# МС-26 Теоретический материал
# Критерии независимости и однородности $\chi^2$-Пирсона, критерии однородности Колмогорова-Смирнова. Сравнение генеральных средних и генеральных дисперсий двух и более нормальных совокупностей

## Критерий независимости $\chi^2$

Имеются две дискретные случайные величины $X$ и $Y$.
Требуется проверить **гипотезу об их независимости**.
Пусть различаются $r$ значений случайной величины $X$ (обозначим их $x_1, \ldots, x_r$) и $s$ значений случайной величины $Y$ (обозначим их $y_1, \ldots, y_s$).
Через $k_{ij}$ обозначим общее количество таких элементов выборки, в которых $X$ принимает значение $x_i$, а $Y$ — значение $y_j$.
Тогда

$$\sum_{i=1}^{r}\sum_{j=1}^{s} k_{ij} = n,$$

где $n$ — объем выборки. Введем также обозначения:

$$\nu_i = \sum_{j=1}^{s} k_{ij}\,;\ \mu_j = \sum_{i=1}^{r} k_{ij}.$$

В рассматриваемом случае результаты наблюдений удобно оформлять в виде таблицы, называемой **таблицей сопряженности признаков**:

| $X/Y$ | $y_1$ | $y_2$ | … | $y_s$ | |
|---|---|---|---|---|---|
| $x_1$ | $k_{11}$ | $k_{12}$ | … | $k_{1s}$ | $\nu_1$ |
| $x_2$ | $k_{21}$ | $k_{22}$ | … | $k_{2s}$ | $\nu_2$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ | $\vdots$ |
| $x_r$ | $k_{r1}$ | $k_{r2}$ | … | $k_{rs}$ | $\nu_r$ |
| | $\mu_1$ | $\mu_2$ | … | $\mu_s$ | $n$ |

Пусть далее $p_{ij} = P(\{X = x_i, Y = y_j\});\ p_i = P(\{X = x_i\}); q_j = P(\{Y = y_j\})$.

$H_0: p_{ij} = p_i q_j, i = 1, \ldots, r,\ j = 1, \ldots, s.$
$H_1: p_{ij} \neq p_i q_j$ для некоторых $i = 1, \ldots, r,\ j = 1, \ldots, s.$
Рассматривается следующая **статистика**:

$$\chi^2 = n\left(\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{k_{ij}^2}{\nu_i\mu_j} - 1\right).$$

**Гипотеза $H_0$ отклоняется**, если вычисленное по выборочным данным значение статистики $\chi^2_{\text{набл}}$ удовлетворяет неравенству:

$$\chi^2_{\text{набл}} > \chi^2_{\alpha;(r-1)(s-1)}.$$

**Критерий однородности $\chi^2$**

Проверяется гипотеза о том, что две выборки принадлежат одной генеральной совокупности.

Данные должны быть представлены в виде интервального статистического ряда.

Имеются выборка объема $\boldsymbol{n_1}$ из генеральной совокупности $\boldsymbol{X_1}$ и выборка объема $\boldsymbol{n_2}$ из генеральной совокупности $\boldsymbol{X_2}$; $\boldsymbol{l}$ — количество интервалов группировки (одинаковое для обеих выборок); $\boldsymbol{\mu_i}$ и $\boldsymbol{\nu_i}$— количество попаданий в $\boldsymbol{i}$-й интервал группирования, соответственно, первой и второй выборок, $\boldsymbol{i = 1, 2, ..., l}$; уровень значимости $\boldsymbol{\alpha}$.

Пусть $F_j(x)$ — функция распределения случайной величины $X_j$, $j = 1, 2$.

Проверяется **гипотеза**

$H_0: F_1(x) = F_2(x),\ x \in \mathbb{R}$,

$H_1: F_1(x) \neq F_2(x)$, для некоторых $x \in \mathbb{R}$.

В случае совпадения объемов выборок: $n_1 = n_2 = n$

**статистика** вычисляется по формуле

$$\chi^2 = \sum_{i=1}^{l} \frac{(\mu_i - \nu_i)^2}{\mu_i + \nu_i}.$$

**Критическое значение статистики**: $\chi^2_{\alpha;l-1}$.

**Гипотеза $\boldsymbol{H_0}$ отклоняется**, если вычисленное по выборочным данным значение статистики $\chi^2_{\text{набл}}$ удовлетворяет неравенству:

$$\chi^2_{\text{набл}} > \chi^2_{\alpha;l-1}.$$

**Статистика критерия** имеет следующий вид:

$$\chi^2 = n_1 n_2 \sum_{i=1}^{l} \frac{\left(\frac{\mu_i}{n_1} - \frac{\nu_i}{n_2}\right)^2}{\mu_i + \nu_i}.$$

**Критерий однородности Колмогорова-Смирнова**

Имеются две выборки – объема $n_1$ из генеральной совокупности $X_1$ и объема $n_2$ из генеральной совокупности $X_2$.

Предполагается, что случайные величины $\mathrm{X}_j$ — непрерывные с функциями распределения $F_j(x), j = 1, 2$.

$H_0: F_1(x) = F_2(x),\ x \in \mathbb{R}$,

$H_1: F_1(x) \neq F_2(x)$, для некоторых $x \in \mathbb{R}$.

Проверка гипотезы производится **по следующей схеме**:

1. По имеющимся выборкам находятся **эмпирические функции распределения** $F_1^*(x)$ и $F_2^*(x)$.
2. Рассматривается **статистика** следующего вида:

$$D = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \cdot \max_x |F_1^*(x) - F_2^*(x)|.$$

3. По таблицам распределения Колмогорова определяется величина $k_\alpha$—$100\alpha$-процентная точка распределения Колмогорова уровня $\alpha$.
4. **Гипотеза $\boldsymbol{H_0}$ отклоняется** на уровне значимости $\alpha$, если вычисленное по выборочным данным значение статистики $D_{\text{набл}}$ удовлетворяет неравенству:

$D_{\text{набл}} > k_\alpha$.

**Замечание**. Критерий Колмогорова–Смирнова применяется при $n_1, n_2 \geq 50$.

**Дисперсионный анализ** – это статистический метод анализа результатов наблюдений, зависящих от различных одновременно действующих факторов, основанный на сравнении оценок дисперсий соответствующих групп выборочных данных.

Под **фактором** понимают различные, независимые, качественные показатели, влияющие на изучаемые признаки.

Признаки, изменяющиеся под воздействием тех или иных факторов, называют **результативными**.

**Сущность метода дисперсионного анализа** заключается в измерении отдельных дисперсий (общая, факторная, остаточная), и дальнейшем определении силы влияния изучаемых факторов (оценки роли каждого из факторов, либо их совместного влияния) на результативный признак.

**Классификация методов дисперсионного анализа**

**По количеству анализируемых признаков**
- **Однофакторный (ANOVA)** (Анализ различий по одному признаку)
- **Многофакторный (MANOVA)** (Анализ различий групп одновременно по двум признакам и более)

*__ANOVA__ - Analysis of Variance
*__MANOVA__ – Multivariate Analysis of Variance

**Предположения для использования однофакторного дисперсионного анализа**:

Чтобы результаты однофакторного дисперсионного анализа были достоверными, должны выполняться следующие допущения:

**1. Нормальность.** Каждая выборка была взята из нормально распределенной популяции.

**2. Равные дисперсии** — дисперсии совокупностей, из которых взяты выборки, равны (можно использовать **тест Бартлетта**, чтобы проверить это предположение).

**3. Независимость**. Наблюдения в каждой группе независимы друг от друга, а наблюдения внутри групп были получены путем случайной выборки.

**Однофакторный дисперсионный анализ** основан на том, что сумму квадратов отклонений статистического комплекса возможно разделить на компоненты:

$$SST = SSA + SSW,$$

где $SST$ - общая сумма квадратов отклонений, $SSA$ - объяснённая влиянием фактора *A* сумма квадратов отклонений, $SSW$ - необъяснённая сумма квадратов отклонений или сумма квадратов отклонений ошибки.

1. **Межгрупповая вариация** $SSA$ – вариация между средним каждой группы и общим средним значением всей выборки
2. **Внутригрупповая вариация** $SSW$– вариация между каждым объектом исследования группы и средним значением соответствующей группы

| **Компоненты дисперсии** | **Сумма квадратов** | **Число степеней свободы** | **Оценки дисперсии** |
|---|---|---|---|
| **Межгрупповая ($SSA$)** **S**um of **S**quares **A**mong Groups | $SSA = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2$ | $k-1$ | $MSA = \frac{SSA}{k-1}$ (**Факторная** дисперсия) |
| **Внутригрупповая ($SSW$)** **S**um of **S**quares **W**ithin Groups | $SSW = \sum_{i}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2$ | $n-k$ | $MSW = \frac{SSW}{n-k}$ (**Остаточная** дисперсия) |
| **Общая ($SST = SSA + SSW$)** **T**otal **S**um of **S**quares | $SST = \sum_{i}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x})^2$ | $n-1$ | $MST = \frac{SST}{n-1}$ |

Чтобы провести однофакторный дисперсионный анализ данных статистического комплекса, нужно найти **фактическое отношение Фишера** - отношение дисперсии, объяснённой влиянием фактора (**межгрупповой, факторной**), и необъяснённой дисперсии (**внутригрупповой, остаточной**).

Тогда **дисперсионное отношение** это $F_{\text{набл}} = \frac{MSA}{MSW}$.

Указанное дисперсионное отношение **тем больше**, **чем сильнее** разнятся уровни фактора по воздействию на групповые средние.

## Проверка гипотезы о совпадении нескольких генеральных средних методом дисперсионного анализа

Пусть $\vec{X}_i = (X_{i1}, \dots, X_{in_i})$ — выборка объема $n_i$ из $N(\mu_i, \sigma^2)$, где $i = 1, \dots, k$.
Предположим также, что $n = n_1 + \dots + n_k$ случайных величин

$$X_{11}, \dots, X_{1n_1};\ X_{21}, \dots, X_{2n_2}; \dots;\ X_{k1}, \dots, X_{kn_k}$$

независимы в совокупности. Таким образом, выборки $\vec{X}_1, \dots, \vec{X}_k$ **независимы** и получены из нормальных распределений с одинаковой дисперсией $\sigma^2$ и, возможно, различными средними $\mu_1, \dots, \mu_k$. **Гипотеза о равенстве всех средних** одновременно записывается как
$H_0: \mu_1 = \dots = \mu_k,\ H_1: (\exists i, j)\ \mu_i \neq \mu_j$.

Заметим, что при верной $H_0$ генеральные распределения совпадают:

$$N(\mu_1, \sigma^2) = \dots = N(\mu_k, \sigma^2).$$

Рассмотрим объединенную выборку объема $n = n_1 + \dots + n_k$

$$\vec{X} = (X_{11}, \dots, X_{1n_1};\ X_{21}, \dots, X_{2n_2}; \dots;\ X_{k1}, \dots, X_{kn_k}).$$

Интерпретируя выборки $\vec{X}_1, \dots, \vec{X}_k$ как группы, на которые разбита совокупность $\vec{X}$, введем обозначения, аналогичные тем, что использовались при изучении межгрупповой дисперсии:

$\bar{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}$ — **выборочное среднее в $\boldsymbol{i}$-й совокупности**;

$\hat{\sigma}_i^2 = \sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2$ — **выборочная дисперсия в той же выборке**;

$\bar{X} = \frac{1}{n}\sum_{i=1}^{k}\bar{X}_i n_i = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}$ — **выборочное среднее в объединенной выборке $\vec{\boldsymbol{X}}$**;

$\bar{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{k}\hat{\sigma}_i^2 n_i$ — **средняя групповая дисперсия ($\boldsymbol{SSW/n}$)**;

$\delta^2 = \frac{1}{n}\sum_{i=1}^{k}(\bar{X}_i - \bar{X})^2 n_i$ — **межгрупповая дисперсия ($\boldsymbol{SSA/n}$)**;

$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X})^2$ — **выборочная дисперсия признака в объединенной выборке $\vec{\boldsymbol{X}}$**.

Известно, что выборочную дисперсию $\hat{\sigma}^2$ можно представить в виде суммы

$$\hat{\sigma}^2 = \bar{\sigma}^2 + \delta^2 = (\boldsymbol{SSA} + \boldsymbol{SSW})/\boldsymbol{n} = \boldsymbol{SST}/\boldsymbol{n},$$

где слагаемое $\bar{\sigma}^2$ характеризует среднюю изменчивость признака в каждой выборке $\vec{X}_1, \dots, \vec{X}_k$, а слагаемое $\delta^2$ характеризует разброс выборочных средних $\bar{X}_1, \dots, \bar{X}_k$.

Критерий проверки $H_0$ против $H_1$ использует так называемое **отношение Фишера**:

$$F = \frac{\frac{1}{k-1}\sum_{i=1}^{k}(\bar{X}_i - \bar{X})^2 n_i}{\frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2} = \frac{\frac{1}{k-1} n\delta^2/s^2}{\frac{1}{n-k} n\bar{\sigma}^2/s^2} = \frac{\boldsymbol{MSA}}{\boldsymbol{MSW}}.$$

Можно доказать, что $F \sim F(k-1, n-k)$, где $F(n-1, n-k)$ — **распределение Фишера с $\boldsymbol{k-1}$ и $\boldsymbol{n-k}$ степенями свободы**.

Для проверки $H_0$ с уровнем значимости $\alpha$ применяется критерий с критической областью $F > F_\alpha(k-1, n-k)$, где $F_\alpha(k-1, n-k)$ — верхняя процентная точка распределения $F(k-1, n-k)$.

**Замечание**. Если $F < 1$, то следует сразу принять гипотезу $H_0$, поскольку $F_{кр}$ всегда больше единицы.

Если **гипотеза о равенстве средних не подтверждается**, имеет смысл оценивать величины $\mu_i$ по отдельности. Получаем для них следующие **доверительные интервалы (с надежностью $\gamma$)**:

$$\bar{X}_i - \delta < \mu_i < \bar{X}_i + \delta,$$

где $\delta = \sqrt{\frac{MSW}{n_i}}\, t_\gamma$, $t_\gamma$ – критическая точка распределения Стьюдента с $n - k$ степенями свободы (для двусторонней области)

**scipy.stats.t.ppf((1+gamma)/2,n-k)**

Одним из условий применения дисперсионного анализа является равенство генеральных групповых дисперсий $\sigma_i^2 = \sigma_0^2$, $i$ = 1,…,$k$.

**Сравнение нескольких дисперсий нормальных генеральных совокупностей по выборкам различного объема. Критерий Бартлетта**

Пусть генеральные совокупности $X_1, X_2, \ldots, X_l$ распределены нормально. Из этих совокупностей извлечены независимые выборки, вообще говоря различных объёмов $n_i$ (некоторые $n_i$ могут быть одинаковыми; если все выборки имеют одинаковые объём, то предпочтительнее пользоваться **критерием Кочрена**).

H$_0$: $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$

1. Найдем исправленные выборочные дисперсии $s_i^2, i = 1, \ldots, l$.
2. Найдём $k_i = n_i - 1$ – число степеней свободы дисперсии $s_i^2, i = 1, \ldots, l$ и сумму степеней свободы $k = \sum_{i=1}^{l} k_i$.
3. Найдем среднюю арифметическую исправленных дисперсий, взвешенную по числам степеней свободы:

$$\overline{s^2} = \frac{\sum_{i=1}^{l} k_i s_i^2}{k}.$$

4. Случайная величина (**критерий Бартлетта**) $B = \frac{V}{C}$,

где

$$V = 2{,}303\left[k \lg \overline{s^2} - \sum_{i=1}^{l} k_i \lg s_i^2\right];\; C = 1 + \frac{1}{3(l-1)}\left[\sum_{i=1}^{l} \frac{1}{k_i} - \frac{1}{k}\right],$$

которая при условии справедливости гипотезы об однородности дисперсий распределена приближенно как $\chi^2$ с $l$— 1 степенями свободы, если объем каждой выборки $n_i > 3, i = 1, \ldots, l$ .

**Правило**. Для того чтобы при заданном уровне значимости $\alpha$ проверить **нулевую гипотезу об однородности дисперсий нормальных совокупностей,** надо вычислить наблюдаемое значение критерия Бартлетта $B_{набл.} = V/C$ , и критическую точку $\chi^2_{кр.}(\alpha; l-1)$ правосторонней критической области. Если $B_{набл.} < \chi^2_{кр.}$ — нет оснований отвергнуть нулевую гипотезу. Если $B_{набл.} > \chi^2_{кр.}$—нулевую гипотезу отвергают.

**Замечание** 1. Не следует торопиться вычислять постоянную $C$. Сначала надо найти $V$ и сравнить с $\chi^2_{\text{кр.}}$ если окажется, что $V < \chi^2_{\text{кр.}}$, то подавно (так как $C > 1$) $B_{\text{набл.}} = V/C < \chi^2_{\text{кр.}}$ и, следовательно, $C$ вычислять не нужно. Если же $V > \chi^2_{\text{кр.}}$, то надо вычислить $C$ и затем сравнить $B_{\text{набл.}}$ с $\chi^2_{\text{кр.}}$

**Замечание 2**. Критерий Бартлетта чувствителен к отклонениям распределений от нормального, поэтому к выводам следует относится с осторожностью.

**Замечание 3**. При условии однородности дисперсий **в качестве оценки генеральной дисперсии** принимают среднюю арифметическую исправленных дисперсий, взвешенную по числам степеней свободы:

**Критерий Бартлетта** также называют **гипотезой об однородности дисперсий**.

https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.bartlett.html

Предположим, что фактор $A$ влияет на результативный признак $X$. Для измерения степени этого влияния используют **выборочный коэффициент детерминации**, равный

$$\eta^2 = \frac{SSA}{SST},$$

который показывает, какую долю выборочной дисперсии составляет дисперсия групповых средних, иначе говоря, какая доля общей дисперсии объясняется зависимостью результативного признака $X$ от фактора $A$.

Для выполнения **однофакторного дисперсионного анализа в Python** применяют функцию

**f_oneway()**

из библиотеки **SciPy.**

https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.f_oneway.html