Практикум по математической статистике

А.Н. Баженов

22 марта 2024 г.

Глава 1

Проверка гипотез

1.1 Проверка гипотезы о законе распределения генеральной совокупности. Метод хи-квадрат

Исчерпывающей характеристикой изучаемой случайной величины является её закон распределения. Поэтому естественно стремление исследователей построить этот закон приближённо на основе статистических данных.

Сначала выдвигается гипотеза о виде закона распределения.

После того как выбран вид закона, возникает задача оценивания его параметров и проверки (тестирования) закона в целом.

Для проверки гипотезы о законе распределения применяются критерии согласия. Таких критериев существует много. Мы рассмотрим наиболее обоснованный и наиболее часто используемый в практике — критерий χ^2 (хи-квадрат), введённый К.Пирсоном (1900 г.) для случая, когда параметры распределения известны. Этот критерий был существенно уточнён Р.Фишером (1924 г.), когда параметры распределения оцениваются по выборке, используемой для проверки.

Мы ограничимся рассмотрением случая одномерного распределе-

ния.

Итак, выдвинута гипотеза H_0 о генеральном законе распределения с функцией распределения $F(x)$.

Рассматриваем случай, когда гипотетическая функция распределения $F(x)$ не содержит неизвестных параметров.

Разобьём генеральную совокупность, т.е. множество значений изучаемой случайной величины X на k непересекающихся подмножеств $\Delta_1, \Delta_2, \ldots, \Delta_k$.

Пусть $p_i = P(X \in \Delta_i), \quad i = 1, \ldots, k.$

Если генеральная совокупность — вся вещественная ось, то подмножества $\Delta_i = (a_{i-1}, a_i]$ — полуоткрытые промежутки $(i = 2, \ldots, k-1)$. Крайние промежутки будут полубесконечными: $\Delta_1 = (-\infty, a_1]$, $\Delta_k = (a_{k-1}, +\infty)$. В этом случае $p_i = F(a_i) - F(a_{i-1})$; $a_0 = -\infty$, $a_k = +\infty$ $(i = 1, \ldots, k)$.

Отметим, что $\sum\limits_{i=1}^{k} p_i = 1$. Будем предполагать, что все $p_i > 0$ $(i = 1, \ldots, k)$.

Пусть, далее, $n_1, n_2, \ldots, n_k$ — частоты попадания выборочных элементов в подмножества $\Delta_1, \Delta_2, \ldots, \Delta_k$ соответственно.

В случае справедливости гипотезы H_0 относительные частоты n_i/n при большом n должны быть близки к вероятностям p_i $(i = 1, \ldots, k)$, поэтому за меру отклонения выборочного распределения от гипотетического с функцией $F(x)$ естественно выбрать величину

$$\mathrm{Z} = \sum_{i=1}^{k} c_i \left(\frac{n_i}{n} - p_i \right)^2, \tag{1.1}$$

где c_i — какие-нибудь положительные числа (веса). К.Пирсоном в качестве весов выбраны числа $c_i = n/p_i$ $(i = 1, \ldots, k)$. Тогда получается статистика критерия хи-квадрат К.Пирсона

$$\chi^2 = \sum_{i=1}^{k} \frac{n}{p_i} \left(\frac{n_i}{n} - p_i \right)^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}, \tag{1.2}$$

которая обозначена тем же символом, что и закон распределения хи-квадрат.

К.Пирсоном доказана теорема об асимптотическом поведении статистики χ^2, указывающая путь её применения.

Теорема К.Пирсона. Статистика критерия χ^2 асимптотически распределена по закону χ^2 с $k-1$ степенями свободы.

Это означает, что независимо от вида проверяемого распределения, т.е. функции $F(x)$, выборочная функция распределения статистики χ^2 при $n \to \infty$ стремится к функции распределения случайной величины с плотностью вероятности

$$f_{k-1}(x) = \begin{cases} 0, & x \le 0; \\ \dfrac{1}{2^{\frac{k-1}{2}} \Gamma\left(\frac{k-1}{2}\right)} x^{\frac{k-3}{2}} e^{-\frac{x}{2}}, & x > 0. \end{cases} \tag{1.3}$$

Для прояснения сущности метода χ^2 сделаем ряд замечаний.

Замечание 1. Выбор подмножеств $\Delta_1, \Delta_2, \ldots, \Delta_k$ и их числа k в принципе ничем не регламентируется, так как $n \to \infty$. Но так как число n хотя и очень большое, но конечное, то k должно быть с ним согласовано. Обычно его берут таким же, как и для построения гистограммы, т.е. можно руководствоваться формулой

$$k \approx 1.72 \sqrt[3]{n} \tag{1.4}$$

или формулой Старджесса

$$k \approx 1 + 3.3 \lg n. \tag{1.5}$$

При этом, если $\Delta_1, \Delta_2, \ldots, \Delta_k$ — промежутки, то их длины удобно сделать равными, за исключением крайних — полубесконечных.

Замечание 2. (о числе степеней свободы).

Числом степеней свободы функции (по старой терминологии) называется число её независимых аргументов. Аргументами статистики χ^2 являются частоты $n_1, n_2, \ldots, n_k$. Эти частоты связаны одним равенством $n_1 + n_2 + \ldots + n_k = n$, а в остальном независимы в силу независимости элементов выборки. Таким образом, функция χ^2 имеет $k-1$ независимых аргументов: число частот минус одна связь. В силу теоремы Пирсона число степеней свободы статистики χ^2 отражается на виде асимптотической плотности $f_{k-1}(x)$.

На основе общей схемы проверки статистических гипотез сформулируем следующее правило.

1.1.1 Правило проверки гипотезы о законе распределения по методу χ^2

1. Выбираем уровень значимости α.
2. По таблице [6, с. 358] находим квантиль $\chi^2_{1-\alpha}(k-1)$ распределения хи-квадрат с $k-1$ степенями свободы порядка $1-\alpha$.
3. С помощью гипотетической функции распределения $F(x)$ вычисляем вероятности $p_i = P(X \in \Delta_i),\ i = 1, \ldots, k$.
4. Находим частоты n_i попадания элементов выборки в подмножества $\Delta_i,\ i = 1, \ldots, k$.
5. Вычисляем выборочное значение статистики критерия χ^2:
$$\chi^2_{\text{В}} = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}.$$
6. Сравниваем $\chi^2_{\text{В}}$ и квантиль $\chi^2_{1-\alpha}(k-1)$.
   а) Если $\chi^2_{\text{В}} < \chi^2_{1-\alpha}(k-1)$, то гипотеза H_0 на данном этапе проверки принимается.
   б) Если $\chi^2_{\text{В}} \geq \chi^2_{1-\alpha}(k-1)$, то гипотеза H_0 отвергается, выбирается одно из альтернативных распределений, и процедура проверки повторяется.

Замечание 3. Из формулы (1.2) видим, что веса $c_i = n/p_i$ пропорциональны n, т.е. с ростом n увеличиваются. Отсюда следует, что если выдвинутая гипотеза неверна, то относительные частоты n_i/n не будут близки к вероятностям p_i, и с ростом n величина χ^2_{B} будет увеличиваться. При фиксированном уровне значимости α будет фиксировано пороговое число — квантиль $\chi^2_{1-\alpha}(k-1)$, поэтому, увеличивая n, мы придём к неравенству $\chi^2_{\text{B}} > \chi^2_{1-\alpha}(k-1)$, т.е. с увеличением объёма выборки неверная гипотеза будет отвергнута.

Отсюда следует, что при сомнительной ситуации, когда $\chi^2_{\text{B}} \approx \chi^2_{1-\alpha}(k-1)$, можно попытаться увеличить объём выборки (например, в 2 раза), чтобы требуемое неравенство было более чётким.

Замечание 4. Теория и практика применения критерия χ^2 указывают, что если для каких-либо подмножеств Δ_i $(i = 1, \ldots, k)$ условие $np_i \geq 5$ не выполняется, то следует объединить соседние подмножества (промежутки).

Это условие выдвигается требованием близости величин

$$(n_i - np_i)/\sqrt{np_i},$$

квадраты которых являются слагаемыми χ^2 к нормальным $N(0,1)$. Тогда случайная величина в формуле (1.2) будет распределена по закону, близкому к хи-квадрат. Такая близость обеспечивается достаточной численностью элементов в подмножествах Δ_i [2, с. 481-485].

1.1.2 Задание

Мощность распределения n=20, 100.
Распределения: нормальное, Стьюдента (k=3), равномерное.
Провести исследование по методике 1.1.1.
Результаты вычислений оформить в виде таблицы.

1.2 Проверка гипотезы однородности выборки

Если выборка близка к нормальной, можно проверить, насколько она однородна.

1.2.1 F-тест или критерий Фишера

F-тест или критерий Фишера (F-критерий, $\varphi^\star$-критерий) — статистический критерий, тестовая статистика которого при выполнении нулевой гипотезы имеет распределение Фишера (F-распределение).

В общм случае сравниваются две выборки A и B и проверяется равенство их дисперсий:

$$F = \frac{\sigma_A^2}{\sigma_D^2}. \tag{1.6}$$

Нулевая гипотеза H_0: $\sigma_A^2 = \sigma_B^2$.

Статистика теста так или иначе сводится к отношению выборочных дисперсий (сумм квадратов, деленных на «степени свободы»). Чтобы статистика имела распределение Фишера, необходимо, чтобы числитель и знаменатель были независимыми случайными величинами и соответствующие суммы квадратов имели распределение χ^2. Для этого требуется, чтобы данные имели нормальное распределение.

Статистика критерия Фишера (1.6) имеет *распределение Фишера* с $n-1$ и $m-1$ степенями свободы $F(n-1, m-1)$. Обычно в числителе ставится большая из двух сравниваемых дисперсий.

Применим тест Фишера для выяснения однородности конкретной выборки. Для этого дисперсию выборки можно оценить двумя способами.
Во-первых, дисперсия, вычисленная для каждой группы — это оценка дисперсии совокупности. Поэтому дисперсию совокупности можно оценить на основании групповых дисперсий. Такая оценка не будет зависеть от различий групповых средних.
Во-вторых, разброс выборочных средних тоже позволяет оценить дисперсию совокупности. Понятно, что такая оценка дисперсии зависит от различий выборочных средних.

В качестве оценки дисперсии совокупности возьмем среднее выборочных дисперсий. Эта оценка называется *внутригрупповой* дисперсией. Обозначим ее s^2_{in}

$$s^2_{in} = \frac{1}{k}\sum_{i=1}^{k} s_i^2 = \frac{1}{k}\sum_{i=1}^{k} \frac{\sum_{j=1}^{n}(x_{ij} - \hat{X})^2}{k-1} \tag{1.7}$$

где $\hat{X}$ — среднее для части выборки, k — количество частей, на которое делим сходную выборку, n — количество элементов в подвыборке.

Также нужно вычислить *межгрупповую* дисперсию. Обозначается она s^2_{out}.

Вычисление межгрупповой дисперсии происходит в несколько этапов:
1. Вычисление среднего значения для всех выбранных подвыборок $(\overline{X}_1, \ldots, \overline{X}_k)$
2. Вычисление среднего этих средних: $\overline{X} = \frac{1}{k}\sum_{i=1}^{k} \overline{X}_i$
3. Вычисление межгрупповой дисперсии

$$s^2_{out} = k \cdot \frac{\sum_{j=1}^{n}(\overline{X}_i - \overline{X})^2}{k-1}. \tag{1.8}$$

Окончательно имеем для проверки гипотезы

$$F = \frac{s^2_{out}}{s^2_{in}}. \tag{1.9}$$

Если значение $F \approx 1$ (1.9), выборку можно считать однородной.

1.2.2 Правило Стерджесса

Данное правило используется для определения оптимального количества интервалов, на которые разбивается наблюдаемый диапазон изменения случайной величины при изучении ее распределения. По этому правилу число интервалов считается по формуле:

$$n = 1 + \log_2 N, \tag{1.10}$$

где n — число интервалов, N — общее число наблюдений.

1.2.3 Пример

На рисунке 1.1 представлены экспериментальные данные по калибровке прибора.

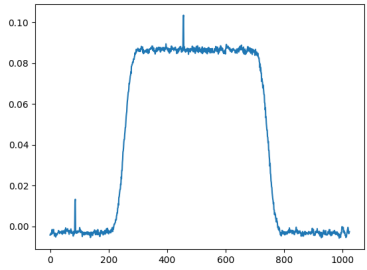

Рис. 1.1. Исходные данные

Необходимо разделить сданные на содержательные подобласти: сигнал, фон, переходные процессы.

Предварительная обработка.

Известно, что наибольшее количество данных приходится на фон (базовая линия). Остальная часть представляет собой относительно длинный сигнал и более короткие переходные процессы.

Выделение областей производится с помощью построения гистограммы, которая представлена на рисунке 1.2.

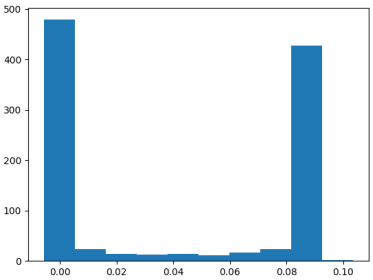

Рис. 1.2. Гистограмма данных

Столбец гистограммы, в котором наибольшее число значений, отвечает за принадлежность к фону, следующий по величине столбец – за

сигнал, а остальное – переходные процессы.

Прежде чем разделить сигнал на области однородности, необходимо определить наличие выбросов и сгладить их. Известно, что выбросы представляют собой оцень короткие (1 отсчет) резкие изменения уровня сигнала.

Поэтому фильтрацию можно провести с помощью медианного фильтра, при этом значение выброса становится равным среднему арифметическому его соседних элементов.

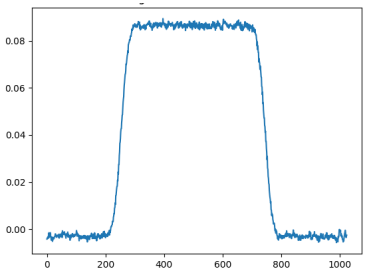

Рис. 1.3. Данные после медианного фильтра

Результат сглаживания представлен на рисунке 1.3.

С помощью гистограммы теперь можно разделить сигнал на разные области, которые представлены на рисунке 1.4.

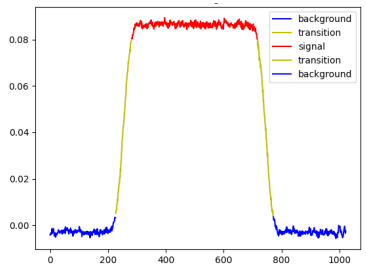

Рис. 1.4. Данные после медианного фильтра с разметкой

Исходный сигнал после разделения на области представлен на рисунке 1.5.

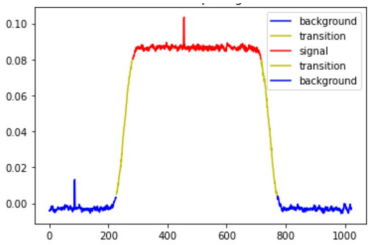

Рис. 1.5. Данные с разметкой

Номер	Тип	k (1.10)	Значение (1.8)	Значение (1.7)	Значение F (1.9)
1	фон				
2	переходный процесс				
3	сигнал				
4	переходный процесс				
5	фон				

Таблица 1.1. Вычисления для теста Фишера

Вычисление теста Фишера.

Чтобы определить однородность каждой части сигнала, необходимо применить критерий Фишера.

Результаты вычисления представлены в таблице.

Значение критерия Фишера для левого фона, правого фона и сигнала . . . , значит эти части общего сигнала однородны, а значение критерия Фишера для переходных процессов больше . . . , то есть эти части общего сигнала неоднородны.

В проделанной работе был рассмотрен один из сигналов и проанализирован на однородность с помощью критерия Фишера.

1.2.4 Задание

Из файла `wave_ampl.txt` выбрать случайно часть данных длиной 1024 отсчёта (всего в файле 800 таких массивов).

Провести исследование по методике §1.2.3.

Результаты вычислений оформить в виде таблицы 1.1.

Литература

[1] Histogram. URL: `https://en.wikipedia.org/wiki/Histogram`

[2] Вероятностные разделы математики. Учебник для бакалавров технических направлений.//Под ред. Максимова Ю.Д. — Спб.: «Иван Федоров», 2001. — 592 с., илл.

[3] Box plot. URL: `https://en.wikipedia.org/wiki/Box_plot`

[4] Анатольев, Станислав (2009) «Непараметрическая регрессия», Квантиль, №7, стр. 37-52.

[5] Вентцель Е.С. Теория вероятностей: Учеб. для вузов. — 6-е изд. стер. — М.: Высш. шк., 1999.— 576 с.

[6] Максимов Ю.Д. Математика. Теория и практика по математической статистике. Конспект-справочник по теории вероятностей : учеб. пособие / Ю.Д. Максимов; под ред. В.И. Антонова. — СПб. : Изд-во Политехн. ун-та, 2009. — 395 с. (Математика в политехническом университете).

[7] Максимов ...

[8] Обработка и анализ данных с интервальной неопределённостью

[9] Gardeñes E., Trepat A., Janer J.M. Approaches to simulation and to the linear problem in the SIGLA system // Freiburger Intervall-Berichte. – 1981. – No. 8. – S. 1–28.

[10] Нестеров В.М. Твинные арифметики и их применение в методах и алгоритмах двустороннего интервального оценивания. дисс. ... д.ф.-м.н. Санкт-Петербург, Санкт-Петербургский институт информатики и автоматизации РАН, 1999, 234 с.

[11] арифметика твинов Нестерова
`twin` — А.Жаворонкова https://github.com/Zhavoronkova-Alina/twin

[12] Н.Мордовин Библиотека FuzzyNumbers https://github.com/MordovinNik/FuzzyNumbers