Лекция 4

Числовые характеристики случайных величин

Вся информация о СВ содержится в ее функции распределения. Источниками информации могут быть априорные знания (например, равновозможные угловые положения проецируемого стержня) или статистические данные. На практике не всегда можно быть уверенным в априорной информация (когда летящий стержень имеет аэродинамическое качество) или нецелесообразно идти на большие расходы, чтобы накопить статистику для построения эмпирической функции распределения. Для оценки СВ часто достаточно знать ее числовые характеристики (среднее значение, степень разброса и т.п.). Кстати, именно *неслучайные характеристики* случайных величин используются в качестве показателей эффективности.

Величины, в сжатой форме выражающие наиболее существенные черты распределения СВ, называются числовыми характеристиками распределения. Все числовые характеристики (ЧХ) определяются законом распределения, если он известен. При экспериментальном изучении СВ оценки ЧХ могут быть вычислены непосредственно по результатам наблюдений без построения законов распределения.

Числовые характеристики и их свойства

Математическое ожидание СВ

Чаще всего используется ЧХ, как бы заменяющая саму СВ. *Математическое ожидание* (МО) представляет собой взвешенную по вероятностям сумму возможных значений СВ, аналогичное по смыслу интегральное выражение для непрерывной СВ (весами выступают элементы вероятностей) или комбинация этих выражений для дискретно-непрерывной СВ:

$$M[X] \equiv m_x = \begin{cases} \sum_{i=1}^{n} x_i p_i, \text{ или } \lim\limits_{\mathrm{n}\to\infty} \sum_{i=1}^{n} x_i p_i, \text{ если ряд сходится,} & (4.1) \\ \int_{-\infty}^{\infty} x f(x) dx, \quad \text{если интеграл сходится,} & (4.2) \\ \int_{-\infty}^{\infty} x f(x) dx + \sum_{i=1}^{n} x_i p_i \quad \text{для смешанной СВ.} & (4.3) \end{cases}$$

Если бесконечный ряд или интеграл не сходятся, МО для такой СВ не существует.

Различие между МО и средним арифметическим

Как следует из формулы (3.9), среднее арифметическое реализаций СВ (выборки) можно заменить статистической оценкой, использующей частость попадания реализаций в достаточно мелкие разряды диапазона возможных значений:

$$m_{\text{ср}} = \frac{1}{N} \sum_{i=1}^{N} X_i \approx \sum_{j=1}^{n} x_j p_j^* \approx \sum_{j=1}^{n} x_j p_j = m_x \,. \qquad (4.4)$$

Вычисленное таким образом среднее является приближенной оценкой МО m_x, как и частоты p_j^* для вероятностей p_j.

Дисперсия

МО тем лучше заменяет саму СВ, чем меньше в среднем отклоняются от него возможные значения. Отклонения СВ X от среднего значения – это *центрированная* СВ $\dot{X} = X - m_x$, ее МО равно нулю, но МО квадратов отклонений положительно и тем больше, чем вероятнее большие отклонения. МО квадрата центрированной СВ называется *дисперсией*:

$$D[X] \equiv D_x = M[(X - m_x)^2]. \tag{4.5}$$

Выражения для дисперсии дискретной и непрерывной СВ получим подстановкой формул (4.1) и (4.2) в (4.5):

$$D[X] = \begin{cases} \sum_{i=1}^{n} (x_i - m_x)^2 p_i, \quad \text{или } \lim_{n\to\infty} \sum_{i=1}^{n} (x_i - m_x)^2 p_i, & (4.6) \\ \int_{-\infty}^{\infty} (x - m_x)^2 f(x) dx. & (4.7) \end{cases}$$

Условия существования дисперсии вытекают из условий существования МО.

Среднеквадратическое отклонение

Среднеквадратическое отклонение (СКО) $\sigma_x = \sqrt{D_x}$ имеет размерность самой СВ. Обычно выражениями вида $m_x \pm k\sigma_x$ задают интервал возможных значений СВ с определенной степенью отклонений, задаваемой коэффициентом k. Так, почти все реализации СВ, подчиненной нормальному закону, находятся в интервале $m_x \pm 3\sigma_x$ (правило «трех сигм»).

Начальные и центральные моменты

МО и дисперсия отражают наиболее важные свойства распределений, но используются и другие ЧХ, определяемые через МО целых степеней СВ. *Начальным моментом* k-о порядка СВ X называется МО k-й степени X:

$$\alpha_k[X] = M[X^k]. \tag{4.8}$$

Центральным моментом k-о порядка СВ X называется МО k-й степени центрированной СВ X:

$$\mu_k[X] = M[\dot{X}^k]. \tag{4.9}$$

Асимметрия распределения

Все центральные моменты нечетных порядков равны нулю для симметричных распределений, у которых одинаковые противоположные отклонения от m_x равновероятны. Так как $\mu_1[X] \equiv 0$ для любых распределений, в качестве меры асимметрии принимается $\mu_3[X]$, точнее, безразмерная величина *As – скошенность* или *асимметрия* СВ X:

$$As = \frac{\mu_3[X]}{\sigma_x^3}. \tag{4.10}$$

Эксцесс распределения

Центральный момент четвертого порядка характеризует «островершинность» распределения по сравнению с нормальным законом распределения, у которого безразмерная величина $\mu_4[X]/\sigma_x^4$ равна 3:

$$Ex = \frac{\mu_4[X]}{\sigma_x^4} - 3. \tag{4.11}$$

Все перечисленные ЧХ называются интегральными или моментными, они определяются через соответствующие начальные и центральные моменты. Еще две ЧХ выделяют характерные значения СВ, так называемые *характеристики положения – мода* (*Mo*) и *медиана* (*Me*).

Мода

Мода дискретной СВ – это ее наиболее вероятное значение: $Mo = x_k$, если $p_k \geq p_i, \ \forall i$. Мода непрерывной СВ доставляет максимум функции плотности распределения: $f(Mo) \geq f(x), \ \forall x$. Функция плотности $f(x)$ может иметь один или несколько максимумов, может иметь минимум или быть постоянной. Соответственно распределение может быть унимодальными, полимодальным, антимодальным или немодальным.

Медиана

Медиана СВ – такое возможное значение *Me*, что события ($X < Me$) и ($X > Me$) равновероятны. Если СВ непрерывна, $F(Me) = 1/2$, но в общем случае медиана обладает свойством $F(Me) \leq 1/2$, $F(Me+0) \geq 1/2$, так как $P(X < Me) \neq P(X > Me) + P(X = Me)$, если медиана приходится на точку разрыва функции распределения.

У непрерывных симметричных распределений мода, медиана и МО совпадают $Mo = Me = m_x = F^{-1}(1/2)$. Медиана смешанных СВ имеет следующий смысл: $F(Me) \leq 1/2$, $F(Me + 0) \geq 1/2$.

Срединное (вероятное) отклонение

Возможное значение x_p $(0 < p < 1)$, для которого выполняется $F(x_p) \leq p$, $F(x_p + 0) \geq p$, называется *квантилью* порядка p (медиана – квантиль порядка 1/2). Величина $E = (x_{3/4} - x_{1/4})/2$ называется *срединным (вероятным) отклонением*. Она так же как и СКО характеризует разброс случайной величины, но может быть легко вычислена (непосредственно из функции распределения) даже для таких распределений, у которых второй центральный момент не существует.

Основные свойства МО

Все интегральные характеристики определены через МО, так что все их важные свойства вытекают из свойств МО:

1. $M[c] = c$, если c – неслучайная величина (имеет единственное возможное значение);
2. $M[cX] = cM[X]$, так как умножение СВ на константу означает умножение всех возможных значений на эту константу;
3. $M[X+Y] = M[X] + M[Y]$ для любых X, Y;
4. $M[XY] = M[X]\,M[Y]$, если X, Y независимы.

Дискретные СВ *независимы*, если независимы все пары событий $(X = x_i)$ и $(Y = y_j)$, $i = 1, \ldots, n$, $j = 1, \ldots, m$. В этом случае $p_{ij} = P(X = x_i \cap Y = y_j) = P(X = x_i)P(Y = y_j) = p_i g_j$, откуда следует свойство 4:

$$M[XY] = \sum_{i=1}^{n}\sum_{j=1}^{m} x_i y_j p_{ij} = \sum_{i=1}^{n}\sum_{j=1}^{m} x_i y_j p_i g_j = \sum_{i=1}^{n} x_i p_i \sum_{j=1}^{m} y_j g_j = M[X]M[Y].$$

Сумма $X + Y$ имеет nm возможных значений $x_i + y_j$ с вероятностями $P(X+Y = x_i + y_j) = P(X = x_i \cap Y = y_j) = p_{ij}$. Без предположения о независимости X, Y получим:

$$M[X+Y]=\sum_{i=1}^{n}\sum_{j=1}^{m}(x_i+y_j)p_{ij}=$$

$$=\sum_{i=1}^{n}x_i\sum_{j=1}^{m}p_{ij}+\sum_{j=1}^{m}y_j\sum_{i=1}^{n}p_{ij}=\sum_{i=1}^{n}x_ip_i+\sum_{j=1}^{m}y_jg_j=M[X]+M[Y].$$

Здесь учтено, что $\sum_{j=1}^{m}p_{ij}=\sum_{j=1}^{m}P((X=x_i)\cap(Y=y_j))=P(X=x_i)=p_i$, так как события $(Y=y_j)$, образуют полную группу. Аналогично, $\sum_{i=1}^{n}p_{ij}=g_j$.

Таким образом, МО суммы любых СВ равно сумме МО слагаемых, а МО произведения независимых СВ равно произведению их МО (свойство 3).

Корреляционный момент

В общем случае МО произведения двух СВ отличается от произведения их МО на величину $M[\dot{X}\dot{Y}]$, равную нулю, если X, Y независимы:

$$M[XY]=M[(\dot{X}+m_x)(\dot{Y}+m_y)]=M[\dot{X}\dot{Y}+m_x\dot{Y}+m_y\dot{X}+m_xm_y]=$$

$$=M[\dot{X}\dot{Y}]+M[X]M[Y].$$

МО произведения двух центрированных СВ является характеристикой их совместного распределения и называется *корреляционным моментом*:

$$K_{xy}=M[\dot{X}\dot{Y}]. \qquad (4.12)$$

Итак, в общем случае МО произведения двух СВ равно произведению их МО, увеличенному на корреляционный момент:

$$M[XY]=M[X]M[Y]+K_{xy}. \qquad (4.13)$$

Из выражения для корреляционного момента дискретных СВ

$$M[\dot{X}\dot{Y}]=\sum_{i=1}^{n}\sum_{j=1}^{m}(x_i-m_x)(y_j-m_y)p_{ij}$$

следует, что величина корреляционного момента тем больше, чем вероятнее одноименные отклонения обеих СВ от своих МО. Если же с увеличением одной СВ вероятнее уменьшение другой (вместе с положительными отклонениями одной СВ чаще реализуются отрицательные отклонения другой), и наоборот, значение корреляционного момента будет отрицательным. Когда любая закономерность в реализациях пары (X, Y) отсутствует, положительные и отрицательные отклонения компенсируются и дают нулевую сумму. Таким образом, положительные, отрицательные или нулевые значения корреляционного момента указывают на наличие прямой или обратной зависимости между случайными реализациями СВ, или отсутствие таковой. Но судить о степени зависимости по абсолютной величине корреляционного момента нельзя, так как она отражает еще и степень разброса отдельных СВ.

Коэффициент корреляции

МО произведения безразмерных центрированных СВ отражает только степень зависимости и называется *коэффициентом корреляции*:

$$r_{xy}=M\left[\frac{\dot{X}}{\sigma_x}\frac{\dot{Y}}{\sigma_y}\right]=\frac{K_{xy}}{\sigma_x\sigma_y}=\frac{K_{xy}}{\sqrt{D_xD_y}}. \qquad (4.14)$$

Можно доказать, что $-1 \le r_{xy} \le 1$, причем, предельные значения коэффициента корреляции соответствуют самой сильной (функциональной) зависимости, прямой или обратной.

Основные свойства дисперсии

Свойства дисперсии – это свойства МО квадрата центрированной СВ:

1. $D[c] = 0$, если c – неслучайная величина;
2. $D[cX] = c^2 D[X]$, так как $M[(c\dot{X})^2] = c^2 M[\dot{X}^2]$;
3. $D[X \pm Y] = D[X] + D[Y] \pm 2K_{xy}$, что следует из $D[X \pm Y] = M[(\dot{X} \pm \dot{Y})^2]$.

Следствие основных свойств МО и дисперсии

МО линейной функции n произвольных (не обязательно независимых) СВ $X_1, \ldots, X_n$ равно той же функции от МО этих СВ:

$$M[c_0 + \sum_{i=1}^{n} c_i X_i] = c_0 + \sum_{i=1}^{n} c_i M[X_i]\,. \tag{4.15}$$

Если $X_1, \ldots, X_n$ взаимно *независимы,* дисперсия их суммы равна сумме дисперсий слагаемых:

$$D\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} D[X_i], \tag{4.16}$$

а дисперсия линейной функции выражается следующим образом:

$$D[c_0 + \sum_{i=1}^{n} c_i X_i] = \sum_{i=1}^{n} c_i^2 D[X_i]. \tag{4.17}$$

Связь между начальными и центральными моментами

Вычисление центральных моментов можно существенно упростить, выразив их через начальные моменты:

$$\mu_k[X] = M[(X - m_x)^k] = M\left[\sum_{i=0}^{k} C_k^i X^i (-m_x)^{k-i}\right] = \sum_{i=0}^{k} (-1)^{k-i} C_k^i m_x^{k-i} \alpha_i[X].$$

В частности, для центральных моментов до 4-о порядка имеем (обозначив $m_x \equiv m$, $\alpha_k[\mathrm{X}] \equiv \alpha_k$):

$$\mu_2[X] = M[(X\text{-}m_x)^2] = \alpha_2 - m^2, \tag{4.18}$$

$$\mu_3[X] = \alpha_3 - 3m\mu_2 - m^3, \tag{4.19}$$

$$\mu_4[X] = \alpha_4 - 4m\mu_3 - 6m^2\mu_2 - m^4. \tag{4.20}$$

Примеры вычисления ЧХ по общим формулам

Вычислим по формулам (4.1), (4.8), (4.18) МО и дисперсию числа попаданий в 10-и независимых выстрелах с вероятностями попадания, меняющимися линейно от $p_1 = 0{,}6$ до $p_{10} = 0{,}3$, используя для вычисления вероятностей $p_{m,10}$ электронную формулу RptTrial:

```
>> n=10;p=linspace(0.6,0.3,n); m=dot(RptTrial(p),0:n),D=dot(RptTrial(p),[0:n].^2)-m^2
m =  4.5000    D =  2.3833
```

Теперь вычислим МО и дисперсию проекции вращающегося стержня на экран, распределенной по синусоидальному закону, построенному в Лекции 3:

```
>> L=10;x=linspace(0,0.999,100)*L;f=2/pi./sqrt(L^2-x.^2);Trap(f,x)
ans =   1.0001
>> m=Trap(x.*f,x), mt = 2*L/pi,  a2=Trap(x.^2.*f,x), D=a2 - m^2, sigma=sqrt(D)
m =   6.3680    mt =  6.3662    a2 =  50.0242    D =  9.4729    sigma =  3.0778
```

Результат численного интегрирования m = 6.37 практически совпадает с точным значением M[X] = 2*l*/π при том, что бесконечная плотность на правом конце интервала не корректироваалась как в Лекции 3 и основное свойство плотности выполнилось приближенно. Дис-

персия D вычислена по формуле (4.18), второй начальный момент, полученный численным интегрированием a2 = 50.02, практически совпадает с результатом интегрирования $l^2/2 = 50$.

Производящая функция для вычисления начальных моментов

Вычисление моментных характеристик с помощью электронных формул не составляет проблемы, если не считать затруднений, связанных с отбрасыванием «хвостов» бесконечных дискретных распределений (геометрического, Пуассона) или с выбором шагов дискретизации непрерывных СВ вблизи особых точек. Но если закон распределения задан параметрической функцией, то ЧХ должны определяться этими параметрами (а распределений Пуассона и геометрического – единственным параметром). Для получения ЧХ непрерывных СВ нужно применить правила интегрирования. Начальные моменты любой дискретной СВ X: $P(X = k) = p_k$, $k = 0, 1, \ldots$ удобно определять с помощью *производящей функции*

$$\varphi(z) = \sum_{k=0}^{\infty} p_k z^k \text{ , где } 0 < z \le 1, \tag{4.21}$$

благодаря ее свойствам:

$$\varphi'(z) = \sum_k p_k k z^{k-1}, \quad \varphi'(z)\big|_{z=1} = \sum_k k p_k = m_x,$$
$$\varphi''(z)\big|_{z=1} = \sum_k k^2 p_k - \sum_k k p_k = \alpha_2 - m_x,$$
$$\varphi'''(z)\big|_{z=1} = \alpha_3 - 3\alpha_2 + 2m_x,$$
$$\varphi''''(z)\big|_{z=1} = \alpha_4 - 6\alpha_3 + 11\alpha_2 - 6m_x.$$

Из этих выражений можно получить все начальные моменты, если известны производные при $z = 1$:

$$\left.\begin{aligned} &\alpha_1 = \varphi'(1), \ \Rightarrow \ (m_x = \varphi'(1)),\\ &\alpha_2 = \varphi''(1) + m_x,\\ &\alpha_3 = \varphi'''(1) + 3\alpha_2 - 2m_x,\\ &\alpha_4 = \varphi''''(1) + 6\alpha_3 - 11\alpha_2 + 6m_x. \end{aligned}\right\} \tag{4.22}$$

Числовые характеристики некоторых дискретных распределений

Индикатор случайного события

Характеристическая СВ для случайного события замечательна тем, что все ее начальные моменты равны вероятности этого события $p = P(A)$:

$$\alpha_k[X] = M[X^k] = 0^k\, P(X{=}0) + 1^k\, P(X{=}1) = P(A) = p,$$

следовательно,

$$M[X] = p,$$
$$D[X] = \alpha_2[X] - m_x{}^2 = p - p^2 = p(1 - p).$$

Биномиальное распределение

Число успехов X в n испытаниях Бернулли можно подсчитать как сумму индикаторов X_i событий A_i, обозначающих успех в i-м испытании: $X = \sum_{i=1}^{n} X_i$. Так как в условиях испытания Бернулли индикаторы X_i независимы, можно применить формулы (4.15), (4.16):

$$M[X] = M\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} M[X_i] = \sum_{i=1}^{n} p_i = np, \tag{4.23}$$

$$D[X] = D\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} D[X_i] = \sum_{i=1}^{n} p_i q_i = npq, \tag{4.24}$$

$$\sigma_x = \sqrt{npq}\,. \tag{4.24}$$

Из полученных выражений видно, что дисперсия биномиального распределения тем больше, чем ближе к 0,5 вероятность успеха в одном испытании, а $n/2$ – наибольшее значение СКО. С помощью производящей функции

$$\varphi(z) = \sum_{k=0}^{n} C_k^n p^k (1-p)^{n-k} z^k = (pz+q)^n$$

можно получить те же результаты для МО и дисперсии, а также старшие моменты. Определим асимметрию биномиального распределения:

$$\mu_3[X] = npq(q-p), \quad As = \frac{\mu_3}{\sigma_3} = \frac{npq(q-p)}{\left(\sqrt{npq}\right)^3}.$$

Мода дискретной СВ вида $P(X=k) = p_k$ – то наименьшее значение k, для которого выполняется неравенство $p_{k+1} < p_k$. В биномиальном распределении

$$p_{k+1} = C_n^{k+1} p^{k+1} q^{n-k-1} = \frac{n!}{(k+1)!(n-k-1)!} p^{k+1} q^{n-k-1} = \frac{n-k}{k+1}\frac{p}{q} C_n^k p^k q^{n-k},$$

поэтому условие

$$\frac{p_{k+1}}{p_k} = \frac{n-k}{k+1}\frac{p}{q} < 1$$

выполняется при $k \geq np - q$. Мода биномиального распределения – это округленная в большую сторону величина $(np - q)$. Она отличается от МО не более, чем на единицу, то есть среднее значение биномиального распределения совпадает или близко к наивероятнейшему.

Интересно, что среднее число попаданий в 10-и выстрелах с вероятностями попадания, меняющимися от $p_1 = 0{,}6$ до $p_{10} = 0{,}3$, полученное ранее с помощью функции RptTrial, совпадает с $np_{ср} = 10 \cdot (0{,}6 + 0{,}3) / 2 = 4{,}5$, где $p_{ср}$ – среднее арифметическое вероятностей попаданий во всех выстрелах. Дисперсия числа попаданий в стрельбе с одинаковыми вероятностями $p_{ср}$, как и следует ожидать, меньше истинной $4{,}5 \cdot (1 - 0{,}45) = 2{,}925 < 3{,}0778$.

Распределение Пуассона

Производящая функция распределения Пуассона

$$\varphi(z) = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} z^k = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda z)^k}{k!} = e^{\lambda(z-1)}$$

такова, что ее производные при z = 1 равны соответствующим степеням параметра λ:

$$\left.\frac{d^k \varphi}{dz^k}\right|_{z=1} = \lambda^k,\ k = 1, 2, 3, 4, \ldots$$

По формулам (4.22) получим все начальные моменты, а с учетом соотношений (4.18) – (4.20) – центральные моменты

$$\alpha_1 = \mu_2 = \mu_3 = \lambda, \quad \mu_4 = 3\lambda^2 + \lambda$$

и все моментные характеристики:

$$m_x = D_x = \lambda, \tag{4.26}$$

$$\sigma_x = \sqrt{\lambda}, \tag{4.27}$$

$$As = \frac{\mu_3}{\sigma^3} = \frac{\lambda}{\lambda^{3/2}} = \frac{1}{\sqrt{\lambda}}, \tag{4.28}$$

$$Ex = \frac{\mu_4}{\sigma^4} - 3 = \frac{1}{\lambda}. \tag{4.29}$$

Дисперсия распределения Пуассона растет вместе с параметром λ, т.е. отклонения от среднего тем больше, чем больше λ, но *относительные* отклонения растут с уменьшением λ:

$$\frac{m_x - k\sigma_x}{m_x} = \frac{\lambda - k\sqrt{\lambda}}{\lambda} = 1 - \frac{k}{\sqrt{\lambda}}. \tag{4.30}$$

Из этого следует, что замена СВ, распределенной по закону Пуассона, ее средним значением особенно опасна при малых параметрах (см. пример ниже). Мода распределения Пуассона – наименьшее k, при котором выполняется условие:

$$\frac{p_{k+1}}{p_k} = \frac{k!}{(k+1)!} \cdot \frac{\lambda^{k+1}}{\lambda^k} = \frac{\lambda}{k+1} < 1.$$

Этому условию удовлетворяет целая часть параметра λ:

$$Mo = k^* = \text{int}(\lambda),$$

причем, если λ целое, $k^* = \lambda$, имеется два модальных значения λ и λ –1

$$\frac{p_{k^*}}{p_{k^*-1}} = \frac{\lambda}{k^*} = 1.$$

Заметим, что формулы для ЧХ распределения Пуассона вытекают из соответствующих формул для биномиального распределения при $\lambda = np$ и $q \approx 1$.

Иллюстрация особенностей ЧХ закона Пуассона

При полной заправке топливом самолет имеет ресурс полета $T = 6$ ч, но каждая взятая на борт бомба сокращает его на 1 час. Самолет обнаруживает цели случайным образом, в среднем S целей за 1 час. Сколько нужно брать бомб, чтобы число атакованных целей за вылет было наибольшим?

Бомб нужно столько же, сколько обнаружено целей. Но число обнаруженных целей за время τ случайно, подчиняется закону Пуассона с параметром $\lambda = \tau S$, зависящем по условию от количества бомб m ($\tau = 6 - m$). Оперируя средним значением числа обнаруженных целей λ как самой СВ, получим условие для оптимального количества бомб $(6 - m)S = m$, откуда следует приближенное решение $m^* = 6S/(1+S)$. Вычислим его для нескольких значений S:

```
>> S=[2,1,0.5,0.2];M1=6*S./(1+S)
M1 =    4.0000    3.0000    2.0000    1.0000
```

Решение по равенству «в среднем» занижает результат, так как не учитывает возможность обнаружения большего числа целей, которые не были бы атакованы из-за отсутствия бомб. Согласно условию (4.30) относительная погрешность возрастает с уменьшением параметра. Чтобы оценить величину погрешности, сравним полученный приближенный результат с точным. Построим распределение числа атакованных целей для каждого из возможных значений m от 1 до 5, найдем МО этих распределений и выберем то значение N, при котором среднее число атакованных целей наибольшее. Если N – число обнаруженных целей, число атакованных целей $L = \min\{N, m\}$. СВ N распределена по закону Пуассона с параметром $\lambda = (6 - m)S$, СВ L принимает свои возможные значения 0, 1, …, m с вероятностями:

$$P(L=k)=\begin{cases} P(N=k)=p_k,\ k<m, \\ P(N\ge m)=1-P(N<m)=1-\sum_{i=0}^{m-1}p_i,\ k=m. \end{cases}$$

По общей формуле (4.1) запишем параметрическое выражение для $M[L]$:

$$M_L(m)=\sum_{k=0}^{m-1}k\cdot p_k+m\left(1-\sum_{k=0}^{m-1}p_k\right)=m-\sum_{k=0}^{m-1}(m-k)p_k.$$

В следующей команде внешний цикл перебирает четыре значения S, внутренний – m:

```
>> for s=S for k=1:5 M(k)=k-dot(k-[0:(k-1)],p_Poisson(s*(6-k),0:k-1));end,M,end

M =   1.0000   1.9966   2.9182   3.2185   1.9775
M =   0.9933   1.8901   2.3279   1.9249   0.9993
M =   0.9179   1.4587   1.4102   0.9957   0.5000
M =   0.6321   0.7419   0.5962   0.3999   0.2000
```

В первой строке результата (при $S=2$) наибольшее среднее число атакованных целей получилось при $m=4$, как и в приближенном решении. В следующих двух строках ($S=1$ и 0,5) оптимальное число бомб 3 и 2 также совпадает с приближенной оценкой. При $S=0{,}2$ приближенная оценка на 1 меньше точного решения.

Геометрическое распределение

Производящая функция распределения $P(X=k)=p_k=q^kp,\ k=0,1,\ldots$ и ее производные при $z=1$:

$$\varphi(z)=\sum_{k=0}^{\infty}pq^kz^k=\frac{p}{1-qz},$$

$$\varphi'(1)=\frac{q}{p},\ \varphi''(1)=\frac{2q^2}{p^2},\ \varphi'''(1)=\frac{6q^3}{p^3},\ \varphi''''(1)=\frac{24q^4}{p^4}.$$

Используя соотношения (4.17) – (4.19), (4.21), получим моментные ЧХ:

$$m_x=\frac{q}{p},\ D_x=\frac{q}{p^2},\ \sigma_x=\frac{\sqrt{q}}{p},\ \mathrm{As}=\frac{\mu_3}{\sigma_x^3}=\frac{1+q}{\sqrt{q}}.$$

Сдвинутое геометрическое распределение

Распределение $P(Y=k)=q^{k-1}p,\ k=1,2,\ldots$ отличается от геометрического тем, что его возможные значения увеличены на 1 по сравнению с геометрическим распределением X: $Y=X+1$. Согласно свойствам МО и дисперсии

$$M[Y]=M[X+1]=M[X]+1=\frac{q}{p}+1=\frac{1}{p}, \qquad (4.31)$$

$$D[Y]=D[X+1]=D[X]=\frac{q}{p^2}. \qquad (4.32)$$

Гипергеометрическое распределение

В предыдущей лекции было показано, что при определенных условиях гипергеометрическое распределение почти не отличается от биномиального. Следовательно, ЧХ этого распределения при тех же условиях можно вычислять по формулам, выведенным для биномиального распределения.

Проверим правильность этого утверждения в условиях выборки из большой и малой партии, сравнив результаты вычисления МО и дисперсии непосредственно по ряду распределения, полученному электронной формулой Sampling, или с помощью универсальной электронной формулы MDS для вычисления ЧХ (Листинг 4.1)

```
>>N=1000;R=100;M=20;k=0:M;P=Sampling(N,R,M,k);m=sum(P.*k), D=sum(P.*(k-m).^2)
 m =    2.0000  D =    1.7657
>> N=50;R=5; [M,D]=MDS('hipergeo',50,5,20)
 m =    2.0000  D =    1.1020
```

В первом случае $n=20$ раз повторяется выбор из большого числа ($N=1000$) изделий с практически одинаковой вероятностью $p=100/1000=0{,}1$ выбрать бракованное. В условиях испытаний Бернулли с такими параметрами среднее число бракованных изделий в контроль-

ной партии составляло бы $np = 20 \cdot 0{,}1 = 2$, дисперсия – $npq = 20 \cdot 0{,}1 \cdot 0{,}9 = 1{,}8$, что практически совпадает с результатами вычисленными значениями тех же ЧХ по гипергеометрическому распределению. В случае большого отличия между многоугольниками истинного распределения и биномиального приближения (при малом объеме партии $N = 50$) МО не изменилось, но истинная дисперсия оказалась существенно меньше, что согласуется с характером различия многоугольников распределения на рис. 3.4: симметрично уменьшились вероятности отклонений от среднего значения.

Статистическое оценивание числовых характеристик

Числовые характеристики определяются параметрами закона распределения, но бывает так, что закон распределения известен из теоретических соображений, а его параметры – нет. Например, то, что число попаданий фрагментов в площадку, ориентированную перепендикулярно направлению разлета, подчиняется закону Пуассона, вытекает из свойств пуассоновского поля, но среднее число попаданий (параметр закона) подлежит экспериментальному определению. Из наблюдаемых в N экспериментах чисел попаданий $X_1, \ldots, X_N$ нужно получить оценку среднего числа попаданий и дисперсии.

Требования к статистическим оценкам

Статистическая оценка неизвестного параметра теоретического распределения как функция от наблюдаемых реализаций СВ сама является случайной величиной и также характеризуется МО и дисперсией. Вид функции для статистических оценок каждого параметра распределения выбирают так, чтобы МО оценки совпадало с оцениваемым параметром при любом объеме выборки (*несмещенность*), а ее дисперсия была минимально возможной при заданном объеме выборки (*эффективность*). *Состоятельной* называют оценку, которая при $N \to \infty$ стремится по вероятности к нулю. Свойства ЧХ функций СВ, каковыми являются статистические оценки, будут подробно изучены в Лекции 9.

Статистическая оценка МО

Согласно (4.0) среднее арифметическое результатов наблюдений по вероятностному смыслу полностью соответствует МО наблюдаемой СВ. В Лекции 9 будет показано, что среднее арифметическое является несмещенной, эффективной и состоятельной статистической оценкой МО:

$$\overline{m}_x = \frac{1}{N}\sum_{i=1}^{N} X_i \quad \text{и} \quad M[\overline{m}_x] = m_x. \tag{4.33}$$

Разности $X_i - \overline{m}_x$ между наблюдаемыми значениями и средним называют *отклонением*. Из (4.33) следует, что сумма всех отклонений равна нулю. Сумма квадратов отклонений всегда положительна и характеризует рассеяние наблюдаемой СВ.

Оценки дисперсии

Из того, что дисперсия определена как МО квадратов центрированной СВ, а оценкой МО является среднее арифметическое, можно заключить, что в качестве оценки дисперсии $\overline{D}_x$ следует принять среднее арифметическое квадратов отклонений:

$$D_x^* = \frac{1}{N}\sum_{i=1}^{N} (X_i - \overline{m}_x)^2 , \tag{4.34}$$

однако, как будет показано в Лекции 9, эта оценка смещенная, хотя смещение стремится к нулю при $N \to \infty$ (такая оценка называется *асимптотически несмещенной*). Несмещенной, эффективной и состоятельной является так называемая *исправленная дисперсия*, отличающаяся от D_x^* при небольших объемах выборки:

$$\widetilde{D}_x = D_x^* \frac{N}{N-1} = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \overline{m}_x)^2 \ . \qquad (4.35)$$

Для удобства вычислений эту формулу удобно преобразовать к виду:

$$\widetilde{D}_x = \frac{1}{N-1}\left[\sum_{i=1}^{N} X_i^2 - \overline{m}_x^2 N\right]. \qquad (4.36)$$

Оценки корреляции

Оценку корреляционного момента и коэффициента корреляции вычисляют по формулам:

$$\widetilde{K}_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \overline{m}_x)(Y_i - \overline{m}_y), \qquad (4.37)$$

$$\widetilde{r} = \frac{\widetilde{K}_{xy}}{\sqrt{\widetilde{D}_x \widetilde{D}_y}} \qquad (4.38)$$

Вычисления оценок параметров распределений в MATLAB

Точечные оценки результатов наблюдений вычисляет файл_функция Ocenki_m_D (Листинг 4.2). В библиотеке MATLAB среднее значение массива вычисляет функция mean, стандартное отклонение – функция std. Вычисления по формулам (4.37), (4.38) удобнее выполнять файл-функцией CorrelCoef (Листинг 4.3), чем аналогичной функцией corrcoef из библиотеки MATLAB, которая возвращает матрицу коэффициентов корреляции. Сформируем два массива случайных значений X и Y, и вычислим оценки СКО и корреляции:

```
>> N=10000;X=rand(1,N);Y=randn(1,N);sX=std(X),sY=std(Y),[r,K]=CorrelCoef([X;Y])
sX =  0.2881    sY =  1.0015   r =  0.0005   K =  0.0014
```

Так как объем статистики очень большой, полученные оценки близки к истинным σ[X] = 1/√12 = 0.2887, σ[Y] = 1, r = K = 0 (X, Y независимы). Если образовать функционально зависимый вектор Z(X), то оценка коэффициента корреляции этих векторов должна иметь максимальное значение 1, а корреляция между Z и Y должна быть нулевой:

```
>> Z=X*2; rXZ=CorrelCoef([X;Z]),rYZ=CorrelCoef([Y;Z])
rXZ =  1.0000    rYZ =  0.0050
```

Точность оценок

При очень большом объеме статистического материала оценки параметров распределения очень близки к истинным значениям. Если же команды из предыдущего примера повторить при N<100, что более приемлемо для реальных экспериментов, расхождение станет существенным. Естественно, возникает вопрос о доверии к оценкам и необходимом объеме статистики. При достаточном объеме и не высоких требованиях к достоверности полученные оценки принимают за приближенные значения соответствующих параметров распределения (*точечное* оценивание). Более полную информацию о качестве оценивания дают *доверительные границы* для оцениваемого параметра, вычисленные при заданной доверительной вероятности. Если, например, $m_н$, $m_в$ и $[m_{1н}, m_{1в}]$ соответственно нижняя и верхняя доверительные границы МО, это значит, что интервалы $[-\infty, m_в]$, $[m_н, \infty]$ или $[m_н, m_в]$ с заданной вероятностью накроют неизвестное значение МО.

Доверительный интервал для оценок

Обычно используют два разных подхода для определения доверительных границ. Первый подход, основанный на точном распределения, будет обоснован в Лекции 9. Второй подход основан на том, что для несмещенных или асимптотически несмещенной оценки $\widetilde{\theta}$ параметра θ должно выполняться неравенство

$$P\left(|\ \widetilde{\theta} - \theta\ | < \varepsilon\right) > \beta$$

при заданной доверительной вероятности β и некотором значении ε. Минимальное значение $\varepsilon = \varepsilon_\beta$, при котором выполняется это неравенство и опре-

деляет границы доверительного интервала, содержащего точечную оценку: $[\widetilde{\theta} - \varepsilon_\beta, \widetilde{\theta} + \varepsilon_\beta]$. Поскольку ни закон распределения СВ $\widetilde{\theta}$, ни его параметры (в том числе и оцениваемый параметр θ) неизвестны, приходится вносить упрощения. Если $\widetilde{\theta}$ – оценка МО, то на основании центральной предельной теоремы закон распределения этой СВ как суммы достаточно большого числа N независимых слагаемых (результатов наблюдений) близок к нормальному с параметрами m и D/N. Заменив неизвестные параметры точечными оценками $\overline{m}$ и $\widetilde{D}$, получим выражение для искомого ε_β:

$$\varepsilon_\beta = \sqrt{\frac{\widetilde{D}}{N}}\Phi^{-1}\left(\frac{\beta}{2}\right).$$

Используя ArgLaplas для вычисления обратной функции Лапласа, составим программу IntForM (Листинг 4.4) для вычисления границ доверительного интервала относительно оценки МО $\overline{m}$: $[\overline{m} - \varepsilon_\beta,\ \overline{m} + \varepsilon_\beta]$.

Влияние объема выборки и доверительной вероятности на качество оценок параметров

Генерируем N =1000 реализаций СВ с параметрами m = 5 и σ = 3, а затем вычислим по этим данным точечные оценки этих же параметров и доверительные границы для среднего:

```
>> N=1000;X=randn(1,N)*3+5;[m,D,s]=Ocenki_m_D(X);m,s,[m1,m2]=IntForM(0.9,X)
m =  5.0683    s = 2.9249    m1 =  4.9162   m2 =  5.2205
```

Повторим ту же команду при N =50:

```
>> N=50;X=randn(1,N)*3+5;[m,D,s]=Ocenki_m_D(X);m,s,[m1,m2]=IntForM(0.9,X)
m =  4.4154   s =  2.2215    m1 = 3.8986     m2 = 4.9321
```

Если при N =1000 оценки близки к истинным значениям параметров, то при небольшом объеме статистики N =50 разница значительна, причем истинное значение МО даже не попадает в доверительный интервал. Исследуем влияние объема выборки и доверительной вероятности на качество интервальных оценок. Для этого получим 100 доверительных интервалов по N =50 реализациям той же СВ и и выведем их на график так, чтобы доверительные интервалы, не содержащие истинного МО, выделялись красным цветом. Составим выражение setcolor, определяющее цвет линии (синий или красный), а также выражение для цикла розыгрышей N случайных чисел со средним значением 5 и построения доверительных интервалов для оценки среднего в каждом розыгрыше:

```
>> plt='c="b";if m1>5|m2<5 c="r";end,plot([m1,m2],[i,i],c),hold on'
>> loop = 'for i=1:100 X=randn(1,N)*3+5;[m1,m2]=IntForM(P,X);L(i)=m2-m1;eval(plt),end'
```

В цикле loop запоминаются длины доверительных интервалов для каждого из 100 розыгрышей. Выполнив цикл при небольшом числе испытаний N =50 и доверительной вероятности 0.9, получим 100 разных доверительных интервалов (они зависят от исходной статистики), причем, примерно десятая часть из них не содержит истинного (рис. 4.1, *а*):

```
>> N=50;P=0.9; eval(loop), m=mean(L)
m =   1.3723
```

Увеличение объема выборки до N =1000 сокращает доверительные интервалы, но не повышает надежности интервальной оценки (рис. 4.1, *б*):

```
>> N=1000;P=0.9; eval(loop), m=mean(L)
m =   0.3124
```

Увеличение доверительной вероятности до 0,99 при том же числе испытаний N =1000 несколько расширяет доверительный интервал, но сильно повышает надежность оценки, хотя и не исключает «красных» интервалов (рис. 4.1, *а*):

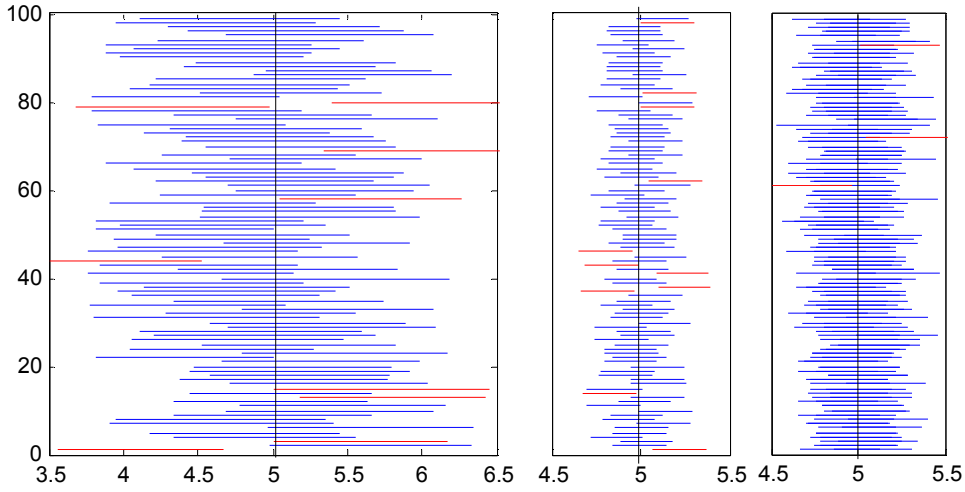
```
>> N=1000;P=0.99; eval(loop), m=mean(L)
m =   0.4822
```

Генерируем N релизаций СВ, распределенной по закону Пуассона с параметром 3 и вычислим оценки параметра при N=100 и N=10:

```
>> X=Gen( 'Poi', 3, 100);[m,D]=Ocenki_m_D(X),[m1,m2]=IntForM(0.8,X)
m =    2.9800    D =  3.2521   m1 =  2.7489   m2 =  3.2111

>> X=Gen( 'Poi', 3, 10);[m,D,s]=Ocenki_m_D(X), [m1,m2]=IntForM(0.8,X)
m =  3.7000   D =  7.3444   m1 =  2.6017   m2 =  4.7983
```

Если при N=100 оценка близка к параметру, то 10-и испытаний слишком мало для полцчения удовлетворительной точности. То, что оценка дисперсии в 2 раза больше оценки МО, не характерно для закона Пуассона.



а *б* *в*

Рис. 4.1. Доверительные интервалы МО для СВ с m = 5 и σ = 3 при:

а) N=50;P=0,9; *б*) N=1000; P=0,9; *в*) N=1000; P=0,99.

Программа верификации кода MATLAB

```
clear all
n=10;p=linspace(0.6,0.3,n);
m=dot(RptTrial(p),0:n),D=dot(RptTrial(p),[0:n].^2)-m^2
L=10;x=linspace(0,0.999,100)*L;f=2/pi./sqrt(L^2-x.^2);Trap(f,x)
m=Trap(x.*f,x), mt = 2*L/pi,  a2=Trap(x.^2.*f,x), D=a2 - m^2, sigma=sqrt(D)

clear all
S=[2,1,0.5,0.2];M1=6*S./(1+S)
for s=S for k=1:5 M(k)=k-dot(k-[0:(k-1)],p_Poisson(s*(6-k),0:k-1));end,M,end
N=100;R=100;M=20;k=0:M;P=Sampling(N,R,M,k);m=sum(P.*k), D=sum(P.*(k-m).^2)
N=50;R=5; [M,D]=MDS('hipergeo',50,5,20)

clear all
N=1000;X=rand(1,N);Y=randn(1,N);sX=std(X),sY=std(Y),[r,K]=CorrelCoef([X;Y])
Z=X*2; rXZ=CorrelCoef([X;Z]),rYZ=CorrelCoef([Y;Z])
N=1000;X=randn(1,N)*3+5;[m,D,s]=Ocenki_m_D(X);m,s,[m1,m2]=IntForM(0.9,X)
N=50;X=randn(1,N)*3+5;[m,D,s]=Ocenki_m_D(X);m,s,[m1,m2]=IntForM(0.9,X)
plt='c=''b'';if m1>5|m2<5 c=''r'';end,plot([m1,m2],[i,i],c),hold on'
loop = 'for i=1:100 X=randn(1,N)*3+5;[m1,m2]=IntForM(P,X);L(i)=m2-
m1;eval(plt),end'
N=50;P=0.9; eval(loop), m=mean(L)
N=1000;P=0.9; eval(loop), m=mean(L)
N=1000;P=0.99; eval(loop), m=mean(L)
X=Gen( 'Poi', 3, 100);[m,D]=Ocenki_m_D(X),[m1,m2]=IntForM(0.8,X)
X=Gen( 'Poi', 3, 10);[m,D,s]=Ocenki_m_D(X), [m1,m2]=IntForM(0.8,X)
```

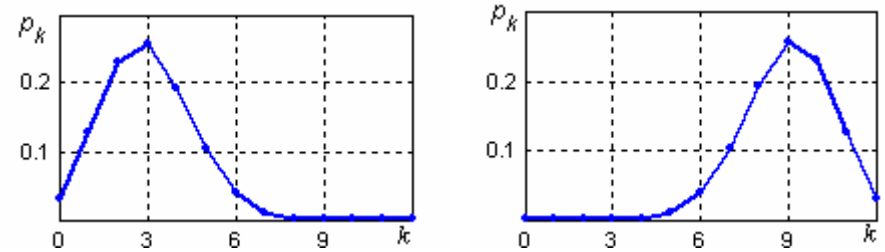
Контрольные вопросы и задачи

1. Объясните вероятностный смысл МО и СКО.
2. Назовите числовые характеристики положения. Как они характеризуют дискретное и непрерывное распределение?
3. Что характеризует СВ срединное отклонение? Чем оно отличается от аналогичной моментной характеристики?
4. Почему дисперсия оценки МО при выводе формулы (4.37) принята равной D/N, где D – дисперсия оцениваемой СВ, N – число наблюдений?
5. Объясните вероятностный смысл корреляционного момента и коэффициента корреляции.
6. Как связаны второй центральный и второй начальный моменты?
7. Как выражаются МО и СКО через параметры биномиального распределения?.
8. Каковы особенности МО, СКО и моды закона Пуассона?
9. Сравните МО и дисперсию геометрического и «геометрического + 1» распределений. Как вычислить МО и дисперсию расхода снарядов в стрельбе до первого попадания?
10. Как вычислить числовые характеристики гипергеометрического распределения?
11. Вычислить числовые характеристики СВ X, распределенной по биномиальному закону с параметрами $n = 12$, $p = 0{,}25$. Правильно ли, что МО этой СВ равно 3, а дисперсия составляет ¾ от этой величины?
12. Положительна или отрицательна асимметрия распределения в предыдущем вопросе?

Чтобы выяснить это, вычислим моментные ЧХ для СВ X, распределенной по биномиальному закону с параметрами $n = 12$, $p = 0{,}25$, используя формулы (4.1) – (4.10) и электронную формулу p_Binom:

```
>> n=12;p=0.25; X=p_Binom(p,n); k=0:n; plot(k,X), M=sum(X.*k), D=sum(X.*(k-M).^2)
M =  3   D =  2.2500
>> sigma = sqrt(D), As=sum(X.*(k-M).^3)/sigma^3, Ex=sum(Y.*(k-M).^4)/sigma^4-3
sigma =  1.5000   As =  0.3333   Ex =  -0.0556
```

Сопоставив результаты с многоугольником распределения (рис. 4.2, *а*) обнаружим, что МО $m_x = 3$ совпало в данном случае с модой $Mo = 3$, дисперсия $D_x = 2{,}25$ такова, что СКО $\sigma_x = 1{,}5$ больше, чем разница между ближайшими возможными значениями, асимметрия $As = 1/3$ распределения, более пологого справа от МО, положительна, эксцесс Ex = -0,056 близок к нулю. Многоугольник распределения с параметрами $n = 12$, $p = 0{,}75$ на рис. 4.2, *б* представляет собой зеркальное отражение предыдущего графика, МО и мода, соответственно смещены вправо, асимметрия отличается знаком:



а *б*

Рис. 4.2. Многоугольники биномиального распределения с параметрами:
а) $n = 12$, $p = 0{,}25$; *б*) $n = 12$, $p = 0{,}75$.