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 Геометрическое 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("Geom{}_{}.txt".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: 3.6
Выборочное среднее при n  = 5: 3.4
Выборочное среднее при n  = 5: 2.6
Выборочное среднее при n  = 5: 7.6
Выборочное среднее при n  = 5: 3.4
Выборочное среднее при n  = 10: 4.3
Выборочное среднее при n  = 10: 3.3
Выборочное среднее при n  = 10: 5.2
Выборочное среднее при n  = 10: 5.1
Выборочное среднее при n  = 10: 4.7
Выборочное среднее при n  = 100: 4.09
Выборочное среднее при n  = 100: 5.11
Выборочное среднее при n  = 100: 5.03
Выборочное среднее при n  = 100: 4.16
Выборочное среднее при n  = 100: 4.46
Выборочное среднее при n  = 1000: 4.427
Выборочное среднее при n  = 1000: 4.697
Выборочное среднее при n  = 1000: 4.44
Выборочное среднее при n  = 1000: 4.416
Выборочное среднее при n  = 1000: 4.368
Выборочное среднее при n  = 10000: 4.4551
Выборочное среднее при n  = 10000: 4.4885
Выборочное среднее при n  = 10000: 4.4878
Выборочное среднее при n  = 10000: 4.4852
Выборочное среднее при n  = 10000: 4.5282
```

Код для нахождения выборочной дисперсии:

```
num = [5, 10, 100, 1000, 10000]
for i in range(5):
    disp = []
    mean = []
    data = []
    for j in range(len(num)):
        with open("Geom{}_{}.txt".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])
        #rint("Выборочное среднее при n  = {}:".format(len(data[k])),mean)
        mean.append(sum(data[k]) / len(data[k]))
    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 = 10: 26.889999999999997
Дисперсия при n = 10: 19.01
Дисперсия при n = 100: 11.8619
Дисперсия при n = 100: 22.077900000000003
Дисперсия при n = 100: 19.5091
Дисперсия при n = 100: 17.9544
Дисперсия при n = 100: 19.928400000000003
Дисперсия при n = 1000: 20.420671000000016
Дисперсия при n = 1000: 19.585190999999963
Дисперсия при n = 1000: 20.306400000000004
Дисперсия при n = 1000: 20.74894400000002
Дисперсия при n = 1000: 19.54457600000001
Дисперсия при n = 10000: 20.335383990000537
Дисперсия при n = 10000: 20.272667749999954
Дисперсия при n = 10000: 20.580851160000417
Дисперсия при n = 10000: 20.14478095999955
Дисперсия при n = 10000: 20.94180475999781
```

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:

$$Mx = \frac{q}{p} = \frac{0{,}8}{0{,}2} = 4$$

Уже при выборке в 1000, выборочное среднее стремится к математическому ожиданию. Однако, видимо из-за неверного моделирования есть отклонение.

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

$$D(\zeta) = \frac{q}{p^2} = \frac{0{,}8}{(0{,}2)^2} = 20$$

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

2. Нахождение параметров распределений событий.

2.1 Геометрическое распределение.

$P(X = x_i) = (1-p)^{x_i-1}p$

Оценку параметра p найдем с помощью метода наибольшего правдоподобия.

Для начала составим функцию правдоподобия:

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

Тогда:

$$\ln\big(L(p)\big) = \ln\left[p^n\prod_{i=1}^{n}(1-p)^{x_i-1}\right] = \ln[p^n] + \ln\left[\prod_{i=1}^{n}(1-p)^{x_i-1}\right] =$$

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

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

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

Условие экстремума:

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

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

$$n\frac{1}{p} + \frac{1}{p-1}\sum_{i=1}^{n} x_i - n\frac{1}{p-1} = 0$$

Преобразуем:

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

$$-\frac{p-1}{p} = \left(\sum_{i=1}^{n} x_i - n\right)/n,$$

$$\frac{1}{p} - 1 = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

$$\frac{1}{p} = \frac{1}{n}\sum_{i=1}^{n} x_i + 1,$$

$$p = 1\Big/\left(1 + \frac{1}{n}\sum_{i=1}^{n} x_i\right).$$

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:

```
num = [5, 10, 100, 1000, 10000]
for i in range(5):
    disp = []
    data = []
    for j in range(len(num)):
        with open("Geom{}_{}.txt".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("Разница при n = {}:".format(len(data[k])), abs(0.2 - 1/(1+mean)))
```

```
Разница при n = 5: 0.017391304347826098
Разница при n = 5: 0.027272727272727254
Разница при n = 5: 0.07777777777777778
Разница при n = 5: 0.08372093023255815
Разница при n = 5: 0.027272727272727254
Разница при n = 10: 0.01132075471698113
Разница при n = 10: 0.03255813953488371
Разница при n = 10: 0.038709677419354854
Разница при n = 10: 0.0360655737704918
Разница при n = 10: 0.02456140350877195
Разница при n = 100: 0.00353634577603143
Разница при n = 100: 0.0363338788870704
Разница при n = 100: 0.034162520729684925
Разница при n = 100: 0.006201550387596927
Разница при n = 100: 0.016849816849816873
Разница при n = 1000: 0.015736134144094333
Разница при n = 1000: 0.02446901878181501
Разница при n = 1000: 0.01617647058823532
Разница при n = 1000: 0.015361890694239322
Разница при n = 1000: 0.013710879284649785
Разница при n = 10000: 0.016685303660794487
Разница при n = 10000: 0.017800856335975224
Разница при n = 10000: 0.017777615802325175
Разница при n = 10000: 0.01769124188726026
Разница при n = 10000: 0.01910929416446583
```