Лабораторная работа № 4. Критерии согласия Колмогорова и Пирсона (хи-квадрат)

Данная лабораторная работа посвящена проверке простой гипотезы согласия о принадлежности статистических данных заданному закону распределения.

Пусть имеется выборка $X_1, X_2, \ldots, X_N$ независимых наблюдений из генеральной совокупности случайной величины ξ с неизвестной функцией распределения F.

Выдвигаемая гипотеза имеет вид

$$H_0: \qquad F(x) \equiv F_0(x), \tag{1}$$

где F_0 – функция распределения, называемая *гипотетической*.

Альтернатива, которая в этом случае явно не формулируется, состоит в том, что не выполняется нулевая гипотеза H_0.

Для проверки гипотез о согласии вида (1) в математической статистике разработано множество критериев, среди которых наибольшее распространение на практике получили критерии Колмогорова и Пирсона (хи-квадрат).

1. Критерий Колмогорова

Пусть, как прежде, $\widehat{F}_N$ обозначает эмпирическую функцию распределения,

$$D_N = \sup_{x \in \mathbb{R}} |\widehat{F}_N(x) - F_0(x)|$$

– равномерное расстояние Колмогорова.

Критерий основан на следующей теореме.

Теорема (*А. Н. Колмогорова*). *Предположим, что функция распределения F непрерывна. Тогда*

$$\mathbf{P}(\sqrt{N} D_N < z) \longrightarrow_{N\to\infty} K(z) = \begin{cases} 0, & z \le 0, \\ \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2 z^2}, & z > 0. \end{cases} \tag{2}$$

Квантили функции Колмогорова $K(z)$ табулированы.

Проверка гипотезы H_0 по критерию Колмогорова (если распределение наблюдаемой с.в. ξ абсолютно непрерывно) происходит следующим образом.

1. Назначается уровень значимости α.
2. По таблицам функции Колмогорова находим квантиль $z_{1-\alpha} = K^{-1}(1-\alpha)$.
3. Вычисляем значение расстояния D_N. Для этого находим

$$D_{N,1} = \max_{1 \leqslant k \leqslant n} \left| \frac{k}{n} - F(X_{k:N}) \right| \quad \text{и} \quad D_{N,2} = \max_{1 \leqslant k \leqslant n} \left| F(X_{k:N}) - \frac{k-1}{n} \right|,$$

где $X_{k:N}$ – k-й член вариационного ряда.

Очевидно, что

$$D_N = \max\left[D_{N,1}, D_{N,2}\right].$$

4. Если $D_N > z_{1-\alpha}/\sqrt{N}$, то гипотеза H_0 отвергается (с вероятностью ошибки, равной α). В противном случае делаем вывод, что гипотеза не противоречит опытным данным.

*2. **Критерий хи-квадрат (Пирсона)***

Основной недостаток критерия Колмогорова состоит в том, что его применение для проверки гипотезы (1) возможно только если ф.р. F наблюдаемой с.в.

Критерий Пирсона свободен от этого ограничения. Статистика критерия Пирсона строится следующим образом. Вначале область $\mathcal{X}$ возможных значений случайной величины разбивается на конечное число непересекающихся подобластей: $\Delta_j,\ 1 \le j \le r$, $\mathcal{X} = \bigcup_{j=1}^{r} \Delta_j$, $r \in N$.

Для каждой из r областей вычисляются следующие величины:
$p_j^0 = \int_{\Delta_j} dF_0(x)$ – гипотетическая вероятность события $\{\xi \in \Delta_j\}$;
Np_j^0 – гипотетическое среднее значение числа наблюдений X_i, попавших в область Δ_j (в случае справедливости гипотезы H_0);
$\nu_j = \sharp\{i : \ X_i \in \Delta_j\}$ – количество наблюдений, попавших в j-ю область,
$\nu = (\nu_1 \dots, \nu_r)$ – вектор частот.

Статистика Пирсона для проверки гипотезы H_0 имеет следующий вид:

$$P_N(\nu) = \sum_{j=1}^{r} \frac{(\nu_j - Np_j^0)^2}{Np_j^0}. \tag{3}$$

В случае скалярных наблюдений Δ_j – это, как правило, интервалы, определяемые точками $a_1 < a_2 < \ldots < a_{r-1}$, т.е. $\Delta_j = [a_{j-1}, a_j)$, $1 \le j \le r$, где мы формально полагаем $a_0 = -\infty$, $a_r = \infty$. Для каждого из r интервалов вычисляются следующие величины:
$p_j^0 = F_0(a_j) - F_0(a_{j-1})$ – гипотетическая вероятность события $\{\xi \in \Delta_j\}$;
Np_j^0 – гипотетическое среднее значение числа наблюдений X_i, попавших в интервал Δ_j (в случае справедливости гипотезы H_0);
$\nu_j = \sharp\{i : \ X_i \in \Delta_j\}$ – количество наблюдений, попавших в j-й интервал,

Проверка гипотезы H_0 по критерию Пирсона основана на следующей теореме.

Теорема (К. Пирсона)*. В случае справедливости гипотезы H_0 распределение статистики $P_N(\nu)$ сходится к распределению χ^2_{r-1} (хи-квадрат с $(r-1)$ степенями свободы) при $N \to \infty$, это означает, что для любого $z > 0$*

$$\mathbf{P}(P_N(\nu) < z) \longrightarrow_{N\to\infty} \int_0^z k_{r-1}(x)\, dx,$$

где $k_{r-1}(x)$ – плотность распределения случайной величины χ^2_{r-1}.

Квантили распределения χ^2_m для небольших значений m (не превосходящих 30) табулированы. В пакете $\mathbb{R}$ квантили хи-квадрат распределения можно получить путем обращения к функции

qchisq(p, df = . . .),

где $p \in (0, 1)$ – уровень требуемой квантили, параметр df — число степеней свободы (без ограничений на df)

Проверка гипотезы по критерию Пирсона производится в соответствии со следующей схемой.
1. На основе предварительного анализа данных выдвигается гипотеза H_0 о распределении наблюдаемой случайной величины ξ.
2. Производится разбиение пространство возможных значений с.в. ξ на подобласти Δ_j, $1 \le j \le r$.

3. Вычисляются частоты ν_j и гипотетические вероятности p_j^0.
4. Вычисляется значение статистики Пирсона $P_N(\nu)$.
5. Назначается уровень значимости α.
6. Определяется квантиль $k_{1-\alpha}(r-1)$ уровня $(1-\alpha)$ распределения хи-квадрат с $(r-1)$ степенями свободы.
7. Если $P_N(\nu) > k_{1-\alpha}(r-1)$, то гипотеза отвергается. Вероятность ошибки при этом равна α. В противном случае считаем, что гипотеза не противоречит опытным данным.

Замечание 1. Поскольку сходимость распределения статистики Пирсона $P_N(\nu)$ к предельному закону χ^2_{r-1} связана с тем, что слагаемые, образующие статистику, распределены в пределе, как квадраты стандартных нормальных случайных величин, при конечных N рекомендуется производить разбиение таким образом, чтобы для всех $1 \le j \le r$ выполнялось условие $Np_j^0 \ge 10$ (условие приемлемой по точности аппроксимации распределения нормированной биномиальной случайной величины стандартным нормальным законом), иначе критерий может привести к некорректным результатам (вместо этого на практике разбиение часто делают так, чтобы в каждый из интервалов попало не менее $8 - 10$ наблюдений).

Замечание 2. Первоначально критерий хи-квадрат был предложен Карлом Пирсоном для проверки простой гипотезы в случае дискретного распределения наблюдений: им было рассмотрена ситуация, когда наблюдаемая с.в. ξ принимает r возможных значений $a_1, a_2, \ldots, a_r$. Вектор распределения вероятностей $\boldsymbol{\bar{p} = (p_1, p_2, \ldots, p_r)}$, где $\boldsymbol{p_i = P\{\xi = a_i\}}$, неизвестен.

Выдвигается гипотеза

$$H_0: \quad \boldsymbol{\bar{p} = \bar{p}_0}, \tag{2}$$

где $\boldsymbol{\bar{p}_0}$ – заданный вектор значений гипотетического распределения.

Фактически проверяется гипотеза о том, что выборка X_i, $i = 1, \ldots, N$, имеет **мультиномиальное** распределение $M(\boldsymbol{\bar{p}_0}, N)$ с параметром $(\boldsymbol{\bar{p}_0}, N)$.

Однако вскоре ученые осознали, что критерий хи-квадрат универсален, он может быть применен и в случае абсолютно непрерывно распределенных наблюдений, если предварительно выполнить их группировку, т.е. заменить наблюдения частотами попадания в подобласти Δ_j, где $\mathcal{X} = \bigcup_{j=1}^{r} \Delta_j$, $r \in N$, и $\mathcal{X}$ – область возможных значений с.в.

Следует отметить, что в случае абсолютно непрерывных распределений проверка гипотезы (1) по сути подменяется проверкой гипотезы (2) о значении вектора вероятностей мультиномиального распределения. Поэтому (в результате потери части информации, связанной с группировкой наблюдений) критерий хи-квадрат оказывается несостоятельным против всех альтернатив $F(x)$, имеющих тот же вектор вероятностей попадания с.в. в подобласти Δ_j, что и $F_0(x)$. Однако возможностью существования таких альтернатив на практике, очевидно, можно пренебречь.

Замечание 3. В случае сложной гипотезы H_0: $F \in \{F_0(x,\theta),\ \theta \in \Theta\}$, где F_0 – полностью известная функция, зависящая от неизвестного параметра $\theta = (\theta_1, \ldots, \theta_k)$ ($\Theta \subseteq \mathbb{R}^k$), проверка гипотезы о виде распределения также может осуществляться по критерию Пирсона, но в этом случае гипотетические вероятности $p_j^0(\theta)$ зависят от неизвестного параметра θ. Для этого параметра находят оценку $\widehat{\theta}$, затем вычисляют гипотетические вероятности $p_j^0(\widehat{\theta})$, которые теперь, как и вектор частот ν, являются случайными величинами, затем вычисляют статистику Пирсона по формуле (3), подставляя в нее $p_j^0(\widehat{\theta})$ вместо p^0. Как было показано Фишером, если выполнены некоторые условия регулярности и если оценка $\widehat{\theta}$ как параметра мультиномиального распределения случайного вектора $\nu = (\nu_1, \ldots, \nu_r)$ получена методом максимального правдоподобия (т. е. $\widehat{\theta}$ – это то значение, при котором достигает максимума по $\theta \in \Theta$ функция правдоподобия: $C \prod_{j=1}^{r} [p_j^0(\theta)]^{\nu_j}$, где $C = n!/(\nu_1!\nu_2! \cdots \nu_r!)$), то предельное распределение статистики Пирсона – это снова распределение хи-квадрат, но с числом степеней свободы $(r - 1 - k)$, где $k = \dim \Theta$ – количество скалярных параметров, оцененных по выборке.

Задание к лабораторной работе №4

Задание 1

1. Смоделировать выборку из биномиального распределения (функция rbinom()) с заданными параметрами. Оценить по выборке математическое ожидание, т.е. вычислить $\bar{X}$. Используя критерий Пирсона (хи-квадрат) проверить гипотезу о том, что выборка действительно соответствует моделируемому распределению.
2. Проверить гипотезу о том, что полученные при моделировании данные имеют распределение Пуассона с параметром $\lambda = \bar{X}$ (предположительно вторая гипотеза должна быть отвергнута критерием).

Задание 2

1. Для своего распределения (из лаб. работ 1-2) на уровне значимости $\alpha = 0.05$ проверить гипотезу о том, что выборка действительно соответствует моделируемому распределению (т. е. проверить качество моделирования). Использовать критерии Колмогорова и Пирсона (хи-квадрат).
3. Используя эти же два критерия, проверить гипотезу о принадлежности данной выборки нормальному закону (предположительно последняя гипотеза должна быть отвергнута обоими критериями). (В $\mathbb{R}$ есть функции chisq.test, norm.test(), проверяющие гипотезу о нормальности)

* *

Пример выполнения задания 2[1] **(*экспоненциальное распределение*)** (в *MATLAB*)

clc % Очистка командного окна
$N = 500$; % Задание объема выборки
$L = 0.5$; % Задание параметра $\lambda > 0$ показательного распределения
$U = rand(1, N)$; % Герерирование выборки из равномерного на $[0, 1]$
% распределения
$X = -log(U)/L$; % преобразование в выборку из показательного закона
% **Проверка гипотезы согласия с помощью критерия Колмогорова**
$Y = sort(X)$; % Получение вариационного ряда (упорядочиваем данные)
$F = 1 - exp(-L * Y)$; %З начения функции $F(x)$ в точках $X_{k:N}$
$Z = 1 : N$; % Формируем вектор (1,2,. . . ,N)
$Z_1 = abs(Z/N - F)$; % Вектор разностей
% $Z_1 = (|\frac{1}{N} - F(X_{(1)})|, \ldots, |\frac{N}{N} - F(X_{(N)}|)$

[1]В приводимом примере пункт 3 (проверка гипотезы о нормальности выборки) выполнен с использованием критерия хи-квадрат. Проверку с помощью критерия Колмогорова произвести самостоятельно.

% Формируем вектор разностей $Z_2 = (|F(X_{(1)}) - \frac{0}{N}|, \ldots, |F(X_{(N)}) - \frac{N-1}{N}|):$
$Z_2 = abs(F - (Z-1)/N);$
$D_{N1} = max(Z_1);\quad D_{N2} = max(Z_2);$ % Вычисление значений D_{N1} и D_{N2}
$W = [D_{N1}, D_{N2}];$ % Создаем вектор $W = (D_{N1}, D_{N2})$
$D_N = max(W);$ % Вычисляем статистику Колмогорова D_N
$\alpha = 0.05$ % Назначаем уровень значимости α
$u_{1-\alpha} = 1.35$ % Квантиль уровня $1-\alpha$ находим по таблицам распределения Колмогорова
$if\quad (sqrt(N) * D_N < u_{1-\alpha})$
disp('Гипотеза H_0 о принадлежности данных к показательному распределению принимается')
else
disp('Гипотеза H_0 о принадлежности данных к показательному распределению отвергается')
end
% **Проверка гипотезы H_0 с помощью критерия хи-квадрат**
$k = fix(1.72 * N^{1/3})$ % Число интервалов разбиения
$V_k = (1:(k-1))/k;$ % На отрезке $[0,1]$ создаем вектор $(1/k, 2/k, \ldots, (k-1)/k)$
% Разбиваем область значений случайной величины ξ на k промежутков,
% вероятность попадания значения ξ в каждый из которых равна $1/k:$
$A_k = -\log(1 - V_k)/L;$
% Вычисляем количество X_i, попавших в каждый из интервалов разбиения
$Ch = zeros(1,k);$ % Начальное значение вектора частот полагаем нулевым
$for\quad i = 2:(k-1)$
$Vec = (Y >= A_k(i-1)\quad \&\quad Y < A_k(i));$ % вектор Vec имеет ту же длину,
% что и Y, его элементы равны 0, если указанное условие не выполнено
% и 1, если оно выполнно
$Ch(i) = sum(Vec);$ % Вычисляем число наблюдений в i-м интервале
end;
$Ch(1) = sum(Y < A_k(1));$ % Число наблюдений в 1-м интервале
$Ch(k) = N - sum(Ch);$ % Число наблюдений в последнем k-м интервале
$P = ones(1,k)/k;$ % Вектор $(1/k, \ldots, 1/k)$ гипотетических вероятностей
% попадания в интервалы (согласно построению разбиения)
% Вычисление статистики Пирсона
$Pirson = sum((Ch - N * P).\hat{}(2)./(N * P))$
$kvantil = input$('Введите квантиль хи-квадрат ='); %Запрос из программы на введение квантили уровня $1-\alpha$ распределения хи-квадрат с числом степеней свободы $(k-1)$ (найти по таблицам и ввести по запросу *input*)
$if\quad (Pirson >= kvantil)$
disp('Гипотеза о принадлежности данных показательному распределению на уровне значимости $\alpha = 0.05$ отвергается')
else
disp('Гипотеза о принадлежности данных показательному распределению не противоречит опытным данным')
end
% **Проверка гипотезы о нормальности распределения**[2]
% Инверсия нормального закона доступна только при установке *MatLAB* со статистическим ***toolbox***, поэтому мы будем использовать другое разбиение.

[2] Данная гипотеза не должна подтвердиться, поскольку моделируемое распределение отличается от нормального.

$h = (Y(N) - Y(1))/k$; % h – шаг разбиения, k – то же, что прежде
$A_k = Y(1) : h : Y(N)$; % Вектор границ интервалов разбиения
% Из A_k формируем векторы $A_{k1} = (a_1, \ldots, a_k)$ и $A_{k2} = (a_2, \ldots, a_{k+1})$
$A1 = A_k(1 : k)$; % Вектор координат нижних границ интервалов разбиения
$A2 = A_k(2 : (k + 1))$; % Вектор координат верхних границ интервалов
$mu = mean(X)$; % Оценка математического ожидания
$sigma = std(X))$; % Оценка среднего квадратического отклонения
% Проверяем гипотезу H_0: $F(x) \equiv \Phi((x - \mu)/\sigma)$, где $\mu = mu$, $\sigma = sigma$.
% Для вычисления гипотетических вероятностей попадания в i-й интервал $1 \leq i \leq k$ по формуле $\Phi(\frac{a_i - \mu}{\sigma}) - \Phi(\frac{a_{i-1} - \mu}{\sigma})$ нормируем координаты границ
$An1 = (A1 - mu)/sigma$; % Нормирование нижних границ
$An2 = (A2 - mu)/sigma$; % Нормирование верхних границ
% Формируем вектор гипотетических вероятностей[3] $P = (p_1, \ldots, p_k)$:
$P = (erf(An2/sqrt(2)) - erf(An1/sqrt(2)))/2$;
% Вычисляем количество наблюдений X_i, попавших в каждый из интервалов
$Ch = zeros(1, k)$;
$for \quad i = 1 : k$
$Vec = (Y >= A1(i) \;\; \& \;\; Y < A2(i))$;
$Ch(i) = sum(V)$;
end
$Ch(k) = Ch(k) + 1$; % Учитываем максимальное наблюдение $Y(N) = X_{N:N}$
% Вычисление статистики Пирсона
$Pirson = sum((Ch - N * P).\hat{}(2)./(N * P))$
$kvantil = input$('Введите квантиль уровня $1 - \alpha$ распределения хи-квадрат
с $k - 1$ степенями свободы');
$if \quad (Pirson >= kvantil)$
disp('Гипотеза о принадлежности данных нормальному закону на данном уровне значимости отвергается')
$else$
disp('Гипотеза о принадлежности данных нормальному закону не противоречит опытным данным')
end

[3] При вычислении разности значений гипотетической (нормальной с параметрами μ и σ^2) функции распределения на концах интервалов разбиения мы используем функцию $erf(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt$, которая входит в стандартный набор функций *MatLAB*, чтобы вычислить требующиеся нам значения функции Гаусса $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp(-x^2/2)\, dx$. В программе используется соотношение: $\Phi(x) = \frac{1}{2}\left(1 + erf\left(\frac{x}{\sqrt{2}}\right)\right)$.