Математическая статистика

Домашняя работа № 3

Оценки

Попов Юрий, СКБ-172

ОГЛАВЛЕНИЕ

Предисловие

Все графики, которые в дальнейшем будут вставлены в эту работу, были сконструированы с помощью различных библиотек, основные которые - это matplotlib и numpy в Jupyter Notebook

К работе приложены 2 основных файла: "Geom_Dz_3.ipynb"и "Expon_Dz_3.ipynb в которых указаны расчеты соотвественно геометрического и экспоненциального распределения

Подробности работа с данными находятся в двух других юпитерских файлах: "Database_Geom_
"Database_Expon_3_New.ipynb в которых находятся расчеты для геометрически и экспоненциально, распределенных данных

Все фотографии, использованные в работе лежат в папке *fotos*

Когда я начинал третье домашнее задание, обновленного файла с домашней работы еще не было(или я не знал о его существовании), поэтому номер 3.2 сделан из старого файла

Большая часть определений, которые представлены в этой работы взять с лекций нашего курса.

Также некоторые определения взяты из источника Г.И. Ивченко, Ю.И. Медведев "Ведение в математическую статистику"

Задание 3.1 Нахождение выборочного среднего и выборочной дисперсии

Выборочные моменты

Наиболее важными характеристиками случайной величины ξ являются ее моменты $\alpha_k = E\xi^k$, а также центральные моменты $\mu_k = E(\xi - \alpha_1)^k$(когда они существуют). Их статическими аналогами, вычисляемыми по соответствующей выборке $X = (X_1, \ldots, X_n)$, являются *выборочные моменты* соответственно *обычные*

$$\hat{\alpha}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

и центральные

$$\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\alpha}_1)^k$$

Особенно важны моменты первого и второго порядков.

При $k = 1$ величину $\hat{\alpha}_1$ называют *выборочным средним* и обозначают стандартным символом $\bar{X}$:

$$\bar{X} = \hat{\alpha}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$$

При $k = 2$ величину $\hat{\mu}_2$ называют *выборочной дисперсией* и также обозначают стандартным символом $S^2 = S^2(X)$:

$$S^2 = \hat{\mu}_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

3.1.1 Экспоненциальное распределение

Посчитаем теоритические значения выборочного среднего и выборочной дисперсии. Расчеты будут проводится для $\lambda = 0.8$.

Выборочное среднее:

$$MX = \frac{1}{\lambda} = \frac{1}{0.8} = 1,25$$

$n = 5$
$\overline{X} = 1,36$
$\Delta = |MX - \overline{X}| = 0,11$

$n = 10$
$\overline{X} = 1,85$
$\Delta = |MX - \overline{X}| = 0,6$

$n = 100$
$\overline{X} = 1,21$
$\Delta = |MX - \overline{X}| = 0,08$

$n = 1000$
$\overline{X} = 1,33$
$\Delta = |MX - \overline{X}| = 0,04$

$n = 10^5$
$\overline{X} = 1,248$
$\Delta = |MX - \overline{X}| = 0,002$

Выборочная дисперсия:

$$DX = \frac{1}{\lambda^2} = \frac{1}{0.8^2} = 1,5625$$

$n = 5$
$\hat{\sigma^2} = 0,7$
$\Delta = |DX - \hat{\sigma^2}| = 0,8624$

$n = 10$
$\hat{\sigma^2} = 1,035$
$\Delta = |DX - \hat{\sigma^2}| = 0,5275$

$n = 100$
$\hat{\sigma^2} = 1,153$
$\Delta = |DX - \hat{\sigma^2}| = 0,4$

$n = 1000$
$\hat{\sigma^2} = 1,515$
$\Delta = |DX - \hat{\sigma^2}| = 0,0475$

$n = 10^5$
$\hat{\sigma^2} = 1,563$
$\Delta = |DX - \hat{\sigma^2}| = 0,0005$

Из приведенных данных прекрасно видно, что при увлечении объема выборки, то есть при увеличении количества рассматриваемых величин, разница между истинными значениями параметров и значениями их параметров стремится к 0.

Это еще раз доказало свойство выборочных моментов - *асимптотическое поведение при неограниченном возрастании объема выборки*.

На основании неравенства Чебышева следует, что для любого $\varepsilon > 0$ при $n \to \infty$

$$P\{|\hat{\alpha_{nk}} - \alpha_k| < \varepsilon\} \to 1$$

3.1.2 Геометрическое распределение

Посчитаем теоритические значения выборочного среднего и выборочной дисперсии. Расчеты будут проводится для $p = 0.2$.

Выборочное среднее:

$$MX = \frac{1}{p} = \frac{1}{0.25} = 4$$

$n = 5$

$\overline{X} = 1,4$
$\Delta = |MX - \overline{X}| = 2,6$

$n = 10$
$\overline{X} = 2,8$
$\Delta = |MX - \overline{X}| = 1,2$

$n = 100$
$\overline{X} = 3,64$
$\Delta = |MX - \overline{X}| = 0,36$

$n = 1000$
$\overline{X} = 3,893$
$\Delta = |MX - \overline{X}| = 0,107$

$n = 10^5$
$\overline{X} = 3,979$
$\Delta = |MX - \overline{X}| = 0,021$

Выборочная дисперсия:

$$DX = \frac{1}{p^2} = \frac{1}{0.25^2} = 16$$

$n = 5$
$\hat{\sigma^2} = 5,04$
$\Delta = |DX - \hat{\sigma^2}| = 10.96$

$n = 10$
$\hat{\sigma^2} = 8,29$
$\Delta = |DX - \hat{\sigma^2}| = 7,71$

$n = 100$
$\hat{\sigma^2} = 12,248$
$\Delta = |DX - \hat{\sigma^2}| = 3,752$

$n = 1000$

$\hat{\sigma^2} = 15,297$
$\Delta = |DX - \hat{\sigma^2}| = 0,703$

$n = 10^5$
$\hat{\sigma^2} = 15,989$
$\Delta = |DX - \hat{\sigma^2}| = 0,011$

Объяснение такое же (написано выше)

Задание 3.2 Построение доверительного интервала для выборочного среднего и выборочной дисперсии

3.2.1 Геометрическое распределение

Для всех выборок построим доверительные интервалы.

```
Для 1 реализации выборки объема 5 доверительный интервал равен: (0.442 <= a <=  7.648)
Для 2 реализации выборки объема 5 доверительный интервал равен: (-0.106 <= a <=  8.196)
Для 3 реализации выборки объема 5 доверительный интервал равен: (-0.327 <= a <=  8.417)
Для 4 реализации выборки объема 5 доверительный интервал равен: (-0.445 <= a <=  8.535)
Для 5 реализации выборки объема 5 доверительный интервал равен: (0.31 <= a <=  7.78)
```

$n = 5$

```
Для 1 реализации выборки объема 10 доверительный интервал равен: (1.895 <= a <=  6.195)
Для 2 реализации выборки объема 10 доверительный интервал равен: (0.713 <= a <=  7.377)
Для 3 реализации выборки объема 10 доверительный интервал равен: (-2.915 <= a <=  11.005)
Для 4 реализации выборки объема 10 доверительный интервал равен: (-0.79 <= a <=  8.88)
Для 5 реализации выборки объема 10 доверительный интервал равен: (2.036 <= a <=  6.054)
```

$n = 10$

```
Для 1 реализации выборки объема 100 доверительный интервал равен: (3.25 <= a <=  4.84)
Для 2 реализации выборки объема 100 доверительный интервал равен: (3.077 <= a <=  5.013)
Для 3 реализации выборки объема 100 доверительный интервал равен: (3.246 <= a <=  4.844)
Для 4 реализации выборки объема 100 доверительный интервал равен: (3.188 <= a <=  4.902)
Для 5 реализации выборки объема 100 доверительный интервал равен: (3.115 <= a <=  4.975)
```

$n = 100$

```
Для 1 реализации выборки объема 1000 доверительный интервал равен: (3.765 <= a <=  4.325)
Для 2 реализации выборки объема 1000 доверительный интервал равен: (3.762 <= a <=  4.328)
Для 3 реализации выборки объема 1000 доверительный интервал равен: (3.782 <= a <=  4.308)
Для 4 реализации выборки объема 1000 доверительный интервал равен: (3.76 <= a <=  4.33)
Для 5 реализации выборки объема 1000 доверительный интервал равен: (3.728 <= a <=  4.362)
```

$n = 1000$

```
Для 1 реализации выборки объема 100000 доверительный интервал равен: (4.017 <= a <=  4.073)
Для 2 реализации выборки объема 100000 доверительный интервал равен: (4.017 <= a <=  4.073)
Для 3 реализации выборки объема 100000 доверительный интервал равен: (4.017 <= a <=  4.073)
Для 4 реализации выборки объема 100000 доверительный интервал равен: (4.017 <= a <=  4.073)
Для 5 реализации выборки объема 100000 доверительный интервал равен: (4.017 <= a <=  4.073)
```

$n = 100000$

3.2.2 Экспоненциальное распределение

Для всех выборок построим доверительные интервалы.

```
Для 1 реализации выборки объема 5 доверительный интервал равен: (0.442 <= a <=  7.648)
Для 2 реализации выборки объема 5 доверительный интервал равен: (-0.106 <= a <=  8.196)
Для 3 реализации выборки объема 5 доверительный интервал равен: (-0.327 <= a <=  8.417)
Для 4 реализации выборки объема 5 доверительный интервал равен: (-0.445 <= a <=  8.535)
Для 5 реализации выборки объема 5 доверительный интервал равен: (0.31 <= a <=  7.78)
```

$n = 5$

```
Для 1 реализации выборки объема 10 доверительный интервал равен: (1.895 <= a <=  6.195)
Для 2 реализации выборки объема 10 доверительный интервал равен: (0.713 <= a <=  7.377)
Для 3 реализации выборки объема 10 доверительный интервал равен: (-2.915 <= a <=  11.005)
Для 4 реализации выборки объема 10 доверительный интервал равен: (-0.79 <= a <=  8.88)
Для 5 реализации выборки объема 10 доверительный интервал равен: (2.036 <= a <=  6.054)
```

$n = 10$

```
Для 1 реализации выборки объема 100 доверительный интервал равен: (3.25 <= a <=  4.84)
Для 2 реализации выборки объема 100 доверительный интервал равен: (3.077 <= a <=  5.013)
Для 3 реализации выборки объема 100 доверительный интервал равен: (3.246 <= a <=  4.844)
Для 4 реализации выборки объема 100 доверительный интервал равен: (3.188 <= a <=  4.902)
Для 5 реализации выборки объема 100 доверительный интервал равен: (3.115 <= a <=  4.975)
```

$n = 100$

```
Для 1 реализации выборки объема 1000 доверительный интервал равен: (3.765 <= a <=  4.325)
Для 2 реализации выборки объема 1000 доверительный интервал равен: (3.762 <= a <=  4.328)
Для 3 реализации выборки объема 1000 доверительный интервал равен: (3.782 <= a <=  4.308)
Для 4 реализации выборки объема 1000 доверительный интервал равен: (3.76 <= a <=  4.33)
Для 5 реализации выборки объема 1000 доверительный интервал равен: (3.728 <= a <=  4.362)
```

$n = 1000$

```
Для 1 реализации выборки объема 100000 доверительный интервал равен: (4.017 <= a <=  4.073)
Для 2 реализации выборки объема 100000 доверительный интервал равен: (4.017 <= a <=  4.073)
Для 3 реализации выборки объема 100000 доверительный интервал равен: (4.017 <= a <=  4.073)
Для 4 реализации выборки объема 100000 доверительный интервал равен: (4.017 <= a <=  4.073)
Для 5 реализации выборки объема 100000 доверительный интервал равен: (4.017 <= a <=  4.073)
```

$n = 100000$

Задание 3.3 Нахождение параметров распределений событий

Определение $\mathcal{F} = F_\theta(x|\theta \in \Theta)$ называется экспоненциальной, если

$$f(x;\theta) = exp(A(\theta) \star B(x) + C(\theta) + D(x)).$$

Для того, чтобы в модели существовала эффективная оценка, необходимо и достаточно, чтобы модель принадлежала экспоненциальному семейству.

Вклад выборки для экспоненциальной модели равен:

$$V(X;\theta) = A'(\theta)\sum_{i=1}^{n} B(X_i) + nC'(\theta) = nA'(\theta)[\frac{1}{n}\sum_{i=1}^{n} B(X_i) + \frac{C'(\theta)}{A'(\theta)}]$$

Это также можно записать в виде:

$$T(X) - \tau(\theta) = \alpha(\theta)V(x;\theta)$$

При этом:

$$\tau(\theta) = -\frac{C'(\theta)}{A'(\theta)}$$

$$\alpha(\theta) = \frac{1}{nA'(\theta)}$$

$$T^* = T^*(X) = \frac{1}{n}\sum_{i=1}^{n} B(X_i)$$

По критерию Рао-Крамера заключаем, что статистика T^* является эффективной оценкой для параметрической функции $\tau(\theta)$

Оба распределения относятся к экспоненциальному семейству

3.3.1 Геометрическое распределение

Найдем оценку для геометрического распределения

$$P(X = k) = p(1-p)^{x-1}, x \in N, p - \text{оцениваемый параметр}$$

$$E[X] = \frac{1}{p}$$

$$D(X) = \frac{1-p}{p^2}$$

Функция правдоподобия:

$$L_\theta =$$

$$\ln(L_\theta) = n\ln(\theta) + n\overline{x}\ln(1-\theta)$$

Уравнение правдоподобия имеет вид:

$$\frac{\partial \ln(L_\theta)}{\partial \theta} = 0$$

Продифференцировав, получаем:

$$\frac{n}{\theta} + \frac{n\overline{x}}{1-\theta} = 0$$

$$\frac{\theta\overline{x}}{\theta - 1} = 1$$

Получаем оценку максимального правдоподобия для $\theta : \hat{\theta} = \frac{1}{\overline{x}}$

Так как оценка получена методом максимального правдоподобия, то она является состоятельной, асимптотически нормальной и эффективной

3.3.2 Экспоненциальное распределение

Данное распределение тоже относится к экспоненциальному семейству.

Плотность вероятности:

$$f(x, \theta) = e^{-\theta x + \ln \theta}$$

Найдем все коэффициенты:

$$A(\theta) = -\theta$$

$$B(x) = x$$

$$C(\theta) = \ln \theta$$

$$D(x) = 0$$

$$T^* = T^*(X) = \frac{1}{n} \sum_{i=1}^{n} B(X_i) = \overline{X}$$

$$\tau(\theta) = -\frac{C'(\theta)}{A'(\theta)} = -\frac{(\ln \theta)'}{(-\theta)'} = \frac{1}{\theta}$$

И наконец получаем оценку параметра θ:

$$\hat{\theta} = \frac{n}{\overline{X}} = \frac{n}{\sum\limits_{i=1}^{n} X_i}$$

Проверим несмещенность:

$$M\hat{\Theta} = M\frac{n}{\sum_{i=1}^{n} X_i} = nM\frac{1}{\sum_{i=1}^{n} X_i}$$

Воспользуемся связью между распределениями, и так как X_i распределена экспоненциально, то сумма X_i будет иметь распределение Эрланга с параметрами n и θ. Получаем:

$$M\hat{\theta} = \frac{n\theta}{n-1}$$

Итог: оценка является смещенной.

Если оценка смещенная, то она не является эффективной. Построим несмещенную оценку θ:

$$\hat{\theta} = \frac{n-1}{n}\frac{n}{\sum_{i=1}^{n} X_i} = \frac{n-1}{\sum_{i=1}^{n} X_i}$$

$$M\tilde{\theta} = \theta$$

Проверим состоятельность:

$$D\tilde{\theta} = M\tilde{\theta^2} - (M\tilde{\theta})^2$$

$$M\tilde{\theta^2} = (n-1)^2 M(\frac{1}{\sum_{i=1}^{n} X_i})^2 = (n-1)^2\frac{(n-2)}{(n)}\theta^2 = \frac{n-1}{n-2}\theta^2$$

$$D\tilde{\theta} = M\tilde{\theta^2} - (M\tilde{\theta})^2 = \frac{n-1}{n-2}\theta^2 - \theta^2 = \frac{\theta^2}{(n-2)}$$

Итог: так как $D\tilde{\theta} \to 0$ при ∞, то оценка будет состоятельной.

Так как распределение принадлежит экспоненциальному семейству, то достаточная статистика для параметра является полной

Проверим на достаточную статистику

Докажем, что $(\overline{X})$ является достаточной статистикой для θ

$$L(x,\theta) = \ln(\prod_{i=1}^{n} \theta e^{-\theta x_i}) = \theta^n \ln(\prod_{i=1}^{n} \theta e^{-\theta n\overline{X}}) = n\ln\theta - \theta n\overline{X} = n(\ln\theta - \theta\overline{X})$$

Соотвественно, так как выполняется критерий факторизации, то $T(X) = \overline{X}$ является достаточной статистикой и значит полной статистикой.

Проверим на оптимальность

По теореме Рао-Блэкулла-Колмагорова - оптимальная оценка, если существует является функцией от достаточной статистики. Если статистика полная, то можно сделать и обратное утверждение. Функция от полной и достаточной статистики является оптимальной.

Оценка $T(\frac{n-1}{\sum_{i=1}^{n} X_i}) = \frac{1-1/n}{\overline{X}}$ является функцией от достаточной полной статистики, а значит по теореме является оптимальной.

Задание 3.4 Работа с данными

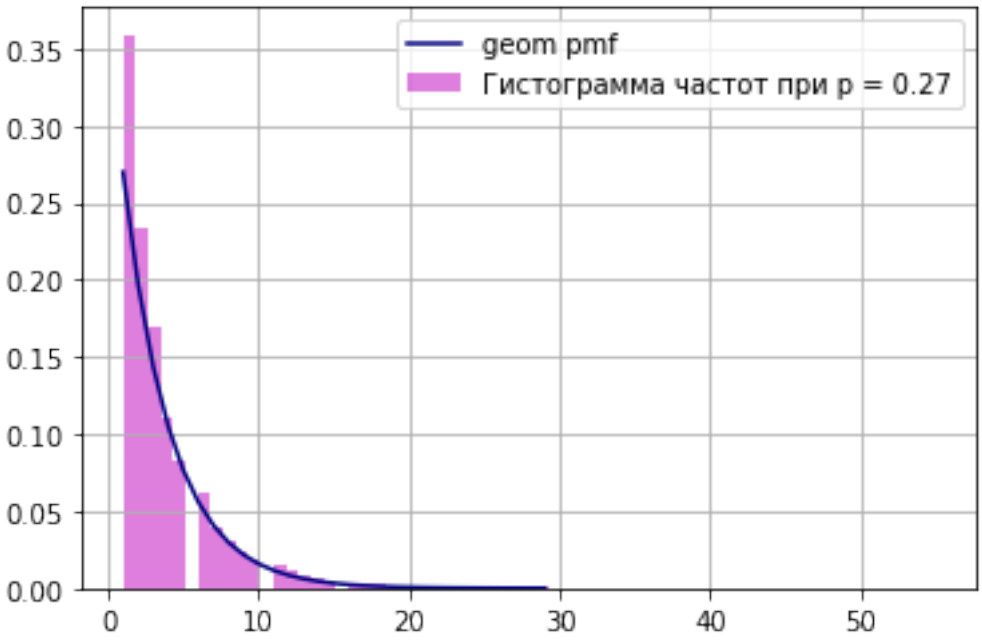
3.4.1 Геометрическое распределение

В перовой домашней работе в качестве нетипичной интерпретации геометрического распределения была выбрана модель, в которой рассматривается ДНК, то есть последовательность 4 видов нуклеотидов, и, например, перед исследователями стоит задача изучить частоту встречаемости какого-то определенного вида. Соответственно, без особых трудностей мной была найдена база данных, показывающая весь геном после применения метода дробовика.[3] Метод дробовика - метод, используемый для секвенирования длинных участков ДНК. Суть метода состоит в получении случайной массированной выборки клонированных фрагментов ДНК данного организма, на основе которых может быть восстановлена исходная последовательность ДНК.

Итак, моя выборка состоит из последовательности нуклеотидов 'A', 'C', 'G','T'. Анализируя эту выборку, я считал расстояние между двумя соседними нуклеотидами 'T'. Я начал двигаться по ДНК, установив при этом счетчик на значение 1. Соответственно, при каждой встрече нужного нуклеотида, значение счетчика заносилось в отдельный массив, затем счетчик устанавливался на значение 1 и движение продолжалось.

В результате анализа, я построил функцию вероятности моей выборки. Синим цветом - теоретически построена функция вероятности с подобранным коэффициентом.

Затем были посчитаны выборочное среднее и выборочная дисперсия. И согласно, полученной мною ранее оптимальной оценки, получена оценка параметра. Они приблизительно оказались равны!!!




```
Выборчоное среднее равно:  3.796
Выборочная дисперсия равна:  14.77
```

```
Оценка параметра равна: 0.263
```

3.4.2 Экспоненциальное распределение

В первой домашней работе в качетсве нетипичной интерпретации для экспоненциального распределения была выбрана теория надежности. Эта интерпретация позволяет исследовать срок службы прибора или механизма, и в нужный момент принять соответствующие меры по ремонту или замене деталей.

Я нашел базу данных, касающуюся безработицы. В этой базе данных

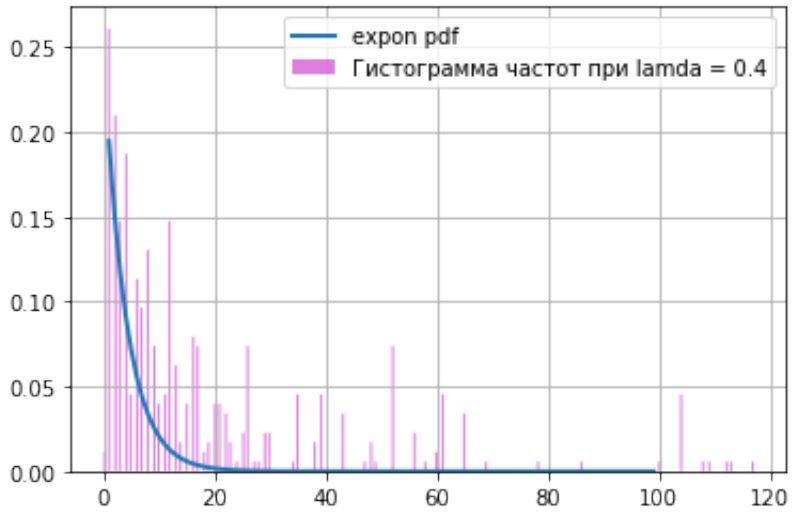
главным фактором была приведена продолжить безработицы, то есть количество дней, которое она продолжалась

Почему, я считаю, что данная интерпретация соответствует экспоненциальному распределению? Работа, или другой похожий способ заработка, определяет доход человека, и соотвественно, его возможности. Если человек, по какой либо причине потеряет работу, то ему будет нечем себя кормить и не на что будет существовать. Но так, как, возможно, что у него(нее) есть накопления, то он сможет спокойно жить без работы, но так как иной прибыли денег нет, то в какой-то момент накопления будут заканчиваться, и жизнь заставит искать работу. Так что со временем, человек перестанет быть безработным

Этот процесс очень схож с жизнью приборов. Чем дольше прибор находится в эксплотации, тем выше вероятность, что он выйдет из строя, а в случае человека - найдет работу.

Для этих данных я проделал тоже самое, что и для предыдущих.

Получил вот такие результаты:



Они приблизительно оказались равны!!!

```
Выборчоное среднее равно:  18.511
Выборочная дисперсия равна:  531.732
```

```
Оценка параметра равна: 0.388
```

Регулярность

Статистическая модель называется *регулярной* (по Рао-Крамеру), если выполнены следующие условия, которые в дальнейшем будем называть условиями регулярности:

1. $L(\bar{x};\theta) > 0$ и дифференцируема по θ, $\forall\theta \in \Theta$ (Θ — параметрическое множество).

2. Случайная величина $V(x;\theta)$, называемая функцией вклада выборки и определенная равенством

$$V(x;\theta) = \frac{\partial \ln(L(\bar{x};\theta))}{\partial\theta} = \sum_{i=1}^{\infty} \frac{\partial \ln(f_i(x_i;\theta))}{\partial\theta},$$

   имеет ограниченную дисперсию:

$$0 < EV^2(X;\theta) < \infty.$$

   При этом значение $\dfrac{\partial \ln(f_i(x_i;\theta))}{\partial\theta}$ будем называть вкладом выборки.

3. $\forall\theta \in \Theta$ $\forall$ статистики $T(X)$ верно равенство:

$$\frac{\partial}{\partial\theta}\int T(x)L(\bar{x};\theta)dx = \int T(x)\frac{\partial L(\bar{x};\theta)}{\partial\theta}dx.$$

Оба распределения являются регулярными моделями.

Так как перед нами степенной ряд, то мы можем спокойно менять знаки интегрирования и дифференцирования под знаком суммы.

Литература

[1]

[2] Ссылка на базу данных для экспоненциального распределения,Unemployment, Unemployment Duration

[3] Ссылка на базу данных для геометрического распределения, NCBI - National Center of Biotechnology Information

[4] // ссылка3

[5] // ссылка4