Яндекс

Skoltech
Skolkovo Institute of Science and Technology

# Множественный дисперсионный анализ

ПМИ ФКН ВШЭ, 9 ноября 2019 г.

Денис Деркач, Алексей Артёмов

ФКН ВШЭ

# Оглавление

Двух и многофакторный дисперсионные анализы

Случайные эффекты

Анализ ковариаций

MANOVA

Дисперсионный анализ с повторными измерениями

Неполные (гнездовые) анализы дисперсии

# Двух и многофакторный дисперсионные анализы

# Напоминание

Ранее мы рассматривали однофакторный дисперсионный анализ (ANOVA):

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

# Двухфакторный анализ

Двухфакторный анализ проводится в случае наличия зависимости от двух категориальных переменных.
Фактически, мы рассматриваем таблицу:

- Наблюдения $Y_{ijk}, 1 \leq i \leq r, 1 \leq j \leq m, 1 \leq k \leq n_{ij}$, для $r$ делений (уровней) по первому фактору, $j$ делений (уровней) по второму фактору, $n_{ij}$ - количество наблюдений в ячейке $i, j$.
- $Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}, \varepsilon_{ijk} \sim \mathcal{N}(0, 1)$.
- связи:

$$\sum_i \alpha_i = 0, \quad \sum_j \beta_j = 0, \quad \sum_j (\alpha\beta)_{ij} = 0, 1 \leq i \leq r, \quad \sum_i (\alpha\beta)_{ij} = 0, 1 \leq j \leq m.$$

# Независимые переменные

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.

# Fixed Effect Two-ways ANOVA

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

- Среднее отклика не зависит от фактора 1.
- Среднее отклика не зависит от фактора 2.
- Фактор 1 и фактор 2 не взаимодействуют.

# Модель

- Средние:

$$\hat{Y}_{ijk} = \bar{Y}_{ij\cdot} = \frac{1}{n_{ij}} \sum_{k=1}^{n_ij} Y_{ijk}$$

- Дисперсии для комбинации:

$$\mathbb{V}ar(\sum_{i=1}^{r}\sum_{j=1}^{m} a_{ij}\bar{Y}_{ij\cdot}) = \sigma^2 \sum_{i=1}^{r}\sum_{j=1}^{m} \frac{a_{ij}^2}{n_{ij}}.$$

# 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:

$$(\alpha\beta)_{ij} - \alpha_i - \beta_j - \mu = 0.$$

NB: Взаимодействие между факторами это не корреляция между ними!

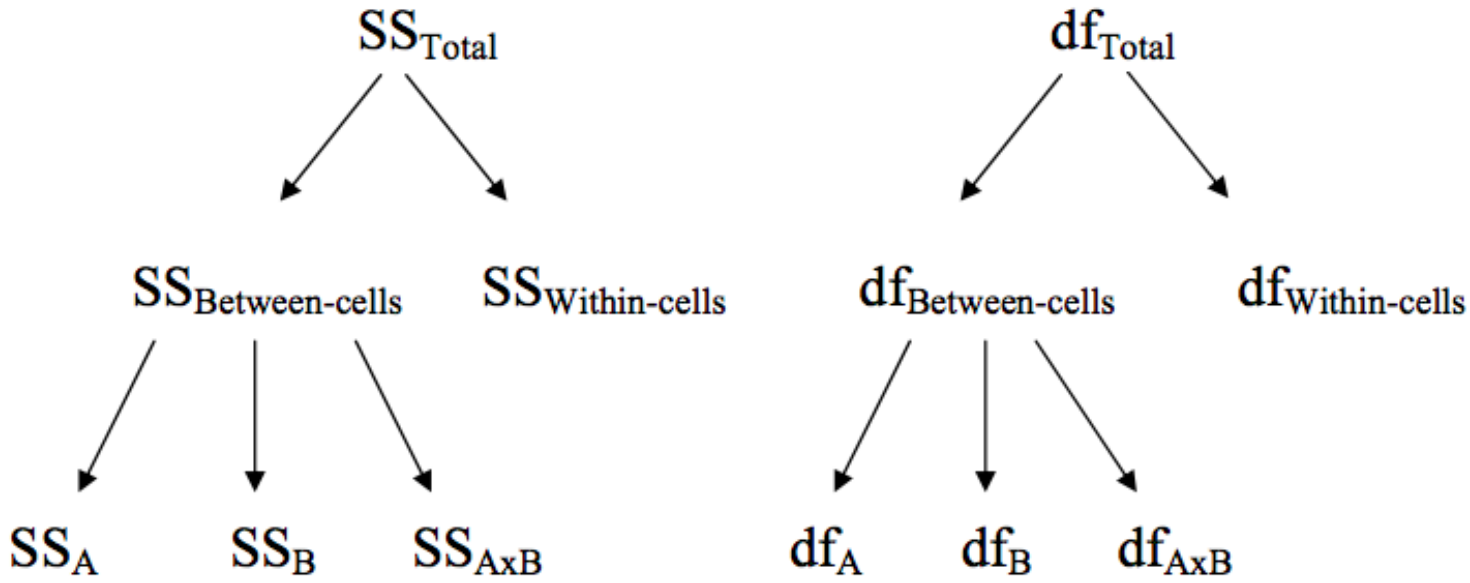
# Суммы квадратов

| Эффект | Сумма |
|---|---|
| A | $\mathrm{SSA} = nm \sum_{i=1}^{r} (\bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot\cdot\cdot})^2$ |
| B | $\mathrm{SSB} = nr \sum_{j=1}^{m} (\bar{Y}_{\cdot j\cdot} - \bar{Y}_{\cdot\cdot\cdot})^2$ |
| AB | $\mathrm{SSAB} = n \sum_{j=1}^{m} \sum_{i=1}^{r} (\bar{Y}_{ij\cdot} - \bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot j\cdot} + \bar{Y}_{\cdot\cdot\cdot})^2$ |
| Error | $\mathrm{SSE} = \sum_{k=1}^{n} \sum_{j=1}^{m} \sum_{i=1}^{r} (Y_{ijk} - \bar{Y}_{ij\cdot})^2$ |

Считая $n_{ij} = n$, иначе можно очевидным образом добавить эту информацию.

# Графическое представление

Очевиден также график суммы величин:



# Two-ways ANOVA таблица

| SS | df | E(MS) |
|---|---|---|
| SSA | $r-1$ | $\sigma^2 + nm\frac{\sum_{i=1}^{r}\alpha_i^2}{r-1}$ |
| SSB | $m-1$ | $\sigma^2 + nr\frac{\sum_{j=1}^{m}\beta_j^2}{m-1}$ |
| SSAB | $(r-1)(m-1)$ | $\sigma^2 + n\frac{\sum_{i=1}^{r}\sum_{j=1}^{m}(\alpha\beta)_{ij}^2}{(r-1)(m-1)}$ |
| SSE | $mr(n-1)$ | $\sigma^2$ |

- $H_0 : (\alpha\beta)_{ij} = 0, \forall i, j$, то есть SSAB=SSE=$\sigma^2$, таким образом:

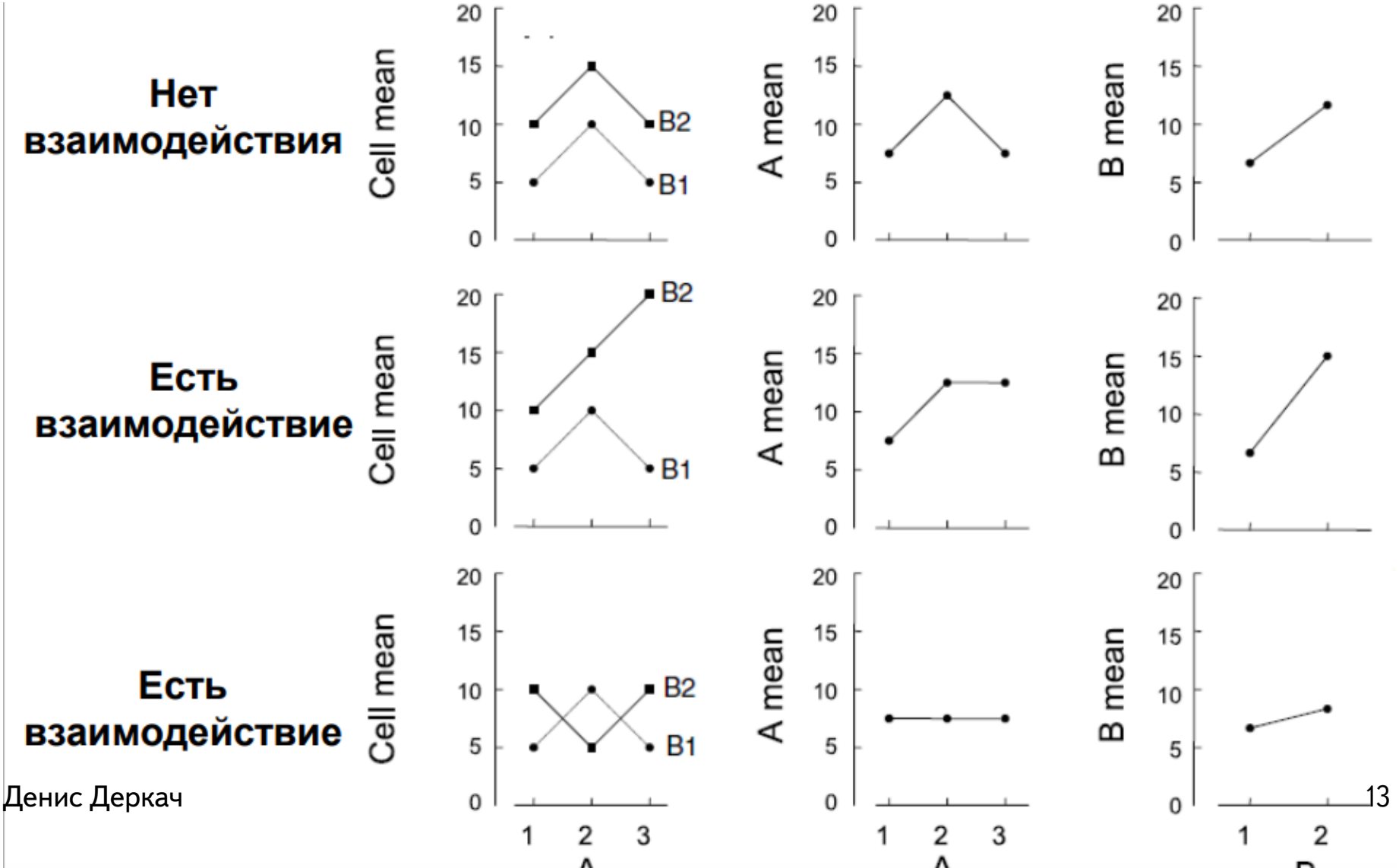
$$\frac{\text{MSAB}}{\text{MSE}} = \frac{\text{SSAB/dfAB}}{\text{SSE/dfE}} \sim F((m-1)(r-1), (n-1)mr);$$

- $H_0 : \alpha_i = 0, \forall i$, то есть SSAB=SSE=$\sigma^2$, таким образом:

$$\frac{\text{MSA}}{\text{MSE}} = \frac{\text{SSA/dfA}}{\text{SSE/dfE}} \sim F(r-1, (n-1)mr);$$

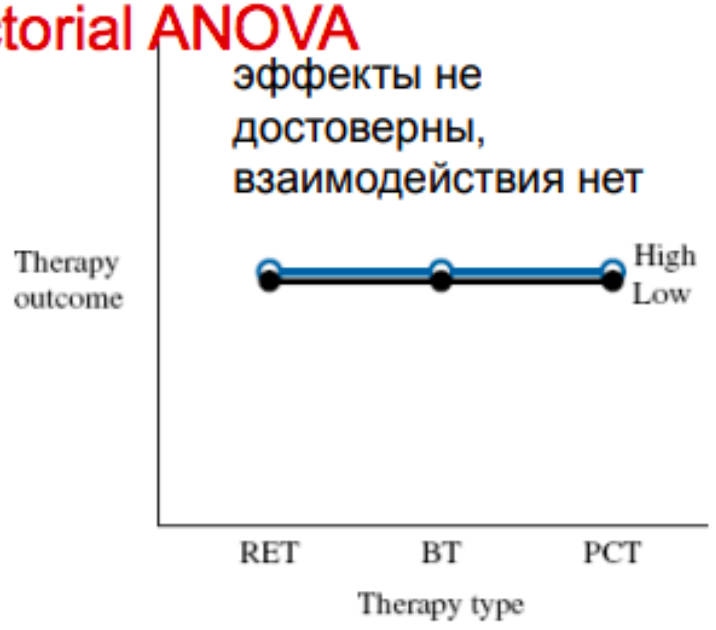
- Аналог для $\beta$: $\sim F(m-1, (n-1)mr)$.

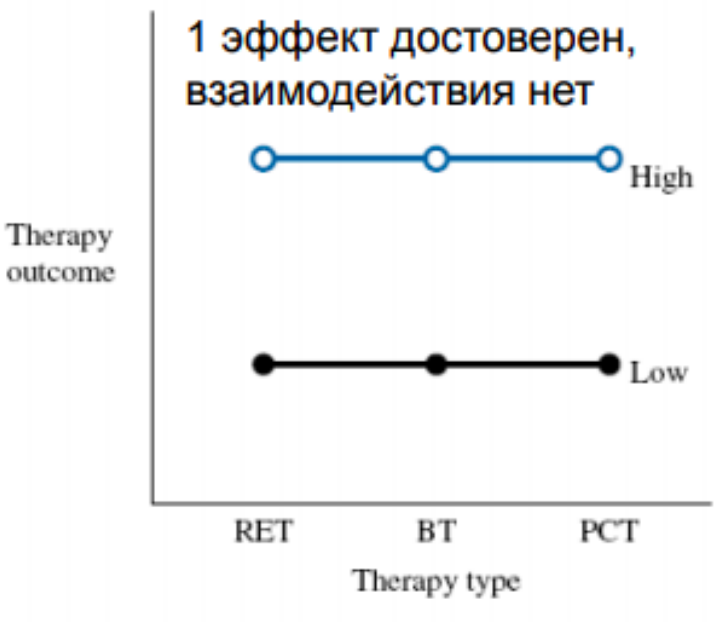
# Графическое представление взаимодействия

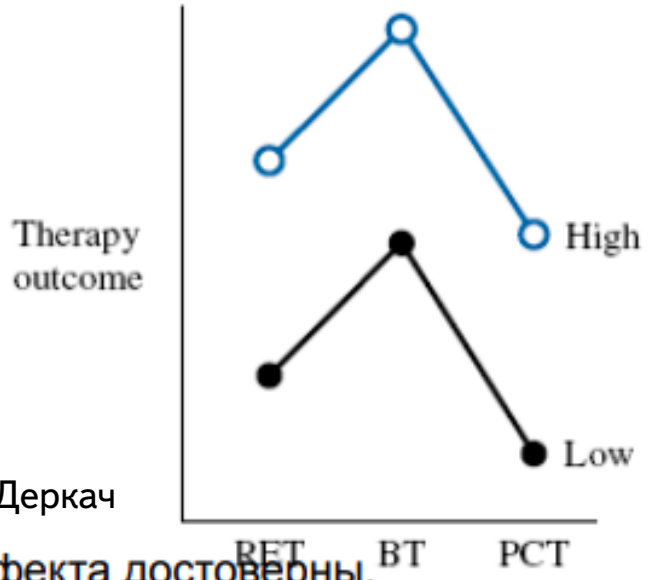


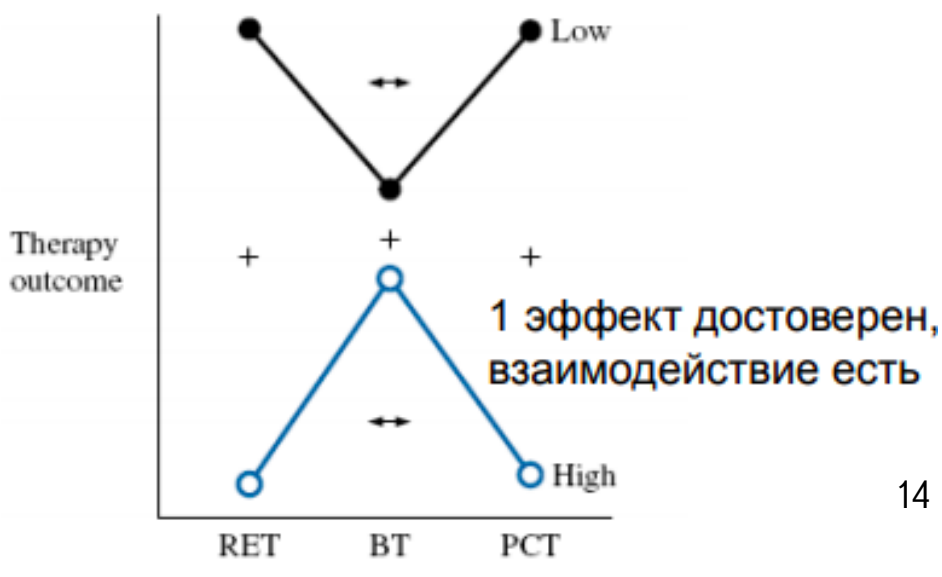
# Достоверность эффектов

Factorial ANOVA



# 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.

# Случайные эффекты

# Случайные эффекты

- Термин случайные эффекты в контексте дисперсионного анализа используется для обозначения факторов плана ANOVA, уровни которых не фиксируются заранее (факторы с фиксированными заранее уровнями называются фиксированными эффектами), а получаются из выборки в ходе эксперимента.
- Пример: исследование разных медицинских центров.
- Пример: исследование в кампусах разных университетов.

# Однофакторная ANOVA со случайными эффектами

Если у нас есть $n$ измерений $r$ источников.

- $Y_{ij} \sim \mu + \alpha_i + \varepsilon_{ij}, 1 \leq i \leq r, 1 \leq j \leq n$
- $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2), 1 \leq i \leq r, 1 \leq j \leq n$
- $\alpha_i \sim \mathcal{N}(0, \sigma_\alpha^2), 1 \leq i \leq r$
- возможные вопросы:
  - среднее в популяции ($\mu$): доверительные интервалы, совпадает ли оно с 0;
  - вариативность в семплах ($\sigma_\alpha$): доверительные интервалы, совпадает ли с 0.

  NB: в модели со случайными эффектами $\mathrm{Cov}(Y_{ij}, Y_{i',j'}) = \sigma_\alpha^2 \delta_{i,i'} + \sigma^2 \delta_{j,j'}$, то есть недиагональные элементы тоже не 0.

# Однофакторная ANOVA со случайными эффектами: таблица

| Источник | SS | df | E(MS) |
|---|---|---|---|
| фактор | $SSTR = \sum_{i=1}^{r} n(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$ | $r-1$ | $\sigma^2 + n\sigma_\alpha^2$ |
| ошибка | $SSE = \sum_{i=1}^{r}\sum_{j=1}^{n}(Y_{ij} - \bar{Y}_{i\cdot})^2$ | $r(n-1)$ | $\sigma^2$ |

Заметим, что результат не сильно отличается от фиксированного дизайна.

Для $H_0 : \sigma_\alpha^2 = 0$ мы можем записать:

$$\frac{MSTR}{MSE} \sim F(r-1, (n-1)r).$$

# Оценка $\mu$

Мы знаем, что $\mathbb{E}(\bar{Y}_{..}) = \mu$, также понятно, что

$$\mathbb{V}ar(\bar{Y}) = \frac{n\sigma_\alpha^2 + \sigma^2}{rn}$$

Таким образом:

$$\frac{\bar{Y}_{..} - \mu}{\sqrt{\frac{SSTR}{(r-1)rn}}} \sim t_{r-1}.$$

То есть, если мы используем $r$ источников, мы не сможем повысить точность семплируя из каждого источника, мы должны увеличивать $r$.

# Оценка $\sigma^2_\alpha$

$$\sigma^2_\alpha = \frac{\mathbb{E}(SSTR/(r-1)) - \mathbb{E}(SSE/((n-1)r))}{n}.$$

Таким образом:

$$S^2_\alpha = \frac{SSTR/(r-1) - SSE/((n-1)r)}{n}$$

В некоторых случаях, оценка будет отрицательной.

# Двухфакторный анализ со случайными эффектами

- $Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}, 1 \leq i \leq r, 1 \leq j \leq m,$ $1 \leq k \leq n_{ij}$;
- $\varepsilon_{ijk} \sim \mathcal{N}(0, 1)$;
- $\alpha_i \sim \mathcal{N}(0, \sigma^2_\alpha), 1 \leq i \leq r$
- $\beta_j \sim \mathcal{N}(0, \sigma^2_\beta), 1 \leq j \leq m$
- $\alpha\beta_{ij} \sim \mathcal{N}(0, \sigma^2_{\alpha\beta}), 1 \leq i \leq r, 1 \leq j \leq m$
- $\mathsf{Cov}(Y_{ijk}, Y_{i',j',k'}) =$ $\sigma^2_\alpha \delta_{i,i'} + \sigma^2_\beta \delta_{j,j'} + \sigma^2_{\alpha\beta} \delta_{i,i'} \delta_{j,j'} + \sigma^2 \delta_{i,i'} \delta_{j,j'} \delta_{k,k'}$

# Таблица

| SS | df | E(MS) |
|---|---|---|
| $SSA = nm\sum_{i=1}^{r}(\bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot\cdot\cdot})^2$ | $r-1$ | $\sigma^2 + nm\sigma^2_{\alpha} + n\sigma^2_{\alpha\beta}$ |
| $SSB = nr\sum_{j=1}^{m}(\bar{Y}_{\cdot j\cdot} - \bar{Y}_{\cdot\cdot\cdot})^2$ | $m-1$ | $\sigma^2 + nr\sigma^2_{\beta} + n\sigma^2_{\alpha\beta}$ |
| $SSAB = n\sum_{i=1}^{r}\sum_{j=1}^{m}(\bar{Y}_{ij\cdot} - \bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot j\cdot} + \bar{Y}_{\cdot\cdot\cdot})^2$ | $(m-1)(r-1)$ | $\sigma^2 + n\sigma^2_{\alpha\beta}$ |
| $SSE = \sum_{i=1}^{r}\sum_{j=1}^{m}\sum_{k=1}^{n}(\bar{Y}_{ijk} - \bar{Y}_{ij\cdot})^2$ | $rm(n-1)$ | $\sigma^2$ |

Аналогично предыдущим случаям:

- $H_0 : \sigma^2_{\alpha}$ с использованием SSA и SSAB.
- $H_0 : \sigma^2_{\alpha\beta}$ с использованием SSAB и SSE.

# Оценка $\sigma_\beta$

В случае двухфактороного анализа со случайными эффектами, естественный выбор оценки:

$$\hat{\sigma}^2_\beta = nr(MSB - MSAB)$$

Как оценить доверительный интервал? Напрямую его оценить трудно, так как $s^2 \sim \chi^2$. Идея - использовать подход Уэлча-Саттерсвейта (Welch—Satterthwaite).

# Подход Уэлча-Саттерсвейта

› Для $k$ назависимых $MS$:

$$\hat{L} \sim \sum_{i=1}^{k} c_i MS_i$$

›

$$\frac{df_T \hat{L}}{\mathbb{E}(\hat{L})} \sim \chi^2_{dfT},$$

где приближённая степень свободы:

$$df_T = \frac{(\sum_{i=1}^{k} c_i MS_i)^2}{\sum_{i=1}^{k} c_i^2 MS_i^2 / df_i}$$

› то есть, $1 - \alpha$ доверительный интервал для $\mathbb{E}(L)$:

$$\left[ \frac{df_T \hat{L}}{\chi^2_{dfT;1-\alpha/2}}; \frac{df_T \hat{L}}{\chi^2_{dfT;\alpha/2}} \right].$$

# Смешанные модели

Иногда требуется, чтобы один фактор был случайным, а один фиксированным. В этом случае строится смешанная модель, а проводимый анализ похож на соответсвтующие одномерные.

# 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 $\mu_{ij}$ 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 (Wilks' lambda)

$$\Lambda = \frac{\det(W)}{det(B+W)}$$

  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.

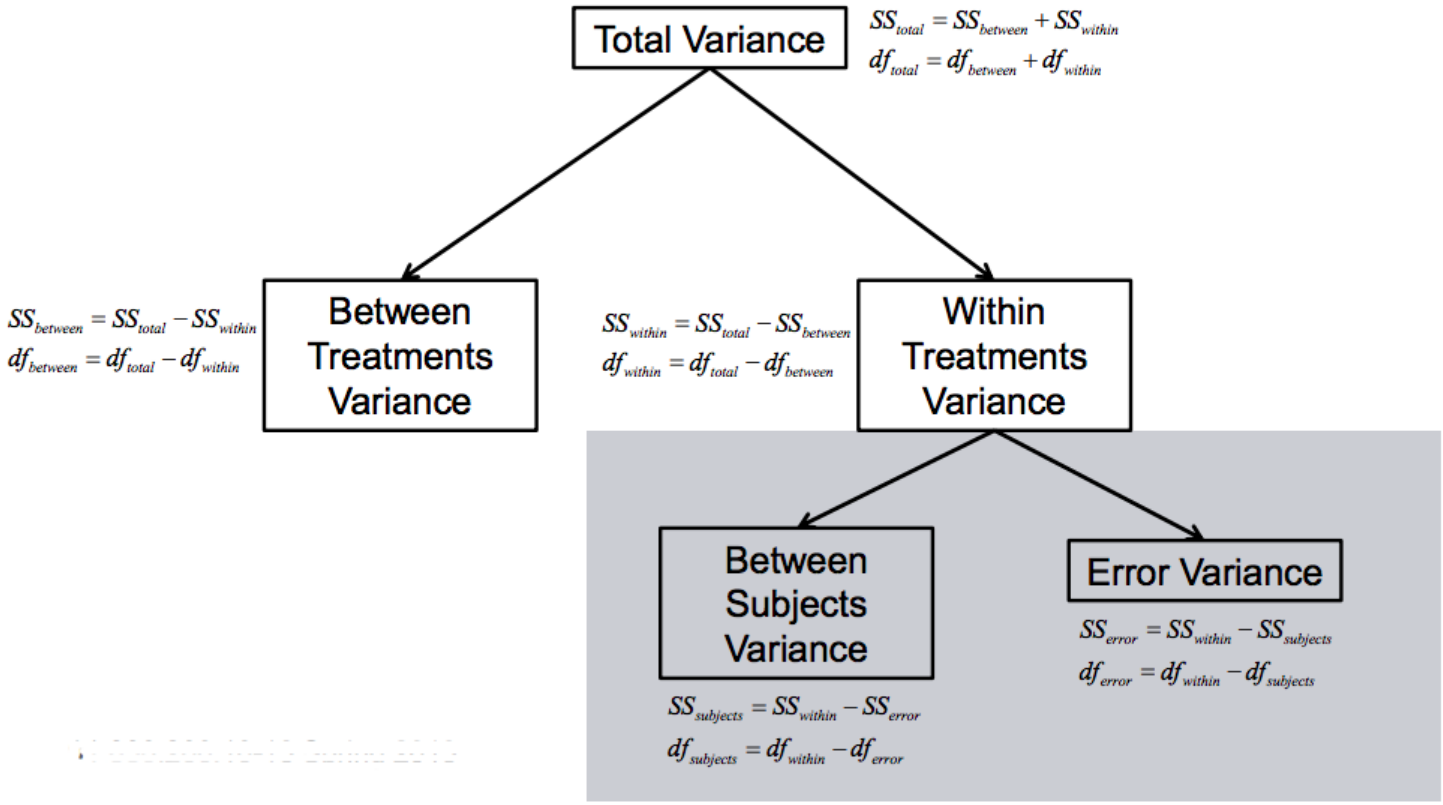
# Гипотезы ANOVArm

- $H_0 : \mu_1 = \ldots = \mu_k$.
- $H_A$ : 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: тест Мошли (Mauchly)

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 |

$$H_0 : \sigma^2_{\text{Tx A}-\text{Tx B}} = \sigma^2_{\text{Tx A}-\text{Tx C}} = \sigma^2_{\text{Tx B}-\text{Tx C}}$$

Сравнивает различные дисперсии с распределением $\chi^2$

NB: крайне ненадёжный тест.

# Тест Фридмана

В случае невыполненных предположений ANOVArm следует применять критерий Фридмана для $c$ выборок и $n$ измерений:

$$S = \frac{12}{nc(c+1)} \sum_{i=1}^{c} R_i^2 - 3n(c+1),$$

где $R_i = \sum_{j=1}^{n} r_{ij}$, для рангов в измерениях. При больших $n$ и $c$:

$$S_\alpha(n, c) \approx \chi^2_\alpha(c-1)$$

# Неполные (гнездовые) анализы дисперсии

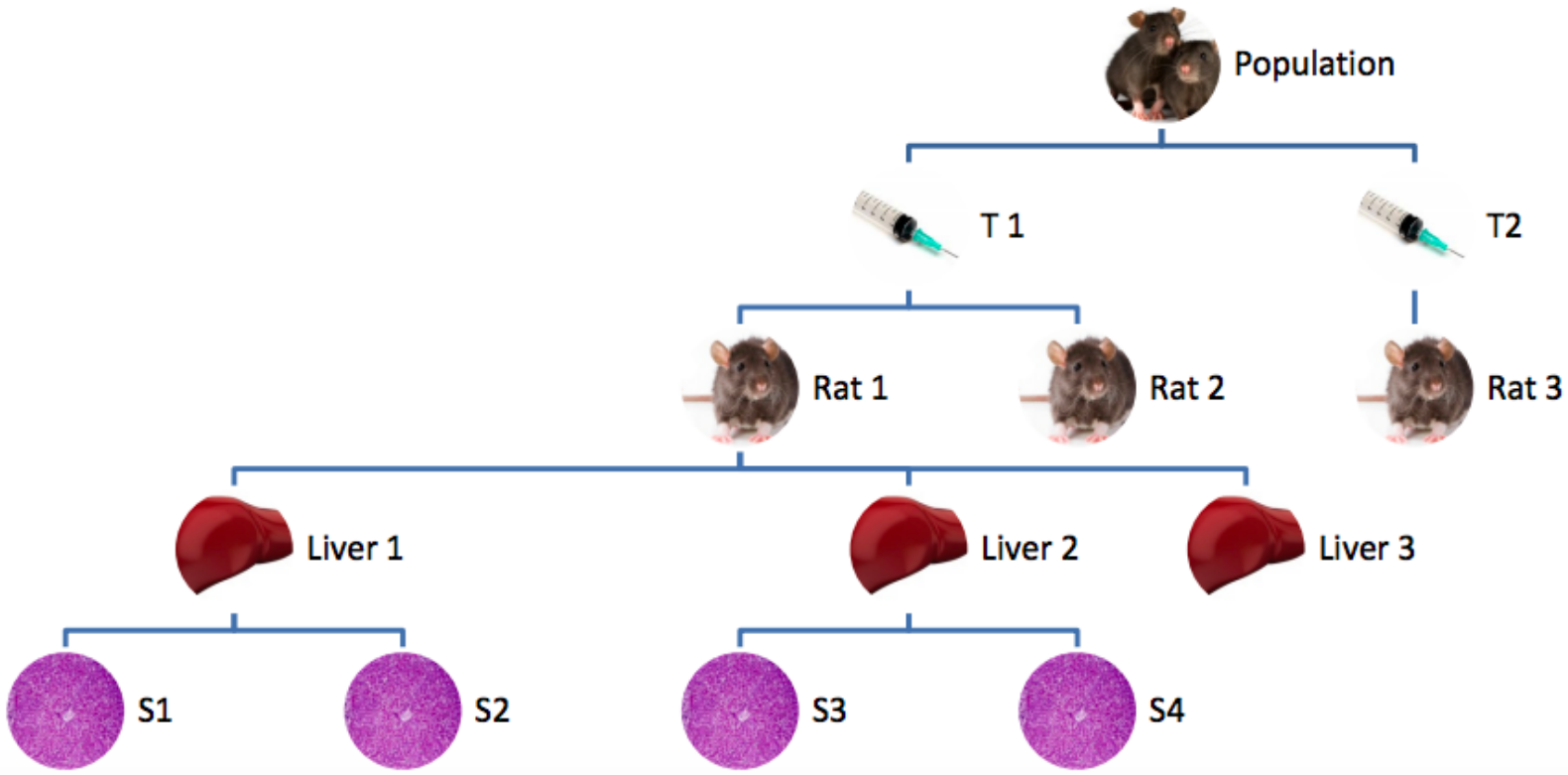
# Мотивация

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.

# Заключение

- ANOVA разнообразна по дизайну эксперимента
- Многие предположения ANOVA можно обойти, использовав соответствующий непараметрический тест.