Идентификация распределений

Тестирование гипотезы о подчинении выборки одному из случайных распределений играет важную роль в
прикладных задачах статистики и анализа данных. Важную роль распределения играют:

1. В имитационных моделях деятельности системы, т.к. позволяют моделировать очень сложные
процессы на основе теоретических распределений вероятности отказа оборудования, времени ответа
на звонок покупателя, времени обслуживания клиента, интенсивности поступающих задач и т.п.

2. В предиктивной аналитике, на основе принадлежности остатков реальных данных и модели можно
судить о полноте предсказательной модели ввиду нормальной распределенности остатков вокруг
моделируемой зависимости.

3. В статистике важна процедура проверки гипотез для ситуаций

принятия решений на основе данных.

Для решения задачи проверки гипотезы о подчинении выборки ранее известным теоретическим
распределениям необходимо знать статистические критерии проверки гипотез с заранее заданным
уровнем надежности. Среди множества процедур можно выделить две универсальные:
1. Проверка гипотез на основе критерия χ2-Пирсона.
2. Проверка распределений на основе применения процедуры спрямления функциональной
зависимости в декартовых координатах с помощью аппарата **анаморфоз**.
Также существуют частные случаи проверки гипотез для конкретных распределений, например для
проверки **нормальности распределения**. Такие критерии важны в угоду большого числа статистических
процедур, требующих нормальности распределения данных.

Постановка задачи

- Скачать папку с исходными данными. Открыть папку, соответствующую своей группе. Далее папка с вариантом (номер в списке)

В папке 4 файла с данными. Данные имеют следующие распределения:

1 и 4 файлы – нормальное распределение

3 и 6 файлы – показательное распределение

Постановка задачи

Необходимо идентифицировать распределения в каждом файле двумя способами:

- с помощью критерия согласия Пирсона
- методом анаморфоз

Постановка задачи

- с помощью критерия согласия Пирсона

Методом Пирсона для каждого файла нужно проверить истинное распределение и ложное. То есть все выборки проверить на нормальное распределение и на показательное.

При расчете теоретических частот в качестве параметров распределений брать их точечные несмещенные оценки.

Постановка задачи

- методом анаморфоз

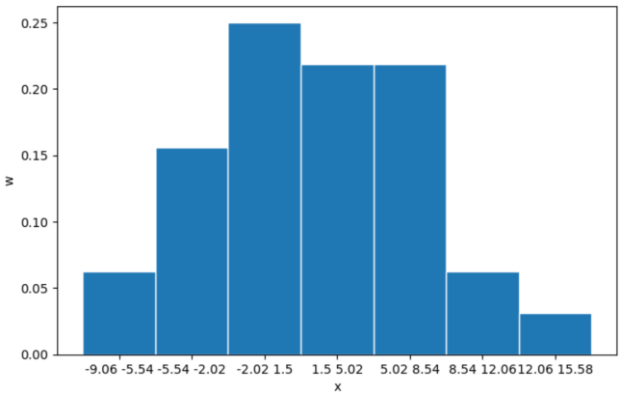
При проверке распределения методом анаморфоз нужно построить 2 графика для каждой выборки. Каждый график представлен в координатах соответствующей анаморфозы (всего 4*2=8 графиков). Тот график, на котором достигнуто спрямление, соответствует истинному распределению. Для проверки качества спрямления необходимо построить линейный тренд (провести линейную регрессию) и показать значения коэффициента детерминации (R^2).

По параметрам прямой найти параметры распределения. Для построения анаморфозы нормального распределения мат. ожидание заменить его несмещенной точечной оценкой.

Пример

Перед применением тестов необходимо сформировать гистограммы данных и рассчитать число интервалов, границы интервалов, абсолютные и относительные частоты. Построить статистический интервальный ряд.

0	1	ni	wi
-9.06	-5.54	2	0.06250
-5.54	-2.02	5	0.15625
-2.02	1.50	8	0.25000
1.50	5.02	7	0.21875
5.02	8.54	7	0.21875
8.54	12.06	2	0.06250
12.06	15.58	1	0.03125



Будем от начала первого интервала отнимать не h/2, а 0.01.
Если делаем -h/2, то получаются другие частоты, это искажает распределение

Пример

Критерий согласия Пирсона

$$\chi^2 = \sum_{i=1}^{m} \frac{(n_i - \hat{n}_i)^2}{\hat{n}_i}$$

$\hat{n}_i = n \cdot P_i$

$\hat{n}_i -$ теоретическая частота
$n_i -$ эмпирическая частота
$m -$ количество интервалов
$n -$ объем выборки

Для проверки гипотезы о нормальном законе распределения

$$P_i = \int_{x_i}^{x_{i+1}} f(t)dt = F(x_{i+1}) - F(x_i) = \Phi\left(\frac{x_{i+1} - \overline{x_в}}{s}\right) - \Phi\left(\frac{x_i - \overline{x_в}}{s}\right)$$

$\overline{x_в} -$ выборочное среднее

x_i и $x_{i+1} -$ левая и правая границы i – го интервала

$s -$ несмещенное стандартное отклонение

Критерий согласия – критерий проверки гипотезы о предполагаемом законе неизвестного распределения
Критерий согласия позволяет ответить на вопрос о том, является ли различие между выборочными и теоретическим распределениями столь незначительными, что они могут быть приписаны лишь случайным факторам.

Теоретические частоты $\hat{n}_i$ вычисляются для заданного закона распределения как количества элементов выборки, которые должны были попасть в каждый интервал, если бы случайная величина имела выбранный закон распределения, параметры которого совпадают с их точечными оценками по выборке.

При верности нулевой гипотезы данный критерий имеет распределение χ^2 с $k = m - r - 1$ степенями свободы, где m – число интервалов, r – число параметров распределения.

Критическая область принимается правосторонней, граница:

$$\chi^2_{\text{кр}} = \chi^2(\alpha, k)$$

Пример

Проверка гипотезы о нормальности распределения $\alpha = 0.05$

$$\chi^2 = \sum_{i=1}^{m} \frac{(n_i - \hat{n}_i)^2}{\hat{n}_i} \qquad \hat{n}_i = n \cdot P_i \qquad P_i = \Phi\left(\frac{x_{i+1} - \overline{x_в}}{s}\right) - \Phi\left(\frac{x_i - \overline{x_в}}{s}\right)$$

0	1	ni	wi	int	Pi	nP
-9.06	-5.54	2	0.06250	-9.06 -5.54	0.061418	1.965381
-5.54	-2.02	5	0.15625	-5.54 -2.02	0.151690	4.854076
-2.02	1.50	8	0.25000	-2.02 1.5	0.243132	7.780213
1.50	5.02	7	0.21875	1.5 5.02	0.252988	8.095618
5.02	8.54	7	0.21875	5.02 8.54	0.170901	5.468837
8.54	12.06	2	0.06250	8.54 12.06	0.074929	2.397739
12.06	15.58	1	0.03125	12.06 15.58	0.021309	0.681900

scipy.stats.chi2.ppf(1-α, число степеней свободы)

Полученное значение критерия Пирсона $\chi^2 = 0.802$

Критическое значение $\chi^2_{1-\alpha,k} = 9.487$

Если $\boldsymbol{\chi^2 > \chi^2_{1-\alpha,k}}$, то **нулевая гипотеза отвергается**

В данном случае нулевая гипотеза принимается. Выборка распределена по нормальному закону.

Пример

Проверка гипотезы о показательном распределении

$$\chi^2 = \sum_{\mathrm{i}=1}^{m} \frac{(n_i - \hat{n}_i)^2}{\hat{n}_i}$$

$$\hat{n}_i = n \cdot P_i$$

$$P_i = \mathrm{e}^{-\lambda x_i} - \mathrm{e}^{-\lambda x_{i+1}}$$

$$\lambda = \frac{1}{\bar{x}_{\mathrm{в}}}$$

Показательное распределение определяется одним параметром, поэтому число степеней свободы k = n – 1 – 1.

Анаморфоза – нелинейное преобразование, которое приводит данные в линейную зависимость, если модель процесса соответствует преобразованию

Анаморфоза для нормального распределения

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \xrightarrow{\text{Возьмем ln}} \ln(f(x)) = ln\frac{1}{\sigma\sqrt{2\pi}} - \frac{(x-\mu)^2}{2\sigma^2}$$

$$\ln\big(f(x)\big) = -\ln(\sigma\sqrt{2\pi}) - \frac{(x-\mu)^2}{2\sigma^2}$$

$$y = b - kx$$

$$b = -\ln(\sigma\sqrt{2\pi})$$

$$k = -\frac{1}{2\sigma^2}$$

$$x = (x-\mu)^2$$

Построение данных, соответствующих нормальному распределению, получает спрямление в координатах

$$\boldsymbol{ln(w) \sim (x - \overline{x}_в)^2}$$

$\boldsymbol{x}$ – середины интервалов

Математическое ожидание $\boldsymbol{\mu}$ заменяем на точечную оценку

За оценку плотности $\boldsymbol{f(x)}$ берем относительные частоты **w**

Путем перебора анаморфоз, соответствующих различным вероятностным распределениям, можно выбрать то распределение, которое линеаризует исходные данные.

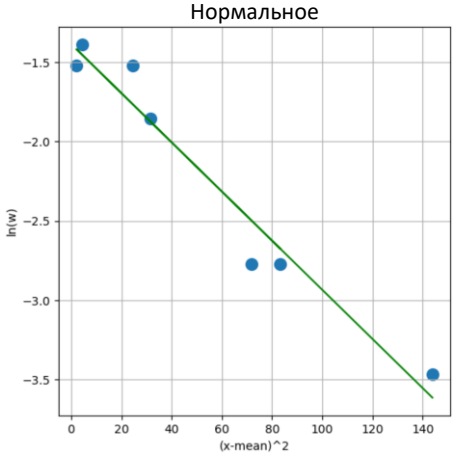
Анаморфоза, которая спрямляет данные на большем интервале, чем другие, будет иметь наивысший ранг значимости.
Построение анаморфоз основано, в частности, на следующих преобразованиях координат: логарифмирование,
сдвиг и растяжение аргумента, инверсия функции или аргумента.

Анаморфоза нормального распределения может быть получена из уравнения для плотности нормального
распределения путем логарифмирования правой и левой частей.

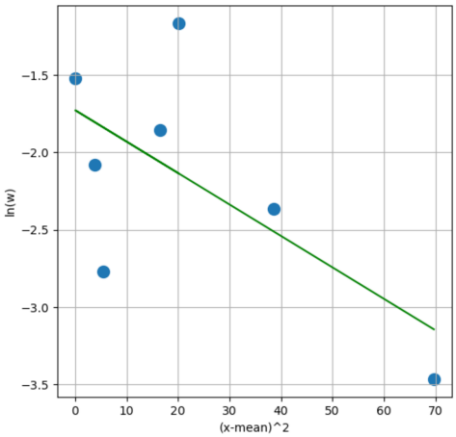
Пример

Анаморфоза для нормального распределения

Нормальное



Показательное



$R^2 = 0.954$

sklearn.linear_model.LinearRegression
метод score

$R^2 = 0.419$

Если поварьируем значение среднего, то можем добиться лучшего спрямления.

Анаморфоза для показательного распределения

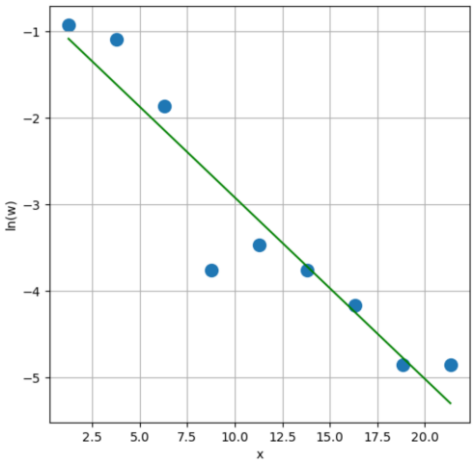
$$f(x) = \begin{cases} \lambda e^{-\lambda x}, x \geq 0 \\ 0 \quad\quad , x < 0 \end{cases}$$

$$\ln(f(x)) = \ln(\lambda) - \lambda x$$
$$y \quad = \quad b \; - \; kx$$

Построение данных, соответствующих показательному распределению, получает спрямление в координатах

$$\ln(f(x)) \sim x$$

$$\ln(w) \sim x$$



$R^2 = 0.9$

Поиск параметров распределения по построенным прямым

Для нормального распределения

Угол наклона прямой $k = -\frac{1}{2\sigma^2} \Rightarrow \sigma = \sqrt{-\frac{1}{2k}}$

Для показательного распределения

$$\lambda = -k$$

Определение качества спрямления

- Построить линейную регрессию
- Посчитать коэффициент детерминации R^2
- Если он превышает 0.8, то данные спрямились хорошо

В отчете составить таблицу

№ выборки	R^2	Вывод
1	0.9	$R^2 > 0.8$, H_0 принимается, выборка распределена нормально
2	0.4	$R^2 < 0.8$, H_0 отвергается, выборка распределена не по нормальному закону

Структура отчета

- 3.1 Постановка задачи (постановка задачи + описать данные)
- 3.2 Ход выполнения работы (правило Стерджесса, статистические ряды для каждой выборки, гистограммы)
- 3.2.1 Проверка распределения с помощью критерия Пирсона (сформулировать гипотезы, формулы, результаты проверки распределений, расчет точечных несмещенных оценок (дать ссылки на формулы), разделить на 4 части, последовательно рассмотреть каждую выборку, в конце в виде таблицы представить результаты)
- 3.3.2 Проверка распределения с помощью метода анаморфоз (формулы анаморфоз, по 2 графика на каждую выборку, расчет R^2). В конце привести таблицу с результатами.
- 3.3.3 Расчёт параметров распределений (формулы, расчеты)
- 3.3 Выводы