

# Беседы о прикладной статистике

*Семинар 10. Дисперсионный анализ для  
сравнения средних. Тест Крускала-Уоллиса*

**Фастовец И.**

**А.**

# Сравнение двух средних

- На предыдущих семинарах мы обсуждали сравнение двух средних значений
- В случае нормального распределения применяют, например, t-тест
- Если распределение не описывается нормальной кривой, для сравнения двух распределений используют, например, тест суммы рангов Уилкоксона (Манна-Уитни)

# Сравнение нескольких средних

- Если сравниваемых групп 3 и более, можно попарно сравнить группы друг с другом, например, при помощи t-теста. В таком случае количество сравнений  $\frac{N(N-1)}{2}$ , где N - количество групп, которые нужно сравнить между собой
- Недостаток такого подхода в том, что теряется статистическая информация из других групп. Это приводит к падению *статистической мощности теста* (1-ошибка второго рода)
- Одним из способов решения проблемы является однофакторный дисперсионный анализ (one-way ANOVA)

# Однофакторный дисперсионный анализ

- $H_0$ : средние всех групп равны
- $H_a$ : хотя бы два средних различаются между собой
- Дисперсии сравниваемых генеральных совокупностей равны
- Задача сводится к построению линейной модели вида  $x_{ij} = \mu_i + \epsilon_{ij}$ , где  $i$  – количество групп,  $j$  – количество наблюдений в группе.
- Параметры модели – средние значения сравниваемых генеральных совокупностей  $\mu_i$  и общее стандартное отклонение  $\sigma$
- Оценка  $\mu_i$  производится при помощи средних выборок по группам:

$$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$$

# Объединенная оценка дисперсии

- Остатки  $e_{ij} = x_{ij} - \bar{x}_i$  отражают разброс данных вокруг средних значений по группам
- Модель ANOVA предполагает, что распределение признака во всех группах нормальное и имеет одинаковую дисперсию
- Объединенная (усредненная) оценка дисперсии по  $I$  группам будет иметь вид:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \dots + (n_I - 1)s_I^2}{(n_1 - 1) + (n_2 - 1) + \dots + (n_I - 1)}$$

- Тогда несмещенная оценка  $s_p = \sqrt{s_p^2}$
- Группы с большим количеством наблюдений будут иметь больший вес

# Регрессия и ANOVA: одно и то же

- Из модели множественной регрессии мы помним, что:

$$SST = \sum (y_i - \bar{y})^2$$

$$SST = SSM + SSE \quad DFT = DFM + DFE$$

$$SSM = \sum (\hat{y}_i - \bar{y})^2$$

$$SSE = \sum (y_i - \hat{y}_i)^2$$

$$MS = \frac{\text{sum of squares}}{\text{degrees of freedom}}$$

- Модель ANOVA аналогична регрессионной модели, где роль линии регрессии выполняют средние по группам
- Поэтому SSM записывают как SSG, что означает сумма квадратов отклонений каждого среднего от генерального среднего
- Аналогично регрессии: SSE – сумма квадратов отклонений значений от внутригрупповых средних, SST – сумма квадратов отклонений каждого значения от генерального среднего

# F-тест для дисперсионного анализа

• Несложно догадаться, что  $s_p^2 = \text{MSE} = \frac{\text{SSE}}{\text{DFE}}$        $s_p = \sqrt{\text{MSE}}$

• Степени свободы для всех отклонений и F-тест :

$$\text{DFT} = N - 1 \qquad R^2 = \frac{\text{SSG}}{\text{SST}} \quad (\text{Аналогично регрессии})$$

$$\text{DFG} = I - 1$$

$$\text{DFE} = N - I$$

$$F = \frac{\text{MSG}}{\text{MSE}}$$

Подчиняется распределению F(I-1, N-I)

Source	Degrees of freedom	Sum of squares	Mean square	<i>F</i>
Groups	$I - 1$	$\sum_{\text{groups}} n_i (\bar{x}_i - \bar{x})^2$	SSG/DFG	MSG/MSE
Error	$N - I$	$\sum_{\text{groups}} (n_i - 1) s_i^2$	SSE/DFE	
Total	$N - 1$	$\sum_{\text{obs}} (x_{ij} - \bar{x})^2$		

# Пример

- Имеем 3 переменных, в каждой 3 наблюдения:

	A	B	C
	3	5	7
	1	3	6
	2	4	5
	2	4	6
	1	1	1

- $\bar{X} = \frac{3+1+2+5+3+4+7+6+5}{9} = \frac{36}{9} = 4$

- $SST = (3-4)^2 + (1-4)^2 + (2-4)^2 + (5-4)^2 + (3-4)^2 + (4-4)^2 + (7-4)^2 + (6-4)^2 + (5-4)^2 = 30$

- $DFT = N - 1 = 9 - 1 = 8$

- $SSE = SSW = (3-2)^2 + (1-2)^2 + (2-2)^2 + (5-4)^2 + (3-4)^2 + (4-4)^2 + (7-6)^2 + (6-6)^2 + (5-6)^2 = 6$

- $SSG = SSB = 3 \times (2-4)^2 + 3 \times (4-4)^2 + 3 \times (6-4)^2 = 12 + 0 + 12 = 24$
- $DFG = DFB = I - 1 = 2$
- $MSG = \frac{SSG}{DFG} = \frac{24}{2} = 12$

- $DFE = N - I = 9 - 3 = 6$

- $MSE = \frac{SSE}{DFE} = \frac{6}{6} = 1$

- $MSE = \frac{(3-1) \times 1 + (3-1) \times 1 + (3-1) \times 1}{(3-1) + (3-1) + (3-1)} = \frac{2+2+2}{2+2+2} = \frac{6}{6} = 1$

$$F(I-1, N-I) = \frac{MSG}{MSE} = \frac{12}{1} = 12$$

$$P=0.008$$

$$! SST = SSG + SSE$$



# Индивидуальные сравнения. Контрасты

- Контраст – это комбинация средних генеральной совокупности вида  $\psi = \sum a_i \mu_i$ , ему соответствует выборочный контраст  $c = \sum a_i \bar{x}_i$
- При этом сумма коэффициентов  $a$  равна 0:  $\sum a_i = 0$
- В ANOVA контраст – это линейная комбинация независимых нормально распределенных величин, таким образом, он имеет нормальное распределение
- Стандартная ошибка выборочного контраста:  $SE_c = s_p \sqrt{\sum \frac{a_i^2}{n_i}}$
- Тест  $H_0: \psi = 0$   $t = \frac{c}{SE_c}$  и доверительный интервал  $c \pm t^* SE_c$   
– уже знакомые нам из предыдущих семинаров, где используем распределение  $t(DFE)$

# Пример расчета контрастов

- Посчитаем значимость различия средних

А-С в нашем примере. Контраст:  $c_1 = 1 \times 2 + (-1) \times 6 = -4$

- $SE_{c_1} = 1 \times \sqrt{\frac{1^2}{3} + \frac{-1^2}{3}} = \sqrt{\frac{2}{3}} = 0.816$  ;  $t(6) = \frac{-4}{0.816} = -4.902$  ;  $p = 0.002705$

- Теперь посчитаем, отличается ли среднее С от среднего средних А-В:

- Контраст:  $c_2 = 1 \times (2 + 4) + (-2) \times 6 = -6$

- $SE_{c_2} = 1 \times \sqrt{\frac{1^2}{3} + \frac{1^2}{3} + \frac{-2^2}{3}} = \sqrt{\frac{6}{3}} = \sqrt{2} = 1.414$  ;  $t(6) = \frac{-6}{1.414} = -4.243$  ;  $p = 0.00542$

- Контрасты можно использовать, даже если общий F-тест не значимый, т.к. в некоторых случаях контрасты мощнее

- Нельзя определять индивидуальные сравнения, глядя на данные! Такие сравнения планируются изначально (устранение ошибки III рода)

	A	B	C
	3	5	7
	1	3	6
	2	4	5
	2	4	6
	1	1	1

$$s_p = 1$$
$$DFE = 6$$

# Множественные сравнения

- Используются только после отвержения  $H_0$  при помощи F-теста!
- Тесты множественных сравнений представляют из себя парные t-тесты с использованием объединенной оценки дисперсии  $s_p$  из ANOVA :

$$t_{ij} = \frac{\bar{x}_i - \bar{x}_j}{s_p \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}} \quad |t_{ij}| \geq t^{**}$$

- Метод подбора  $t^{**}$  зависит от используемой процедуры сравнения
- Тест НСР (Fisher's LSD) не использует поправку на множественные сравнения, и поэтому не является корректным
- Простейшее решение – использовать поправку Бонферрони
- Огромное количество поправок на любой вкус!

# Что делать, если допущения нарушаются

- Если распределения остаются предположительно нормально распределенными, но дисперсия в группах гетерогенна
- Если наибольшее и наименьшее стандартные отклонения различаются менее чем в 2 раза, то можно ничего не делать
- Если различия дисперсий резкие, рекомендуется использовать F-тест Уэлча для разных дисперсий
- Далее для множественных сравнений можно применить тест Геймса-Хоуэлла (Games-Howell test)
- Эти методы менее мощные, чем классические, однако применимы даже при очень малых выборках

# Ранговый ANOVA

- Если резко нарушаются допущения, можно обратиться к непараметрическим методам оценки
- Самый неприятный случай – когда возможны резкие выбросы, которые нельзя объяснить и убрать
- Простые и примитивные непараметрические тесты – ранговые
- На предыдущих семинарах мы рассматривали ранговые корреляции Спирмена и тесты попарных сравнений Уилкоксона
- Дисперсионный анализ также можно произвести ранговыми методами. В этом случае мы тестируем общую нулевую гипотезу не F-тестом, а тестом Крускала-Уоллиса (Kruskal-Wallis test)

# Тест Крускала-Уоллиса

- Проранжируем все наблюдения (общее ранжирование), рассчитаем суммы рангов  $R_i$  в  $i$  группах объемом  $n_i$  и общим количеством наблюдений  $N$ :

- $H$  статистика Крускала-Уоллиса имеет вид:

$$H = \frac{12}{N(N+1)} \sum \frac{R_i^2}{n_i} - 3(N+1)$$

- Когда объемы выборок большие и во всех группах примерно одинаковое распределение,  $H$  – статистика распределяется в соответствии с  $\chi^2(I-1)$
- В большинстве случаев (но не всегда) асимптотический  $H$  тест дает надежные результаты и при малых выборках

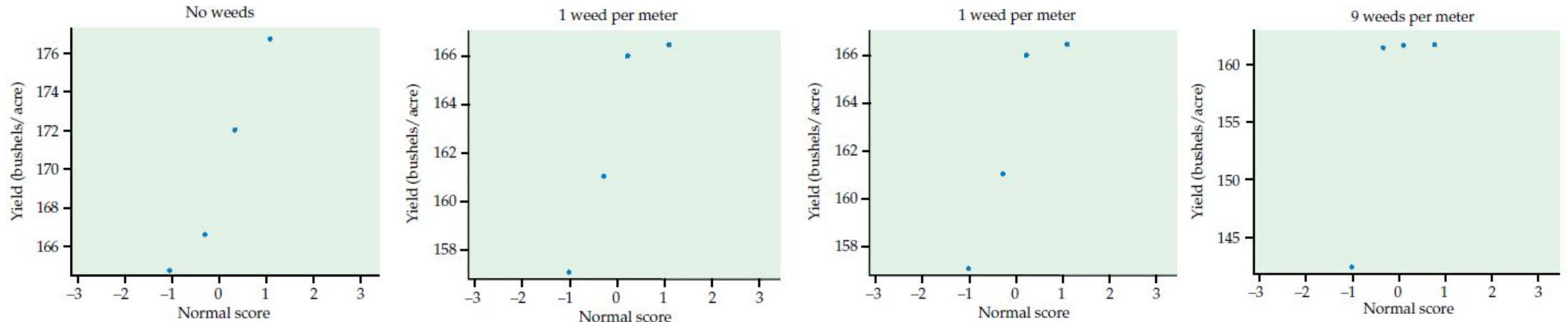
# Тест Крускала-Уоллиса

- Рассмотрим урожаи культуры при разном количестве сорняков:

Weeds per meter	Corn yield	Weeds per meter	Corn yield	Weeds per meter	Corn yield	Weeds per meter	Corn yield
0	166.7	1	166.2	3	158.6	9	162.8
0	172.2	1	157.3	3	176.4	9	142.4
0	165.0	1	166.7	3	153.1	9	162.7
0	176.9	1	161.1	3	156.0	9	162.4

Weeds	<i>n</i>	Mean	Std. dev.
0	4	170.200	5.422
1	4	162.825	4.469
3	4	161.025	10.493
9	4	157.575	10.118

- Графики нормальных квантилей по группам:



# Тест Крускала-Уоллиса

- Ранги наблюдений и суммы рангов по группам

Yield	142.4	153.1	156.0	157.3	158.6	161.1	162.4	162.7
Rank	1	2	3	4	5	6	7	8
Yield	162.8	165.0	166.2	166.7	166.7	172.2	176.4	176.9
Rank	9	10	11	12.5	12.5	14	15	16

Weeds		Ranks			Rank sums
0	10	12.5	14	16	52.5
1	4	6	11	12.5	33.5
3	2	3	5	15	25.0
9	1	7	8	9	25.0

- Статистика Крускала-Уоллиса

$$\begin{aligned}
 H &= \frac{12}{N(N+1)} \sum \frac{R_i^2}{n_i} - 3(N+1) \\
 &= \frac{12}{(16)(17)} \left( \frac{52.5^2}{4} + \frac{33.5^2}{4} + \frac{25^2}{4} + \frac{25^2}{4} \right) - (3)(17) \\
 &= \frac{12}{272} (1282.125) - 51 \\
 &= 5.56
 \end{aligned}$$

$$\begin{aligned}
 P &= \\
 &0.1344
 \end{aligned}$$



# Многофакторный дисперсионный анализ

- Как и регрессия, дисперсионный анализ может быть многофакторным
- Кроме того, существуют различные модификации регрессии и дисперсионного анализа, входящие в класс общих линейных моделей (GLM)
- Многофакторный анализ мощнее, чем однофакторный по каждому фактору
- Особый интерес представляет возможность нахождения и тестирование значимости взаимодействия между факторами