Домашнее задание 3

Юрасов Никита Андреевич

Обновлено 14 ноября 2019 г.

Содержание

1 Нахождение выборочного среднего и выборочной дисперсии **2**
1.1 Распределение Пуассона 2
1.2 Распределение Эрланга 2

2 Нахождение параметров распределений событий **3**
2.1 Распределение Пуассона 3
2.2 Распределение Эрланга 5

3 Работа с данными **8**
3.1 Распределение Пуассона 8
3.2 Распределение Эрланга 10

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*

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

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

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

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

Данные генерировались с параметром $\lambda = 2$

```
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

Данные генерировались с параметрами $m = 2, \lambda = 0.2$

```
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:

$$L(x,\theta) = \prod_{i=1}^{n} f(x,\theta) = \prod_{i=1}^{n} \left(\frac{\theta^{x_i} e^{-\theta}}{x_i!} \right) = e^{\theta n} \frac{\theta^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}$$

Возьмем от функции правдоподобия натуральный логарифм:

$$lnL(x,\theta) = -\theta n + \sum_{i=1}^{n} x_i \cdot ln\theta - ln \prod_{i=1}^{n} x_i!$$

Продифференцируем полученное выражение по θ и приравняем к нулю:

$$\frac{\partial lnL(x,\theta)}{\partial \theta} = -n + \frac{\sum_{i=1}^{n} x_i}{\theta} = 0$$

Будем решать это уравнение относительно θ:

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

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

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

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

Состоятельность
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
Проверим эффективность предложенной оценки:

Определение 2.1. *Эффективной оценкой называют ту оценку, для которой выполнена нижняя граница неравенства Рао-Крамера:*

$$D_\theta T \geqslant \frac{[\tau'(\theta)]^2}{ni(\theta)}$$

Это условие выполняется тогда и только когда, когда

$$T(x) - \tau(\theta) = a(\theta)V(x,\theta),$$

где $V(x,\theta)$ есть вклад выборки.

Для начала найдем вклад выборки:

$$V(x,\theta) = \frac{\partial lnL(x,\theta)}{\partial\theta} = -n + \frac{\sum\limits_{i=1}^{n} x_i}{\theta}$$

$$V(x,\theta)(\frac{\theta}{n}) = \overline{X} - \theta = T(x) - \tau(\theta),$$

а $a(\theta) = \frac{\theta}{n}$

Найдем нижнюю границу неравенства Рао-Крамера:

$$D\overline{X} = \frac{[\tau'(\theta)]^2}{ni(\theta)} = \frac{[\theta']^2}{ni(\theta)} = \frac{\theta}{n}$$

так как

$$i(\theta) = -M\frac{\partial^2 lnf(x,\theta)}{\partial\theta^2} = -M(-\frac{x}{\theta^2}) = \sum_{i=0}^{n}\frac{i}{\theta^2}\frac{\theta^i e^{-\theta}}{i!} = \frac{e^{-\theta}}{\theta}\sum_{i=0}^{n}\frac{\theta^{i-1}}{(i-1)!} = \frac{1}{\theta}$$

Теперь найдем дисперсию выбранной статистики $T(x)$:

$$Dx_i = \theta, \quad \text{так как } x_i \sim Poiss(\theta)$$

$$D\sum_{i=1}^{n} x_i = n\theta$$

$$D\overline{X} = D\frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n^2}n\theta = \frac{\theta}{n}$$

Следовательно, можно сделать вывод, что выбранная оценка является эффективной. Эта оценка является также оптимальной, так как для нее выполнено требование наименьшей дисперсии.

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:

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

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

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

Продифференцируем полученное по λ и приравняем к нулю:

$$\frac{\partial lnL(x,m,\lambda)}{\partial \lambda} = \frac{mn}{\lambda} - \sum_{i=1}^{n} x_i = 0$$

Решая относительно λ, получим:

$$\hat{\lambda} = \frac{mn}{\sum\limits_{i=1}^{n} x_i} = \frac{m}{\hat{\alpha_1}} \tag{2}$$

Несмещенность
Такая оценка оказывается смещенной:
Надо найти $M(\frac{m}{\overline{X}})$

$$x_i \sim \Gamma(x,m,\lambda)$$

$$\sum_{i=1}^{n} x_i \sim \Gamma(x,mn,\lambda)$$

$$f(x,mn,\lambda) = \frac{\lambda^{mn} x^{mn-1}}{\Gamma(mn-1)} e^{-\lambda x}$$

Пусть $\sum_{i=1}^{n} x_i = l$, тогда найдем $M(\frac{1}{l})$:

$$M\frac{1}{l} = \int\limits_0^{+\infty} \frac{1}{x} \frac{\lambda^{mn} x^{mn-1}}{\Gamma(mn-1)} e^{-\lambda x} dx = \frac{\lambda^{mn}}{\Gamma(mn-1)} \int\limits_0^{+\infty} x^{mn-2} e^{-\lambda x} dx = \boxed{\text{замена } z = \lambda x} =$$

$$= \frac{\lambda}{\Gamma(mn-1)} \int\limits_0^{+\infty} z^{mn-2} e^{-z} dz = \lambda$$

То есть,

$$M\left(\frac{mn}{\sum_{i=1}^{n} x_i}\right) = mn\lambda$$

Эффективность
Найдем вклад выборки:

$$V(x,\lambda) = -\sum_{i=1}^{n} x_i + \frac{nm}{\lambda}$$

$$V(x,\lambda)(-\frac{1}{n}) = \overline{X} - \frac{m}{\lambda}$$

где

$$a(\lambda) = -\frac{1}{n}, T(x) = \overline{X}, \tau(\lambda) = \frac{nm}{\lambda}$$

Из определения эффективной оценки и неравенства Рао-Крамера (см. 2.1) найдем нижнюю границу:

$$DT(x) = \frac{\frac{m^2}{\lambda^4}}{ni(\lambda)} = \frac{m}{n\lambda^2},$$

так как

$$i(\lambda) = -M\frac{\partial^2 lnf(x,m,\lambda)}{\partial\lambda^2} = M(\frac{m}{\lambda^2}) = m\int\limits_0^{+\infty} \frac{1}{\lambda^2}\frac{\lambda^m x^{m-1}}{\Gamma(m)}e^{-\lambda x}dx = \boxed{\text{замена } z=\lambda x} =$$

$$= \frac{m}{\Gamma(m)\lambda^2}\int\limits_0^{+\infty} z^{m-1}e^{-z}dz = \frac{m}{\lambda}$$

Посчитаем дисперсию $T(x)$:

$$Dx_i = \frac{m}{\lambda^2}, \text{ так как } x_i \sim \Gamma(m,\lambda)$$

$$D\sum_{i=1}^{n} x_i = \frac{mn}{\lambda^2}$$

$$D\overline{X} = \frac{1}{n^2}D\sum_{i=1}^{n} x_i = \frac{m}{n\lambda^2}$$

Следовательно оценка эффективна, так как эта оценка имеет нижнюю границу неравенства Рао-Крамера, что подтверждается нахождением $D\overline{X}$. Отсюда следует, что такая оценка является и оптимальной (минимальная дисперсия).

Состоятельность Для проверки состоятельности воспользуемся критерием состоятельности, который вытекает из неравенства Чебышева.

Теорема 2.1. *Критерий состоятельности.*
Для проверки состоятельности некоторой несмещенной оценки T_n для g достаточно убедиться, что ее дисперсия $DT_n \to 0$ при $n \to \infty$, так как в этом случае по неравенству Чебышева

$$P\{|T_n - g| > \epsilon\} \leqslant \frac{DT_n}{\epsilon^2} \to 0, \quad \forall\epsilon$$

Проверим состоятельность:

$$\lim_{n\to\infty} DT_n = \lim_{n\to\infty}\frac{m}{n\lambda^2} = 0$$

Следовательно, оценка состоятельна.

Оценка второго параметра

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

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

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

Выразим из первого уравнения λ:

$$\lambda = \frac{mn}{\sum\limits_{i=1}^{n} x_i}$$

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

$$\hat{m} = \frac{\hat{\alpha_1}^2}{n^4\hat{\alpha_2}} \tag{3}$$

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 на странице 9 (значения были округлены до 5 знака после запятой).

Характеристики для выбранных данных следующие:
`Выборочное среднее всех голов: 2.8210526315789473`
`Выборочная дисперсия всех голов: 2.5574515235457063`

Оценка 1 совпадает по значению с выборочным средним, то есть

$$\hat{\theta} = 2.8210526315789473$$

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

3.2 Распределение Эрланга

Обширной базой данных по более чем 100 видам рака за последние 20 лет обладает Министерство здравоохранения и социальных служб США (U.S. Department of Health and Human Services), который предоставляет информацию в открытом доступе на их сайте.

До начала анализа данных была проведена работа по обработке и сортировке данных, так как все файлы можно получить только в виде обычных текстовых документов. Также я выбрал для более детального анализа рака поджелудочной железы, как одного из самых распространенных видов рака. Ниже будут представлены таблицы со сводными показателями.

Для этих данных нужно определить свои выборочное среднее и выборочную дисперсию, так как данные представлены в количестве человек в каждой возрастной группе, а для характеристик нужно учитывать и сами временные группы.
Пусть выборочное среднее выглядит следующим образом:

$$\overline{X} = \frac{1}{N} \sum_{i=1}^{n} a_i * n_i,$$

где N–сумма всех онкобольных, a_i–номер возрастной группы (индексы в таблицах), n_i–количество онкобольных в каждой возрастной группе.

$$S_n^2 = \frac{1}{N} \sum_{i=1}^{n} n_i \left(a_i - \overline{X}\right)^2$$

Таблица 2: Male Cancer

	Sex	Cancer Sites	Age Group	Count
1	Male	All Invasive Cancer Sites Combined	< 1 year	7446
2	Male	All Invasive Cancer Sites Combined	1-4 years	25803
3	Male	All Invasive Cancer Sites Combined	5-9 years	18507
4	Male	All Invasive Cancer Sites Combined	10-14 years	20008
5	Male	All Invasive Cancer Sites Combined	15-19 years	33580
6	Male	All Invasive Cancer Sites Combined	20-24 years	48242
7	Male	All Invasive Cancer Sites Combined	25-29 years	67302
8	Male	All Invasive Cancer Sites Combined	30-34 years	91123
9	Male	All Invasive Cancer Sites Combined	35-39 years	134169
10	Male	All Invasive Cancer Sites Combined	40-44 years	233607
11	Male	All Invasive Cancer Sites Combined	45-49 years	429675
12	Male	All Invasive Cancer Sites Combined	50-54 years	771615
13	Male	All Invasive Cancer Sites Combined	55-59 years	1117184
14	Male	All Invasive Cancer Sites Combined	60-64 years	1384025
15	Male	All Invasive Cancer Sites Combined	65-69 years	1570122
16	Male	All Invasive Cancer Sites Combined	70-74 years	1525995
17	Male	All Invasive Cancer Sites Combined	75-79 years	1343389
18	Male	All Invasive Cancer Sites Combined	80-84 years	940474
19	Male	All Invasive Cancer Sites Combined	85+ years	669930

Таблица 3: TableName

	Sex	Cancer Sites	Age Group	Count
1	Female	All Invasive Cancer Sites Combined	< 1 year	6620
2	Female	All Invasive Cancer Sites Combined	1-4 years	21708
3	Female	All Invasive Cancer Sites Combined	5-9 years	14840
4	Female	All Invasive Cancer Sites Combined	10-14 years	17964
5	Female	All Invasive Cancer Sites Combined	15-19 years	29577
6	Female	All Invasive Cancer Sites Combined	20-24 years	51307
7	Female	All Invasive Cancer Sites Combined	25-29 years	89907
8	Female	All Invasive Cancer Sites Combined	30-34 years	153644
9	Female	All Invasive Cancer Sites Combined	35-39 years	256177
10	Female	All Invasive Cancer Sites Combined	40-44 years	436685
11	Female	All Invasive Cancer Sites Combined	45-49 years	655748
12	Female	All Invasive Cancer Sites Combined	50-54 years	843441
13	Female	All Invasive Cancer Sites Combined	55-59 years	968075
14	Female	All Invasive Cancer Sites Combined	60-64 years	1054342
15	Female	All Invasive Cancer Sites Combined	65-69 years	1113436
16	Female	All Invasive Cancer Sites Combined	70-74 years	1116917
17	Female	All Invasive Cancer Sites Combined	75-79 years	1104209
18	Female	All Invasive Cancer Sites Combined	80-84 years	923852
19	Female	All Invasive Cancer Sites Combined	85+ years	884462

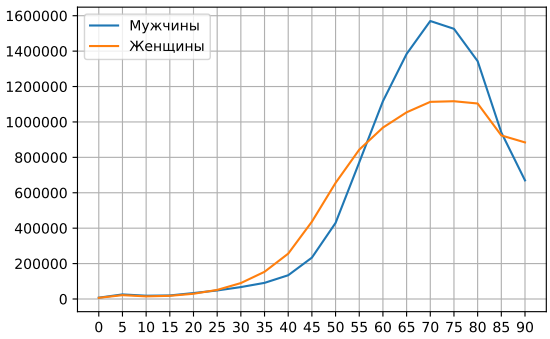
Таблица 4: Рак поджелудочной железы

Таблица 5: Мужчины

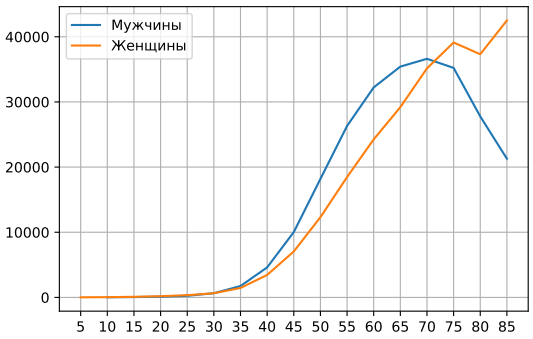
	Sex	Cancer	Age Group	Count
4	Male	Pancreas	10-14 years	20
5	Male	Pancreas	15-19 years	53
6	Male	Pancreas	20-24 years	114
7	Male	Pancreas	25-29 years	250
8	Male	Pancreas	30-34 years	644
9	Male	Pancreas	35-39 years	1745
10	Male	Pancreas	40-44 years	4607
11	Male	Pancreas	45-49 years	10024
12	Male	Pancreas	50-54 years	18191
13	Male	Pancreas	55-59 years	26309
14	Male	Pancreas	60-64 years	32225
15	Male	Pancreas	65-69 years	35420
16	Male	Pancreas	70-74 years	36618
17	Male	Pancreas	75-79 years	35223
18	Male	Pancreas	80-84 years	27779
19	Male	Pancreas	85+ years	21291

Таблица 6: Женщины

	Sex	Cancer	Age Group	Count
3	Female	Pancreas	5-9 years	17
4	Female	Pancreas	10-14 years	38
5	Female	Pancreas	15-19 years	92
6	Female	Pancreas	20-24 years	185
7	Female	Pancreas	25-29 years	345
8	Female	Pancreas	30-34 years	636
9	Female	Pancreas	35-39 years	1453
10	Female	Pancreas	40-44 years	3448
11	Female	Pancreas	45-49 years	7050
12	Female	Pancreas	50-54 years	12311
13	Female	Pancreas	55-59 years	18467
14	Female	Pancreas	60-64 years	24257
15	Female	Pancreas	65-69 years	29196
16	Female	Pancreas	70-74 years	35149
17	Female	Pancreas	75-79 years	39120
18	Female	Pancreas	80-84 years	37319
19	Female	Pancreas	85+ years	42498



Общее количество онкобольных



Рак поджелудочной железы

Рис. 1: Графики по 4 выборкам

```
Выборочное среднее мужчин, болеющих раком: 14
Выборочное среднее женщин, болеющих раком: 13
Выборочное среднее мужчин, болеющих раком поджелудочной железы: 11
Выборочное среднее женщин, болеющих раком поджелудочной железы: 13
```

```
Выборочная дисперсия мужчин, болеющих раком: 8
Выборочная дисперсия женщин, болеющих раком: 10
Выборочная дисперсия мужчин, болеющих раком поджелудочной железы: 6
Выборочная дисперсия женщин, болеющих раком поджелудочной железы: 6
```

О полученных характеристиках нужно сделать оговорку, что числа означают не количество человек, а ближайшую возрастную группу.

Значение оценки на данных

$$\hat{\lambda} = \frac{m}{\overline{X}}$$

Пусть m еще известна до начала сбора данных, а именно $m \approx 15$.
Тогда значение для:

- всех мужчин: $\lambda \approx 8.28571$
- всех женщин: $\lambda \approx 8.92308$
- женского рака поджелудочной железы: $\lambda \approx 1.36364$
- мужского рака поджелудочной железы: $\lambda \approx 1.15385$