Прикладная статистика 5. Дисперсионный анализ.

Рябенко Евгений
riabenko.e@gmail.com

17 марта 2014 г.

Однофакторный дисперсионный анализ (one-way ANOVA)

Пусть имеется K выборок:

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

Эквивалентная запись:
фактор $f\colon X \to \{1, \ldots, K\}$

f	1	$\ldots$	k	$\ldots$	K
X^N	X_{11} $\vdots$ X_{1n_1}	$\ldots$	X_{k1} $\vdots$ X_{kn_k}	$\ldots$	X_{K1} $\vdots$ 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 (one-way ANOVA)

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

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

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

Возведём в квадрат и просуммируем:

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

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

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

Разновидности однофакторного дисперсионного анализа

- По типу выборок: независимые (between-subjects) 1 2 3, связанные (within-subjects, repeated measurements) 4 5 6.
- По используемым предположениям: нормальный 1 4, непараметрический 2 3 5 6.
- По типу альтернативы: общая 1 2 4 5, тренда 3 6.
- По объёму выборок: одинаковый (balanced), различный (unbalanced).
- По типу эффектов: случайные (random-effects) 7, фиксированные (fixed-effects) 8.

Однофакторный дисперсионный анализ для независимых выборок

Линейная модель:

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

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

μ — глобальное среднее значение признака X,
α_k — отклонение от μ, вызванное влиянием k-го уровня фактора f,
ε_{ki} — случайные независимые одинаково распределённые ошибки.

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

Критерий Фишера

выборки: $X^N = X_1^{n_1} \bigcup \ldots \bigcup X_K^{n_K}, \quad X_{ki} \sim N\left(\mu_k, \sigma^2\right);$

нулевая гипотеза: $H_0: \mu_1 = \mu_2 = \ldots = \mu_K;$

альтернатива: $H_1: H_0$ неверна;

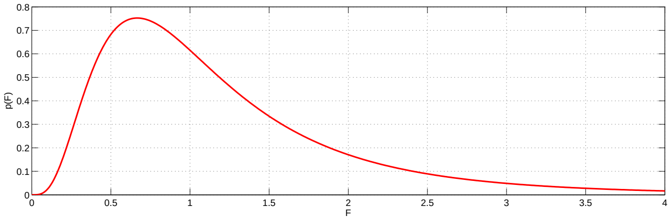
статистика:

$$F\left(X^N\right) = \frac{SS_{bg}/(K-1)}{SS_{wg}/(N-K)},$$

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

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

$F\left(X^N\right) \sim F(K-1, N-K)$ при H_0;



достигаемый уровень значимости:

$$p(f) = 1 - fcdf(f, K-1, N-K).$$

1 Критерий Фишера

Предположения метода:

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

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.

1 Критерий Фишера

Пример: топливная компания тестирует влияние трёх видов присадок на потребление бензина. Выборка получена на 12 одинаковых автомобилях, на каждом из которых использовалась одна из трёх присадок.

$H_0:$ все три вида присадок одинаково влияют на среднее потребление бензина.
$H_1:$ между средними уровнями потребления бензина с разными присадками есть различия $\Rightarrow p = 2.1717 \times 10^{-5}$.

2 Критерий Краскела-Уоллиса

выборки: $X^N = X_1^{n_1} \bigcup \ldots \bigcup X_K^{n_K}, \quad X_{ki} \sim F(x + \Delta_k)$;
нулевая гипотеза: $H_0\colon \Delta_1 = \Delta_2 = \ldots = \Delta_K$;
альтернатива: $H_1\colon H_0$ неверна;
статистика: $K\left(X^N\right) = (N-1)\frac{\sum\limits_{k=1}^{K} n_k(\bar{r}_k - \bar{r})^2}{\sum\limits_{k=1}^{K}\sum\limits_{i=1}^{n_k}(r_{ki}-\bar{r})^2}$,
$K\left(X^N\right)$ имеет табличное распределение при H_0.

Если нет связок, то:

$$\bar{r} = \frac{N-1}{2},$$

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

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

Аппроксимация для $n_k > 5$:

$$K\left(X^N\right) \sim \chi^2_{K-1}.$$

2 Критерий Краскела-Уоллиса

Пример: дегустаторы оценивают торты по совокупности факторов — вкус, внешний вид, запах и фактура. Итоговая оценка выставляется в баллах от 0 до 100. Сравниваются оценки трёх видов тортов, представленных каждый отдельной команде дегустаторов.

H_0: оценки трёх видов тортов в среднем одинаковы.
H_1: между оценками разных видов тортов есть различия $\Rightarrow p = 0.6587$.

3 Критерий Джонкхиера

выборки: $X^N = X_1^{n_1} \bigcup \ldots \bigcup X_K^{n_K}, \quad X_{ki} \sim F(x + \Delta_k);$

нулевая гипотеза: $H_0\colon \Delta_1 = \Delta_2 = \ldots = \Delta_K$

$\Rightarrow \ \mathrm{med}\, X_1 = \ldots = \mathrm{med}\, X_K;$

альтернатива: $H_1\colon \ \mathrm{med}\, X_1 \leq \ldots \leq \mathrm{med}\, X_K;$

статистика: $S\left(X^N\right) = \sum\limits_{k=1}^{K} \sum\limits_{i=1}^{n_k} a_{ki},$

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

$S\left(X^N\right)$ имеет табличное распределение при H_0.

Аппроксимация для $n_k > 10$:

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

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

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

3 Критерий Джонкхиера

Пример: исследуется влияние информированности (знания цели работы) на выполнение монотонных производственных операций. 18 рабочих были случайным образом разделены на 3 группы. Попавшие в группу 1 не имели информации о требуемой производительности, в группу 2 — получили общее представление о том, что нужно делать, в группу 3 — точную информацию о задании и график выполнения работ.

$H_0:$ информированность не влияет на производительность.
$H_1:$ информированность влияет на производительность $\Rightarrow p = 0.113$.
$H_1:$ информированность повышает производительность $\Rightarrow p = 0.022$.

Однофакторный дисперсионный анализ для связанных выборок

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

Линейная модель:

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

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

μ — глобальное среднее значение признака X,
α_k — отклонение от $\mu + \beta_i$, вызванное влиянием k-го уровня фактора f,
β_i — отклонение от μ, вызванное влиянием особенностей i-го объекта,
ε_{ki} — случайные независимые одинаково распределённые ошибки.

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

4 Критерий Фишера

выборки: $X^N = X_1^{n_1} \bigcup \ldots \bigcup X_K^{n_K}, \quad X_{ki} \sim N\left(\mu_k, \sigma^2\right);$

нулевая гипотеза: $H_0\colon \mu_1 = \mu_2 = \ldots = \mu_K;$

альтернатива: $H_1\colon H_0$ неверна;

статистика:

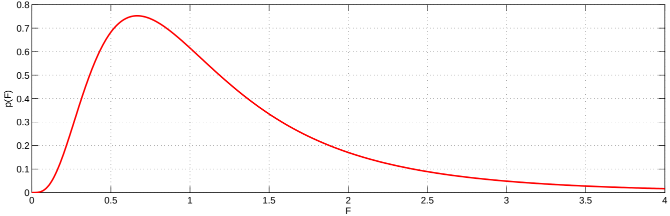
$$F\left(X^N\right) = \frac{SS_{bg}/(K-1)}{(SS_{wg}-SS_s)/(n-1)(K-1)},$$

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

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

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

$F\left(X^N\right) \sim F(K-1, (n-1)(K-1))$ при H_0;



достигаемый уровень значимости:

$$p(f) = 1 - fcdf(f, K-1, (n-1)(K-1)).$$

4 Критерий Фишера

Предположения метода:

1. значения признака во всех группах нормально распределены;
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.

4 Критерий Фишера

Пример: (Pearson et al, 2003, Treatment effects of methylphenidate on behavioral adjustment in children with mental retardation and ADHD) 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
$\Rightarrow p = 0.004.$

5 Критерий Фридмана

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

нулевая гипотеза: $H_0 \colon \alpha_1 = \ldots = \alpha_K;$

альтернатива: $H_1 \colon H_0$ неверна;

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

$R_k = \sum\limits_{i=1}^{n} r_{ki},$

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

$S(X)$ имеет табличное распределение при H_0.

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

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

Более точная аппроксимация:

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

5 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
$\Rightarrow p = 0.356.$

6 Критерий Пейджа

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

нулевая гипотеза: $H_0\colon \alpha_1 = \ldots = \alpha_K;$

альтернатива: $H_1\colon \alpha_1 \le \ldots \le \alpha_K;$

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

$R_k = \sum\limits_{i=1}^{n} r_{ki},$

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

$L(X)$ имеет табличное распределение при H_0.

Аппроксимация для $n > 15, K > 10$:

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

6 Критерий Пейджа

Пример: на 20 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 $\Rightarrow p = 0.126$.
$H_1:$ с ростом дозировки удобрений прочность хлопка увеличивается $\Rightarrow p = 0.046$.

7 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 горбаток 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.

7 Модель со случайным эффектом (random-effects model ANOVA)

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

Результат — доля межгрупповой дисперсии в общей дисперсии X^N.

8 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.

8 Модель с фиксированным эффектом (fixed-effects model ANOVA)

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.

Fisher's LSD (Least Significant Difference)

Если $\mu_i = \mu_j$, то

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

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

Рассмотрим величину

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

где t_α — α-квантиль распределения Стьюдента с $n_i + n_j - 2$ степенями свободы.

Если $\left|\bar{X}_i - \bar{X}_j\right| > LSD_{ij}$, то частная нулевая гипотеза $H_0\colon \mu_i = \mu_j$ отклоняется против двусторонней альтернативы.

Метод LSD можно использовать только в случае отвержения общей гипотезы однородности.

Tukey's HSD (Honest Significant Difference)

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

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

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 стьюдентизированного размаха с $N-K$ степенями свободы.

Если $\left|\bar{X}_i - \bar{X}_j\right| > HSD$, то частная нулевая гипотеза $H_0\colon \mu_i = \mu_j$ отклоняется против двусторонней альтернативы.

Метод HSD можно использовать независимо от справедливости общей гипотезы однородности.

Критерий Неменьи

Непараметрический аналог процедуры HSD.

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

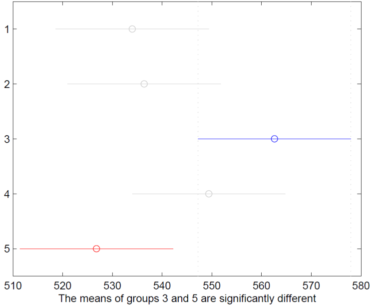
где q'_{α} — критическое значение статистики критерия, основанное на распределении стьюдентизированного размаха.

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

Пример

Овсяная мука пяти видов помола расфасовывается при помощи одного диспенсера. Стандартный объём упаковки — 500 г, но диспенсер обычно насыпает больше. Производитель подозревает, что объём упаковки может зависеть от помола муки.

Метод LSD: вес в группах 3 и 5 значимо отличается.



Метод HSD: значимых различий между средними не обнаружено.

Критерий Бартлета

выборки: $X^N = X_1^{n_1} \bigcup \ldots \bigcup X_K^{n_K}, \quad X_{ki} \sim N\left(\mu_k, \sigma_k^2\right);$

нулевая гипотеза: $H_0\colon \sigma_1 = \sigma_2 = \ldots = \sigma_K;$

альтернатива: $H_1\colon H_0$ неверна;

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

$S^2 = \frac{1}{N-K}\sum\limits_{k=1}^{K}\left(n_k-1\right)S_k^2,$

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

$B\left(X^N\right)$ имеет табличное распределение при H_0.

Аппроксимация для $n_k > 6$:

$$B\left(X^N\right) \sim \chi^2_{K-1}.$$

Критерий Бартлета

Пример: четыре шпиндельные головки сравниваются по вариабельности размеров деталей, которые выточены с их помощью. Контролёром качества собраны выборки из 31, 15, 20 и 42 деталей.

$H_0:$ дисперсия размеров деталей, выточенных с помощью различных головок, одинакова.
$H_1:$ дисперсия размеров деталей, выточенных с помощью различных головок, неодинакова $\Rightarrow p = 0.0626$.

Критерий квадратов рангов

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

нулевая гипотеза: $H_0\colon \sigma_1 = \sigma_2 = \ldots = \sigma_K;$

альтернатива: $H_1\colon H_0$ неверна;

статистика: $T_2\left(X^N\right) = \frac{1}{D^2}\left(\sum\limits_{k=1}^{K} \frac{S_k^2}{n_k} - N\bar{S}^2\right),$

$S_k = \sum\limits_{i=1}^{n_k} r\left(\left|X_{ki} - \bar{X}_k\right|\right)^2,$

$\bar{S} = \frac{1}{N}\sum\limits_{k=1}^{K} S_k,$

$D^2 = \frac{1}{N-1}\left(\sum\limits_{i=1}^{N} r_i^4 - N\bar{S}^2\right);$

$T_2\left(X^N\right)$ имеет табличное распределение при H_0.

Если нет связок, то:

$$\bar{S} = \frac{1}{6}(N+1)(2N+1),$$

$$D^2 = \frac{1}{180}N(N+1)(2N+1)(8N+11).$$

Аппроксимация для $n_k > 10$:

$$T_2\left(X^N\right) \sim \chi^2_{K-1}.$$

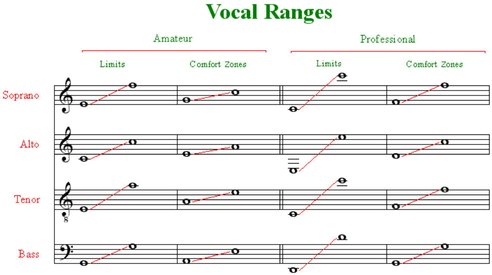
Критерий квадратов рангов

Пример: четыре шпиндельные головки сравниваются по вариабельности размеров деталей, которые выточены с их помощью. Контролёром качества собраны выборки из 31, 15, 20 и 42 деталей.

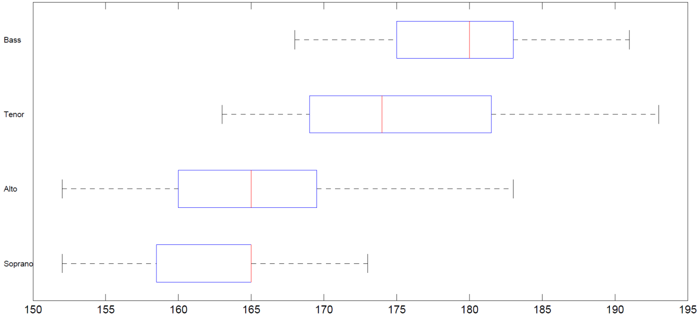
$H_0:$ дисперсия размеров деталей, выточенных с помощью различных головок, одинакова.
$H_1:$ дисперсия размеров деталей, выточенных с помощью различных головок, неодинакова $\Rightarrow p = 0.0856$.

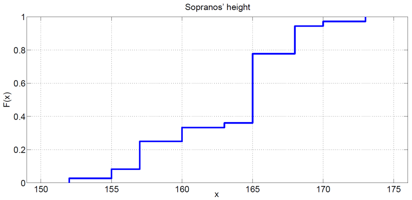
Рост певцов хора

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?



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.

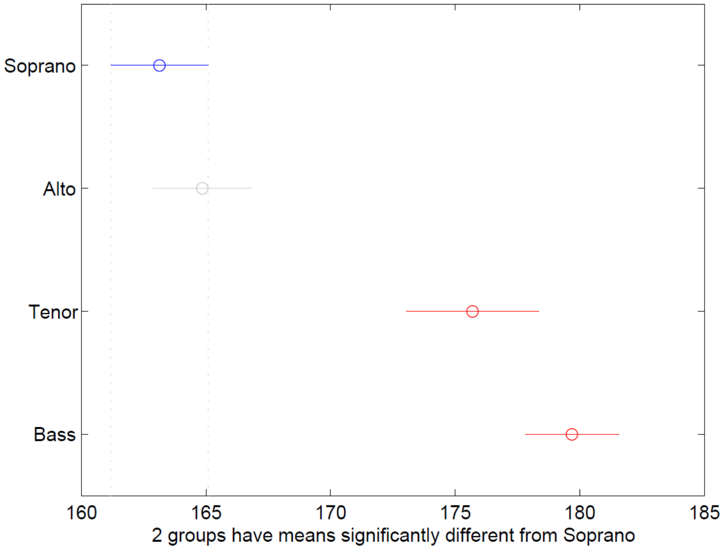
Рост певцов хора

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

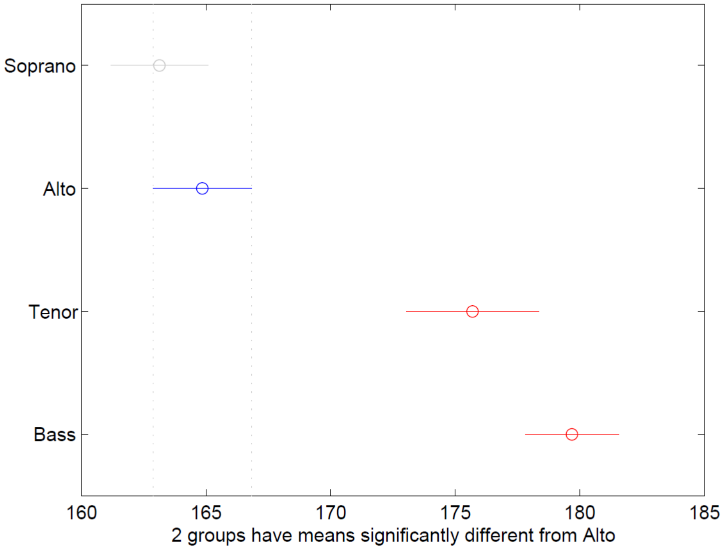
Критерий Стьюдента для проверки гипотезы равенства роста певцов с тенором и басом: $p = 0.0597$ — против двусторонней альтернативы, $p = 0.0298$ — против односторонней альтернативы.

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

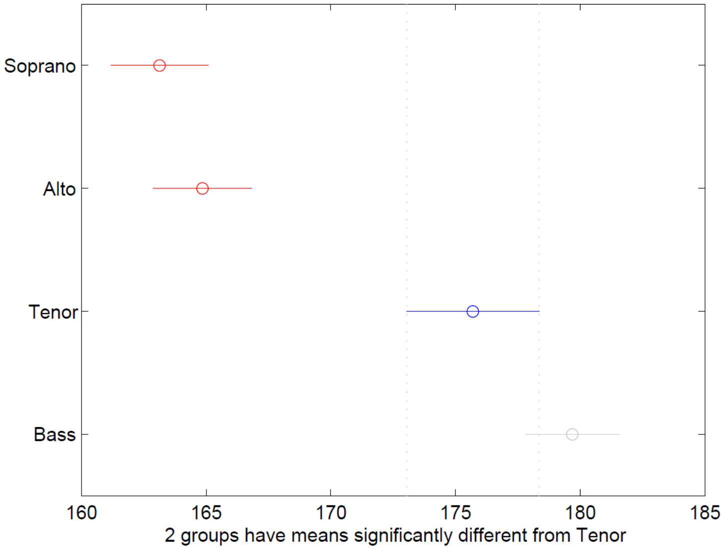
Рост певцов хора



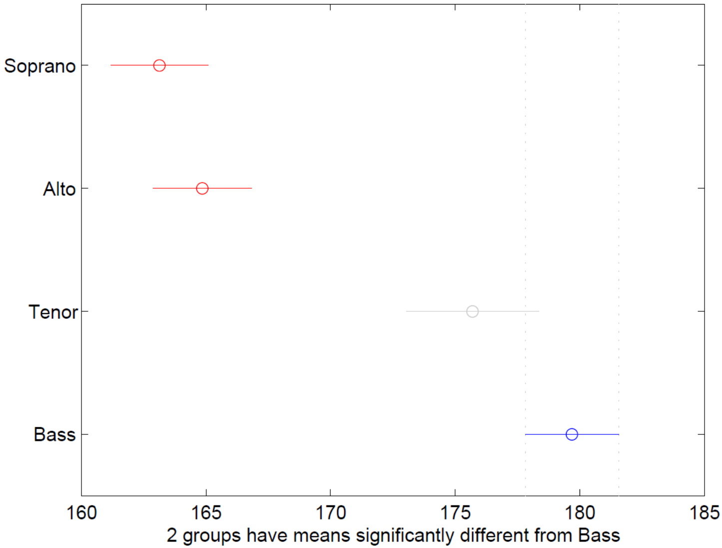
Рост певцов хора



Рост певцов хора



Рост певцов хора



Двухфакторный дисперсионный анализ (two-way ANOVA)

$$f_1\colon X \to \{1,\ldots,K_1\}\,,\quad f_2\colon X \to \{1,\ldots,K_2\}$$

$f_1 \backslash f_2$	1	$\ldots$	j	$\ldots$	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.

Двухфакторный дисперсионный анализ (two-way ANOVA)

Линейная модель:

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

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

μ — общее среднее значение признака,
α_i — воздействие уровня i фактора f_1,
β_j — воздействие уровня j фактора f_2,
γ_{ij} — дополнительное воздействие комбинации уровней i и j факторов f_1 и f_2,
ε_{ijk} — случайные независимые одинаково распределённые ошибки.

Двухфакторный дисперсионный анализ (two-way ANOVA)

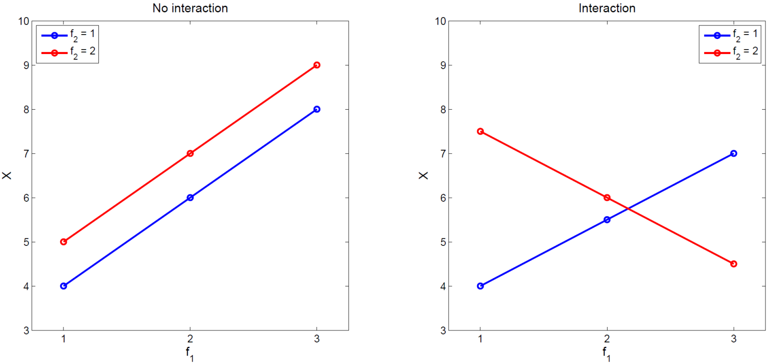
$H_0^1:$ фактор f_1 не влияет на значение признака $X \Leftrightarrow$
$\alpha_i = 0 \ \ \forall i,$
$H_1^1:$ f_1 влияет на значение $X;$

$H_0^2:$ фактор f_2 не влияет на значение признака $X \Leftrightarrow$
$\beta_j = 0 \ \ \forall j,$
$H_1^2:$ f_2 влияет на значение $X;$

$H_0^{12}:$ между факторами f_1, f_2 нет взаимодействия $\Leftrightarrow$
$\gamma_{ij} = 0 \ \ \forall i, j,$
$H_1^{12}:$ между факторами f_1, f_2 есть взаимодействие.

Двухфакторный дисперсионный анализ (two-way ANOVA)

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).



Дополнительные разновидности двухфакторного дисперсионного анализа

- По типу факторов: независимые, вложенные (nested), латинский квадрат (latin square).

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

Нормальный двухфакторный дисперсионный анализ

Предположим, что $X_{ijk} \sim N\left(\mu_{ij}, \sigma^2\right) \Leftrightarrow \varepsilon_{ijk} \sim N\left(0, \sigma^2\right)$.
$\bar{X}_{ij}$ — 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:

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

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

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

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

Нормальный двухфакторный дисперсионный анализ

Проверка значимости факторов и их взаимодействия:

- $n > 1$:

$$F_1 = \frac{S_1^2}{S_{res}^2} \sim F\left(K_1 - 1, K_1 K_2 (n-1)\right) \text{ при } H_0^1,$$
$$F_2 = \frac{S_2^2}{S_{res}^2} \sim F\left(K_2 - 1, K_1 K_2 (n-1)\right) \text{ при } H_0^2,$$
$$F_{12} = \frac{S_{12}^2}{S_{res}^2} \sim F\left((K_1 - 1)(K_2 - 1), K_1 K_2 (n-1)\right) \text{ при } H_0^{12};$$

- $n = 1$:

$$F_1 = \frac{S_1^2}{S_{12}^2} \sim F\left(K_1 - 1, (K_1 - 1)(K_2 - 1)\right) \text{ при } H_0^1,$$
$$F_2 = \frac{S_2^2}{S_{12}^2} \sim F\left(K_2 - 1, (K_1 - 1)(K_2 - 1)\right) \text{ при } H_0^2.$$

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

Марихуана и скорость реакции

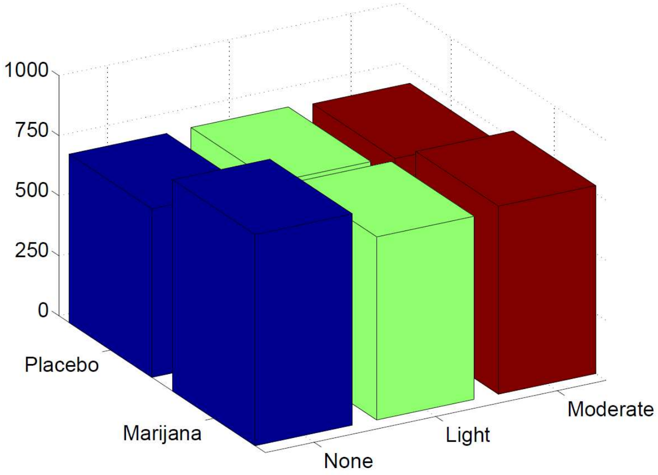
Изучалось воздействие марихуаны на скорость реакции. В качестве испытуемых были выбраны по 12 человек из каждой категории:

- 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.

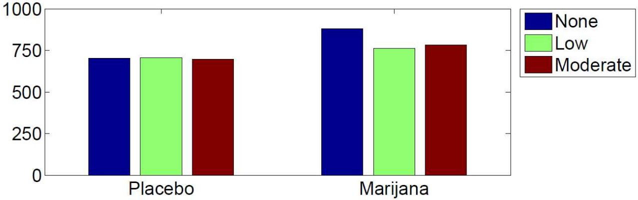
Марихуана и скорость реакции

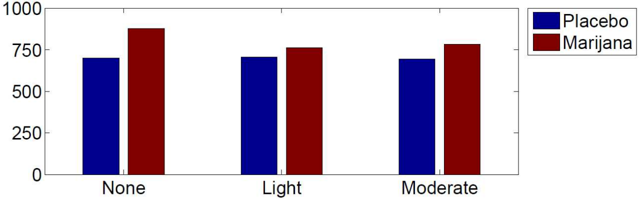
Плохой график:



Марихуана и скорость реакции

Хорошие графики:





Марихуана 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.

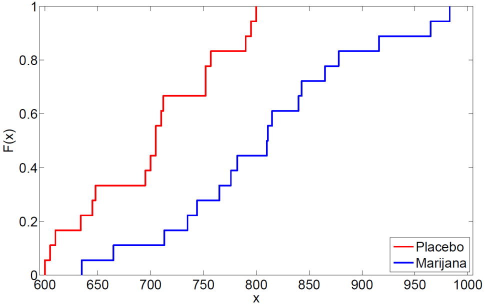
Source	SS	df	MS	F	Prob>F
Group	103041	1	103041	17.58	0.0002
Past use	23634.5	2	11817.2	2.02	0.1508
Interaction	23642.2	2	11821.1	2.02	0.1507
Error	175796.3	30	5859.9		
Total	326114	35			

Марихуана и скорость реакции

Вывод: гипотеза о том, что предыдущий опыт употребления не влияет на скорость реакции, не отклоняется $\Rightarrow$ 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: $p = 0.00018, \quad ci = (61.3, \infty)$;



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).

CBI чернобрюхой дрозофилы

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

CBI чернобрюхой дрозофилы

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			

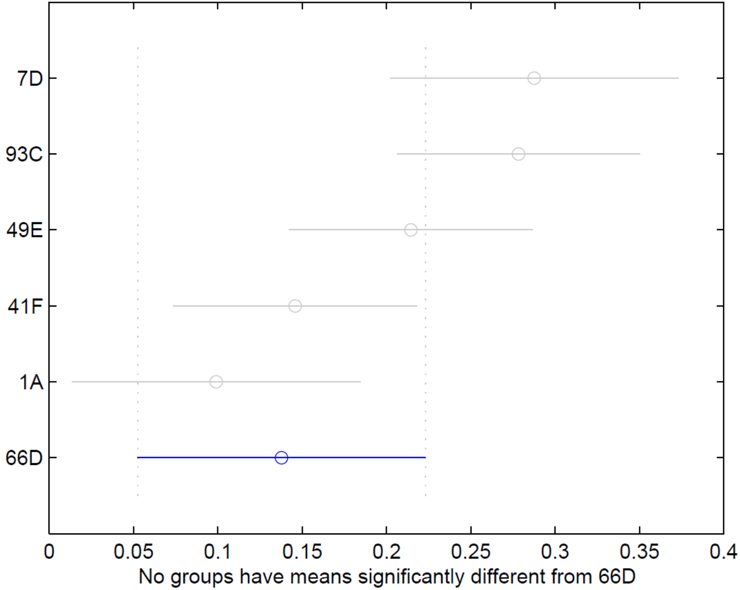
Есть различия между регионами, нет различий между хромосомами.

CBI чернобрюхой дрозофилы

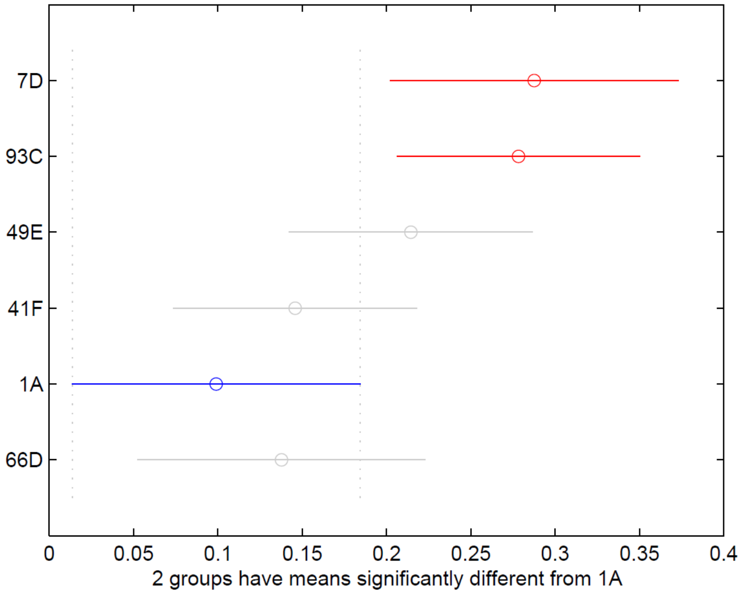
Для уточнения различий применим метод HSD:

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

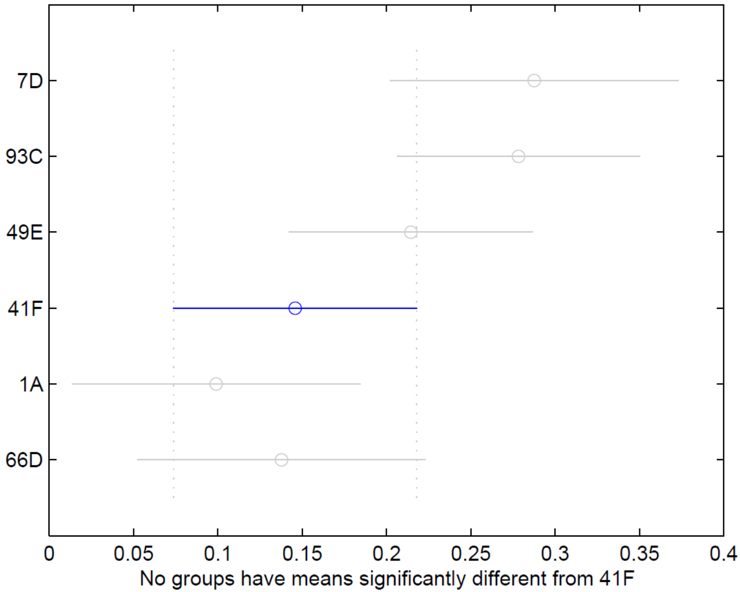
CBI чернобрюхой дрозофилы



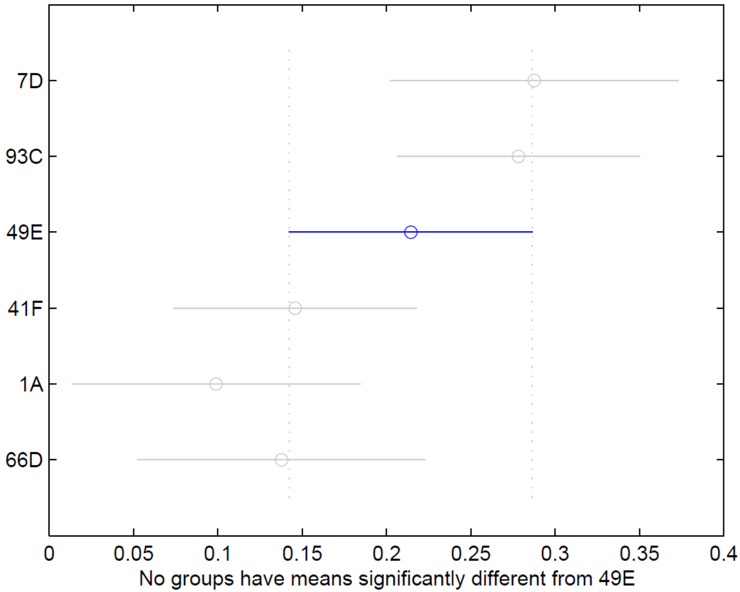
CBI чернобрюхой дрозофилы



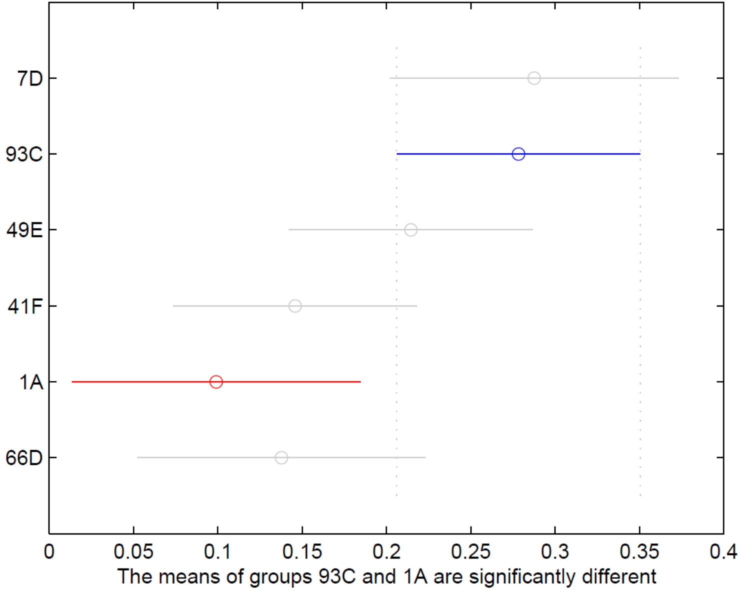
CBI чернобрюхой дрозофилы



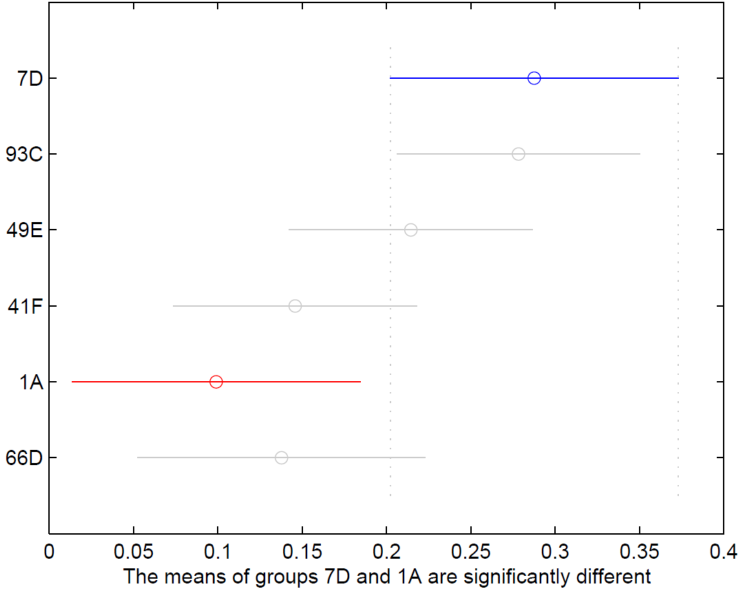
CBI чернобрюхой дрозофилы



CBI чернобрюхой дрозофилы



CBI чернобрюхой 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.

Литература

- разновидности ANOVA — Tabachnick, 3.2;
- применение в R — Chang, `http://www.cookbook-r.com/Statistical_analysis/ANOVA/`;
- 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`;
- 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;
- непараметрический двухфакторных дисперсионный анализ — Wilcox, 7.9.

Tabachnick B.G., Fidell L.S. *Using Multivariate Statistics*. — Boston: Pearson Education, 2012.
Chang W. *Cookbook for R*. — `http://www.cookbook-r.com/`.
Лагутин М.Б. *Наглядная математическая статистика*. — Москва: Бином, 2007.
Кобзарь А.И. *Прикладная математическая статистика*. — М.: Физматлит, 2006.
Wilcox R.R. *Introduction to Robust Estimation and Hypothesis Testing*. — Academic Press, 2012.