Дисперсионный анализ

Центр биоэлектрических интерфейсов, 13 февраля 2019 г.

Денис Деркач, Влад Белавин

Оглавление

Мотивация

- ранее, мы рассматривали одно- и двухвыборочные тесты (например, t-test);
- что делать, если мы хотим сравнить сразу несколько выборок?

Мотивирующий пример

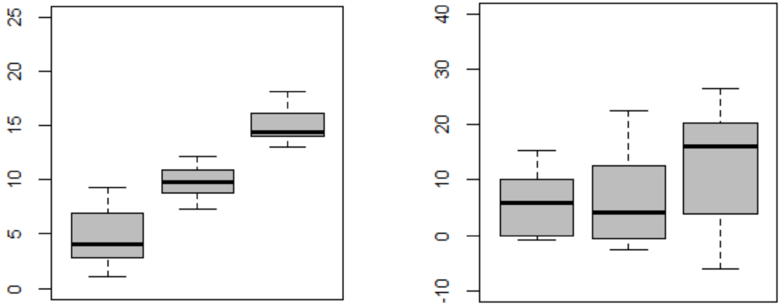
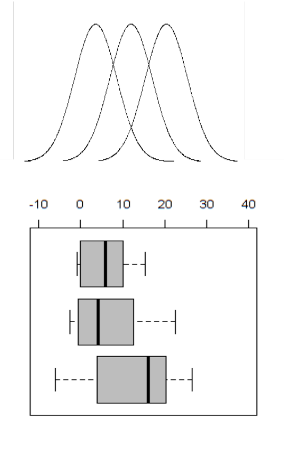


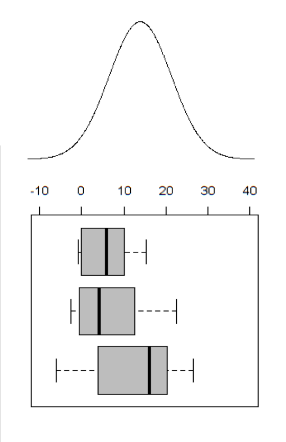
График слева: мы можем сказать, что средние, видимо, отличаются.
Что мы можем сказать про график справа?

Мотивирующий пример

Иными словами, пришли ли семплы из одного распределения или из разных?



Или



Обсуждение мотивирующего примера

- Эти два случая отличаются так как в первом случае данные внутри семпла не сильно вариьруются. Большая дисперсия отражает большую неопределенность в отношении значений истинных неизвестных средних.
- Необходимо сравнить дисперсию внутри группы с дисперсией между группами, для того, чтобы получить вывод о равенстве средних.

Тестирование гипотез: нулевая гипотеза

Сформулируем нулевую гипотезу:

- H_0: $\mu_1 = \mu_2 = \mu_3$.

Это сложная гипотеза, то есть она содержит в себе много парных и непарных простых (например $\mu_1 = \mu_2$ или $\mu_2 = \frac{\mu_1+\mu_3}{2}$).
NB: есть другие способы сформулировать нулевую гипотезу ANOVA.

Тестирование гипотез: альтернативная гипотеза

Сформулируем альтернативную гипотезу:

- H_1: $\mu_1 \neq \mu_2$ или $\mu_2 \neq \mu_3$ или $\mu_1 \neq \mu_3$.

Заметим, что H_0 отвергается, если верна хотя бы одна из маленьких частных альтернативных гипотез (парных или комплексных).
NB: ANOVA не говорит какая.

Какой тест предпочесть?

Мы можем взять несколько попарных t-тестов, проверяя:

- H_0: $\mu_1 = \mu_2$;
  H_1: $\mu_1 \neq \mu_2$.
- H_0: $\mu_2 = \mu_3$;
  H_1: $\mu_2 \neq \mu_3$.
- и т.д.

Проблема. вероятность ошибки первого рода резко увеличивается/

TABLE 1: Probability of Committing at Least One Type I Error by Using Two-Sample *t* Tests for All *C* Pairwise Comparisons of *k* Means*

		Level of Significance, α, Used in the *t* Tests				
k	C	0.10	0.05	0.01	0.005	0.001
2	1	0.10	0.05	0.01	0.005	0.001
3	3	0.27	0.14	0.03	0.015	0.003
4	6	0.47	0.26	0.06	0.030	0.006
5	10	0.65	0.40	0.10	0.049	0.010
6	15	0.79	0.54	0.14	0.072	0.015
10	45	0.99	0.90	0.36	0.202	0.044
	∞	1.00	1.00	1.00	1.000	1.000

*There are $C = k(k-1)/2$ pairwise comparisons of k means. This is the number of combinations of k items taken two at a time.

Идея ANOVA

- Заметим, что при верной $_0$ все группы получены из популяций с одинаковыми средним μ и дисперсией σ^2.
- Давайте оценим дисперсию разными независимыми способами и сравним!
- Можем оценить исходя из вариативности внутри группы и вариативности между групп.

Вариативность между группами

Оценим σ^2 на основе дисперсии средних между группами (посчитаем ошибку среднего, как будто это выборочные средние, и из неё вычислим дисперсию):

$$s_{\bar{x}} = \frac{\sum_j (\bar{X}_j - \bar{\bar{X}})}{k-1}$$

Тогда mean square between groups (MS_B):

$$MS_B = \frac{\sum_j (\bar{X}_j - \bar{\bar{X}}) n_j}{k-1}$$

Количество степеней свободы при этом:

$$DF_B = k - 1,$$

где k — число групп.

Вариативность внутри группы

Mean square within groups = error MS

$$MS_W = \frac{s_1^2 + \ldots + s_k^2}{k}$$

Количество степеней свободы при этом:

$$DF_B = N - k,$$

где k – число групп, N - полное число семплов.

F-статистика

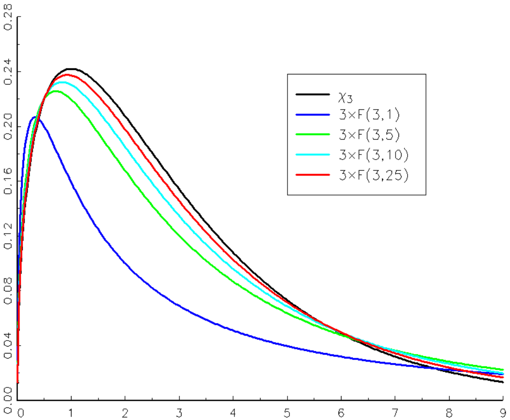
$$F = \frac{\text{оценка дисперсии между группами}}{\text{оценка дисперсии внутри групп}} = \frac{MS_B}{MS_W}.$$

Тестирование H_0

- для заданных df рассчитывается критическое значение F;
- на основе групп считается F и сравнивается с критическим значением;
- если F больше критического — $_0$ о равенстве средних в группах отвергается;
- F - это отношение дисперсий, оно имеет особое распределение, оно всегда положительно; ANOVA — принципиально односторонний тест.

F-статистика

$$F = \frac{\text{оценка дисперсии между группами}}{\text{оценка дисперсии внутри групп}} = \frac{MS_B}{MS_W}.$$



Sum of squares

SS - это суммы квадратов отклонений (sum of squared deviations):

- SSBetween - сумма квадратов отклонений каждого среднего в группе от общего среднего = Effect;
- SSWithin – сумма квадратов отклонений каждого измерения от среднего в соответствующей группе = Error;
- SSTotal – сумма квадратов отклонений каждого измерения от общего среднего = Total.

При этом:

$$SS_T = SS_W + SS_B.$$

ANOVA достигнутый эффект

Для того, чтобы понять насколько значим полученный результат в тесте строят два типа переменных:

- $R^2 = \eta^2 = \frac{SS_B}{SS_T}$, чем выше R^2, тем больше полученный эффект.
- $f = \frac{s_{\bar{x}}}{\sqrt{MS_W}}$, чем выше, тем больше полученный эффект.

Переменные ANOVA

Типы переменных:

- Группирующая переменная, фактор (factor, predictor).
- Зависимая переменная (dependent variable, response).

Мы пока разбираем случай с одним фактором (one-way). В ANOVA одна зависимая переменная, а факторов может быть несколько, и они могут составлять довольно сложные конструкции.

Дизайн эксперимента

Факторы могут быть двух видов:

- fixed. Рассматриваются именно эти значения фактора. Другие значения не существуют или не интересуют. Пример: пол, время суток и тд.
- random. Рассматриваются случайно выбранные значения фактора из многих возможных. За пределами исследования существуют другие значения фактора. Пример: происхождение семпла, процент лекарства.

Для этих типов факторов по-разному оценивается межгрупповая изменчивость. Когда фактор один, это не важно, но в сложных моделях с несколькими факторами эти различия очень важны!

Допущения ANOVA

- Выборки должны быть случайными, измерения — независимыми.
- Размеры групп должны различаться как можно меньше.
- Нормальность в каждой группе по отдельности.
- Равенство дисперсий в группах.

Требование нормальности

Возможные проверки:

- Сделать тест по методу моментов (Обычно достаточно проверить эксцесс и асимметрию).
- Построить гистограмму распределения остатков ($x_i - \bar{x}$) внутри каждого семпла (и проверить goodness-of-fit тесты).
- Тест Шапиро-Уилка.

Тест Шапиро-Уилка

Критерий Шапиро-Уилка основан на оптимальной линейной несмещённой оценке дисперсии к её обычной оценке методом максимального правдоподобия. Статистика критерия имеет вид:

$$W = \frac{1}{s^2}\left[\sum_{i=1}^{n} a_{n-i+1}(x_{n-i+1} - x_i)\right]^2,$$

где $s^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2, \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, а коэффициенты a_{n-i+1} берутся из таблиц.

На практике, следует проверять применимость таблиц в том или ином софте. Можно также использовать тест Шапиро-Франчиа.

Ненормальные данные

- Если семплы достаточно большие, можно оставить как есть.
- Провести преобразование к нормальным:
  - стандартизовать распределение;
  - метод Бокса-Кокса.
- Использовать непараемтрический ANOVA:
  - односторонний дисперсионный анализ Краскела—Уоллиса.

Метод Бокса-Кокса

Для последовательности: $\{y_1, \dots y_n\}$, $y_i > 0$ однопараметрическое преобразование Бокса-Кокса с параметром λ определяется следующим образом:

$$y_i^\lambda = \begin{cases} \frac{y_i^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0, \\ \log(y_i), & \text{if } \lambda = 0. \end{cases}$$

Где λ - свободный параметр.

Требование равенства дисперсий

Для проверки можно:

- Использовать тест Левена (Ливиня).
- Использовать F-тест.
- Построить зависимость остатков (residuals) от средних.

Тест Ливиня

Проверяет равенство дисперсий всех семплов.

$$W = \frac{(N-k)}{(k-1)} \cdot \frac{\sum_{i=1}^{k} N_i (Z_{i\cdot} - Z_{\cdot\cdot})^2}{\sum_{i=1}^{k} \sum_{j=1}^{N_i} (Z_{ij} - Z_{i\cdot})^2},$$

Здесь Z может центрироваться на:

- среднее выборки (для симметричных распределений);
- медиану выборки (для асимметричных распределений);
- усечённое среднее выборки (для распределений с тяжёлыми хвостами).

В общем случае рекомендуют использовать медиану.
Значение W затем сравнивается с соответсвующим F распределением.

Свойства ANOVA

- Возможно провести one-way ANOVA в случае, если у нас в руках есть только средние значения, показатели разброса (SD, SE, s^2) и размер выборок (например, из какой-нибудь статьи).
- В случае двух выборок ANOVA эквивалентна t-тесту.

Апостериорные (post-hoc) тесты

ANOVA не называет причину, по которой была отвергнута гипотеза H_0. Потому используют апостериорные тесты:

- Сначала сравнить все группы между собой с помощью ANOVA.
- Если различия есть, использовать методы множественного сравнения (сравнивают группы попарно, сохраняя общую $\alpha = 0.05$).
- Если различий нет, анализ следует считать завершённым (и не проводить post-hoc тесты).

NB: проведение апостериорных тестов может испортить весь анализ.

Тест Тьюки

Он же honestly significant difference test (HSD test) or wholly significant difference test (WSD test).

- Выстраиваем средние по выборке по возрастанию.
- Строим статистику $q = \frac{Y_A - Y_B}{SE}$, где Y - среднее (причём $Y_A > Y_B$, SE - стандартное отклонение.
- Ищем значимость $q_{\alpha, N-k, k}$ для нужного α.

Тест Тьюки

- Наиболее распространённый и рекомендуемый в литературе тест;
- строго контролирует α (0.05);
- проверяет все парные гипотезы сразу;
- плохо работает, если размер групп сильно различается;
- увствителен к неравенству дисперсий;
- считает статистику (q) на основе MSwithin и df.

Другие post-hoc тесты

- Тьюки-Крамера, решает проблему теста Тьюки для неравных выборок.
- Критерий Ньюмена-Кейлса. Все средние упорядочивают по возрастанию и пошагово вычисляют статистики; начинают от сравнения наибольшего с наименьшим. Сравнивают с $q_{\alpha,N-k,p}$, где p — диапазон средних. Мощнее теста Тьюки, но плохо контрлирует ошибку 1-го рода.
- Критерий Шеффе (Scheffe test) — очень консервативный, мощность меньше, чем у теста Тьюки (но см. ниже).
- Критерий Даннетта (Dunnett test) — используется для сравнения нескольких групп с контрольной группой, мощнее, чем тест Тьюки. Размер контрольной группы рекомендуется делать больше, чем размеры остальных групп в $\sqrt{k-1}$ раз.

Failed post-hoc

Бывает так, что в ANOVA нулевая гипотеза отвергается, а пост-хок тесты не обнаруживают различий, так как их мощность ниже. В этом случае необходимо увеличивать размер выборки.

Анализ контрастов (planned comparisons)

- Проводится вместо ANOVA.
- Важно: то, какие группы сравнивать, выбирают заранее, до проведения какого-либо анализа. В идеале — ещё при постановке исследования.
- В тесте проверяется только одна гипотеза;
- Можно провести 2-3 таких теста в пределах одного «набора» групп, только надо следить, чтобы сравнения не сильно перекрывались, не были избыточными.
- Мощнее post-hoc тестов.

Пример

У нас 4 группы тигров, их кормят: овощами; фруктами;рыбой; мясом.
Вопрос: отличается ли масса тигров, питающихся животной и растительной едой?

Построение контрастов

Контраст — линейная комбинация средних значений. Коэффициенты сравнения — константы, на которые умножены средние. Таким образом гипотезы формулируются:

$$H_0 : \sum_i C_i \mu_i = 0;$$

$$H_1 : \sum_i C_i \mu_i \neq 0.$$

При этом $\sum_i C_i = 0$.
Если тестируется несколько гипотез: $\sum_i C_{1,i} C_{2,i} = 0$. В этом случае статистика строится: $t = \frac{\sum_i C_i \mu_i}{SE}$ и имеет t распределение.

Пример

У нас 4 группы тигров, их кормят: овощами; фруктами;рыбой; мясом.
Вопрос: отличается ли масса тигров, питающихся животной и растительной едой?
Мы строим контраст: $\frac{1}{2}\mu_1 + \frac{1}{2}\mu_2 - \frac{1}{2}\mu_3 - \frac{1}{2}\mu_4$.