Санкт-Петербургский политехнический университет Петра Великого
Институт прикладной математики и механики
Кафедра «Прикладная математика»

ОТЧЁТ ПО ЛАБОРАТОРНОЙ РАБОТЕ №7 ПО ДИСЦИПЛИНЕ «МАТЕМАТИЧЕСКАЯ СТАТИСТИКА»

Выполнил
студент группы 3630102/70401

Мельникова Анна Николаевна

Проверил
к. ф.-м. н., доцент

Баженов Александр Николаевич

Санкт-Петербург
2020

Содержание

Список таблиц

Список иллюстраций

1 Постановка задачи

Сгенерировать выборку объёмом 100 элементов для нормального распределения $N(x, 0, 1)$. По сгенерированной выборке оценить параметры μ и σ нормального закона методом максимального правдоподобия. В качестве основной гипотезы H_0 будем считать, что сгенерированное распределение имеет вид $N(x, \hat{\mu}, \hat{\sigma})$. Проверить основную гипотезу, используя критерий согласия χ^2. В качестве уровня значимости взять $\alpha = 0.05$. Привести таблицу вычислений χ^2.

2 Теория

2.1 Метод максимального правдоподобия

Одним из универсальных методов оценивания является метод максимального правдоподобия, предложенный Р.Фишером (1921).

Пусть $x_1, ..., x_n$ — случайная выборка из генеральной совокупности с плотностью вероятности $f(x, \theta)$; $L(x_1, ..., x_n, \theta)$ — функция правдоподобия (ФП), представляющая собой совместную плотность вероятности независимых с.в. $x_1, ..., x_n$ и рассматриваемая как функция неизвестного параметра θ:

$$L(x_1, ..., x_n, \theta) = f(x_1, \theta) f(x_2, \theta) ... f(x_n, \theta) \tag{1}$$

Определение. Оценкой максимального правдоподобия (о.м.п.) будем называть такое значение $\hat{\theta}_{ml}$ из множества допустимых значений параметра θ, для которого ФП принимает наибольшее значение при заданных $x_1, ..., x_n$:

$$\hat{\theta}_{ml} = \arg\max_{\theta} L(x_1, ..., x_n, \theta) \tag{2}$$

Если ФП дважды дифференцируема, то её стационарные значения даются корнями уравнения

$$\frac{\partial L(x_1, ..., x_n, \theta)}{\partial \theta} = 0$$

Достаточным условием того, чтобы некоторое стационарное значение $\tilde{\theta}$ было локальным максимумом, является неравенство

$$\frac{\partial^2 L}{\partial \theta^2}(x_1, ..., x_n, \tilde{\theta}) < 0$$

Определив точки локальных максимумов ФП (если их несколько), находят наибольший, который и даёт решение задачи (1). Часто проще искать максимум логарифма ФП, так как он имеет максимум в одной точке с ФП:

$$\frac{\partial \ln L}{\partial \theta} = \frac{1}{L} \frac{\partial L}{\partial \theta}, L > 0$$

и соответственно решать уравнение

$$\frac{\partial \ln L}{\partial \theta} = 0$$

которое называют *уравнением правдоподобия*. В задаче оценивания векторного параметра $\theta = (\theta_1, ..., \theta_m)$ аналогично (2) находится максимум ФП нескольких аргументов:

$$\hat{\theta} = \arg \max_{\theta_1, \theta_2 ... \theta_m} L(x_1, x_2, ... x_n, \theta_1, \theta_2, ... \theta_m)$$

и в случае дифференцируемости ФП выписывается система уравнений правдоподобия

$$\frac{\partial L}{\partial \theta_k} = 0 \text{ или } \frac{\partial \ln L}{\partial \theta_k} = 0, k = 1, ..m$$

Пример: Вывод оценок математического ожидания μ и дисперсии σ^2 нормального распределения $N(\mu, \sigma^2)$. Для этого составим функцию правдоподибия:

$$L(x_1, ..., x_n, \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} exp\left[-\frac{(x_i - \mu)^2}{2\sigma}\right] = (2\pi\sigma^2)^{-n/2} exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right],$$

$$\ln L = -\frac{n}{2}\ln 2\pi - \frac{1}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

Получим такие уравнения правдоподобия:

$$\begin{cases} \frac{\partial \ln L}{\ln \mu} = \frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}(x_i - \hat{\mu}) = \frac{n}{\hat{\sigma}^2}(\overline{x} - \hat{\mu}) = 0, \\ \frac{\partial \ln L}{\partial(\sigma^2)} = -\frac{n}{2\hat{\sigma}^2} + \frac{1}{2(\hat{\sigma}^2)^2}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{n}{2(\hat{\sigma}^2)^2}\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 - \hat{\sigma}^2\right] = 0, \end{cases}$$

откуда следует, что:

- Выборочное среднее $\overline{x}$ - оценка максимального правдоподобия математического ожидания: $\hat{\mu}_{ml} = \overline{x}$,
- Выборочная дисперсия $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$ - оценка максимального правдоподобия генеральной дисперсии: $\hat{\sigma}^2_{ml} = s^2$[1, с. 442 - 444].

Эти результаты мы используем при выполнении задания.

2.2 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 распределения.

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 $\Delta_1, \Delta_2, ..., \Delta_k$.

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

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

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

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

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

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

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

$$\chi^2 = \sum_{i=1}^{k} \frac{n}{p_i} \left(\frac{n_i}{n} - p_i\right)^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}, \qquad (3)$$

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

$$f_{k-1}(x) = \begin{cases} 0\,,\, x \le 0 \\ \frac{1}{2^{\frac{k-1}{2}}\Gamma(\frac{k-1}{2})} x^{\frac{k-3}{2}} e^{-\frac{x}{2}}\,,\, x > 0 \end{cases}$$

Для прояснения сущности метода χ^2 сделаем ряд замечаний.

Замечание 1. Выбор подмножеств $\Delta_1, \Delta_2, ..., \Delta_k$ 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

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

или формулой Старджесса

$$k \approx 1 + 3.3lgn$$

При этом, если $\Delta_1, \Delta_2, ..., \Delta_k$ — 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 $n_1, n_2, ..., n_k$. Эти частоты связаны одним равенством $n_1 + n_2 + ... + n_k = n$, 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.

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

1. Выбираем уровень значимости α.

2. По таблице [3, 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 $p_i = P(X \in \Delta_i), i = 1, ..., k$.

4. Находим частоты n_i попадания элементов выборки в подмножества $\Delta_i, i = 1, ..., k$.

5. Вычисляем выборочное значение статистики критерия χ^2:

$$\chi^2_B = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}.$$

6. Сравниваем χ^2_B и квантиль $\chi^2_{1-\alpha}(k-1)$:

   - Если $\chi^2_B < \chi^2_{1-\alpha}$(k − 1), то гипотеза H_0 на данном этапе проверки принимается.
   - Если $\chi^2_B \geq \chi^2_{1-\alpha}(k-1)$, то гипотеза H_0 отвергается, выбирается одно из альтернативных распределений, и процедура проверки повторяется.

Замечание 3. Из формулы (3) 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 $\chi^2_B > \chi^2_{1-\alpha}(k-1)$, т.е. с увеличением объёма выборки неверная гипотеза будет отвергнута.
Отсюда следует, что при сомнительной ситуации, когда $\chi^2_B \approx \chi^2_{1-\alpha}(k-1)$, 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 $\Delta_i (i = 1, ..., k)$ условие $np_i \geq 5$ не выполняется, то следует объединить соседние подмножества (промежутки).
Это условие выдвигается требованием близости величин $\frac{(n_i - np_i)}{\sqrt{np_i}}$, квадраты которых являются слагаемыми χ^2 к нормальным $N(0, 1)$. Тогда случайная величина в формуле (3) будет распределена по закону, близкому к хи-квадрат. Такая близость обеспечивается достаточной численностью элементов в подмножествах Δ_i [1, с. 481-485].

3 Реализация

Лабораторная работа выполнена с помощью встроенных средств языка программирования Python в среде разработки PyCharm (для визуализации использовался пакет matplotlib, для вычислений - numpy). Исходный код лабораторной работы и tex-файлы отчета приведёны в приложении в виде ссылки на репозиторий GitHub.

4 Результаты и описание выполненной работы

Для нормального распределения согласно выведенным в примере(2.1) оценкам максимального правдоподобия:

- $\hat{\mu} = \overline{x}$
- $\hat{\sigma}^2 = s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$

Тогда $\hat{\mu} \approx 0.07, \hat{\sigma} \approx 0.9$.
И проверяем гопотезу H_0, что сгенерированная по закону $N(x, 0, 1)$ 100-элементная выборка подчиняется закону $N(x, \hat{\mu}, \hat{\sigma})$. Используем критерий согласия χ^2:

- $\alpha = 0.05$ - уровень значимости,
- $n = 100$ - размер выборки,
- $k := \lfloor 1 + 3.3 \lg 100 \rfloor = \lfloor 7.6 \rfloor = 7$ - количество промежутков,
- Квантиль $\chi^2_{1-\alpha}(k-1) = \chi^2_{0.95}(6) = //$ из таблицы[3, с. 358]// ≈ 12.59

i	$\Delta_i = [a_{i-1}, a_i)$	n_i	p_i	np_i	$n_i - np_i$	$(n_i - np_i)^2/(np_i)$
1	($-\infty$, -1.5)	3	0.0401	4.01	-1.01	0.25
2	[-1.5, -0.9)	11	0.0996	9.96	1.04	0.11
3	[-0.9, -0.3)	16	0.1998	19.98	-3.98	0.79
4	[-0.3, 0.3)	33	0.2608	26.08	6.92	1.84
5	[0.3, 0.9)	20	0.2215	22.15	-2.15	0.21
6	[0.9, 1.5)	10	0.1224	12.24	-2.24	0.41
7	[1.5, $+\infty$)	7	0.0559	5.59	1.41	0.35
$\sum$	($-\infty$, $+\infty$)	100	1	100	0	3.96 = χ^2_B

Таблица 1: Вычисление χ^2_B при проверке гипотезы H_0 о нормальном законе распределения $N(x, \hat{\mu}, \hat{\sigma})$.

Сравним $\chi^2_{0.95}(6) \approx 12.59$ и найденное $\chi^2_B \approx 3.96 : 12.59 > 3.96$. Следовательно гипотезу H_0 на данном этапе проверки можно принять.

5 Обсуждение

По результатам проверки на близость с помощью критерия хи-квадрат можно принять гипотезу H_0 о нормальном распределении $N(x, \hat{\mu}, \hat{\sigma})$ на уровне значимости $\alpha = 0.05$ для выборки, сгенерированной согласно $N(x, 0, 1)$. То есть если взять для гипотезы нормальное распределение с параметрами сдвига и масштаба равными оценкам максимального правдоподобия для μ, σ, вычисленным по выборке $\sim N(x, 0, 1)$, то критерий отразит эту согласованность.

6 Приложения

Код программы - GitHub URL: https://github.com/kaustika/Statistics2020

Список литературы

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

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

[3] 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).