Домашнее задание 4

Юрасов Никита Андреевич

Обновлено 26 ноября 2019 г.

Содержание

1 Проверка гипотез о виде распределения

1.1 Критерий согласия Колмогорова-Смирнова

Введем статистику, которая представляет собой максимальной отклонение эмпирической функции распределения $\hat{F}(x)$, построенной по выборке X, от гипотетической функции распределения $F(x)$:

$$D_n = D_n(X) = \sup_{-\infty<x<\infty} |\hat{F}(x) - F(x)|$$

Пусть существует $X = (X_1, ..., X_n)$ – выборка из $\mathcal{L}(\xi)$ с неизвестной функцией распределения $F_\xi(x)$, и пусть выдвинута гипотеза $H_0 : F_\xi(x) = F(x)$, где функция $F(x)$ полностью задана.

Для принятия или отвержения гипотезы H_0 необходимо по критерию Колмогорова сравнить $\sqrt{n}D_n$ с λ_α, которая определяется следующим равенством:

$$K(\lambda_\alpha) = 1 - \alpha,$$

где $K(x)$ – распределение Колмогорова.

На практике статистику D_n удобнее вычислять в следующем виде $D_n = max(D_n^+, D_n^-)$, где

$$D_n^+ = \max_{1\leqslant k\leqslant n} \left(\frac{k}{n} - F(X_{(k)})\right), \quad D_n^- = \max_{1\leqslant k\leqslant n} \left(F(X_{(k)}) - \frac{k-1}{n}\right)$$

Ответ на вопрос о виде распределения дает следующее сравнение:

- Если $\sqrt{n}D_n \geqslant \lambda_\alpha$, то гипотеза H_0 отвергается;
- Если $\sqrt{n}D_n \leqslant \lambda_\alpha$, то гипотеза H_0 принимается.

В неравенстве можно воспользоваться поправкой Большева о статистике $S(D_n)$, которая быстрее сходится к распределению Колмогорова:

$$S = \frac{6nD_n + 1}{6\sqrt{n}}$$

Преимущества
Критерий согласия Колмогорова начинает эффективно работать при выборке объемом $n \geqslant 20$, что допускает использование его при достаточно малых выборках данных.

Недостатки
Критерий Колмогорова-Смирнова применяется только для непрерывных распределениях. Также вычисление статистики D_n предполагает достаточно большие аналитические вычисления, что затрудняет проверку.

Реализация для непрерывного распределения
В приложенном `Jupyter Notebook` написана функция `simple_kolmogorov_text`, которая по заданной выборке и уровне значимости проверяет критерий согласия Колмогорова-Смирнова. Далее будут представлены только результаты, а саму выборку размера 1000 можно будет в переменной `erlang_sample_for_KSTest`. Выборка генерировалась с параметрами $m = 2, \lambda = 0.2$

```
S_Bolshev = 1.1012256532624491 < 1.2238478702170825 and K-S test accepts with
alpha=0.1
S_Bolshev = 1.1012256532624491 < 1.3580986393225505 and K-S test accepts with
alpha=0.05
```

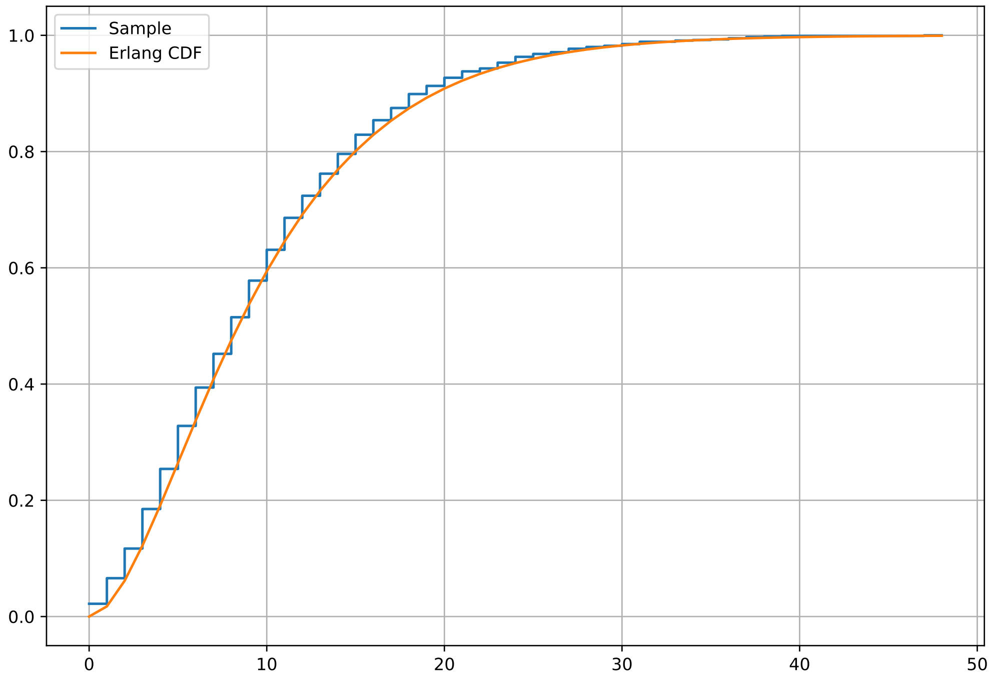

Рис. 1: Сравнение эмпирической функции и функции распределения Эрланга

1.2 Критерий согласия χ^2

Введем статистику $\mathring{X}$, введенную К. Пирсоном, которая будет показывать отклонение эмпирических данных от гипотетических значений и которая называется мера *хи-квадрат*:

$$\mathring{X}_n^2 = \mathring{X}_n^2(\nu) = \sum_{i=1}^{N} \frac{(\nu_i - n\mathring{p}_i)^2}{n\mathring{p}_i}, \quad \text{где}$$

- N – количество принимаемых значений в эксперименте;
- $\nu = (\nu_1, ..., \nu_N)$ – частоты появления каждого результата эксперимента;
- $n = \sum_{i=1}^{N} \nu_i$ – общий объем выборки;
- $\mathring{p} = (\mathring{p}_1, ..., \mathring{p}_N)$ – вероятность появления $i-$го события;

После подсчета *меры хи-квадрат* необходимо сравнить ее с критическим значением распределения хи-квадрат на уровне значимости α с $N-1$ степенями свободы:

$$\chi^2_{1-\alpha, N-1}$$

Сравнение:

- Если $\mathring{X}_n^2 > \chi^2_{1-\alpha,N-1}$, то говорят, что гипотеза H_0 отклоняется;
- Если $\mathring{X}_n^2 \leqslant \chi^2_{1-\alpha,N-1}$, то говорят, что гипотеза H_0 принимается.

где гипотеза H_0 определена так же, как и в критерии согласия Колмогорова-Смирнова (см. страницу 2).

Преимущества

Критерий работает первоначально только с дискретными данными, но так как любые данные можно свести к дискретным методом группировки (см. правило Стёрджеса: Википедия). Также, этот критерий можно использовать для расчетов с хорошим приближением уже при $n \geqslant 50$.

Недостатки

Критерий χ^2 ошибается на выборках с низкочастотными (редкими) событиями. Решить эту проблему можно отбросив низкочастотные события, либо объединив их с другими событиями. Этот способ называется *коррекцией Йетса*.

Указания для проверки непрерывных распределений

Так как вероятность попадания в одну конкретную точку (в случае непрерывных распределений) равна 0, воспользуемся отмеченным ранее правилом Стёрджеса для разбиения отрезка на k не пересекающихся интервалов:

$$k = 1 + \lfloor log_2 N \rfloor$$

Также необходимо вместо вектора гипотетических вероятностей в каждой точек использовать вероятность попадания в каждый из полученных интервалов. Для этого нужно вычислить значение интеграла:

$$\int\limits_{x_i}^{x_{i+1}} f(x)dx, \ \text{где}$$

$f(x)$ – плотность распределения, а x_i – точки разбиения отрезка.

Реализация для дискретного распределения

Сгенерируем выборку (распределение Пуассона) с параметром $\lambda = 2$ и размером 1000, которую будем хранить в переменной `poisson_sample_for_Chi2Test`. Результаты для двух уровней значимости выглядят следующим образом:

```
S = 9.849679407558991 <= 15.50731305586545 and Chi2 test accepts with alpha=0.05
S = 9.849679407558991 <= 13.36156613651173 and Chi2 test accepts with alpha=0.1
```

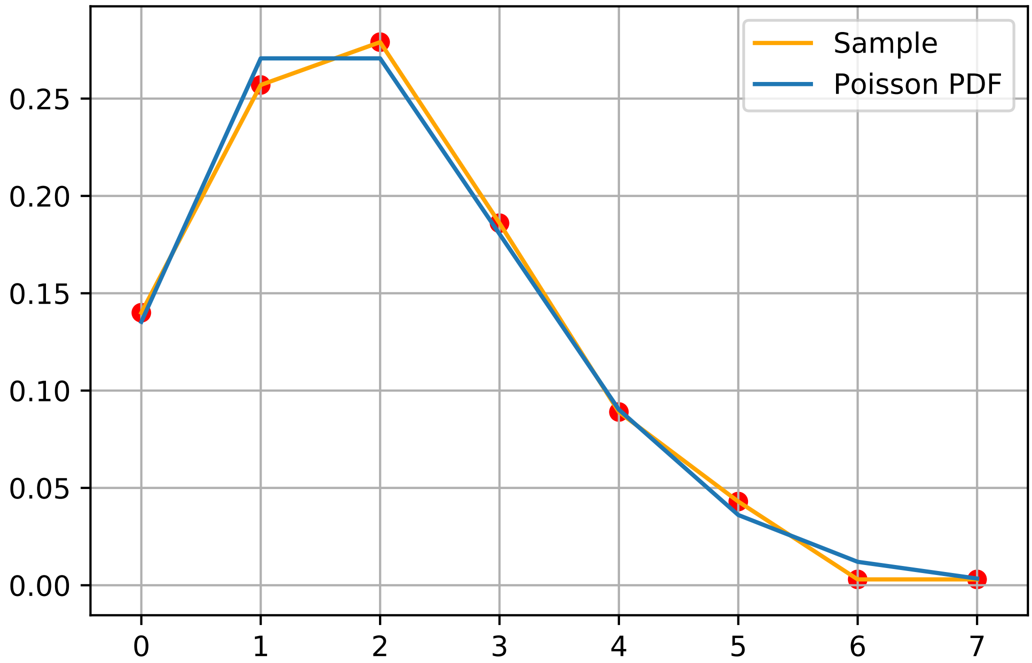

Рис. 2: Сравнение полигона частот и распределения Пуассона

Реализация для непрерывного распределения

Выборка, хранящаяся в переменной `erlang_sample_for_Chi2Test`, генерировалась с параметрами $m = 2, \lambda = 0.2$.

```
S = 13.199273296059534 <= 15.50731305586545 and Chi2 test accepts with alpha=0.05
S = 13.199273296059534 <= 13.36156613651173 and Chi2 test accepts with alpha=0.1
```

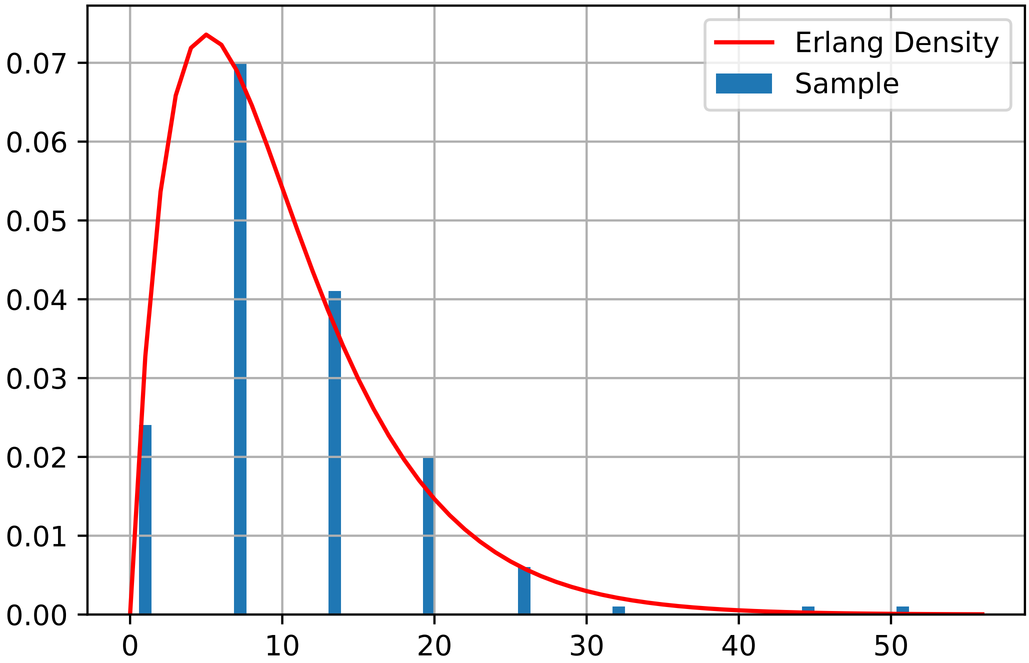

Рис. 3: Сравнение гистограммы частот и плотности распределения Эрланга