## МС-17 Теоретический материал
## Критерии однородности, согласия

**Критерием согласия** называется статистический критерий о предполагаемом законе распределения.

Следует понимать, что проверяется не то, что случайная величина *действительно* имеет определенный закон распределения (например, нормальный), а проверяется лишь, достаточно ли хорошо наблюдаемые данные **согласуются** с некоторым законом распределения, чтобы можно было использовать этот закон для прогнозирования поведения данной случайной величины.

Гипотеза называется **простой**, если проверяется соответствие некоторому закону распределения с заданными параметрами.

Гипотеза называется **сложной**, если проверяется соответствие некоторому закону распределения с произвольными параметрами. (В этом случае параметры оцениваются по выборке.)

I. <u>**Простая гипотеза**</u>.

### Критерий согласия Пирсона

Производится серия повторных независимых испытаний, $\boldsymbol{n}$ – число испытаний, $\boldsymbol{\omega_t}$ – элементарный исход испытания с номером $t = 1, \dots, n$.

Поскольку испытания повторные, в качестве их общей вероятностной модели принимается одно и то же вероятностное пространство $(\Omega, \mathcal{F}, P)$ , очевидно, что все элементарные исходы $\boldsymbol{\omega_t} \in \Omega$.

Предположим, что $\boldsymbol{A_1}$, ..., $\boldsymbol{A_l} \in \boldsymbol{\mathcal{F}}$– попарно несовместные события, такие что $\boldsymbol{A_1}$+...+$\boldsymbol{A_l} \in \Omega$.
В качестве $\boldsymbol{H_0}$ примем гипотезу, состоящую в том, что вероятности событий $\boldsymbol{A_i}$ $(\boldsymbol{i} = \boldsymbol{1}, \dots, \boldsymbol{l})$ заданы таблицей

| Событие | $A_1$ | ... | $A_l$ |
|---|---|---|---|
| Вероятность | $p_1$ | ... | $p_l$ |

Пусть $n_i$ – эмпирическая частота события $A_i$, т.е. число испытаний, в которых $A_i$ наступило.

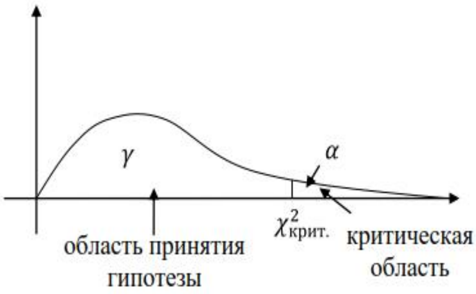
| Событие | $A_1$ | … | $A_l$ |
|---|---|---|---|
| Частота | $n_1$ | … | $n_l$ |

Если основная гипотеза верна, согласно статистическому определению вероятности $\hat{p}_i \approx p_i$, где $\hat{p}_i = n_i/n$ – относительная частота события $A_i$.

В качестве меры одновременной близости $l$ пар чисел $(\hat{p}_i, p_i)$ можно принять любую сумму вида $c_1(\hat{p}_1 - p_1)^2 + \dots + c_l(\hat{p}_l - p_l)^2$, в которой $c_i > 0$ – какие-либо положительные числа. **К.Пирсон** обнаружил, что если придать большие веса маловероятным событиям, положив $c_i = n/p_i$, то при неограниченном увеличении $n$ распределение **статистики**

$$\chi^2 = \sum_{i=1}^{l} \frac{n}{p_i}(\hat{p}_i - p_i)^2 = \sum_{i=1}^{l} \frac{(n_i - np_i)^2}{np_i}$$

перестает зависеть от конкретных значений вероятностей $p_i$ и стремится к распределению хи-квадрат с $l-1$ степенями свободы.
При верной $H_0$, случайные величины $n_i \sim Bin(n, p_i)$), вследствие чего $np_i = E(n_i)$ называется **ожидаемой (теоретической) частотой события $A_i$** .




Критическая область $\chi^2 > \chi^2_\alpha(l-1)$.

**Замечание**. На практике данный критерий Пирсона применяется, если объем выборки $n > 50$ и все ожидаемые частоты $np_i > 5$. Несоблюдение данных условий обычно приводит к значительному отклонению фактического уровня значимости $P_{H_0}(\chi^2 > \chi^2_\alpha(l-1))$ от требуемого уровня $\alpha$.

## II. **Сложная гипотеза**.

Если значения параметров гипотетической функции распределения $F_0(x)$ неизвестны, то имеем **сложную гипотезу**.
Основная гипотеза $H_0$ заключается в том, что функция распределения имеет вид
$$F_0(x) = F(\text{x}, \theta_1, \dots, \theta_k)$$
при некоторых неизвестных значениях параметров $\theta_1, \dots, \theta_k$. В этом случае вероятности $p_1, \dots, p_l$ r также зависят от параметров.
**Статистика хи-квадрат** принимает имеет вид
$$\chi^2 = \sum_{i=1}^{l} \frac{(n_i - np_i(\theta_1, \dots, \theta_k)))^2}{np_i(\theta_1, \dots, \theta_k)}$$
При известных значениях параметров имел бы место первый случай. Но так как истинные значения $\theta_1, \dots, \theta_k$ **неизвестны**, то подставляя их оценки, **найденные методом максимального правдоподобия (методом моментов)**, получаем статистический критерий $\chi^2$ с меньшим числом степеней свободы, а именно $l$– $k$– $1$, где $l$ – число интервалов, на которые разбит весь диапазон наблюдаемых значений, $k$ – число неизвестных параметров гипотетической функции распределения.

Сравнивая наблюдаемое значение статистики $\chi^2$ с критическим значением $\chi^2_\alpha(l-k-1)$, по приведенной схеме, делаем заключение об истинности нулевой гипотезы: гипотеза принимается, если $\chi^2 < \chi^2_\alpha(l-k-1)$, и отвергается в противном случае.

## III. Критерий однородности.

### 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, $\boldsymbol{i = 1, 2, \ldots, l}$; 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 $x \in \mathbb{R}$.
**Статистика критерия** имеет следующий вид:

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

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 $H_0$ отклоняется**, если вычисленное по выборочным данным значение статистики $\chi^2_{\text{набл}}$удовлетворяет неравенству: $\chi^2_{\text{набл}} > \chi^2_{\alpha;l-1}$.

### 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 $H_0$ 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 $n_1, n_2 \geq 50$.

## Критерий согласия Колмогорова

Критерий согласия Колмогорова применяется **для проверки гипотез о законах распределения только непрерывных случайных величин**.

Проверяется гипотеза $H_0$: $F(x) = F_0(x)$ против альтернативной $H_1$: $F(x) \neq F_0(x)$.

Критерий основан на том факте, что распределение супремума разности между теоретической и эмпирической функциями распределения

$$D_n = \sup_x | F\ (x) - F(x) |$$

одинаково при любой $F(x)$. Величину $D_n$ называют **статистикой Колмогорова**.
При малых $n$ для статистики Колмогорова существуют **таблицы критических точек** $D_{\text{кр}}$. Если $D_n < D_{\text{кр}}$, то гипотеза $H_0$ принимается, иначе отвергается. При больших $n$ используют ***предельное распределение Колмогорова***. Имеет место следующая теорема.

**Теорема (Колмогорова)**. $P\left(\sqrt{n}D_n < x\right) \longrightarrow Q(x) = 1 + 2\sum_{k=1}^{\infty}(-1)^k e^{-2k^2x^2},\ n \longrightarrow \infty.$

Для распределения Колмогорова $Q(x)$, предельного для статистики $\lambda_n = \sqrt{n}D_n$, также существуют таблицы критических точек $\lambda_{\text{кр}}$. Практически их используют уже при $n > 20$. Если $\lambda_n < \lambda_{\text{кр}}$, то гипотеза $H_0$ принимается, иначе отвергается.

**Замечание 1.** На **практике** статистику Колмогорова (в предположении, что $F = F_0$ можно вычислить по эквивалентным формулам:

$$D_n = \max_{1 \le i \le n}\left\{\left|F_0\left(x_{(i)}\right) - \frac{i-1}{n}\right|, \left|\frac{i}{n} - F_0\left(x_{(i)}\right)\right|\right\}$$

и

$$D_n = \max_{1 \le i \le n}\left|F_0\left(x_{(i)}\right) - \frac{2i-1}{n}\right| + \frac{1}{2n},$$

где $x_{(i)}$ - члены вариационного ряда.

**Замечание 2.** Критерий Колмогорова, строго говоря, *нельзя* применять в случаях сгруппированных данных при неизвестных параметрах распределения. Тем не менее, он иногда применяется на практике и в подобных ситуациях. Однако при этом статистики критерия получаются *заниженными,* что увеличивает ошибку первого рода. В таких случаях предпочтительней пользоваться **критерием хи-квадрат Пирсона**.

**Критические точки для статистики Колмогорова $D_n$**

| Объем выборки | Уровень значимости α | | | |
|---|---|---|---|---|
| $n$ | 0,10 | 0,05 | 0,02 | 0,01 |
| 1 | 0,95 | 0,98 | 0,99 | 0,995 |
| 2 | 0,78 | 0,84 | 0,90 | 0,93 |
| 3 | 0,64 | 0,71 | 0,78 | 0,83 |
| 4 | 0,57 | 0,62 | 0,69 | 0,73 |
| 5 | 0,51 | 0,56 | 0,62 | 0,67 |
| 6 | 0,47 | 0,52 | 0,58 | 0,62 |
| 7 | 0,44 | 0,48 | 0,54 | 0,58 |
| 8 | 0,41 | 0,45 | 0,51 | 0,54 |
| 9 | 0,39 | 0,43 | 0,48 | 0,51 |
| 10 | 0,37 | 0,41 | 0,46 | 0,49 |
| 11 | 0,35 | 0,39 | 0,44 | 0,47 |
| 12 | 0,34 | 0,38 | 0,42 | 0,45 |
| 13 | 0,33 | 0,36 | 0,40 | 0,43 |
| 14 | 0,31 | 0,35 | 0,39 | 0,42 |
| 15 | 0,30 | 0,34 | 0,38 | 0,40 |
| 16 | 0,29 | 0,33 | 0,37 | 0,39 |
| 17 | 0,29 | 0,32 | 0,36 | 0,38 |
| 18 | 0,28 | 0,31 | 0,34 | 0,37 |
| 19 | 0,27 | 0,30 | 0,34 | 0,36 |
| 20 | 0,26 | 0,29 | 0,33 | 0,35 |

**Критические точки распределения Колмогорова**

$$Q(\lambda) = 1 + 2\sum_{k=1}^{\infty}(-1)^k e^{-2k^2\lambda^2}$$

| α | 0,10 | 0,05 | 0,02 | 0,01 |
|---|---|---|---|---|
| $\lambda_{кр}$ | 1,23 | 1,36 | 1,52 | 1,63 |

Функция **chi2_contingency()** реализует критерий «**Хи-квадрат независимости**»

**Некоторые критерии согласия в Python**

1. **Хи-квадрат Пирсона**.
Функция **scipy.stats.chisquare**
https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chisquare.html

2. **Тест Колмогорова**, предназначенный для проверки простой гипотезы о непрерывном распределении
**scipy.stats.kstest**
**scipy.stats.ks_1samp**

3. **Тест Шапиро-Уилка**. Специальный тест на нормальность (для сложной гипотезы). Один из наиболее мощных тестов нормальности (т.е. очень чувствителен к ненормальности).
Функция **scipy.stats.shapiro**

4. **Комбинированный тест нормальности**.
Функция **scipy.stats.normaltest**

5. **Квантильный график** (Q-Q plot) показывает соотношение между выборочными и теоретическим квантилями. Визуально характеризует близость выборки к заданному (по умолчанию нормальному) распределению.
Функция **scipy.stats.probplot**