

Прикладная статистика. Занятие 5. Дисперсионный анализ.

27 марта 2012 г.

# Модели дисперсионного анализа

## Модель с фиксированным эффектом (fixed-effects model ANOVA):

- разбиение на группы определено до получения данных;
- при повторе эксперимента ожидается, что соотношения между средними групп сохранятся;
- если между средними есть различия, на следующем этапе анализируется, какие именно группы различаются.

## Примеры.

- Продолжительность жизни разноногих раков в морской воде и растворах глюкозы и маннозы.
- Экспрессия определённого гена в тканях мозга, печени, лёгких и мышц; необходимо понять, в какой ткани экспрессия выше.
- Вкусовые качества персиков с 10 различных деревьев; планируется выбрать лучшее дерево для дальнейшего разведения.

## Модели дисперсионного анализа

### Модель со случайным эффектом (random-effects model ANOVA):

- характеристика, определяющая разбиение на группы, не представляет непосредственного интереса;
- группы случайно выбраны из множества возможных групп;
- если между группами есть неоднородность, ожидается, что она сохранится при повторе эксперимента, но соотношения между средними могут измениться.

### Примеры.

- Размеры горбатов в разных семьях, выращенных на одном и том же растении; цель — определить значимость фактора семьи для дальнейших исследований.
- Уровень гликогена в различных образцах икроножной мышцы крысы; если вариация между образцами даёт маленький вклад в общую вариацию, то можно считать, что для измерения уровня достаточно одного образца.
- Вкусовые качества персиков с 10 различных деревьев; планируется сравнить различия во вкусовых качествах персиков с разных деревьев с различиями у персиков с одного дерева. Если последние больше, то бессмысленно выбрать для размножения дерево с лучшей средней оценкой.

## Модели дисперсионного анализа

Если используется **модель с фиксированным эффектом**, то, в случае отвержения гипотезы однородности средних, проводится дополнительное сравнение с целью уточнения характера различий.

Сравнение может быть:

- запланированным, когда группы для дальнейшего сравнения отобраны до сбора данных.
- незапланированным, когда группы для сравнения выбираются по результатам первичного анализа данных.

Если при запланированном сравнении сравниваемые группы ортогональны, то модифицировать получаемые p-value на множественность не нужно.

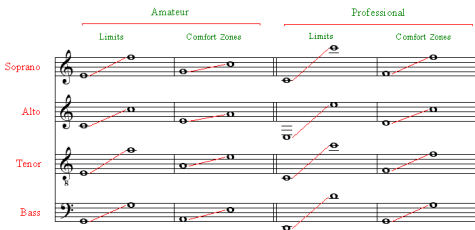
Для незапланированного сравнения можно использовать LSD Фишера, критерий Стьюдента-Ньюмена-Кейлса (Кобзарь, 407), интервалы сравнения Габриэля (Gabriel, A simple method of multiple comparison of means) и т. д.

Если используется **модель со случайным эффектом**, следующий шаг — разделение дисперсий на внутригрупповые и межгрупповые (см. <http://udel.edu/mcdonald/statanovapart.html>).

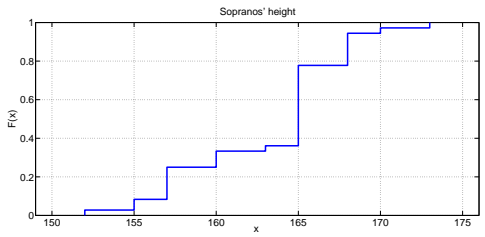
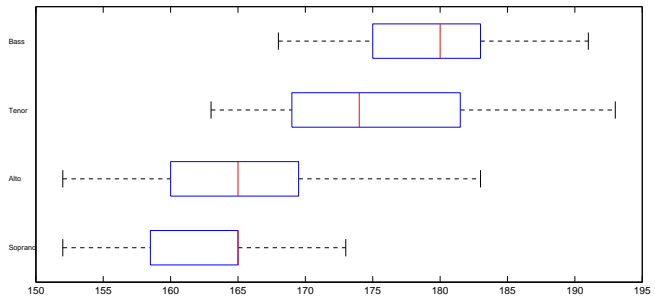
# Рост певцов хора

В 1979 году 130 участников Нью-Йоркской ассоциации хорового пения сообщили данные своего роста; для каждого известен также регистр голоса. Есть ли связь между ростом и регистром?

## Vocal Ranges



# Рост певцов хора



## Рост певцов хора

$H_0$ : рост и регистр голоса не связаны.

$H_1$ : для каких-то видов регистра голоса средний рост отличается.

Source	SS	df	MS	F	Prob>F
Groups	6901.4	3	2300.47	55.73	5.34718e-023
Error	5201.1	126	41.28		
Total	12102.5	129			

SS — сумма квадратов отклонений, df — число степеней свободы, MS — дисперсия, F — статистика критерия;

строка Groups — оценки по выборочным средним, строка Error — оценки по выборочным дисперсиям.

Предположения ANOVA:

- 1 значения признака во всех группах нормально распределены;
- 2 дисперсия значений признака во всех группах одинакова;
- 3 все наблюдения независимы.

Метод устойчив к нарушению первых двух предположений.

## Рост певцов хора

Критерий Стьюдента для проверки гипотезы равенства роста певцов с альтом и сопрано:  $p = 0.2460$  — против двусторонней альтернативы,  $p = 0.1230$  — против односторонней альтернативы.

Критерий Стьюдента для проверки гипотезы равенства роста певцов с тенором и басом:  $p = 0.0597$  — против двусторонней альтернативы,  $p = 0.0298$  — против односторонней альтернативы.

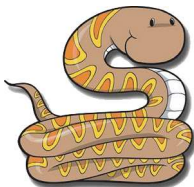
Критерий Джонкхиера для проверки наличия тренда (увеличение роста с понижением регистра голоса):  $p < 0.00001$ .



## Скорость обживания клеток у гремучих змей

Place, Abramson, Habituation of the rattle response in western diamondback rattlesnakes, *Crotalus atrox* (2008): испытания проводились в течение четырёх дней. Каждый день гремучая змея помещалась в клетку, крышка которой автоматически открывалась и закрывалась каждые 5 минут. Первое время при этом змея начинала греметь, но со временем обживалась и переставала реагировать. Для 6 змей известен номер открытия крышки, при котором впервые змея не начинала греметь.

Требуется проверить, отличается ли скорость обживания клетки для различных змей и для различных дней проведения испытаний.

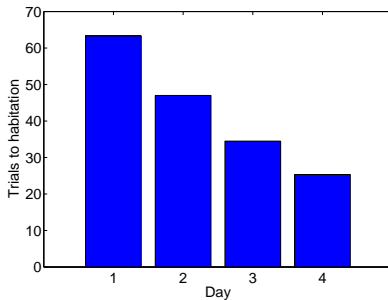


## Скорость обживания клеток у гремучих змей

$H_{01}$ : скорость обживания одинакова во все дни проведения испытания.

$H_{02}$ : скорость обживания одинакова для всех змей.

Source	SS	df	MS	F	Prob>F
Days	4877.79	3	1625.93	3.32	0.0487
Snakes	3042.21	5	608.44	1.24	0.3382
Error	7345.96	15	489.73		
Total	15265.96	23			



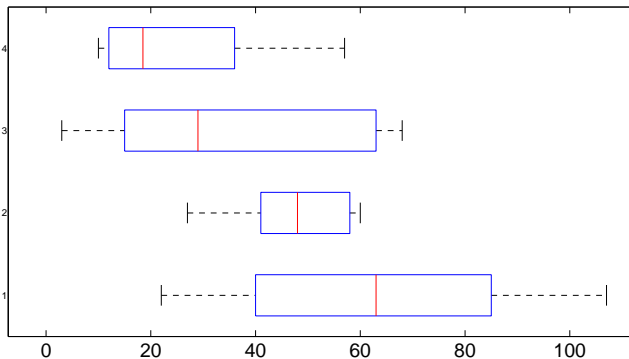
## Скорость обживания клеток у гремучих змей

Критерий Фридмана убирает влияние одного из факторов, оценивает значимость оставшегося:

- если нейтрализовать влияние особи, получаем  $p = 0.0384$ ;
- если нейтрализовать влияние дня, получаем  $p = 0.1643$ .

# Скорость обживания клеток у гремучих змей

Однофакторный дисперсионный анализ с учётом только дня:  $p = 0.0485$ .



Критерий Джонкхиера для проверки наличия тренда (ускорение обживания):  $p = 0.0037$ .

## Марихуана и скорость реакции

Изучалось воздействие марихуаны на скорость реакции. В качестве испытуемых были выбраны по 12 человек из каждой категории:

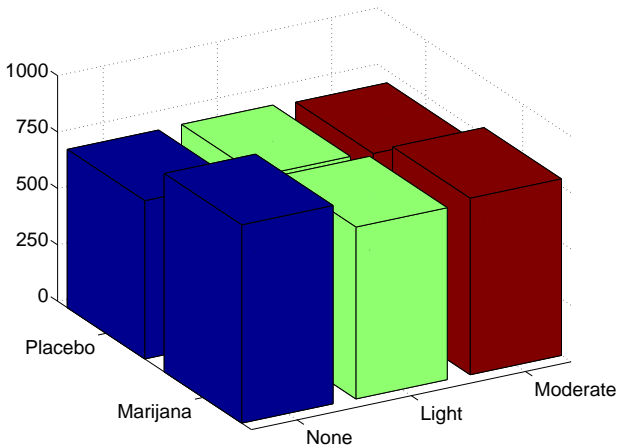
- никогда не пробовали марихуану;
- иногда употребляют марихуану;
- регулярно употребляют марихуану.

Испытуемые были разделены на две равные группы; половине из них дали выкурить две сигареты с марихуаной, вторая половина выкурила две обычные сигареты с запахом и вкусом марихуаны. Сразу после этого все испытуемые прошли тест на скорость реакции.

Требуется оценить влияние марихуаны на скорость реакции, учитывая фактор предыдущего опыта употребления.

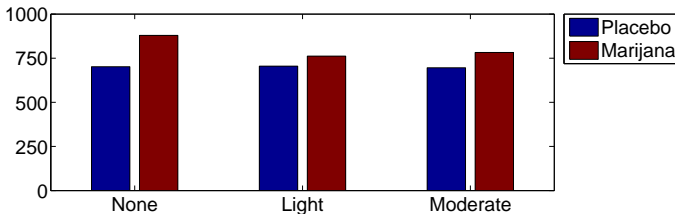
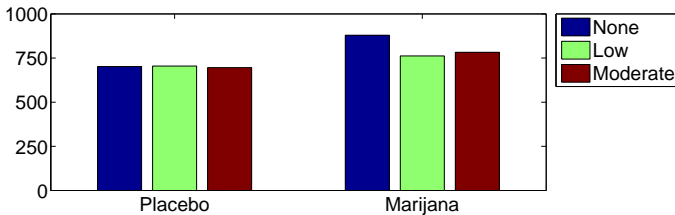
# Мариhuана и скорость реакции

Плохой график:



# Марихуана и скорость реакции

Хорошие графики:



# Марихуана и скорость реакции

$H_{01}$ : средняя скорость реакции одинакова при употреблении и марихуаны, и сигарет.

$H_{02}$ : средняя скорость реакции не зависит от предыдущего опыта употребления марихуаны.

$H_{03}$ : отсутствует межфакторное взаимодействие между употребляемым веществом и предыдущим опытом употребления марихуаны.

Source	SS	df	MS	F	Prob>F
Group	103041	1	103041	17.58	0.0002
Past use	23634.5	2	11817.2	2.02	0.1508
Interaction	23642.2	2	11821.1	2.02	0.1507
Error	175796.3	30	5859.9		
Total	326114	35			

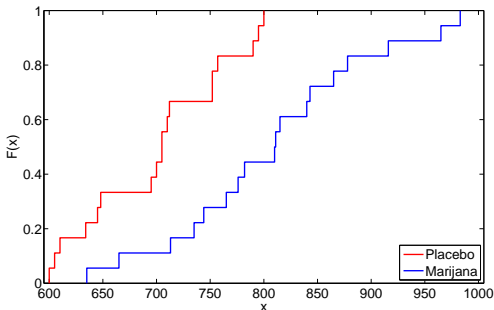


# Марихуана и скорость реакции

Вывод: гипотеза о том, что предыдущий опыт употребления влияет на скорость реакции, не принимается  $\Rightarrow$  данные по группам можно объединить.

Для объединённых данных:

- однофакторный дисперсионный анализ:  $p = 0.00036$ ;
- критерий Уилкоксона, двусторонняя альтернатива:  $p = 0.000596$ ;
- критерий Стьюдента, односторонняя альтернатива:  
 $p = 0.00018$ ,  $ci = (61.3, \infty)$ ;



## Иерархический дизайн

Стандартная постановка двухфакторного дисперсионного анализа предполагает, что уровни факторов в выборке распределены независимо.

Пример, когда это не так: признак — уровень гликогена в икроножной мышце крысы, фактор 1 — уровень стресса крыс, фактор 2 — различия между клетками. Крысы со стрессом живут в клетках 1 и 2, без стресса — 3 и 4.

Решение — иерархический дисперсионный анализ (nested ANOVA).

## СБИ чернотелой дрозофилы

Codon bias index (CBI) — мера случайности использования синонимичных кодонов в геноме — была определена для нескольких регионов двух хромосом чернотелой дрозофилы. Требуется определить, есть ли систематические различия по величине СБИ между разными хромосомами и регионами.



# СВІ чернобрюхой дрозодилы

Source	SS	df	MS	F	Prob>F
Chromosome	0.00496	2	0.00248	0.32	0.7319
Region(Chromosome)	0.16295	3	0.05432	6.92	0.0011
Error	0.23564	30	0.00785		
Total	0.40891	35			

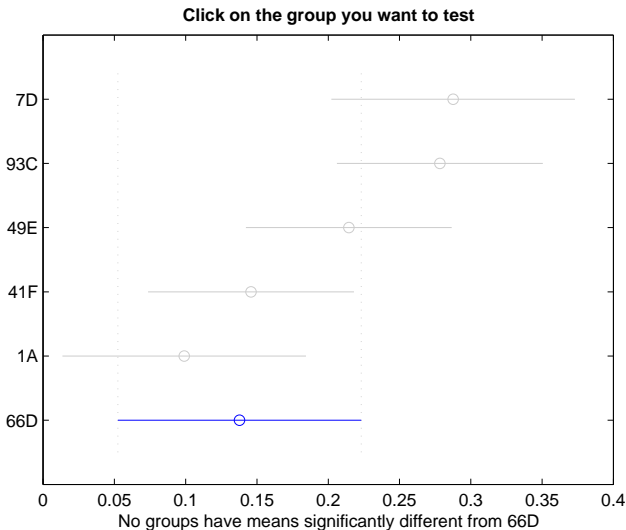
Есть различия между регионами, нет различий между хромосомами.

## СВІ чернобрюхой дрозодилы

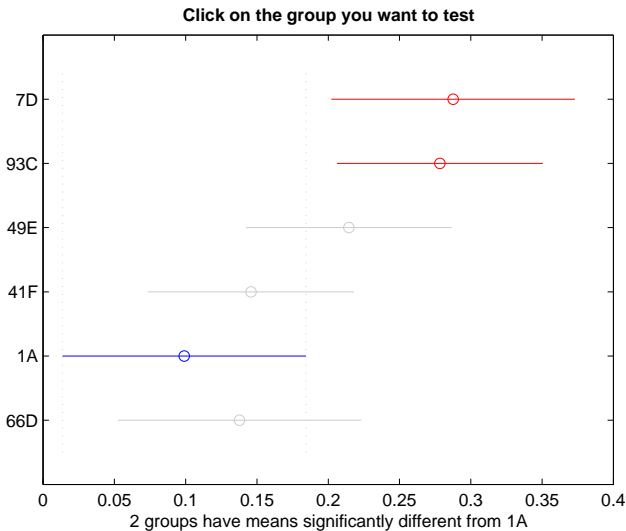
Для уточнения различий применим метод HSD (Tukey's honestly significant difference):

Группа 1	Группа 2	$CI_L$	mean	$CI_U$
7D	93C	-0.1485	0.0093	0.1672
7D	49E	-0.0847	0.0732	0.2310
7D	41F	-0.0161	0.1417	0.2996
7D	1A	0.0181	0.1886	0.3591
7D	66D	-0.0207	0.1498	0.3203
93C	49E	-0.0802	0.0639	0.2079
93C	41F	-0.0117	0.1324	0.2765
93C	1A	0.0214	0.1793	0.3371
93C	66D	-0.0174	0.1405	0.2983
49E	41F	-0.0755	0.0686	0.2127
49E	1A	-0.0424	0.1154	0.2733
49E	66D	-0.0812	0.0766	0.2345
41F	1A	-0.1110	0.0469	0.2047
41F	66D	-0.1498	0.0081	0.1659
1A	66D	-0.2093	-0.0388	0.1317

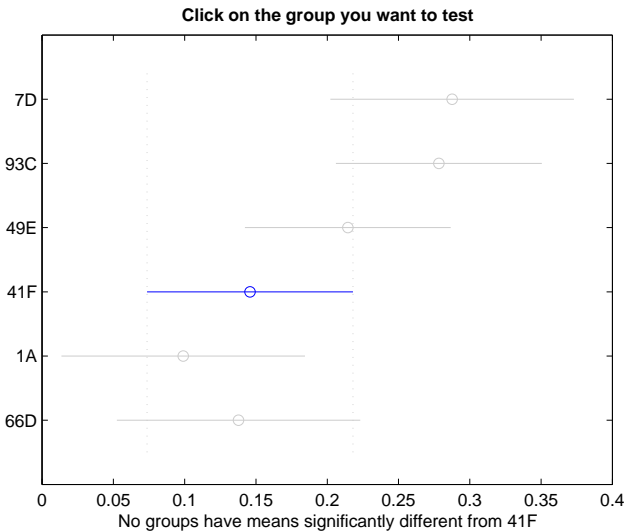
# СВІ чернoбрюхой дрозофилы



# СВІ чернобрюхой дрозозфилы

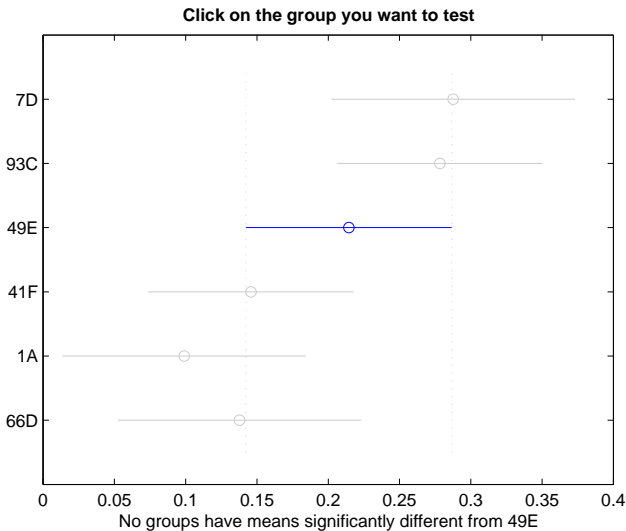


# СВІ чернoбрюхой дрозофилы

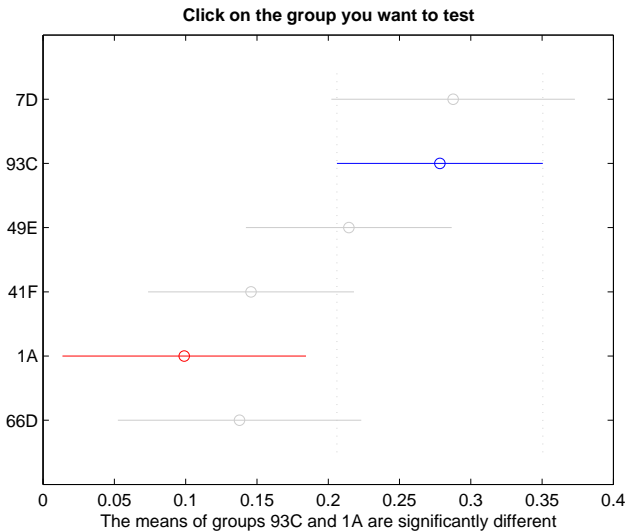




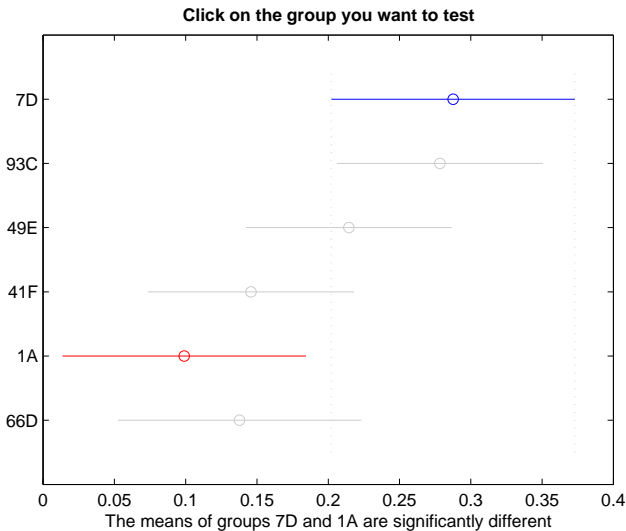
# СВІ чернoбрюхой дрозофилы



# СВІ чернoбрюхой дрозофилы



# СВІ чернoбрюхой дрозофилы



## Лечение гипертонии

72 пациента проходили лечение от гипертонии. Для лечения использовались три вида лекарств, при этом их эффект изучался как при использовании специальной диеты, так и в её отсутствии; кроме того, в ряде случаев применялась психотерапия. Данные — артериальное давление пациента по окончании лечения. Требуется сравнить эффективность методов для лечения гипертонии.

Дизайн  $[3 \times 2 \times 2]$ .

## Лечение гипертонии

Трёхфакторный дисперсионный анализ, все взаимодействия:

Source	SS	df	MS	F	Prob>F
Therapy	2048	1	2048	13.07	0.0006
Diet	5202	1	5202	33.2	0
Drug	3675	2	1837.5	11.73	0.0001
Therapy*Diet	32	1	32	0.2	0.6529
Therapy*Drug	259	2	129.5	0.83	0.4425
Diet*Drug	903	2	451.5	2.88	0.0638
Therapy*Diet*Drug	1075	2	537.5	3.43	0.0388
Error	9400	60	156.67		
Total	22594	71			

## Лечение гипертонии

Значимость многофакторных взаимодействий:

- Diet\*Drug: воздействие диеты различно при различных применяемых препаратах (или наоборот, действие препаратов зависит от диеты);
- Therapy\*Diet\*Drug: воздействие одного из факторов различно при различных комбинациях двух других. Хотя эффект Therapy\*Drug незначим в целом, значимость Therapy\*Diet\*Drug говорит о том, что влияние Therapy\*Drug необходимо оценивать отдельно для пациентов, использующих и не использующих диету.

Прикладная статистика  
Семинар 5. Дисперсионный анализ.

Рябенко Евгений  
riabenko.e@gmail.com