Санкт-Петербургский политехнический университет
Петра Великого

Институт прикладной математики и механики
Кафедра «Прикладная математика»

Отчет по лабораторной работе № 7
по дисциплине: Математическая статика.

Выполнила студентка:
Заболотских Екатерина Дмитриевна
группа: 3630102/70301

Проверил:
к.ф.-м.н., доцент
Баженов Александр Николаевич

Санкт-Петербург
2020 г.

Оглавление

Список таблиц

Постановка задачи

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

Дополнительное задание:
Создать выборку распределения Лапласа мощностью 20 событий и проверь гипотезу ее «нормальности»

Теория

Метод максимального правдоподобия

Пусть $x_1, \dots, x_n$ – случайная выборка из генеральной совокупности с плотностью вероятности $f(x, \theta)$; L ($x1$, ..., xn, θ) – функция правдоподобия (ФП), представляющая собой совместную плотность вероятности независимых с. в. $x_1, \dots, x_n$ и рассматриваемая как функция неизвестного параметра θ:

$$L(x_1, \dots, x_n, \theta) = f(x_1, \theta) f(x_2, \theta) \dots f(x_n, \theta) \quad (1)$$

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

$$\hat{\theta}_{\text{мп}} = \arg\max_{\theta} L(x_1, \dots, x_n, \theta) \quad (2)$$

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

$$\frac{\partial L(x_1, \dots, x_n, \theta)}{\partial \theta} = 0 \quad (3)$$

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

$$\frac{\partial^2 L}{\partial \theta^2}\left(x_1, \dots, x_n, \tilde{\theta}\right) < 0 \quad (4)$$

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:

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

и соответственно решать уравнение

$$\frac{\partial lnL}{\partial \theta} = 0, \quad (5)$$

Которое называют уравнением правдоподобия.

В задаче оценивания векторного параметра θ = (θ1, ..., θm) аналогично (2) находится максимум ФП нескольких аргументов:

$$\hat{\theta}_{\text{мп}} = arg \max_{\theta_1, \dots, \theta_m} L(x_1, \dots, x_n, \theta_1, \dots, \theta_m) \quad (6)$$

и в случае дифференцируемости ФП выписывается система уравнений правдоподобия

$$\frac{\partial L}{\partial \theta_k} = 0 \text{ или } \frac{\partial LnL}{\partial \theta_k} = 0, k = 1, \dots, m \quad (7)$$

Критерий согласия χ^2

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 χ на *k* непересекающихся подмножеств $\Delta_1, \dots, \Delta_k$. Пусть $pi = P(X \in \Delta_i), i = 1, \dots, k.$

Если генеральная совокупность – вся вещественная ось, то подмножества – полуоткрытые промежутки. Крайние промежутки будут полу бесконечными:

$$\Delta_i = (a_{i-1}, a_i], i = 2, \dots, k-1,\ \Delta_1 = (-\infty, a_1],\ \Delta_k = (a_{k-1}, +\infty). \quad (8)$$

В этом случае $pi = F(a_i) - F(a_{i-1}); a_0 = -\infty, a_k = +\infty\ (i = 1, \dots, k)$.

Пусть $n_1, n_2, \dots, n_k$ – частоты попадания выборочных элементов в подмножества $\Delta_1, \dots, \Delta_k$ соответственно.В случае справедливости гипотезы H_0 относительные частоты $\frac{n_i}{n}$ при большом n должны быть близки к вероятностям $p_i\ (i = 1, \dots, k)$, поэтому за меру отклонения выборочного распределения от гипотетического с функцией $F(x)$ естественно выбрать величину

$$Z = \sum_{i=1}^{n} c_i \left(\frac{n_i}{n} - p_i\right)^2, \quad (9)$$

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

$$\chi^2 = \sum_{i=1}^{k} \frac{n}{p_i}\left(\frac{n_i}{n} - p_i\right)^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}, \quad (10)$$

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

$$f_{k-1}(x) = \begin{cases} 0, & x \le 0 \\ \dfrac{1}{2^{\frac{k-1}{2}}\,\Gamma(\frac{k-1}{2})} x^{\frac{k-3}{2}} e^{-\frac{x}{2}}, & x > 0 \end{cases} \quad (11)$$

Для пояснения сущности метода χ^2 сделаем ряд замечаний.

Замечание 2.: Выбор подмножеств $\Delta_1, \dots, \Delta_k$ 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 Райса

$$k \approx 1.72 \sqrt[3]{n} \quad (12)$$

Или формулой Старджесса

$$k \approx 1 + 3.3\, lg\, n \quad (13)$$

При этом, если $\Delta_1, \Delta_2, \dots, \Delta_k$ – промежутки, то их длины удобно сделать равными за исключением крайних – полу бесконечных.

В данной работе применялось правило Скотта для ширины (считаем все интервалы кроме крайних одинаковой ширины):

$$a_i = med\, \mathcal{N}(\widehat{\mu}, \widehat{\sigma}) + \left(i - \frac{k-1}{2}\right) h, \text{где } h = \frac{3.49\widehat{\sigma}}{\sqrt[3]{n}} \quad (14)$$

Замечание 2. (о числе степеней свободы).

Числом степеней свободы функции (по старой терминологии) называется число её независимых аргументов. Аргументами статистики χ^2 являются частоты $n_1, n_2, \dots, n_k$. Эти частоты связаны одним равенством $n_1 + n_2 + \dots + n_k = n$, 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.

Выбираем значимости α.

По таблице находим квантиль $\chi^2_{1-\alpha}(k-1)$ распределения хи-квадрат с k-1 степенями свободы порядка $1-\alpha$.

С помощью гипотетической функции распределения $F(x)$ вычисляем вероятности $pi = P(X \in \Delta_i), i = 1, \dots, k$.

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

Вычисляем выборочное значение статистики критерия χ^2 :

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

Сравниваем χ_B^2 и квантиль $\chi_{1-\alpha}^2(k-1)$.

Если $\chi_B^2 < \chi_{1-\alpha}^2(k-1)$, то гипотеза H_0 на данном этапе проверки принимается.

Если $\chi_B^2 \geq \chi_{1-\alpha}^2(k-1)$, то гипотеза H_0 отвергается, выбирается одно из альтернативных распределений, и процедура проверки повторяется.

Замечание 3. Из формулы (10) 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_B^2 будет увеличиваться. При фиксированном уровне значимости α будет фиксированное пошаговое число – квантиль $\chi_{1-\alpha}^2(k-1)$, поэтому, увеличивая n, мы придём к неравенству $\chi_B^2 > \chi_{1-\alpha}^2(k-1)$, т.е. с увеличением объема выборки неверная гипотеза будет отвергнута.

Отсюда следует, что при сомнительной ситуации, когда $\chi_B^2 \approx \chi_{1-\alpha}^2(k-1)$, можно попытаться увеличить объем 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 $(i = 1, \dots, k)$ 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 (10) будет распределена по закону, близкому к хи-квадрат. Такая близость обеспечивается достаточной численностью элементов в подмножествах Δ_i.

Реализация

Код программы, реализующий данную задачу, был написан на языке Python в интегрированной среде разработке PyCharm.

Были использованы библиотеки:

- **Numpy** – библиотека для работы с данными.
- **SciPy** – модуль “stats” для генерации данных и работы с распределениями

Результаты

$\widehat{\mu} = -0.09$ $\widehat{\sigma} = 0.9$ $\alpha = 0.5$

Число промежутков: $1.72\sqrt[3]{100} = 8$

Таблица вычислений χ^2:

i	Δ_i	n_i	p_i	$n_i - np_i$
1	$(-\infty;\ -2.48]$	0	0.0042	-0.4249
2	$(-2.48;\ -1.8]$	2	0.0258	-0.5823
3	$(-1.8;\ -1.12]$	15	0.0996	5.0379
4	$(-1.12;\ -0.44]$	16	0.2238	-6.3785
5	$(-0.44; 0.25]$	34	0.293	4.6955
6	$(0.25;\ 0.93]$	18	0.2238	-4.3785
7	$(0.93;\ 1.61]$	13	0.0996	3.0379
8	$(1.61;\ \infty)$	2	0.0301	-1.0071

Таблица 1: таблица вычислений χ^2

Имеем $\chi^2_{1-\alpha}(k-1) \approx 14.0671$, а вычисленное $\chi^2_{\text{в}} = 7.7947$. Видно, что $\chi^2_{\text{в}} < \chi^2_{1-\alpha}(k-1)$.

Дополнительное задание:

Для выборки в 20 элементов, сгенерированной по распределению Лапласа $\mathcal{L}(0,\ \frac{1}{\sqrt{2}})$ имеем $\chi^2_{1-\alpha}(k-1) \approx 9.4877$, а вычисленное $\chi^2_{\text{в}} = 1.1365$. Видно, что $\chi^2_{\text{в}} < \chi^2_{1-\alpha}(k-1)$. Критерий дает вывод, что генеральное распределение является нормальным $\mathcal{N}(-0.01, 0.65)$.

Обсуждение

Эксперимент показал, что заданное распределение $\mathcal{N}(\widehat{\mu}, \widehat{\sigma})$ является генеральным законом, по которому построена выборка с уровнем $\alpha = 0.5$, значит оценки максимального правдоподобия состоятельны.

Список литературы

1. Конспекты лекции
2. Википедия: https://ru.wikipedia.org/wiki

Ссылка на github: https://github.com/KateZabolotskih/MathStatLabs