Дисперсионный анализ

Центр биоэлектрических интерфейсов, 20 февраля 2019 г.

Денис Деркач, Влад Белавин

Оглавление

Двух и многофакторный дисперсионные анализы

Напоминание

Ранее мы рассматривали однофакторный дисперсионный анализ (ANOVA):

- необходима для оценки зависимости среднего от одной переменной;
- сравнивает дисперсию внутри групп и между группами;
- использует F распределение.

Двухфакторный анализ

Факторы бывают двух типов:

- случайный (random);
- фиксированный (fixed).

В зависимости от типов факторов дисперсионный анализ может быть fixed-effects, random-effects или mixed-effects.
NB: разделение на fixed и random имеет несколько разных школ.

Two-ways ANOVA

В случае анализа многих факторов, мы не только должны следить за эффектом от каждого из них, но и за взаимодействие между этими факторами. Поэтому тестируем сразу несколько H_0:

- Среднее не зависит от фиксированных факторов.
- Дисперсия не зависит от случайных факторов.
- Фактор 1 и фактор 2 не взаимодействуют.

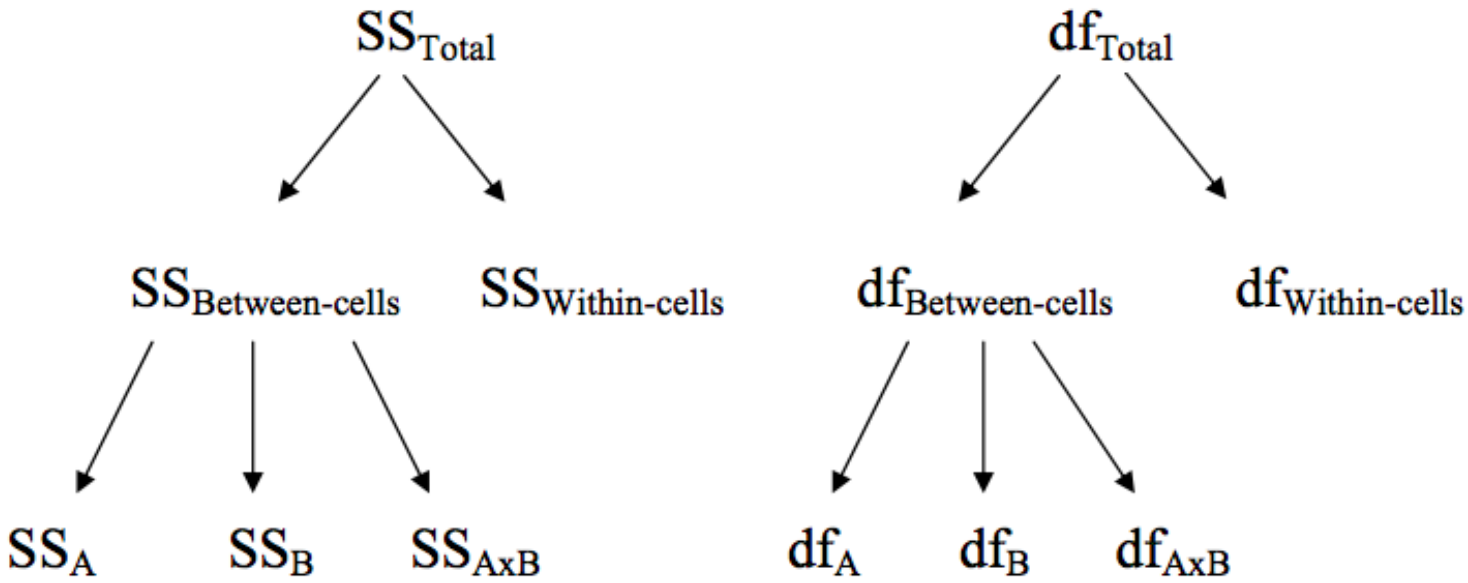
Взаимодействие факторов

В случае отсутствия взаимодействия, каждый фактор отвечает за отклонение от общего среднего в каждой ячейке на строго определённое число, для каждого уровня фактора своё.
Фактически, тестируем:

$$\mu_{ij} - \mu_i - \mu_j - \mu = 0.$$

NB: Взаимодействие между факторами это не корреляция между ними!

Графическое представление



Суммы квадратов

Column	Description	Column Sum
$(Y - \overline{Y}_{\bullet})^2$	Squared deviation of each score from the grand mean	SS_{Total}
$(\overline{AB}_{ij} - \overline{Y}_{\bullet})^2$	Squared deviation of each score's cell mean from the grand mean	$SS_{Between\text{-}cells}$
$(Y - \overline{AB}_{ij})^2$	Squared deviation of each score from its cell mean	$SS_{Within\text{-}cells}$, or SS_{Error}
$(\overline{A}_i - \overline{Y}_{\bullet})^2$	Squared deviation of each score's A-mean from the grand mean	SS_A
$(\overline{B}_j - \overline{Y}_{\bullet})^2$	Squared deviation of each score's B-mean from the grand mean	SS_B
$[(\overline{AB}_{ij} - \overline{Y}_{\bullet}) - (\overline{A}_i - \overline{Y}_{\bullet}) - (\overline{B}_j - \overline{Y}_{\bullet})]^2$	Square of: [(cell mean – grand mean) – (A-mean – grand mean) – (B-mean – grand mean)]	SS_{AB}

Степени свободы

Recall that for the fixed effects model, the test statistics are:

$$ts_{\mathbf{A}} = \frac{\mathbf{SS_A}/(I-1)}{\mathbf{SS_{err}}/(n_{\cdot\cdot\cdot} - IJ)} = \mathbf{MS_A}/\mathbf{MS_{err}}$$

$$ts_{\mathbf{B}} = \frac{\mathbf{SS_B}/(J-1)}{\mathbf{SS_{err}}/(n_{\cdot\cdot\cdot} - IJ)} = \mathbf{MS_B}/\mathbf{MS_{err}}$$

$$ts_{\mathbf{AB}} = \frac{\mathbf{SS_{AB}}/(I-1)(J-1)}{\mathbf{SS_{err}}/(n_{\cdot\cdot\cdot} - IJ)} = \mathbf{MS_{AB}}/\mathbf{MS_{err}}$$

4

In contrast, for the random effects model, the test statistics are

$$ts_{\mathbf{A}} = \frac{\mathbf{SS_A}/(I-1)}{\mathbf{SS_{AB}}/(I-1)(J-1)} = \mathbf{MS_A}/\mathbf{MS_{AB}}$$

$$ts_{\mathbf{B}} = \frac{\mathbf{SS_B}/(J-1)}{\mathbf{SS_{AB}}/(I-1)(J-1)} = \mathbf{MS_B}/\mathbf{MS_{AB}}$$

$$ts_{\mathbf{AB}} = \frac{\mathbf{SS_{AB}}/(I-1)(J-1)}{\mathbf{SS_{err}}/(n_{\cdot\cdot\cdot} - IJ)} = \mathbf{MS_{AB}}/\mathbf{MS_{err}}$$

Степени свободы

For the mixed effects model, assume that factor A is fixed and factor B is random. Then the test statistics are

$$ts_{\mathbf{A}} = \frac{\mathbf{SS_A}/(I-1)}{\mathbf{SS_{AB}}/(I-1)(J-1)} = \mathbf{MS_A}/\mathbf{MS_{AB}}$$

$$ts_{\mathbf{B}} = \frac{\mathbf{SS_B}/(J-1)}{\mathbf{SS_{err}}/(n_{\cdots}-IJ} = \mathbf{MS_B}/\mathbf{MS_{err}}$$

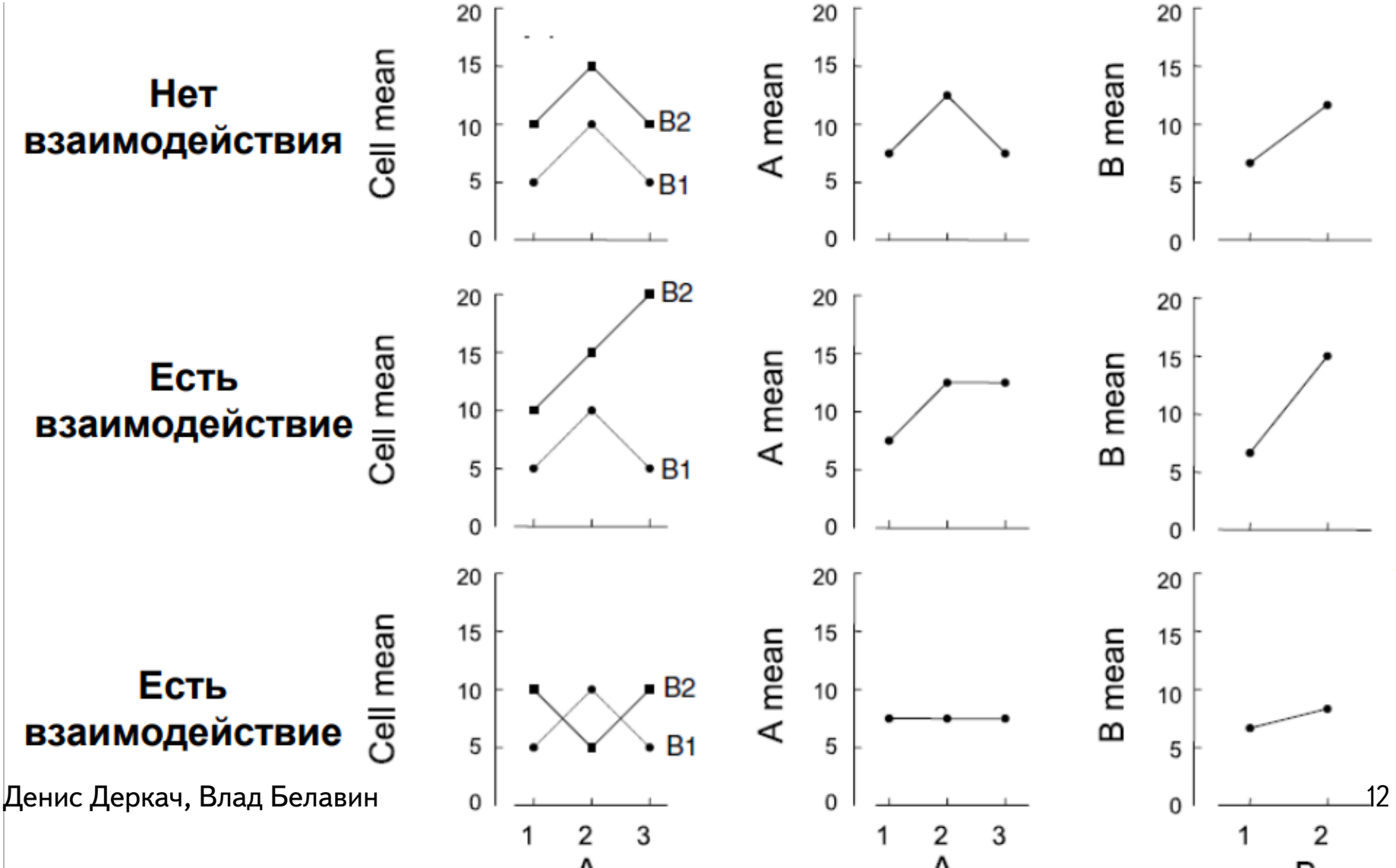
$$ts_{\mathbf{AB}} = \frac{\mathbf{SS_{AB}}/(I-1)(J-1)}{\mathbf{SS_{err}}/(n_{\cdots}-IJ)} = \mathbf{MS_{AB}}/\mathbf{MS_{err}}$$

The interaction test is always the same. For random effect tests, use the mean squared error as the denominator; for fixed effect tests, use the mean square for interaction as the denominator.

The situation is complex when there are more than two factors, or when factors have both random and fixed levels. Satterthwaite's approximation is needed to create synthetic mean squared terms to make the tests.

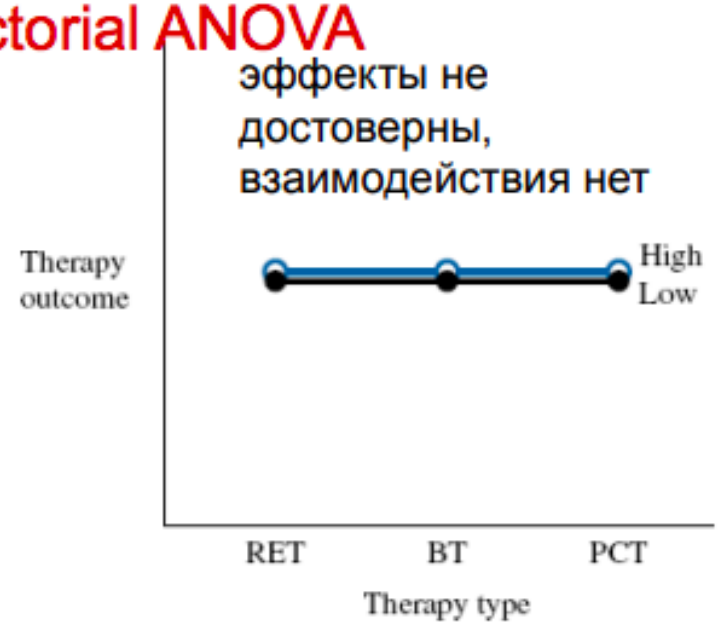
5

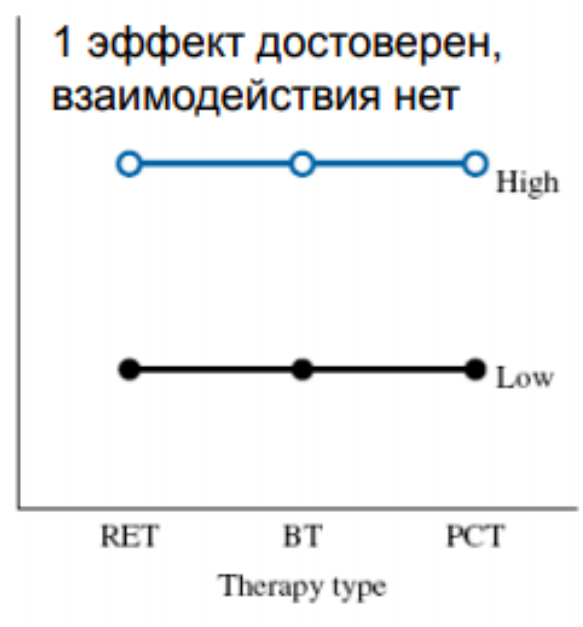
Графическое представление взаимодействия

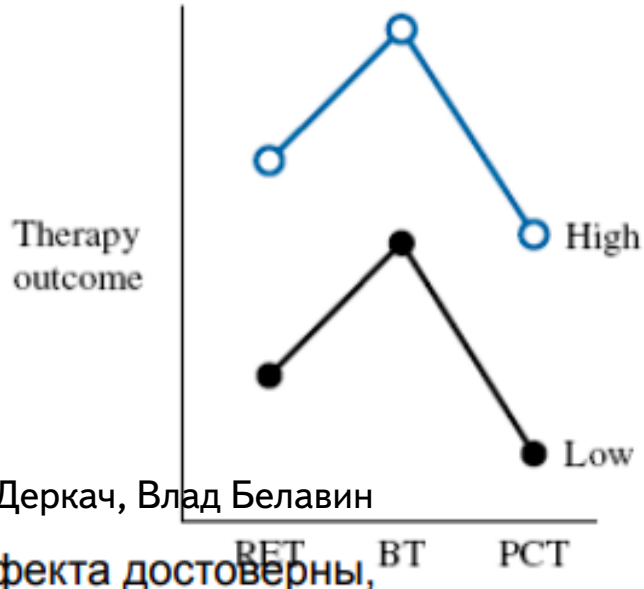


Достоверность эффектов

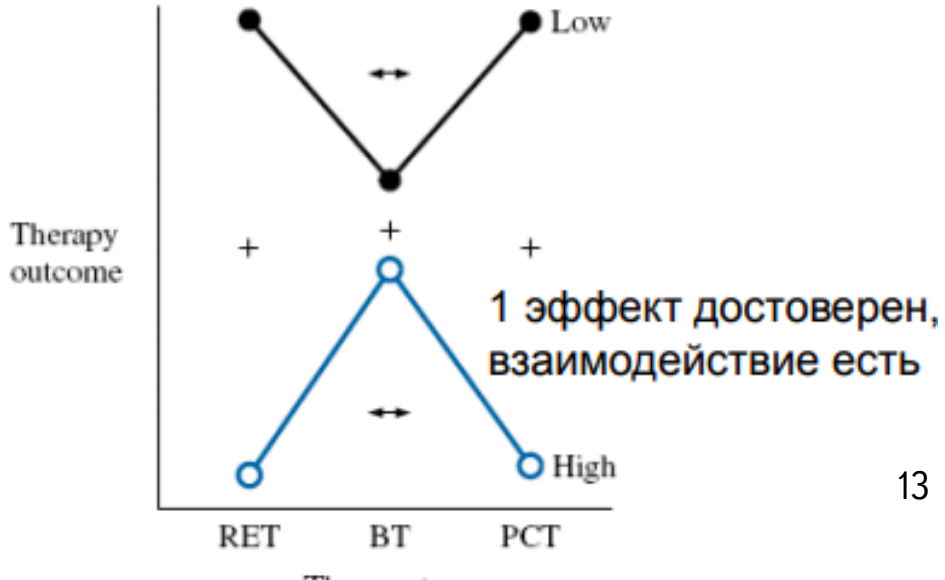
Factorial ANOVA







2 эффекта достоверны,



Апостериорные тесты

- Не используются для random factors;
- Если взаимодействие между факторами достоверно, бессмысленно проводить пост хок тесты для каждого из факторов по отдельности, нужно сравнивать между собой ячейки.

Multiway ANOVA

Если факторов 2 — Two-way ANOVA; если много, а зависимая переменная одна - Multiway ANOVA В этом случае становится много гипотез о взаимодействии факторов (для 3-х факторов 4 гипотезы о взаимодействии). Не рекомендуется исследовать действие более 4-х факторов, так как затрудняется интерпретация результатов.

Анализ ковариаций

Мотивация

До этого мы использовали только категориальные дискретные переменные, а что произойдет, если мы будем использовать также непрерывные переменные?

Пример: Пусть у нас имеется 3 метода обучения арифметики и группа студентов. Группа разбивается случайным образом на 3 подгруппы для обучения одним из методов. В конце курса обучения студенты проходят общий тест, по результатам которого выставляются оценки. Также для каждого студента имеется одна или несколько характеристик (количественных) их общей образованности.

Требуется проверить гипотезу об одинаковой эффективности методик обучения.

Ковариционный анализ - ANCOVA

- Объединение регрессионного и дисперсионного анализов.
- Непрерывная переменная называется ковариата (covariate).
- Проверяемые гипотезы подобны ANOVA.

Таблица ковариационного анализа

Source	Sum of Squares	Degrees of Freedom	Variance Estimate (Mean Square)	*F* Ratio
Covariate	SS_{Cov}	1	MS_{Cov}	$\frac{MS_{Cov}}{MS'_W}$
Between	SS'_B	$K-1$	$MS'_B = \frac{SS'_B}{K-1}$	$\frac{MS'_B}{MS'_W}$
Within	SS'_W	$N-K-1$	$MS'_W = \frac{SS'_W}{N-K-1}$	
Total	SS'_T	$N-1$		

Отличие от ANOVA в количестве степеней свободы.

Предположения для ANCOVA

- Для каждой независимой переменной связь между зависимой переменной (y) и ковариатом (x) линейна.
- Все линейные связи из пункта выше представимы в виде параллельных линейных зависимостей.
- Ковариат и фактор не зависят друг от друга.

MANOVA

Мотивация

Иногда мы сталкиваемся с необходимостью проанализировать несколько зависимых переменных. При этом мы не можем просто применить несколько раз ANCOVA:

- повысится вероятность ошибки 1-го рода;
- таким образом, мы забудем о возможных корреляциях;
- зачастую эффект виден только в многомерном пространстве.

MANOVA гипотезы

Для двух факторов:

- H_0: $\mu_{11} = \mu_{12} = \ldots = \mu_{1k}$ и $\mu_{21} = \mu_{22} = \ldots = \mu_{2k}$, где μ_{ij} обозначает среднее по переменной i в группе j;
- H_A: одно из равенств не соблюдается.

Предположения MANOVA

- многомерная нормальность (хорошо переносит асимметрии, но плохо эксцесс — падает мощность);
- равенство дисперсий и ковариаций;
- примерное равный размер групп;
- мощность падает с ростом числа переменных.

Матрица суммы квадратов и кросс-произведений

Source of variation	Matrix of sum of squares and cross-products (SSP)	Degrees of freedom (d.f.)
Treatment	$B = \sum_\ell n_\ell(\bar{x}_\ell - \bar{x})(\bar{x}_\ell - \bar{x})'$	$g - 1$
Residual	$W = \sum_\ell \sum_j (x_{\ell j} - \bar{x}_\ell)(x_{\ell j} - \bar{x}_\ell)'$	$n - g$
Total corrected	$B + W = \sum_\ell \sum_j (x_{\ell j} - \bar{x})(x_{\ell j} - \bar{x})'$	$n - 1$

Используемые критерии

- Критерий Вилкса (Wilk's lambda)

$$\Lambda = \frac{|W|}{|B + W|}$$

  чем она меньше, тем больше межгрупповые различия;

- Критерий Хотеллинга (Hotelling trace):

$$nT_0^2 = Tr(BW^{-1})$$

  — чем больше, тем больше различия групп;

Используемые критерии

- Критерий Пиллая (Pillai's trace):

$$V = Tr(B(B + W)^{-1}).$$

- Критерий Роя (Roy's maximum root): тестирование наибольшего собственного значения матрицы BW^{-1}.

Выбор критерия

Все эти критерии преобразуют в величину, аппроксимирующуюся F-распределением (и их сравнивают с критическим F-значением).

- Критерий Роя хуже всего приближается F распределением.
- Критерий Пиллая более устойчив к ненормальным данным и наиболее мощен в случае коррелированных данных.
- Критерий Роя наиболее мощен для некоррелированных данных.

Дисперсионный анализ с повторными измерениями

Мотивация

Часто бывает, что данные необходимо собирать, повторяя одни и те же измерения с одними и теми же точками сбора экспериментальных данных, например, в случаях когда

- количество исследуемых ограничено;
- количество времени ограничено;
- постановка эксперимента предполагает изучение зависимости от времени.

Каждый набор связанных измерений называется блок (block). Дизайн эксперимента при этом называется randomised block design.

Источники изменчивости

- между измерениями - уровнями фактора;
- между особями или блоками (дисперсия средних значений блоков);
- ”ошибка” (внутри «исправленных» измерений) — error = residual — после исключения различий между блоками.

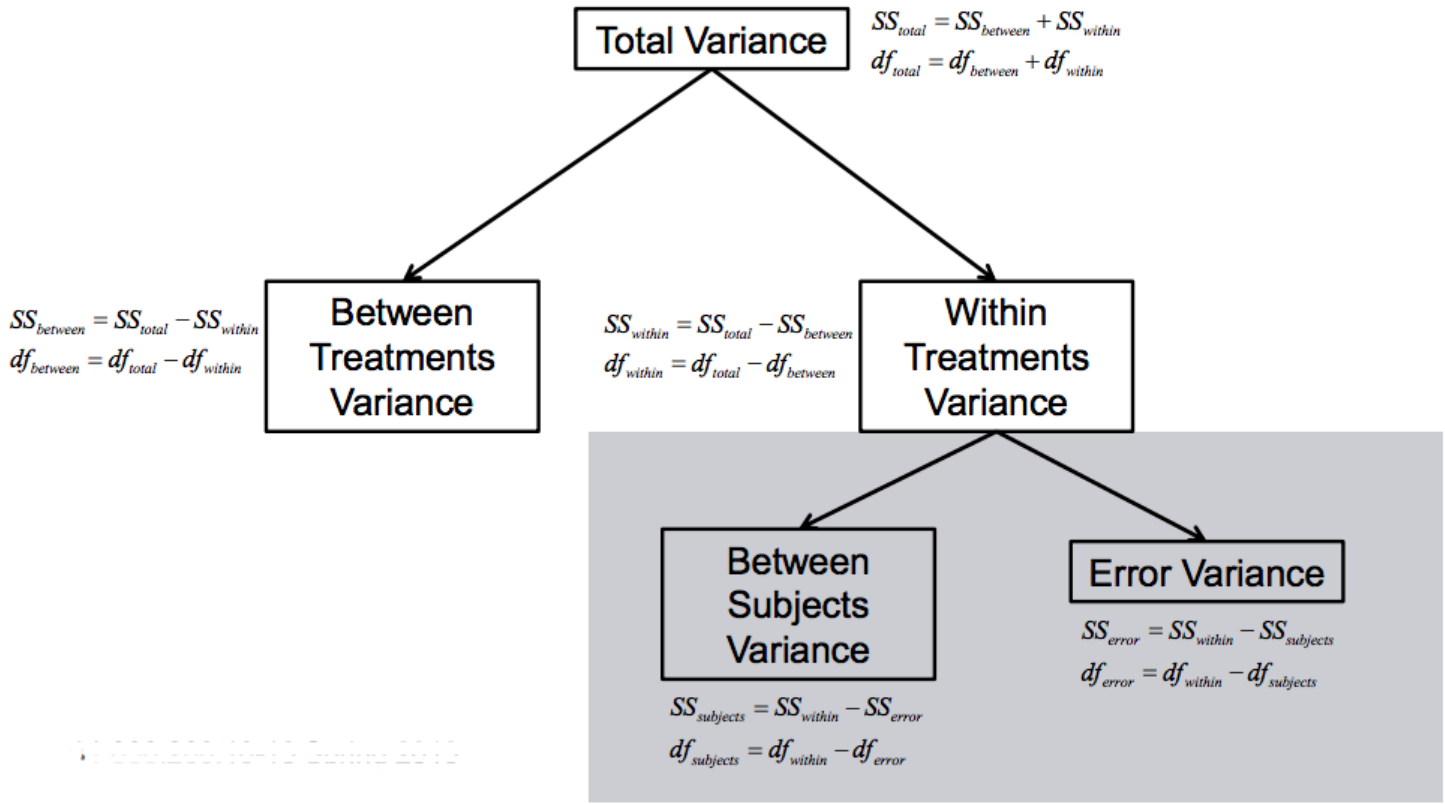
Гипотезы ANOVArm

- $H_0 : \mu_1 = \ldots = \mu_k$.
- H_A : нулевая гипотеза не верна.

При этом F-статистика составляется по-другому:

$$F = \frac{\text{оценка дисперсии между измерениями}}{\text{”ошибка” внутри исправленных измерений}}$$

Разделение ошибок



$$F = \frac{MS_B}{MS_E} = \frac{SS_B/df_B}{SS_E/df_E}$$

Степени свободы

- $df_T = N - 1$ - общее количество степень свободы.
- $df_B = k - 1$ - между группами.
- $df_s = n - 1$ - между объектами в группе.
- $df_E = df_T - df_B - df_s = N - k - n + 1$ - в группе, из-за повторяющихся экспериментов.

Свойства ANOVArm

- Мощность дисперсионного анализа для повторных измерений выше, чем обыкновенного дисперсионного анализа (в случае связанных выборок).
- В случае необходимости добавить фактор, проводят split-plot ANOVA.

Предположения

- нормальное распределение внутри измерений;
- гомоскедастичность (то есть гомогенность дисперсий между измерениями);
- отсутствие пропусков в данных;
- сферичность (spherisity) - дисперсии различий между всеми возможными парами внутрисубъектных условий (то есть уровней независимой переменной) равны.

Сферичность

Пример: реакция пациентов на лекарство.

Patient	Tx A	Tx B	Tx C	Tx A – Tx B	Tx A – Tx C	Tx B – Tx C
1	30	27	20	3	10	7
2	35	30	28	5	7	2
3	25	30	20	–5	5	10
4	15	15	12	0	3	3
5	9	12	7	–3	2	5
			Variance:	17	10.3	10.3

Очевидно, что в случае A-B разницы, дисперсия получается очень большая. Для подтверждения сферичности используют тест Мошли (Mauchly's test), который сравнинвает дисперсии с χ^2.
NB: крайне ненадёжный тест.

Неполные (гнездовые) анализы дисперсии

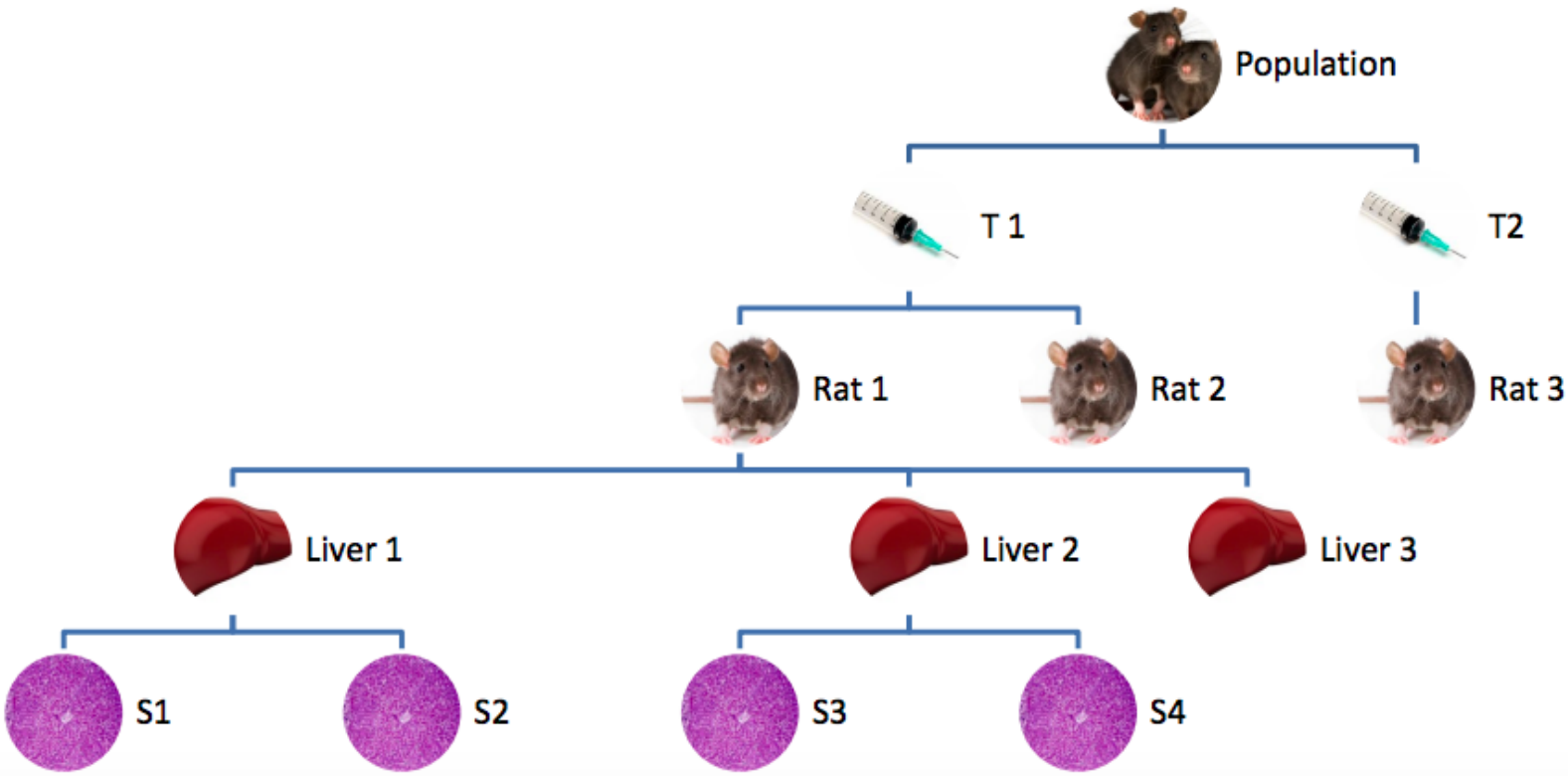
Мотивация

Во многих случаях можно пренебречь эффектом взаимодействия. Это происходит или когда известно, что в популяции эффект взаимодействия отсутствует, или когда осуществление полного факторного плана невозможно.
В случае полностью заполненной таблицы измерений, мы говорим о Crossed design ANOVA, иначе о nested ANOVA.

Пример

- Influence of treatment on rat liver Glycogen content

3 treatments (T1, T2, T3)
2 rats/treatment
3 liver sections
2 preparations of each liver section



Гипотезы

У нас есть два типа факторов: fixed (уровня A) и random (уровня B nested in A).
Нулевые гипотезы:

- H_0: $\mu_1 = \ldots = \mu_k$ (для уровня A).
- H_0: $\sigma_B = 0$ (межгрупповая дисперсия равна 0).

Подсчёт статистик

$$F = \frac{MS_{\text{between groups}}}{MS_{\text{subgroups witin groups}}}$$

Проверка действия основного фактора (если nested фактор - random)

$$F = \frac{MS_{\text{subgroups within groups}}}{MS_{\text{error within subgroups}}}$$

Проверка действия nested фактора

Так обозначают nested фактор

Source	SS	df	MS
A	$nq\sum_{i=1}^{p}(\bar{y}_i - \bar{y})^2$	$p-1$	$\frac{SS_A}{p-1}$
B(A)	$n\sum_{i=1}^{p}\sum_{j=1}^{q}(\bar{y}_{j(i)} - \bar{y}_i)^2$	$p(q-1)$	$\frac{SS_{B(A)}}{p(q-1)}$
Residual	$\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{n}(y_{ijk} - \bar{y}_{j(i)})^2$	$pq(n-1)$	$\frac{SS_{Residual}}{pq(n-1)}$
Total	$\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{n}(y_{ijk} - \bar{y})^2$	$pqn-1$	

Комментарии

- апостериорные тесты только для уровня A;
- можно игнорировать уровень B и провести однофакторный анализ, но это может вылиться в неправильную интерпретацию;
- обычно фактор B случаен, в случае фиксированного используют split plots.

Предположения

- нормальное распределение внутри измерений;
- гомоскедастичность (то есть гомогенность дисперсий между измерениями);
- сбалансированный дизайн.

Анализ Краскела — Уоллиса

Мотивация

В некоторых случаях невозможно добиться данных с нормальным распределением. Потому необходимо использовать непараметрические тесты или линейные модели.

Анализ Краскела — Уоллиса

- непараметрический тест, который не требует нормальности данных (но чувствителен к разным дисперсиям);
- менее мощный, чем параметрические тесты, в случае однофакторного анализа мощность около 95%, в остальных случаях ниже 80%.
- для двух групп эквивалентен тесту Манна-Уитни.

NB: если данные гетероскедастичны, следует использовать дисперсионный анализ Уелча.

Алгоритм

1. выставить ранг r_i согласно значению Y;
2. подсчитать статистику:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} n_i \bar{r}_{i\cdot}^2 - \ 3(N+1),$$

   где N - общее количество семплов, k - количество групп, n_i - количество семплов в группе;
3. найти критическую область для H-статистики.

Наблюдения

- обычно $\sum_i r_i \approx \frac{N(N+1)}{2}$.
- для $k > 5$ можно сравнивать с χ^2_{k-1};
- в случае дополнительных связей необходимо применять коррекцию.
- H очень похожа на сравнение рангов между группами и внутри группы (то есть, анализ дисперсий на рангах);
- для нормальной работы теста рекомендуют $n_i > 5$;
- H_0: все группы происходят из одного распределения (Отличается от обычного анализа!).

Многофакторные непараметрические задачи

В случае бОльшего количества факторов можно использовать улучшения теста Краскела-Уоллеса:

- Scheirer-Ray-Hare Test
- Jonckheere-Terpstra Test

NB: относительная мощность непараметрических тестов быстро падает с ростом количества факторов.