Домашнее задание 3. Оценки.

Шубин Никита СКБ172

1. Нахождение выборочного среднего и выборочной дисперсии.

Выборочное среднее - это приближение теоретического среднего распределения, основанное на выборке из него, и рассчитывается по следующей формуле:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{\infty} x_i = \int_{1}^{\infty} t\, dF_n\,(f)$$

Выборочная дисперсия - это оценка теоретической дисперсии распределения, имеющая вид:

$$s_2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{X})^2 = \int_{-\infty}^{\infty} (t - \bar{X})^2\, dF_n\,(f)$$

1.1 Логарифмическое распределение.

Выборочное среднее геометрического распределения можно найти следующим образом:

```
num = [5, 10, 100, 1000, 10000]
for i in range(5):
    disp = []
    data = []
    for j in range(len(num)):
        with open("Log{}.txt".format("{}_{}".format(num[i], j+1))) as f:
            for line in f:
                data.append(list([float(x) for x in line.split()]))
    for k in range(len(data)):
        mean = sum(data[k])/len(data[k])
        print("Выборочное среднее при n = {}:".format(len(data[k])), mean)
```

Результат выполнения программы:

```
Выборочное среднее при n = 5: 1.0
Выборочное среднее при n = 5: 1.8
Выборочное среднее при n = 5: 2.6
Выборочное среднее при n = 5: 1.6
Выборочное среднее при n = 5: 1.6
Выборочное среднее при n = 10: 1.2
Выборочное среднее при n = 10: 2.1
Выборочное среднее при n = 10: 1.2
Выборочное среднее при n = 10: 1.2
Выборочное среднее при n = 10: 1.3
Выборочное среднее при n = 100: 1.55
Выборочное среднее при n = 100: 1.54
Выборочное среднее при n = 100: 1.49
Выборочное среднее при n = 100: 1.38
Выборочное среднее при n = 100: 1.49
Выборочное среднее при n = 1000: 1.486
Выборочное среднее при n = 1000: 1.517
Выборочное среднее при n = 1000: 1.547
Выборочное среднее при n = 1000: 1.47
Выборочное среднее при n = 1000: 1.543
Выборочное среднее при n = 10000: 1.4927
Выборочное среднее при n = 10000: 1.5021
Выборочное среднее при n = 10000: 1.4843
Выборочное среднее при n = 10000: 1.4929
```

Код для нахождения выборочной дисперсии:

```python
num = [5, 10, 100, 1000, 10000]
for i in range(5):
    disp = []
    mean = []
    data = []
    for j in range(len(num)):
        with open("Log{}.txt".format("{}_{}".format(num[i], j+1))) as f:
            for line in f:
                data.append(list([float(x) for x in line.split()]))
    for k in range(len(data)):
        #mean = sum(data[k])/len(data[k])
        mean.append(sum(data[k]) / len(data[k]))
        #print("Выборочное среднее при n = {}:".format(len(data[k])), mean)
    for m in range(5):
        summ = 0
        for n in range(len(data[m])):
            summ += (data[m][n] - mean[m])**2
        disp.append(summ/len(data[m]))
    for o in range(len(disp)):
        print("Дисперсия при n = {}".format(len(data[1])), disp[o])
```

Результат выполнения программы:

```
Дисперсия при n = 5 0.0
Дисперсия при n = 5 0.16000000000000006
Дисперсия при n = 5 0.64
Дисперсия при n = 5 0.24
Дисперсия при n = 5 0.64
Дисперсия при n = 10 0.16000000000000006
Дисперсия при n = 10 3.090000000000001
Дисперсия при n = 10 0.16
Дисперсия при n = 10 0.16000000000000003
Дисперсия при n = 10 0.21000000000000002
Дисперсия при n = 100 0.967499999999999
Дисперсия при n = 100 0.9284000000000008
Дисперсия при n = 100 0.7298999999999991
Дисперсия при n = 100 0.5555999999999988
Дисперсия при n = 100 0.6298999999999993
Дисперсия при n = 1000 0.8478039999999873
Дисперсия при n = 1000 1.0677110000000063
Дисперсия при n = 1000 1.069791000000008
Дисперсия при n = 1000 0.9250999999999977
Дисперсия при n = 1000 0.9221509999999954
Дисперсия при n = 10000 0.9327467100000438
Дисперсия при n = 10000 0.9853955900000365
Дисперсия при n = 10000 0.8855535100000163
Дисперсия при n = 10000 0.9181495900000024
Дисперсия при n = 10000 0.9119228720664978
```

Свойства выборочного среднего и выборочной дисперсии.

1. Выборочное среднее $\bar{X} = \frac{1}{n}\sum_{i=1}^{\infty} x_i$ является несмещенной оценкой x_i.

Проверим это, найдя математическое ожидание выборочного среднего.

$$M_0(\bar{X}) = M_0 \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{i=1}^{n} M_0(x_i) = \frac{nM_0(x_i)}{n} = M_0(x_i)$$

2. Выборочное среднее при $n \to \infty$, стремится к математическому ожиданию случайной величины. Проверим, так ли это:

$$Mx = \boxed{\frac{p}{(p-1)\ln(1-p)}} = \{\text{при } p = 0.53\} \sim 1.5$$

Уже при выборке в 1000, выборочное среднее стремится к математическому ожиданию.

3. При $n \to \infty$ выборочная дисперсия сходится к дисперсии случайной величины.

$$D(\zeta) = \boxed{-p\,\frac{p+\ln(1-p)}{(p-1)^2\ln^2(1-p)}} = \{\text{при } p = 0.53\} \sim 0.94$$

Так, уже на выборке из 100000 можно заметить стремление выборочной дисперсии к дисперсии случайной величины.

1.2 Распределение Эрланга.

Код для выборочного среднего такой же.

Результат выполнения для параметров 24, 0.8:

```
Выборочное среднее при n = 5: 29.05583781604613
Выборочное среднее при n = 5: 27.139476310566863
Выборочное среднее при n = 5: 34.01254758997517
Выборочное среднее при n = 5: 30.91355820009601
Выборочное среднее при n = 5: 32.414961189766544
Выборочное среднее при n = 10: 30.72622624358497
Выборочное среднее при n = 10: 32.92100169739641
Выборочное среднее при n = 10: 30.468641740421855
Выборочное среднее при n = 10: 28.484073037712328
Выборочное среднее при n = 10: 31.578627584512482
Выборочное среднее при n = 100: 29.80922480017805
Выборочное среднее при n = 100: 30.264882857157886
Выборочное среднее при n = 100: 29.657744427294638
Выборочное среднее при n = 100: 30.030079913190338
Выборочное среднее при n = 100: 30.22347383953637
Выборочное среднее при n = 1000: 29.646153617242685
Выборочное среднее при n = 1000: 30.07277497956174
Выборочное среднее при n = 1000: 30.23604627889212
Выборочное среднее при n = 1000: 29.893029911086774
Выборочное среднее при n = 1000: 30.01369684818322
Выборочное среднее при n = 100000: 30.01065521459933
Выборочное среднее при n = 100000: 29.96761574430949
Выборочное среднее при n = 100000: 29.973053099428743
Выборочное среднее при n = 100000: 30.006810296469382
```

Код для нахождения выборочной дисперсии такой же.

Результат выполнения:

```
Дисперсия при n = 5 0.0
Дисперсия при n = 5 0.16000000000000006
Дисперсия при n = 5 0.64
Дисперсия при n = 5 0.24
Дисперсия при n = 5 0.64
Дисперсия при n = 10 0.16000000000000006
Дисперсия при n = 10 3.090000000000001
Дисперсия при n = 10 0.16
Дисперсия при n = 10 0.16000000000000003
Дисперсия при n = 10 0.21000000000000002
Дисперсия при n = 100 0.9674999999999999
Дисперсия при n = 100 0.9284000000000008
Дисперсия при n = 100 0.7298999999999991
Дисперсия при n = 100 0.5555999999999988
Дисперсия при n = 100 0.6298999999999993
Дисперсия при n = 1000 0.8478039999999873
Дисперсия при n = 1000 1.0677110000000063
Дисперсия при n = 1000 1.069791000000008
Дисперсия при n = 1000 0.9250999999999977
Дисперсия при n = 1000 0.9221509999999954
Дисперсия при n = 10000 0.9327467100000438
Дисперсия при n = 10000 0.9853955900000365
Дисперсия при n = 10000 0.8855535100000163
Дисперсия при n = 10000 0.9181495900000024
Дисперсия при n = 10000 0.9119228720664978
```

2. Нахождение параметров распределений событий.

2.1 Логарифмическое распределение.

$$P(X = x_i) = \frac{-\,p^{x_i}}{x_i \ln(1-p)}$$

Оценку параметра p найдем с помощью метода наибольшего правдоподобия.

Для начала составим функцию правдоподобия:

$$L(p) = \prod_{i=1}^{n} p(x_i, \lambda) = \prod_{i=1}^{n} \frac{-\,p^{x_i}}{x_i \ln(1-p)} = \frac{1}{\mathrm{n} \ln\,(1-p)} \prod_{i=1}^{n} \frac{-p^{x_i}}{x_i}$$

Тогда:

$$\ln\bigl(L(p)\bigr) = \ln\left[\frac{1}{\mathrm{n} \ln\,(1-p)} \prod_{i=1}^{n} \frac{-p^{x_i}}{x_i}\right] = \ln\left[\frac{1}{\mathrm{n} \ln\,(1-p)}\right] + \ln\left[\prod_{i=1}^{n} \frac{-p^{x_i}}{x_i}\right] =$$

$$= \ln\frac{1}{\mathrm{n} \ln\,(1-p)} + \sum_{i=1}^{n} \ln\left[\frac{-p^{x_i}}{x_i}\right] = \ln\frac{1}{\mathrm{n} \ln\,(1-p)} + \sum_{i=1}^{n} x_i \ln(-p) - \ln\,(x_i) =$$

Условие экстремума:

$$\frac{d \ln L}{dp} = \frac{d}{dp}\left(\ln\frac{1}{\mathrm{n} \ln\,(1-p)} + \sum_{i=1}^{n} x_i \ln(-p) - \ln\,(x_i)\right) =$$

$$= \frac{1}{(1-p)\ln\,(1-p)} - \sum_{i=1}^{n} \frac{x_i}{p} = 0,$$

$$\frac{1}{(1-p)\ln\,(1-p)} - \sum_{i=1}^{n} \frac{x_i}{p} = 0$$

Преобразуем:

$$\sum_{i=1}^{n} \frac{x_i}{p} = \frac{1}{(1-p)\ln(1-p)}$$

$$\sum_{i=1}^{n} x_i = \frac{p}{(1-p)\ln(1-p)} = \frac{p}{(1-p)} \frac{1}{\ln(1-p)}$$

(Для геометрического)

Таким образом, в качестве оценки получаем: $p = 1/(1+\bar{x})$

Данная оценка является состоятельной, так как $p = 1/(1+\bar{x})$ есть непрерывная функция.

Проверим оценку на смещение.

Оценка параметра называется несмещенной, если ее математическое ожидание равно истинному значению оцениваемого параметра.

$$M\left(\frac{1}{1+x_1}\right) = \sum_{x_1=0}^{\infty} \frac{1}{1+x_1} p(1-P)^{x_1} = \frac{1}{1} p(1-p)^0 + \sum_{x_1=1}^{\infty} \frac{1}{1+x_1} p(1-p)^{x_1}$$

$$= p + \sum_{x_1=1}^{\infty} \frac{1}{1+x_1} p(1-p)^{x_1} > p$$

Так как мат. ожидание оценки не совпало с параметром, оценка является смещенной, следовательно, эта оценка не является эффективной.

Докажем, что эта оценка достаточна.

Воспользуемся критерием факторизации:

Статистика T достаточна тогда, и только тогда, когда правдоподобие L представимо в виде: $L(x_1, \dots, x_n;\ \theta) = f(T(x_1, \dots, x_n), \theta) h(x_1, \dots, x_n)$, где h и f некоторые борелевские функции.

$$L(x_1, \dots, x_n;\ p) = p^n (1-p)^{\sum_{i=1}^{n} x_i} = p^n (1-p)^{n(\frac{1}{n}\sum_{i=1}^{n} x_i)} = p^n (1-p)^{n(\bar{x})} =$$

$$p^n (1-p)^{n(\bar{x}+1-1)} = p^n (1-p)^{n\left(\frac{1}{T(x)}-1\right)} = f(T(x_1, \dots, x_n), p) h(x_1, \dots, x_n)$$

Получается, что $h(x_1, \dots, x_n) = 1$, а $f(T(x_1, \dots, x_n), p) = p^n (1-p)^{n\left(\frac{1}{T(x)}-1\right)}$.

Следовательно, данная статистика является достаточной. Проверим расхождение параметра и его оценки:

```python
num = [5, 10, 100, 1000, 100000]
for i in range(5):
    disp = []
    #mean = []
    data = []
    for j in range(len(num)):
        with open('Geom{}.txt'.format('{}_{}'.format(num[i], j+1))) as f:
            for line in f:
                data.append(list([float(x) for x in line.split()]))
    for k in range(len(data)):
        mean = sum(data[k]) / len(data[k])
        #mean.append(sum(data[k]) / len(data[k]))
        print('Разница при n = {}:'.format(len(data[k])), abs(0.2 - 1/(1+mean)))
```

```
Разница при n = 5: 0.11249999999999999
Разница при n = 5: 0.027272727272727254
Разница при n = 5: 0.008333333333333331
Разница при n = 5: 0.027272727272727254
Разница при n = 5: 0.05294117647058824
Разница при n = 10: 0.06842105263157897
Разница при n = 10: 0.027586206896551724
Разница при n = 10: 0.021428571428571436
Разница при n = 10: 0.008333333333333331
Разница при n = 10: 0.043902439024390255
Разница при n = 100: 0.026757369614512444
Разница при n = 100: 0.0307952622673435
Разница при n = 100: 0.0003992015968063978
Разница при n = 100: 0.002371541501976271
Разница при n = 100: 0.004687500000000011
Разница при n = 1000: 0.004381846635367781
Разница при n = 1000: 0.005044084478162791
Разница при n = 1000: 0.0013508144616607154
Разница при n = 1000: 0.0003205128205128194
Разница при n = 1000: 0.00353634577603143
Разница при n = 100000: 0.0013863435892685072
Разница при n = 100000: 0.0010694506333948917
Разница при n = 100000: 0.0006565482257948063
Разница при n = 100000: 0.0001000500250124825
Разница при n = 100000: 0.00016774056659477066
```

2.1 Распределение Коши.

Найдем оценки параметров с помощью метода наибольшего правдоподобия.

Функция правдоподобия имеет вид:

$$L(\gamma,\mu) = \prod_{i=1}^{n} \frac{1}{\pi\gamma\left[1+\left(\frac{x-\mu}{\gamma}\right)^2\right]} = \frac{1}{\pi^n}\prod_{i=1}^{n}\frac{1}{\gamma+\gamma\frac{(x-\mu)^2}{\gamma^2}} =$$

$$= \frac{\gamma^n}{\pi^n}\prod_{i=1}^{n}\frac{1}{\gamma^2+(x_i-\mu)^2}$$

Прологарифмируем:

$$\ln L(\gamma, \mu) = n \ln \gamma - n \ln \pi + \sum_{i=1}^{n} \ln \left[\frac{1}{\gamma^2 + (x_i - \mu)^2} \right]$$

Продифференцируем:

$$\frac{d \ln L(\gamma, \mu)}{d\gamma} = \frac{n}{\gamma} - \sum_{i=1}^{n} \frac{2\gamma}{\gamma^2 + (x_i - \mu)^2} = 0$$

$$\frac{d \ln L(\gamma, \mu)}{d\mu} = 2 \sum_{i=1}^{n} \frac{x_i - \mu}{\gamma^2 + (x_i - \mu)^2} = 0$$

Таким образом, оценки параметров γ и μ будут корнями системы уравнений:

$$\begin{cases} \dfrac{n}{\gamma} - \displaystyle\sum_{i=1}^{n} \frac{2\gamma}{\gamma^2 + (x_i - \mu)^2} = 0 \\ 2 \displaystyle\sum_{i=1}^{n} \frac{x_i - \mu}{\gamma^2 + (x_i - \mu)^2} = 0 \end{cases} \leftrightarrow \begin{cases} F(\gamma, \mu) = 0 \\ G(\gamma, \mu) = 0 \end{cases}$$

Система имеет решение при условии, что параметр $\gamma \neq 0$.

Решение можно получить с помощью метода Ньютона, высчитывая якобианы F и G и сами F и G на каждой итерации.