Домашнее задание 3

Юрасов Никита Андреевич

Обновлено 10 ноября 2019 г.

Содержание

Эта работа представляет собой отчет к Домашнему Заданию №3.
Так как в моделировании используется пакет `numpy.random`, то, пусть, для него будет выставлено по умолчанию стартовое значение генератора 12345678.

1 Нахождение выборочного среднего и выборочной дисперсии

Определение 1.1. *Пусть $X_1, X_2, \cdots, X_n$ – выборка из какого-то распределения вероятности. Тогда ее выборочным средним называется случайная величина*

$$\overline{X} = \frac{1}{n}\sum_{i=0}^{n} X_i$$

Определение 1.2. *Пусть $X_1, X_2, \cdots, X_n$ – выборка из какого-то распределения вероятности. Тогда выборочная дисперсия – это случайная величина*

$$S_n^2 = \frac{1}{n}\sum_{i=0}^{n} \left(X_i - \overline{X}\right)^2,$$

где $\overline{X}$ – выборочное среднее.

Для нахождения этих двух значений можно воспользоваться методами `mean` и `var` библиотеки `numpy`, но для реализации была написана собственная функция `sample_mean` и `sample_variance` соответственно. Время работы на больших выборках почти одинаковое.

1.1 Распределение Пуассона

```
Выборочное среднее для выборки [2. 1. 1. 3. 0.] = 1.4
Выборочное среднее для выборки [2. 1. 2. 3. 7. 2. 3. 1. 2. 0.] = 2.3
Выборочная дисперсия для выборки [2. 1. 1. 3. 0.] = 1.04
Выборочная дисперсия для выборки [2. 1. 2. 3. 7. 2. 3. 1. 2. 0.] = 3.21
```

1.2 Распределение Эрланга

```
Выборочное среднее для выборки [ 9.43737905 11.94755981 1.6335522 11.63186523
1.95757948] = 7.321587156076815
Выборочное среднее для выборки [ 8.30297653 17.47684737 5.71182291 2.67860603
19.66877258 7.92660288 5.52776384 8.11891813 9.22277337 16.35132395] = 10.098640758
Выборочная дисперсия для выборки [ 9.43737905 11.94755981 1.6335522 11.63186523
1.95757948] = 21.11620307310885
Выборочная дисперсия для выборки [ 8.30297653 17.47684737 5.71182291 2.67860603
19.66877258 7.92660288 5.52776384 8.11891813 9.22277337 16.35132395] = 29.294400894
```

2 Нахождение параметров распределений событий

Для каждого из двух распределений будем строить оценку максимального правдоподобия.

2.1 Распределение Пуассона

Пусть функция распределения будет выглядеть следующим образом:

$$f(x,\theta) = \frac{\theta^x e^{-\theta}}{x!}, \quad x \geq 0$$

Тогда функция правдоподобия:

$$L(x,\theta) = \prod_{i=0}^{n} f(x,\theta) = \prod_{i=0}^{n} \left(\frac{\theta^{x_i} e^{-\theta}}{x_i!} \right) = e^{\theta n} \frac{\theta^{\sum_{i=0}^{n} x_i}}{\prod_{i=0}^{n} x_i!}$$

Возьмем от функции правдоподобия натуральный логарифм:

$$lnL(x,\theta) = -\theta n + \sum_{i=0}^{n} x_i \cdot ln\theta - ln \prod_{i=0}^{n} x_i!$$

Продифференцируем полученное выражение по θ и приравняем к нулю:

$$\frac{\partial lnL(x,\theta)}{\partial \theta} = -n + \frac{\sum_{i=0}^{n} x_i}{\theta} = 0$$

Будем решать это уравнение относительно θ:

$$\hat{\theta} = \frac{1}{n} \sum_{i=0}^{n} x_i$$

В итоге, получается, что оценка максимального правдоподобия параметра θ распределения Пуассона имеет вид выборочного среднего (см. определение 1.1).

Предложенная оценка $\hat{\theta}$ является несмещенной, так как выборочное среднее является в свою очередь несмещенной оценкой:

$$M(\frac{1}{n} \sum_{i=0}^{n} x_i) = \frac{1}{n} \sum_{i=0}^{n} Mx_i = \frac{1}{n} \sum_{i=0}^{n} x_i$$

Состоятельность можно проверить по утверждению, что выборочные моменты k-го порядка сходятся по вероятности к k-ым моментам X, то есть:

$$\hat{\alpha_k} \xrightarrow{P} MX^k$$

В нашем случае k=1:

$$\hat{\alpha_k} \xrightarrow{P} MX$$

Следовательно оценка $\hat{\theta} = \hat{\alpha_k}$, которая является состоятельной.

Эффективность представленной оценки также подтверждается, так как $\hat{\theta}$ – оценка максимального правдоподобия, а такая оценка эффективна. Такой вывод следует из теоремы Дюге.

2.2 Распределение Эрланга

Рассмотрим функцию распределения, которая зависит от двух параметров:

$$f(x, m, \lambda) = \frac{\lambda^m x^{m-1}}{\Gamma(m)} e^{-\lambda x}, \quad m \in \mathbb{N}, \lambda > 0, x > 0$$

Построим функцию правдоподобия:

$$L(x, m, \lambda) = \prod_{i=0}^{n} f(x, m, \lambda) = \left(\frac{\lambda^m}{\Gamma(m)}\right)^n \prod_{i=0}^{n} x_i^{m-1} e^{-\lambda x_i} = \left(\frac{\lambda^m}{\Gamma(m)}\right)^n e^{-\lambda \sum_{i=0}^{n} x_i} \prod_{i=0}^{n} x_i^{m-1}$$

Возьмем натуральный логарифм от $L(x, m, \lambda)$:

$$lnL(x, m, \lambda) = mnln\lambda + (m-1)\prod_{i=0}^{n} lnx_i - \lambda \sum_{i=0}^{n} x_i - nln\Gamma(m)$$

Продифференцируем полученное по λ и приравняем к нулю:

$$\frac{\partial lnL(x, m, \lambda)}{\partial \lambda} = \frac{mn}{\lambda} - \sum_{i=0}^{n} x_i = 0$$

Решая относительно λ, получим:

$$\hat{\lambda} = \frac{mn}{\sum\limits_{i=0}^{n} x_i} = \frac{m}{\hat{\alpha_1}} \tag{1}$$

Относительно оценки $\hat{\lambda}$ (см. оценку 1) можно сказать, пользуясь теоремой Дюге, что такая оценка будет состоятельной и эффективной. И подобная оценка окажется несмещенной:

$$M\left(\frac{m}{\frac{1}{n}\sum\limits_{i=0}^{n} x_i}\right) = \frac{m}{MX} = \frac{m}{\frac{m}{\lambda}} = \lambda$$

Также можно предложить оценку параметра m, используя метод моментов.
Используя значения, полученные в работе №1, найдем оценку:

$$E\xi = \frac{m}{\lambda}, \quad D\xi = \frac{m}{\lambda^2}$$

$$\begin{cases} \frac{m}{\lambda} = \frac{1}{n}\sum\limits_{i=0}^{n} x_i \\ \frac{m}{\lambda^2} = \frac{1}{n}\sum\limits_{i=0}^{n} x_i^2 \end{cases}$$

Выразим из первого уравнения λ:

$$\lambda = \frac{mn}{\sum\limits_{i=0}^{n} x_i}$$

Как нетрудно заметить, полученная оценка (методом моментов) совпадает с оценкой, полученной методом правдоподобия.
Подставим полученное λ во второе уравнение:

$$\hat{m} = \frac{\hat{\alpha_1}^2}{n^4 \hat{\alpha_2}} \tag{2}$$

Проверим полученную оценку на смещенность:

$$M\hat{m} = \frac{1}{n^4}\frac{M\left(\frac{1}{n}\sum\limits_{i=0}^{n} x_i\right)^2}{M\left(\frac{1}{n}\sum\limits_{i=0}^{n} x_i^2\right)} = \frac{1}{n^4}\frac{(\frac{m}{\lambda})^2}{\frac{m}{\lambda^2} - \frac{m}{\lambda}} = \frac{m^2}{n^4(m - m^2)} = \frac{m}{n^4(1-m)}$$

Отсюда можно сделать вывод, что полученная оценка является смещенной.
Для проверки на состоятельность должно быть выполнено следующее условие:

$$\lim_{n\to\infty}(|\hat{m} - m| < \epsilon) = 1, \forall\epsilon \geqslant 0$$

Что не выполняется для предложенной оценки (сходимость по вероятности):

$$\lim_{n\to\infty}(\left|\frac{\left(\frac{1}{n}\sum\limits_{i=0}^{n} x_i\right)^2}{n^4\sum\limits_{i=0}^{n} x_i^2} - m\right|) = m \not\leqslant \epsilon, \quad \forall\epsilon$$

Следовательно, оценка 2 не состоятельна, что означает несовпадение оценки и реального параметра при стремлении n к бесконечности.

3 Работа с данными

3.1 Распределение Пуассона

В силу выбранной нетипичной интерпретации задача состояла в том, чтобы найти открытую базу данных с полной статистикой по футбольным матчам за определенный период. Для иллюстрации работы с данными была выбрана Premier League английского футбола за сезон 2018/2019. Данные находятся в открытом доступе по ссылке. В этой работе будем полагать, что данные, представленные на этом сайте, достоверные и отражают полностью футбольную статистику за выбранный сезон.

В этом разделе будут проиллюстрированы лишь выборочное среднее и выборочная дисперсия, но в Jupyter Notebook раскрыта более подробна тема со ставками, а именно, как можно предугадать количество голов в очередном матче и примерные коэффициенты букмекера на этот матч (в выбранной базе данных прикреплены коэффициенты букмекера в Англии на все матчи). Форматированные данные представлены в таблице 3.1.

Характеристики для выбранных данных следующие:

```
Выборочное среднее всех голов: 2.8210526315789473
Выборочная дисперсия всех голов: 2.5574515235457063
```

Таблица 1: Футбольная статистика

	TotIn	TotOut	InHome	InAway	OutHome	OutAway	WMeanInHome	WMeanInAway	WMeanOutHome	WMeanOutAway
Wolves	47.0	46.0	28.0	19.0	21.0	25.0	0.9396	0.79832	0.88235	0.83893
Chelsea	63.0	39.0	39.0	24.0	12.0	27.0	1.30872	1.0084	0.5042	0.90604
Bournemouth	56.0	70.0	30.0	26.0	25.0	45.0	1.00671	1.09244	1.05042	1.51007
Everton	54.0	46.0	30.0	24.0	21.0	25.0	1.00671	1.0084	0.88235	0.83893
Crystal Palace	51.0	53.0	19.0	32.0	23.0	30.0	0.63758	1.34454	0.96639	1.00671
Newcastle	42.0	48.0	24.0	18.0	25.0	23.0	0.80537	0.7563	1.05042	0.77181
Burnley	45.0	68.0	24.0	21.0	32.0	36.0	0.80537	0.88235	1.34454	1.20805
West Ham	52.0	55.0	32.0	20.0	27.0	28.0	1.07383	0.84034	1.13445	0.9396
Cardiff	34.0	69.0	21.0	13.0	38.0	31.0	0.7047	0.54622	1.59664	1.04027
Southampton	45.0	65.0	27.0	18.0	30.0	35.0	0.90604	0.7563	1.2605	1.1745
Fulham	34.0	81.0	22.0	12.0	36.0	45.0	0.73826	0.5042	1.51261	1.51007
Man United	65.0	54.0	33.0	32.0	25.0	29.0	1.10738	1.34454	1.05042	0.97315
Brighton	35.0	60.0	19.0	16.0	28.0	32.0	0.63758	0.67227	1.17647	1.07383
Man City	95.0	23.0	57.0	38.0	12.0	11.0	1.91275	1.59664	0.5042	0.36913
Watford	52.0	59.0	26.0	26.0	28.0	31.0	0.87248	1.09244	1.17647	1.04027
Tottenham	67.0	39.0	34.0	33.0	16.0	23.0	1.14094	1.38655	0.67227	0.77181
Leicester	51.0	48.0	24.0	27.0	20.0	28.0	0.80537	1.13445	0.84034	0.9396
Arsenal	73.0	51.0	42.0	31.0	16.0	35.0	1.4094	1.30252	0.67227	1.1745
Huddersfield	22.0	76.0	10.0	12.0	31.0	45.0	0.33557	0.5042	1.30252	1.51007
Liverpool	89.0	22.0	55.0	34.0	10.0	12.0	1.84564	1.42857	0.42017	0.40268