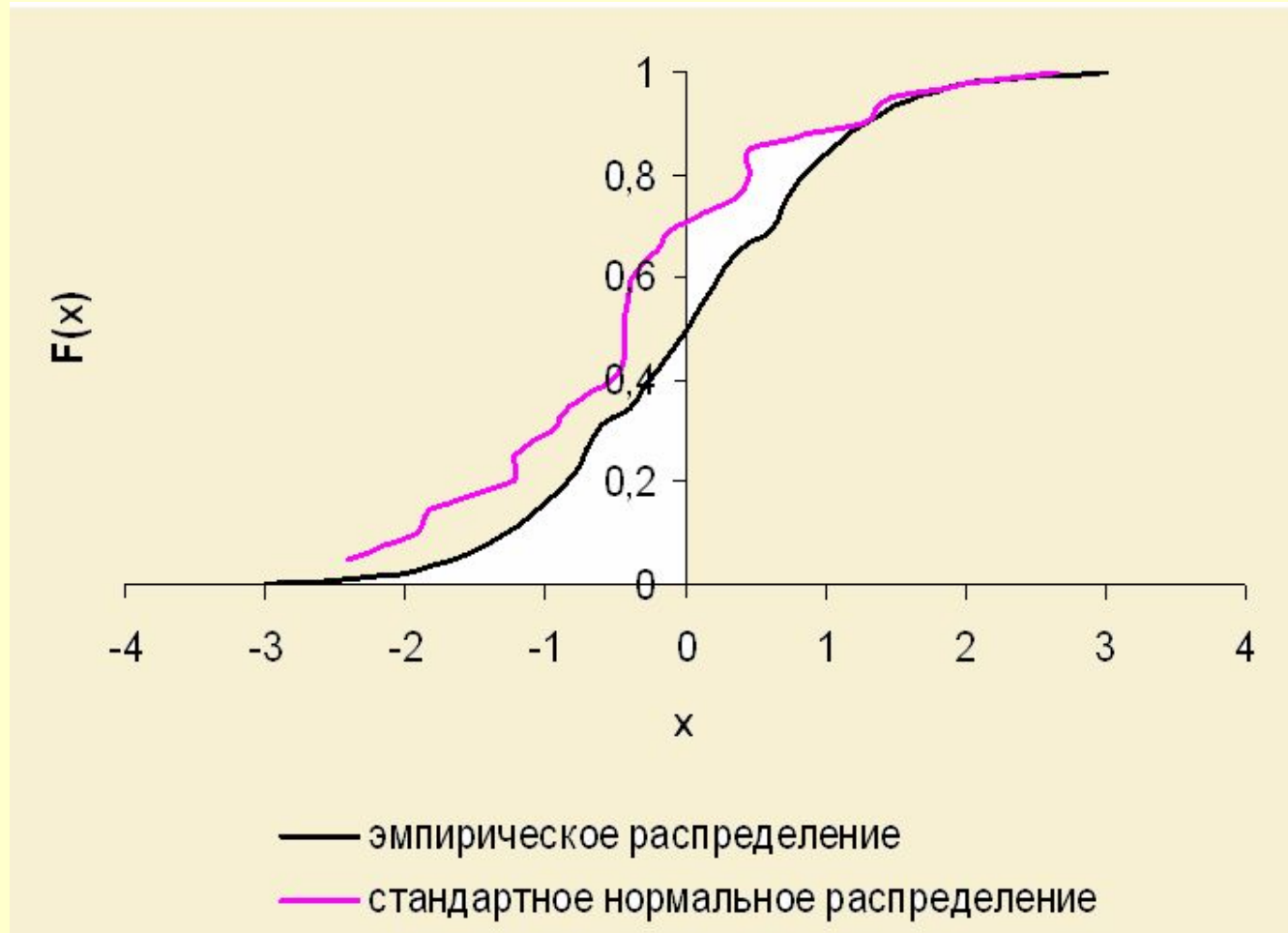
Пример 2: Нормальное распределение

0,33	-0,52	-2,41	-1,93	0,46	-0,44	-0,07
-0,38	0,48	1,29	-1,82	-1,23	-0,21	2,66
-1,22	-0,41	-0,95	1,47	-0,83	-0,43	

Среднее арифметическое = -0,308; дисперсия = 1,47, стандартное отклонение = 1,28.

Нулевая гипотеза: рассматриваемое распределение $F(x)$ является нормальным с нулевым средним и единичной дисперсией.

Функции распределения



Процедура расчетов

1. Данные в выборке ранжируются по возрастанию.

2. Вычисляются кумулятивные разности:

$$D_i = R_i - n\Phi(x_i)$$

3. Находится абсолютное наибольшее значение кумулятивных разностей $|D_i|_{max}$

4. Вычисляется значение D критерия Колмогорова-Смирнова и сравнивается с соответствующим табличным значением.

$$D = |D_i|_{max} / n$$

Наблюде-ние	Ранг (R _i)	Ожидаемо е n · Ф(x _i)	Раз-ность	Наблюде-ние	ранг	Ожида-емое	разность
-2,41	1	0,16	0,84	-0,41	11	6,82	4,18
-1,93	2	0,54	1,46	-0,38	12	7,04	4,96
-1,82	3	0,69	2,31	-0,21	13	8,34	4,66
-1,23	4	2,19	1,81	-0,07	14	9,44	4,56
-1,22	5	2,22	2,78	0,33	15	12,59	2,41
-0,95	6	3,42	2,58	0,46	16	13,54	2,46
-0,83	7	4,07	3,93	0,48	17	13,69	3,31
-0,52	8	6,03	2,97	1,29	18	18,03	-0,03
-0,44	9	6,6	3,4	1,47	19	18,58	0,42
-0,43	10	6,67	3,33	2,66	20	19,92	0,08

$D = 4,96/20 = 0,248 < D_{\text{крит}} = 0,304$ ($\alpha = 0,05$); нулевая гипотеза не отклоняется. Данные подчиняются нормальному закону распределения.


Заключение

Таким образом, нами рассмотрены основы непараметрического дисперсионного анализа, применение критерия Колмогорова-Смирнова

РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА:

Основная литература:

- Попов А.М. Теория вероятней и математическая статистика /А.М. Попов, В.Н. Сотников. – М.: ЮРАЙТ, 2011. – 440 с.
- Герасимов А. Н. Медицинская статистика: учебное пособие / А. Н. Герасимов. – М. : Мед. информ. агентство, 2007. – 480 с.
- Балдин К. В. Основы теории вероятностей и математической статистики : учебник / К. В. Балдин. – М. : Флинта, 2010. – 488с.



БЛАГОДАРЮ ЗА ВНИМАНИЕ