13. Дисперсионный анализ. Критерий Бартлетта (проверяемая гипотеза и используемое распределение).

Дисперсионный анализ

Определение 7.1. *Дисперсионный анализ* — статистический метод, предназначенный для оценки влияния различных факторов на результат эксперимента, а также для планирования аналогичных экспериментов.

По числу факторов, влияние которых исследуется, различают однофакторный и многофакторный анализ. В рамках данного курса будем рассматривать только однофакторный дисперсионный анализ.

Основные понятия:

- **Фактор** – это независимая переменная (например, вид удобрения, метод обучения, тип лекарства), влияние которой на зависимую переменную мы изучаем.
- **Уровни фактора** – это различные значения или категории фактора (например, три разных дозы удобрения, четыре метода обучения, две марки лекарства).
- **Цель дисперсионного анализа** – проверить, оказывают ли разные уровни фактора статистически значимое влияние на среднее значение зависимой переменной.

Зачем нужно требование равенства дисперсий?
Дисперсионный анализ строится на предположении, что **остатки модели (разности между наблюдаемыми и предсказанными значениями) распределены нормально с одинаковой дисперсией**:

$$\varepsilon_{ij} \sim N(0, \sigma^2),$$

где σ^2 – общая дисперсия ошибок для всех групп.

Зачем нужно равенство дисперсий в дисперсионном анализе? Обоснование подробно

Дисперсионный анализ (ANOVA) основан на **модели**:

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij},$$

где:

- y_{ij} — наблюдаемое значение в i-й группе,
- μ — общее среднее,
- τ_i — эффект i-го уровня фактора,
- ε_{ij} — случайная ошибка.

Ключевое предположение: ошибки ε_{ij} независимы и имеют **одинаковое нормальное распределение**:

$$\varepsilon_{ij} \sim N(0, \sigma^2),$$

где σ^2 — одна и та же дисперсия для **всех групп**.

Что происходит, если дисперсии не равны (гетероскедастичность)?

1. Нарушается точность F-критерия

В ANOVA для проверки гипотезы $H_0 : \tau_1 = \tau_2 = \cdots = \tau_k$ используется **F-статистика**:

$$F = \frac{MS_{\text{между}}}{MS_{\text{внутри}}},$$

где $MS_{\text{внутри}}$ — средний квадрат внутри групп (оценка общей дисперсии σ^2).

Если дисперсии разные, то $MS_{\text{внутри}}$ перестаёт быть хорошей оценкой **одной** общей дисперсии. В результате:

- **Распределение F-статистики отклоняется от теоретического F-распределения** (Фишера).
- **Фактический уровень значимости** перестаёт соответствовать номинальному (например, выбранному $\alpha = 0.05$).

2. Возрастает вероятность ошибок I и II рода

- **Ошибка I рода (ложное обнаружение эффекта)** — когда мы отвергаем H_0, хотя на самом деле различий между группами нет.
  - При неравных дисперсиях F-критерий может стать **либеральным**: он будет чаще отвергать H_0, чем следует.
  - **Пример:** Если в группах с маленькой дисперсией случайно оказались крайние значения, F-статистика может искусственно завыситься.
- **Ошибка II рода (пропуск эффекта)** — когда мы не отвергаем H_0, хотя реальные различия есть.
  - При неравных дисперсиях F-критерий может стать **консервативным**: он будет реже обнаруживать реальные эффекты.
  - **Пример:** Если в группах с большим разбросом "шум" маскирует систематические различия, мощность критерия падает.

3. Искажение доверительных интервалов

Доверительные интервалы для средних или разностей средних строятся в предположении общей дисперсии. При гетероскедастичности:

- Интервалы становятся **уже или шире**, чем должны быть.
- **Уровень доверия** (например, 95%) не соблюдается.

Почему это важно для практики?

- Если вы сравниваете эффективность трёх лекарств, но в одной группе пациенты сильно различаются по реакции (большая дисперсия), а в других — нет, то:
  - Вы можете сделать **неверный вывод**, что лекарства действуют по-разному (ошибка I рода).
  - Или наоборот, **не заметить** реальную разницу (ошибка II рода).

Поэтому **перед применением дисперсионного анализа необходимо проверять гипотезу о равенстве дисперсий**, например, с помощью **критерия Бартлетта** (для нормальных данных) или **критерия Левена** (более устойчивого к отклонениям от нормальности).

Критерий Бартлетта используется для проверки этого условия, когда количество групп $k > 2$.

Проверяемая гипотеза

Критерий Бартлетта проверяет гипотезу о равенстве дисперсий нескольких выборок, извлеченных из нормальных популяций.

- **Нулевая гипотеза (H_0):** Дисперсии во всех группах равны между собой:

$$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$$

- **Альтернативная гипотеза (H_1):** Хотя бы две дисперсии значимо различаются (нарушение однородности).

Алгоритм применения

1. Находим несмещенные оценки s_i^2 групповых дисперсий σ_i^2 по формуле

$$s_j^2 = \frac{n_j}{n_j - 1}\,\widehat{\sigma}_j^2,$$

где $\widehat{\sigma}_j^2$ — выборочные групповые дисперсии, n_j — численность наблюдений в группах.

2. Находим значение величины s_0^2 - так называемой обобщенной дисперсии:

$$s_0^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2 + \ldots + (n_\nu - 1)s_\nu^2}{(n_1-1) + (n_2-1)\ldots + (n_\nu - 1)}$$

3. Определяем вспомогательный параметр q:

$$q = \left[1 + \frac{1}{3(\nu-1)}\left(\frac{1}{n_1-1} + \frac{1}{n_2-1} + \ldots + \frac{1}{n_\nu - 1} - \frac{1}{(n_1-1)+\ldots+(n_\nu-1)}\right)\right]^{-1}.$$

4. Вычисляем значение критерия Бартлетта:

$$K_B = q\left[(n_1-1)\ln\frac{s_0^2}{s_1^2} + (n_2-1)\ln\frac{s_0^2}{s_2^2} + \ldots + (n_\nu-1)\ln\frac{s_0^2}{s_\nu^2}\right]$$

при выполнении условия $n_j > 3(j = 1, 2, \ldots, \nu)$ и гипотезы (7.3) K_B будет иметь распределение, близкое к χ^2-распределению с $k = \nu - 1$ степенями свободы.

5. Задавшись уровнем значимости α, находим правостороннюю критическую точку $\chi^2_{\text{кр, пр}}$ по следующей схеме:

$$\left.\begin{array}{lcl} \alpha & \to & \gamma = 1-\alpha \\ \nu & \to & k = \nu - 1 \end{array}\right\} \to \chi^2_\gamma \to \chi^2_{\text{кр, пр}} = \chi^2_\gamma.$$

6. Если K_B попадает в интервал $(\chi^2_{\text{кр, пр}}, +\infty)$, то гипотезу H_0 (7.3) отвергаем. В противном случае считаем, что гипотеза (7.3) не противоречит результатам наблюдений.