

Прикладной статистический анализ данных.  
6. Дисперсионный анализ.

Ольга Кравцова  
Юлиан Сердюк  
cs.msu.psad@gmail.com

26.03.2024

## Разновидности дисперсионного анализа (ANOVA)

- по числу факторов: однофакторный (one-way), двухфакторный (two-way) и т. д
- по типу выборок: независимые (between-subjects), связанные (within-subjects, repeated measurements)
- по типу альтернативы: общая, тренда
- по типу эффектов: случайные (random-effects), фиксированные (fixed-effects)
- по типу уровней факторов: независимые, вложенные (nested), с болтающимся контролем (dangling control group), латинский квадрат (latin square)
- по используемым предположениям: нормальный, непараметрический
- по объёму выборок: одинаковый (balanced), различный (unbalanced)

# Однофакторный дисперсионный анализ

Пусть имеется  $K$  выборок:

$$X^N = X_1^{n_1} \cup X_2^{n_2} \cup \dots \cup X_K^{n_K}, \quad N = \sum_{i=1}^K n_i.$$

Эквивалентная запись в виде псевдотаблицы:

фактор  $f: X \rightarrow \{1, \dots, K\}$

$f$	1	...	$k$	...	$K$
$X^N$	$X_{11}$		$X_{k1}$		$X_{K1}$
	$\vdots$	...	$\vdots$	...	$\vdots$
	$X_{1n_1}$		$X_{kn_k}$		$X_{Kn_K}$

**Задача:** проверить гипотезу об отсутствии влияния фактора  $f$  на среднее значение признака  $X$ , то есть, о равенстве средних значений  $K$  выборок.

# Однофакторный дисперсионный анализ

**Идея:** рассмотрим две компоненты разброса значений  $X_{ki}$  относительно глобального среднего  $\bar{X}$ :

$$X_{ki} - \bar{X} = (X_{ki} - \bar{X}_k) + (\bar{X}_k - \bar{X}),$$

где  $\bar{X}_k$  — среднее в  $k$ -й выборке.

Возведём в квадрат и просуммируем:

$$\sum_{k=1}^K \sum_{i=1}^{n_k} (X_{ki} - \bar{X})^2 = \sum_{k=1}^K \sum_{i=1}^{n_k} (X_{ki} - \bar{X}_k)^2 + \sum_{k=1}^K n_k (\bar{X}_k - \bar{X})^2,$$

$$SS_{total} = SS_{wg} + SS_{bg}.$$

Если средние в группах значительно отличаются, преобладает вторая компонента, если же они одинаковы — первая.

# Однофакторный дисперсионный анализ

Линейная модель:

$$X_{ki} = \mu + \alpha_k + \varepsilon_{ki},$$

$$i = 1, \dots, n_k, \quad k = 1, \dots, K.$$

$\mu$  — глобальное среднее значение признака  $X$ ,

$\alpha_k$  — отклонение от  $\mu$ , вызванное влиянием  $k$ -го уровня фактора  $f$ ,

$\varepsilon_{ki}$  — случайные независимые одинаково распределённые ошибки.

Средние значения  $X$  во всех  $K$  выборках одинаковы  $\Leftrightarrow \alpha_1 = \dots = \alpha_K$ .

# Критерий Фишера

выборки:  $X^N = X_1^{n_1} \cup \dots \cup X_K^{n_K}$

нулевая гипотеза:  $H_0: \alpha_1 = \dots = \alpha_K$

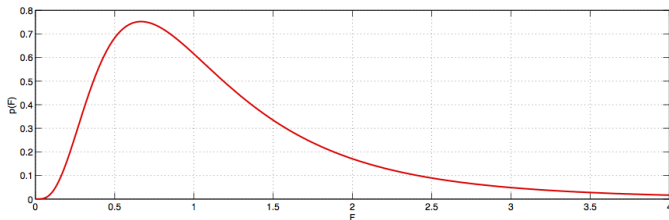
альтернатива:  $H_1: H_0$  неверна

статистика:  $F(X^N) = \frac{SS_{bg}/(K-1)}{SS_{wg}/(N-K)}$

$$SS_{bg} = \sum_{k=1}^K n_k (\bar{X}_k - \bar{X})^2$$

$$SS_{wg} = \sum_{k=1}^K \sum_{i=1}^{n_k} (X_{ki} - \bar{X}_k)^2$$

нулевое распределение:  $F(K-1, N-K)$



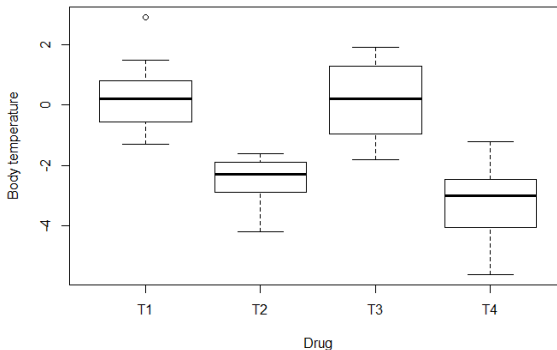
# Критерий Фишера

Предположения метода:

- ❶ выборочные распределения средних значений признака во всех группах нормальны;
  - ❷ дисперсия значений признака во всех группах одинакова;
  - ❸ наблюдения независимы.
- Первое предположение считается выполненным, если распределение признака во всех группах нормально, или если объёмы выборок примерно одинаковы и  $N - K - 1 \geq 20$ .
  - Второе предположение считается выполненным, если отношение наибольшей выборочной дисперсии к наименьшей не превосходит 10.
  - При  $n_1 = \dots = n_K$  метод устойчив к нарушению первых двух предположений.
  - Если объёмы выборок различаются, нарушение предположения о равенстве дисперсий может привести к росту вероятности ошибки первого рода.
  - Выбросы могут оказывать существенное влияние на результат.

# Критерий Фишера

**Пример** (Bonpini, табл. 3.2): исследуется эффективность четырёх жаропонижающих средств, в составе которых один и тот же активный ингредиент присутствует в разных дозировках. Для каждой из четырёх групп из 15 морских свинок известно изменение температуры после введения жаропонижающего. Есть ли различия в действии препаратов?

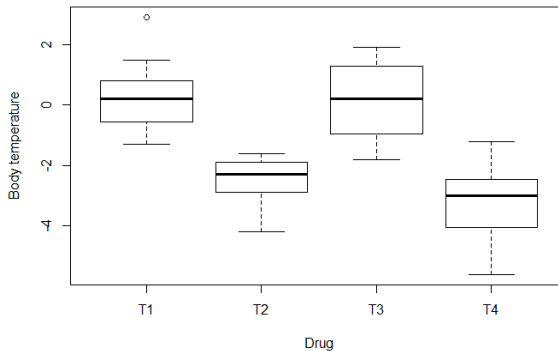


$H_0$ : температура меняется в среднем одинаково.

$H_1$ : для каких-то препаратов среднее изменение температуры отличается от остальных.



# Критерий Фишера



Критерий Фишера:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
drug	3	143.0	47.67	40.21	5.43e-14
Residuals	56	66.4	1.19		

# Критерий Краскела-Уоллиса

выборки:  $X^N = X_1^{n_1} \cup \dots \cup X_K^{n_K}$ ,  $X_k \sim F(x + \Delta_k)$   
 нулевая гипотеза:  $H_0: \Delta_1 = \Delta_2 = \dots = \Delta_K$   
 альтернатива:  $H_1: H_0$  неверна

статистика:  $K(X^N) = (N-1) \frac{\sum_{k=1}^K n_k (\bar{r}_k - \bar{r})^2}{\sum_{k=1}^K \sum_{i=1}^{n_k} (r_{ki} - \bar{r})^2}$ ,  $r_{ki} \equiv \text{rank}(X_{ki})$

нулевое распределение: табличное

Если нет связей, то:

$$\bar{r} = \frac{N-1}{2}, \sum_{k=1}^K \sum_{i=1}^{n_k} (r_{ki} - \bar{r})^2 = \frac{(N-1)N(N+1)}{12},$$

$$K(X^N) = \frac{12}{N(N+1)} \sum_{k=1}^K n_k \bar{r}_k^2 - 3(N+1).$$

Аппроксимация для  $n_k > 5$ :

$$K(X^N) \sim \chi_{K-1}^2.$$

В предыдущем примере:  $p = 1.5 \times 10^{-9}$ .

# Критерий Джонкхиера

выборки:  $X^N = X_1^{n_1} \cup \dots \cup X_K^{n_K}, \quad X_k \sim F(x + \Delta_k)$

нулевая гипотеза:  $H_0: \Delta_1 = \Delta_2 = \dots = \Delta_K$

$$\Rightarrow \text{med } X_1 = \dots = \text{med } X_K$$

альтернатива:  $H_1: \text{med } X_1 \leq \dots \leq \text{med } X_K$

статистика:  $S(X^N) = \sum_{k=1}^K \sum_{i=1}^{n_k} a_{ki}$

$a_{ki}$  — число наблюдений из первых  $k - 1$  выборок меньших, чем  $X_{ki}$

нулевое распределение: табличное

Аппроксимация для  $n_k > 10$ :

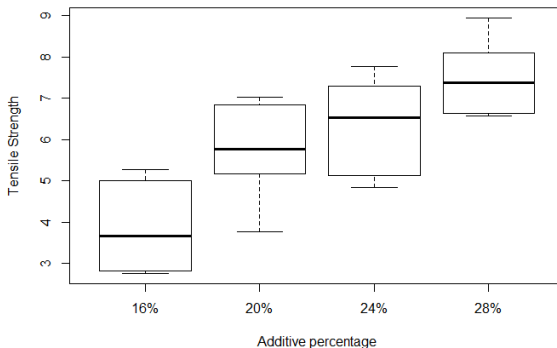
$$S(X^N) \sim N(\mu, \sigma^2),$$

$$\mu = \frac{1}{4} \left( N^2 - \sum_{k=1}^K n_k^2 \right),$$

$$\sigma = \frac{1}{72} \left( N^2 (2N + 3) - \sum_{k=1}^K n_k^2 (2n_k + 3) \right).$$

# Критерий Джонкхиера

**Пример** (Bonini, табл. 3.4): исследуется зависимость предела прочности (в Ньютонах на квадратный метр) армированного бетона с разной концентрацией присадки — 16, 20, 24 и 28%. Меняется ли средний предел прочности вместе с уровнем присадки?



$H_0$ : концентрация присадки не влияет на среднюю прочность.

$H_1$ : концентрация присадки влияет на среднюю прочность  $\Rightarrow p = 0.0042$ .

$H_1$ : увеличение концентрации присадки повышает среднюю прочность  
 $\Rightarrow p = 2.936 \times 10^{-5}$ .

## Модель со случайным эффектом

- 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.

# Модель со случайным эффектом

Если используется **модель со случайным эффектом**, следующий шаг — разделение дисперсий на внутригрупповые и межгрупповые.

Доля межгрупповой дисперсии в общей дисперсии выборки:

$$\eta^2 = \frac{SS_{bg}}{SS_{total}};$$

в популяции:

$$\hat{\omega}^2 = \frac{SS_{bg} - SS_{wg} (K - 1) / (N - K)}{SS_{total} + SS_{wg} / (N - K)}.$$

# Модель с фиксированным эффектом

- 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.

# Модель с фиксированным эффектом

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.

Для запланированного попарного сравнения групп можно просто использовать подходящий двухвыборочный критерий.

Для незапланированного сравнения всё сложнее.



# Критерий Даннета

$$D_i = \frac{\bar{X}_i - \bar{X}_1}{S \sqrt{\frac{1}{n_i} + \frac{1}{n_1}}},$$

$$S^2 = \frac{1}{N - K} \sum_{k=1}^K (n_k - 1) S_k^2,$$

где  $S_k^2$  — дисперсия выборки  $X_k^{n_k}$ .

Если  $X_{i,j} \sim N(\mu_i, \sigma^2)$ , то при  $\mu_1 = \dots = \mu_K$  вектор  $D = (D_2, \dots, D_K)$  имеет многомерное распределение Стьюдента. Кроме того, для  $D$  выполняется свойство subset pivotality, поэтому можно построить процедуру, контролирующую FWER, методом maxT.

Варианты процедуры:

- нисходящая модификация;
- непараметрическая версия.

# LSD Фишера (Least Significant Difference)

Если  $\alpha_i = \alpha_j$ , то

$$\frac{\bar{X}_i - \bar{X}_j}{S \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}} \sim St(n_i + n_j - 2),$$

где  $S^2 = \frac{(n_i-1)S_i^2 + (n_j-1)S_j^2}{n_i + n_j - 2}$ .

Рассмотрим величину

$$LSD_{ij} = \frac{t_\alpha S}{\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}},$$

где  $t_\alpha$  —  $\alpha$ -квантиль распределения Стьюдента с  $n_i + n_j - 2$  степенями свободы.

Если  $|\bar{X}_i - \bar{X}_j| > LSD_{ij}$ , то частная нулевая гипотеза  $H_0: \alpha_i = \alpha_j$  отклоняется против двусторонней альтернативы.

LSD можно использовать только в случае отвержения общей гипотезы однородности.

# HSD Тьюки (Honest Significant Difference)

$$n = \frac{K}{\sum_{k=1}^K \frac{1}{n_k}},$$

$$S^2 = \frac{1}{N - K} \sum_{k=1}^K (n_k - 1) S_k^2,$$

где  $S_k^2$  — дисперсия выборки  $X_k^{n_k}$ ,

$$HSD = \frac{q_\alpha (N - K) S}{\sqrt{n}},$$

где  $q_\alpha (N - K)$  — критическое значение распределения студентизированного размаха с  $N - K$  степенями свободы.

Если  $|\bar{X}_i - \bar{X}_j| > HSD$ , то частная нулевая гипотеза  $H_0: \alpha_i = \alpha_j$  отклоняется против двусторонней альтернативы.

HSD можно использовать независимо от справедливости общей гипотезы однородности.

# Критерий Неме́ньи

Ранговый аналог HSD.

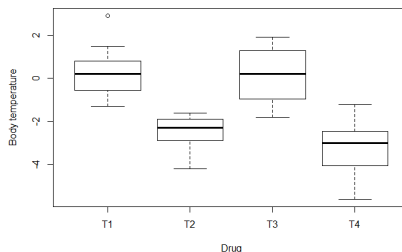
$$CD = q'_\alpha \sqrt{\frac{K(K+1)}{6N}},$$

где  $q'_\alpha$  — критическое значение статистики критерия, основанное на распределении стьюдентизированного размаха.

Если  $|\bar{r}_i - \bar{r}_j| > CD$ , то частная нулевая гипотеза  $H_0: \Delta_i = \Delta_j$  отклоняется против двусторонней альтернативы.

## Пример

Действие жаропонижающих на морских свинок:

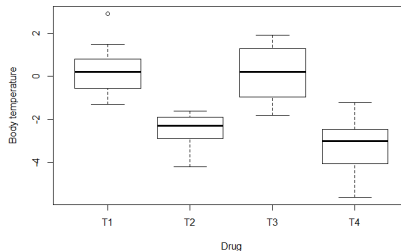


HSD:

	diff	lwr	upr	p adj
T2-T1	-2.72666667	-3.779467	-1.6738668	0.0000000
T3-T1	-0.06666667	-1.119467	0.9861332	0.9983032
T4-T1	-3.43333333	-4.486133	-2.3805335	0.0000000
T3-T2	2.66000000	1.607200	3.7127999	0.0000001
T4-T2	-0.70666667	-1.759467	0.3461332	0.2949015
T4-T3	-3.36666667	-4.419467	-2.3138668	0.0000000

# Пример

Действие жаропонижающих на морских свинок:



Критерий Неменьи:

	T1	T2	T3
T2	0.00016	-	-
T3	0.99999	0.00018	-
T4	$1.9 \times 10^{-6}$	0.79418	$2.2 \times 10^{-6}$

# Критерий Бартлетта

выборки:  $X^N = X_1^{n_1} \cup \dots \cup X_K^{n_K}, \quad X_{ki} \sim N(\mu_k, \sigma_k^2)$

нулевая гипотеза:  $H_0: \sigma_1 = \sigma_2 = \dots = \sigma_K$

альтернатива:  $H_1: H_0$  неверна

статистика:  $B(X^N) = \frac{\ln 10}{C} \left( (N - K) \ln S^2 - \sum_{k=1}^K (n_k - 1) \ln S_k^2 \right)$

$$S^2 = \frac{1}{N-K} \sum_{k=1}^K (n_k - 1) S_k^2$$

$$C = 1 + \frac{1}{3K+1} \left( \sum_{k=1}^K \frac{1}{n_k-1} - \frac{1}{N} \right)$$

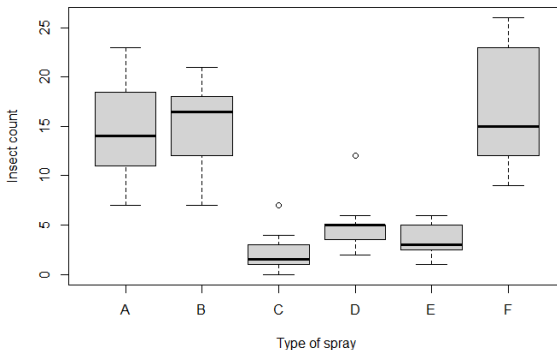
нулевое распределение: табличное

Аппроксимация для  $n_k > 6$ :

$$B(X^N) \sim \chi_{K-1}^2.$$

# Критерий Бартлетта

**Пример** (Beall, 1942): шесть видов инсектицидов тестируется на 12 полях каждый, исследуемый признак — количество насекомых на поле через некоторое время после обработки.



$H_0$ : дисперсия числа насекомых на полях, обрабатываемых разными инсектицидами, одинакова.

$H_1$ : дисперсия числа насекомых на полях, обрабатываемых разными инсектицидами, неодинакова  $\Rightarrow p = 9 \times 10^{-5}$ .



# Критерий Флайнера-Киллиана

выборки:  $X^N = X_1^{n_1} \cup \dots \cup X_K^{n_K}, \quad X_{ki} \sim F(\mu_k + \sigma_k x)$

нулевая гипотеза:  $H_0: \sigma_1 = \sigma_2 = \dots = \sigma_K$

альтернатива:  $H_1: H_0$  неверна

статистика:  $X^2(X^N) = \frac{1}{V^2} \sum_{k=1}^K n_k (\bar{A}_k - \bar{a})$

$$\bar{a} = \frac{1}{N} \sum_{k=1}^K \sum_{i=1}^{n_k} a_{ki}, \quad \bar{A}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} a_{ki}$$

$$a_{ki} = \Phi^{-1} \left( \frac{1 + \frac{r_{ki}}{N+1}}{2} \right)$$

$r_{ki}$  — ранг  $|X_{ki} - \tilde{X}_k|$  в объединённой выборке

$\tilde{X}_k$  — выборочная медиана  $k$ -й выборки

$$V^2 = \frac{1}{N-1} \sum_{k=1}^K \sum_{i=1}^{n_k} (a_{ki} - \bar{a})^2$$

нулевое распределение:  $\chi_{K-1}^2$

В предыдущем примере:  $p = 0.01282$ .

# Двухфакторный дисперсионный анализ

$$f_1: X \rightarrow \{1, \dots, K_1\}, \quad f_2: X \rightarrow \{1, \dots, K_2\}$$

$f_1 \backslash f_2$	1	...	$j$	...	$K_2$
1					
$\vdots$					
$i$			$X_{ij1}$ $\vdots$ $X_{ijn_{ij}}$		
$\vdots$					
$K_1$					

Задача: проверить гипотезу об отсутствии влияния факторов  $f_1$  и  $f_2$  на среднее значение признака  $X$ .

Случай выборок разного размера для двух факторов значительно сложнее, поэтому будем считать, что  $n_{11} = \dots = n_{K_1 K_2} = n$ .

# Двухфакторный дисперсионный анализ

Линейная модель:

$$X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk},$$

$$i = 1, \dots, K_1, j = 1, \dots, K_2, k = 1, \dots, n.$$

$\mu$  — общее среднее значение признака,

$\alpha_i$  — воздействие уровня  $i$  фактора  $f_1$ ,

$\beta_j$  — воздействие уровня  $j$  фактора  $f_2$ ,

$\gamma_{ij}$  — дополнительное воздействие комбинации уровней  $i$  и  $j$  факторов  $f_1$  и  $f_2$ ,

$\varepsilon_{ijk}$  — случайные независимые одинаково распределённые ошибки.

# Двухфакторный дисперсионный анализ

$H_0^1$ : фактор  $f_1$  не влияет на значение признака  $X \Leftrightarrow$   
 $\alpha_i = 0 \quad \forall i,$

$H_1^1$ :  $f_1$  влияет на значение  $X$ ;

$H_0^2$ : фактор  $f_2$  не влияет на значение признака  $X \Leftrightarrow$   
 $\beta_j = 0 \quad \forall j,$

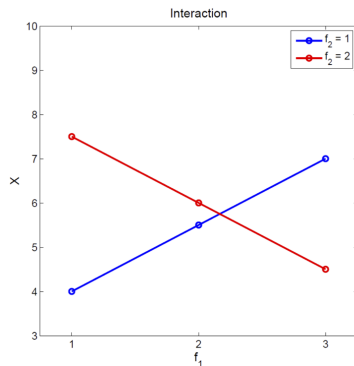
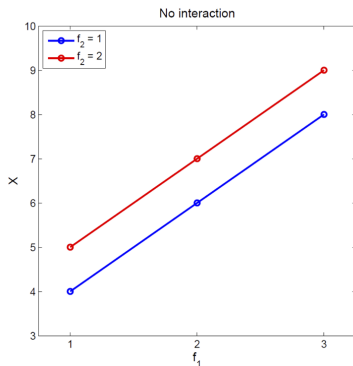
$H_1^2$ :  $f_2$  влияет на значение  $X$ ;

$H_0^{12}$ : между факторами  $f_1, f_2$  нет взаимодействия  $\Leftrightarrow$   
 $\gamma_{ij} = 0 \quad \forall i, j,$

$H_1^{12}$ : между факторами  $f_1, f_2$  есть взаимодействие.

# Двухфакторный дисперсионный анализ

**Пример:**  $X$  — успешность решения задачи (в баллах от 0 до 10),  
 $f_1$  — размер команды (1 — маленькая, 2 — средняя, 3 — большая),  
 $f_2$  — наличие назначенного лидера (1 — нет, 2 — есть).



# Нормальный двухфакторный дисперсионный анализ

Предположим, что  $X_{ijk} \sim N(\mu_{ij}, \sigma^2) \Leftrightarrow \varepsilon_{ijk} \sim N(0, \sigma^2)$ .

$\bar{X}_{ij}$  — среднее в ячейке,

$\bar{X}_{i\bullet}$  — среднее по строке  $i$ ,

$\bar{X}_{\bullet j}$  — среднее по столбцу  $j$ ,

$\bar{X}$  — среднее по всей таблице.

Внутрифакторные дисперсии:

$$S_1^2 = \frac{nK_2}{K_1 - 1} \sum_{i=1}^{K_1} (\bar{X}_{i\bullet} - \bar{X})^2,$$

$$S_2^2 = \frac{nK_1}{K_2 - 1} \sum_{j=1}^{K_2} (\bar{X}_{\bullet j} - \bar{X})^2,$$

$$S_{12}^2 = \frac{n}{(K_1 - 1)(K_2 - 1)} \sum_{i,j} (\bar{X}_{ij} - \bar{X}_{i\bullet} - \bar{X}_{\bullet j} + \bar{X})^2,$$

$$S_{res}^2 = \frac{1}{K_1 K_2 (n - 1)} \sum_{k=1}^n \sum_{i,j} (X_{ijk} - \bar{X}_{ij})^2.$$

# Нормальный двухфакторный дисперсионный анализ

Проверка значимости факторов и их взаимодействия:

- $n > 1$ :

$$F_1 = \frac{S_1^2}{S_{res}^2} \sim F(K_1 - 1, K_1 K_2 (n - 1)) \text{ при } H_0^1,$$

$$F_2 = \frac{S_2^2}{S_{res}^2} \sim F(K_2 - 1, K_1 K_2 (n - 1)) \text{ при } H_0^2,$$

$$F_{12} = \frac{S_{12}^2}{S_{res}^2} \sim F((K_1 - 1)(K_2 - 1), K_1 K_2 (n - 1)) \text{ при } H_0^{12};$$

- $n = 1$ :

$$F_1 = \frac{S_1^2}{S_{12}^2} \sim F(K_1 - 1, (K_1 - 1)(K_2 - 1)) \text{ при } H_0^1,$$

$$F_2 = \frac{S_2^2}{S_{12}^2} \sim F(K_2 - 1, (K_1 - 1)(K_2 - 1)) \text{ при } H_0^2.$$

При этом подразумевается, что  $H_0^{12}$  верна.

# Марихуана и скорость реакции

**Пример** (Pagano, 2012, задача 16.2): изучалось воздействие марихуаны на скорость реакции. В качестве испытуемых были выбраны по 12 человек из каждой категории:

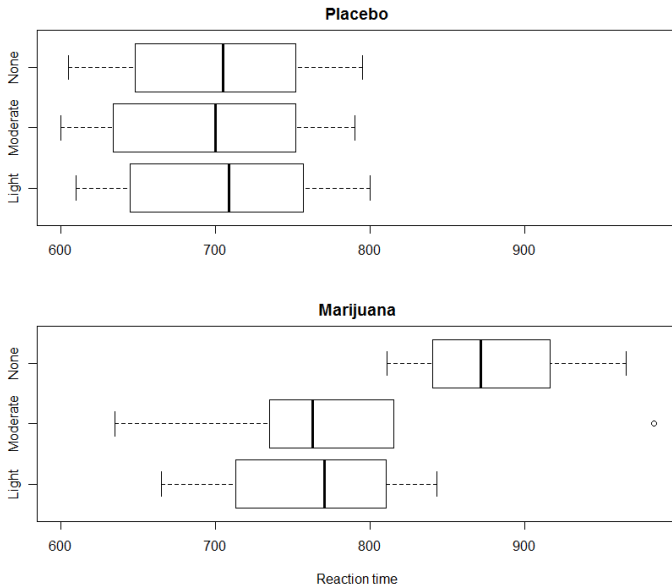
- никогда не пробовали марихуану;
- иногда употребляют марихуану;
- регулярно употребляют марихуану.

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.



# Марихуана и скорость реакции



# Маришуана и скорость реакции

$H_0^1$ : средняя скорость реакции одинакова при употреблении и марихуаны, и сигарет.

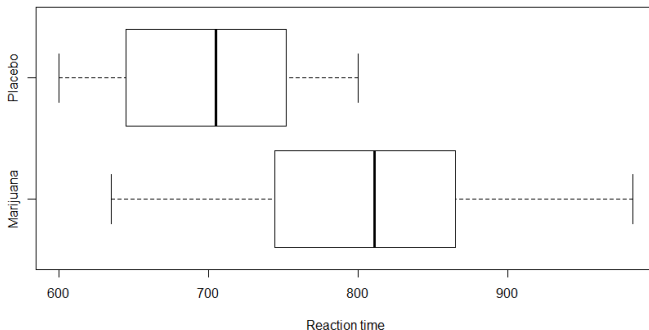
$H_0^2$ : средняя скорость реакции не зависит от предыдущего опыта употребления марихуаны.

$H_0^{12}$ : отсутствует межфакторное взаимодействие между употребляемым веществом и предыдущим опытом употребления марихуаны.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Treatment	1	103041	103041	17.584	0.000224
Past usage	2	23634	11817	2.017	0.150752
Past usage:Treatment	2	23642	11821	2.017	0.150665
Residuals	30	175796	5860		

# Марихуана и скорость реакции

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



Для объединённых данных:

- однофакторный дисперсионный анализ:  $p = 0.0004$ ;
- критерий Стьюдента, односторонняя альтернатива:  $p = 0.0002$ , 95% нижний доверительный предел — 61.2.

# 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.

## CBI чернотрухой дрозофилы

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



## CBI чернотрухой дрозофилы

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Chromosome	2	0.01033	0.00516	0.657	0.52554
Chromosome:Region	3	0.16295	0.05432	6.915	0.00113
Residuals	30	0.23564	0.00785		

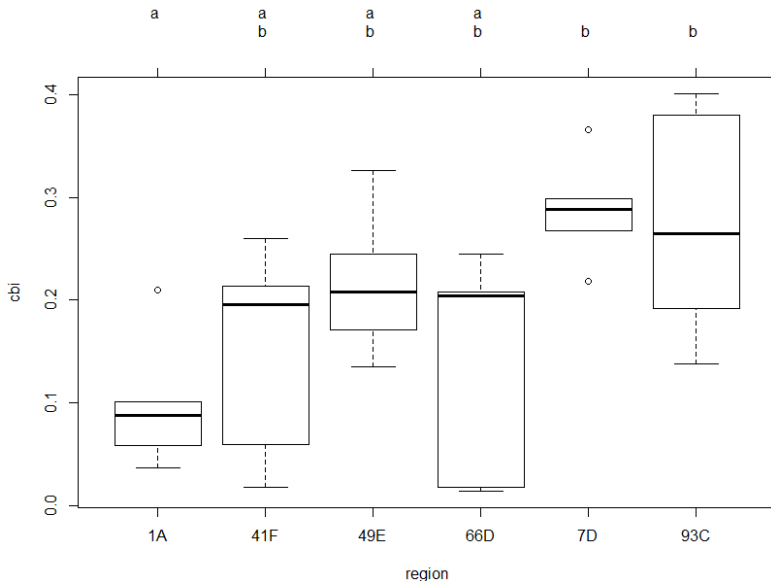
Есть различия между регионами, нет различий между хромосомами.

## СВІ чернобрюхой дрозофилы

Для уточнения различий применим метод HSD:

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 чернотрухой дрозофилы





# Болтающаяся контрольная группа

Лекарство \ Доза	5 мг	10 мг
Препарат А		
Препарат В		

Плацебо, 0 мг

Используется однофакторный дисперсионный анализ с последующими запланированными сравнениями.

# Однофакторный дисперсионный анализ для связанных выборок

Объект \ $f$	1	...	$k$	...	$K$
1	$X_{11}$		$X_{k1}$		$X_{K1}$
$\vdots$	$\vdots$	...	$\vdots$	...	$\vdots$
$n$	$X_{1n}$		$X_{kn}$		$X_{Kn}$

Линейная модель:

$$X_{ki} = \mu + \alpha_k + \beta_i + \varepsilon_{ki},$$

$$i = 1, \dots, n, \quad k = 1, \dots, K.$$

$\mu$  — глобальное среднее значение признака  $X$ ,

$\beta_i$  — отклонение от  $\mu$ , вызванное влиянием особенностей  $i$ -го объекта,

$\alpha_k$  — отклонение от  $\mu + \beta_i$ , вызванное влиянием  $k$ -го уровня фактора  $f$ ,

$\varepsilon_{ki}$  — случайные независимые одинаково распределённые ошибки.

Средние значения  $X$  во всех  $K$  выборках одинаковы  $\Leftrightarrow \alpha_1 = \dots = \alpha_K$ .

# Критерий Фишера

выборки:  $X^N = X_1^{n_1} \cup \dots \cup X_K^{n_K}$   
 нулевая гипотеза:  $H_0: \alpha_1 = \alpha_2 = \dots = \alpha_K$   
 альтернатива:  $H_1: H_0$  неверна

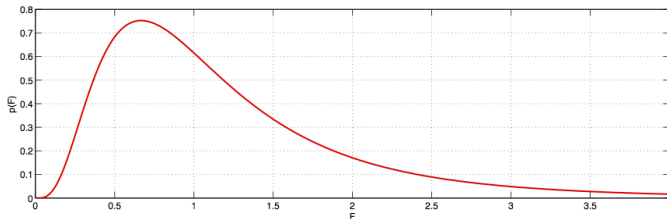
статистика:  $F(X^N) = \frac{SS_{bg}/(K-1)}{(SS_{wg} - SS_s)/(n-1)(K-1)}$

$$SS_{bg} = n \sum_{k=1}^K (\bar{X}_k - \bar{X})^2$$

$$SS_{wg} = \sum_{k=1}^K \sum_{i=1}^n (X_{ki} - \bar{X}_k)^2$$

$$SS_s = K \sum_{i=1}^n (\bar{X}_i - \bar{X})^2$$

нулевое распределение:  $F(K-1, (n-1)(K-1))$



# Критерий Фишера

Предположения метода:

- 1 выборочные распределения средних значений признака во всех группах нормальны;
- 2 для фактора с более чем двумя уровнями: попарные разности признака имеют одинаковую дисперсию для любых уровней фактора (сферичность);
- 3 объекты независимы.

Предположение сферичности на практике нарушается наиболее часто, причём это может привести к росту вероятности ошибки первого рода. Проверить гипотезу сферичности можно с помощью критерия Маухли, если она отвергается, используются модификации числа степеней свободы критерия Фишера.

# Критерий Фишера

**Пример** (Pearson, 2003): исследовалось влияние метилфенидата на способность к отсрочке удовольствия умственно отсталыми детьми с синдромом дефицита внимания и гиперактивности. Каждый испытуемый принимал либо препарат в одной из трёх дозировок, либо плацебо, после чего проходил тест.

$H_0$ : препарат не влияет на среднюю способность к отсрочке удовольствия.

$H_1$ : препарат влияет на среднюю способность к отсрочке удовольствия  
 $\Rightarrow p = 0.004$ .

# Критерий Фридмана

выборки:  $X_{ki} = \mu + \alpha_k + \beta_i + \varepsilon_{ki}, \quad i = 1, \dots, n, \quad k = 1, \dots, K$   
 нулевая гипотеза:  $H_0: \alpha_1 = \dots = \alpha_K$   
 альтернатива:  $H_1: H_0$  неверна

статистика:  $S(X) = \frac{12}{nK(K+1)} \sum_{k=1}^K R_k^2 - 3n(K-1)$

$$R_k = \sum_{i=1}^n r_{ki}$$

$r_{ki}$  — ранг  $k$ -го элемента в  $i$ -й строке

нулевое распределение: табличное

Распространённая аппроксимация для  $n > 15, K > 10$ :

$$S(X) \sim \chi_{K-1}^2.$$

Более точная аппроксимация:

$$\frac{(n-1)S(X)}{n(K-1) - S(X)} \sim F(n-1, (n-1)(K-1)).$$

# Критерий Фридмана

**Пример:** исследуется 5 технологий вытачивания детали. Каждый из 15 рабочих в течение нескольких смен использовал каждую из технологий.  $X_{ki}$  — производительность  $i$ -го рабочего при использовании  $k$ -й технологии.

$H_0$ : выбор технологии не меняет производительности рабочих.

$H_1$ : выбор технологии влияет на производительность рабочих

$\Rightarrow p = 0.356$ .

# Критерий Пейджа

выборки:  $X_{ki} = \mu + \alpha_k + \beta_i + \varepsilon_{ki}, \quad i = 1, \dots, n, \quad k = 1, \dots, K$

нулевая гипотеза:  $H_0: \alpha_1 = \dots = \alpha_K$

альтернатива:  $H_1: \alpha_1 \leq \dots \leq \alpha_K$

статистика:  $L(X) = \sum_{k=1}^K kR_k$

$$R_k = \sum_{i=1}^n r_{ki}$$

$r_{ki}$  — ранг  $k$ -го элемента в  $i$ -й строке

нулевое распределение: табличное

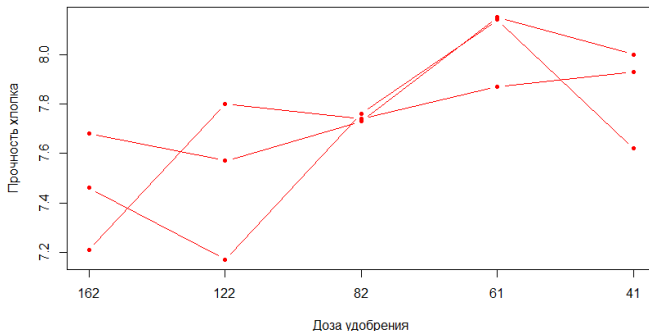
Аппроксимация для  $n > 15, K > 10$ :

$$L(X) \sim N \left( \frac{nK(K+1)^2}{4}, \frac{n(K^3 - K)^2}{144(K-1)} \right).$$



# Критерий Пейджа

**Пример** (Лагутин, 17.3.2): на 3 полях тестируется 5 доз калийных удобрений. Каждое поле поделено на 5 участков, по одному на каждую дозу. Измерена прочность выращенного на каждом участке хлопка.



$H_0$ : дозировка удобрений не влияет на прочность хлопка.

$H_1$ : дозировка удобрений влияет на прочность хлопка  $\Rightarrow p = 0.0663$ .

$H_1$ : с ростом дозировки удобрений прочность хлопка уменьшается  
 $\Rightarrow p < 0.01$ .

# Литература

- разновидности ANOVA — Tabachnick, 3.2;
- unbalanced two-way ANOVA — Tabachnik, 6;
- критерии Краскела-Уоллиса (Kruskal–Wallis) и Джонкхиера (Jonckheere) — Кобзарь, 4.2.1.2.1, 4.2.1.2.9;
- критерии Фридмана (Friedman) и Пейджа (Page) — Лагутин, гл. 17;
- перестановочные аналоги — Bonnini, гл. 3, 4;
- критерий Маухли (Mauchly's sphericity test), поправки при отсутствии сферичности (Huynh-Feldt, Greenhouse-Geisser, lower-bound) — [http://en.wikipedia.org/wiki/Mauchly's\\_sphericity\\_test](http://en.wikipedia.org/wiki/Mauchly's_sphericity_test);
- profile analysis — альтернатива w.s. ANOVA — Davis;
- **примеры проведения дисперсионного анализа в R:**  
1-way b.s., 2-way b.s., 1-way w.s., 2-way w.s.

# Литература

Лагутин М.Б. *Наглядная математическая статистика*, 2007.

Кобзарь А.И. *Прикладная математическая статистика*, 2006.

Beall G. (1942). *The Transformation of data from entomological field experiments*. Biometrika, 29, 243–262.

Bonnini S., Corain L., Marozzi M., Salmaso S. *Nonparametric Hypothesis Testing - Rank and Permutation Methods with Applications in R*, 2014.

Davis C.S. *Statistical Methods for the Analysis of Repeated Measurements*, 2002.

Pagano R.R. *Understanding Statistics in the Behavioral Sciences*, 2012.

Pearson D.A, Santos C.W., Roache J. D., et al. (2003). *Treatment effects of methylphenidate on behavioral adjustment in children with mental retardation and ADHD*. Journal of the American Academy of Child and Adolescent Psychiatry, 42(3), 209-216.

Tabachnick B.G., Fidell L.S. *Using Multivariate Statistics*, 2012.