Характеристики распределений случайных величин, введение в статистику

Лекция №2

Статистические методы в ядерном эксперименте, ИЯФ, 2024 г.

1. Математическое ожидание и дисперсия случайной величины
2. Асимметрия, высшие моменты, производящая функция
3. Примеры
4. Введение в статистику
   - Понятие выборки случайной величины
   - Эмпирическая функция распределения
   - Гистограмма
   - Выборочные характеристики и их свойства
   - Примеры

Характеристики распределений I

Для описания наиболее важных черт распределений используют такие характеристики, как среднее (математическое ожидание), дисперсия, асимметрия и др.

Математическое ожидание случайной величины x есть среднее значение x с учётом вероятности (или плотности вероятности) реализации каждого значения x.

$E(x) = \sum_r r \cdot P(x = r)$ - для дискретного распределения

$E(x) = \int_{-\infty}^{+\infty} x \cdot f(x)dx$ - для непрерывного распределения

Математическое ожидание функции $h(x)$ случайной величины x есть средняя величинв $h(x)$ по всем возможным значениям x.

$E(x) = \sum_r h(r) \cdot P(x = r)$ - для дискретного распределения

$E(x) = \int_{-\infty}^{+\infty} h(x) \cdot f(x)dx$ - для непрерывного распределения

Свойства математического ожидания

Пусть x и y - случайные величины:

- $E(\lambda x) = \lambda E(x)$, λ-числовая константа
- $E(\lambda_1 x + \lambda_2 y) = \lambda_1 E(x) + \lambda_2 E(y)$, $\lambda_{1,2}$-числовые константы
- $E(xy) = E(x)E(y)$, для независимых x и y

Замечание 1: Однако из равенства $E(xy) = E(x)E(y)$ не следует независимости x и y.

Замечание 2: Если $\mu = E(x)$, то $E(h(x)) \neq h(\mu)$.

Пример: Рассмотрим среднее от $h(x) = \frac{1}{1+x}$ для дискретной случайной величины, распределённой по закону Пуассона:

$$E(h(r)) = \sum_{r=0}^{+\infty} \frac{1}{1+r} \frac{\mu^r}{r!} e^{-\mu} = \frac{1}{\mu} \sum_{r=0}^{+\infty} \frac{\mu^{r+1}}{(r+1)!} e^{-\mu} = \frac{1}{\mu} \left(\sum_{k=0}^{+\infty} \frac{\mu^k}{k!} e^{-\mu} - e^{-\mu} \right) =$$

$$= \frac{1}{\mu}(1 - e^{-\mu}) \neq h(\mu) = \frac{1}{1+\mu}$$

Характеристики распределений III

Дисперсией случайной величины x называют среднее квадрата отклонения случайной величины от её математического ожидания $\mu = E(x) = \overline{x}$, её обозначают $D(x)$ или σ_x^2:

$$D(x) = \sigma_x^2 = E[(x-\mu)^2] = \overline{(x-\overline{x})^2}$$

$\sigma_x = \sqrt{D(x)}$ называют среднеквадратичным или стандартным отклонением случайной величины x, она служит мерой разброса x относительно среднего значения $\overline{x}$.

Асимметрия распределения описывается параметром:

$$\gamma_1 = \frac{E[(x-\mu)^3]}{\sigma_x^3} = \frac{\overline{(x-\overline{x})^3}}{\sigma_x^3}$$

- $\gamma_1 < 0$ - $f(x)$ вытянута влево от $\overline{x}$
- $\gamma_1 = 0$ - $f(x)$ симметрична относительно $\overline{x}$
- $\gamma_1 > 0$ - $f(x)$ вытянута вправо от $\overline{x}$

Свойства дисперсии

Пусть x и y - случайные величины:

- $D(x) = \sigma_x^2 \geq 0$, $D(x) = 0$ если x является константой
- $D(x) = \overline{(x - \overline{x})^2} = \overline{x^2} - \overline{x}^2$
- $D(\lambda x) = \lambda^2 D(x)$, λ–числовая константа
- $D(x + y) = D(x) + D(y) + 2\sqrt{D(x)D(y)}\rho(x, y) =$
  $= \sigma_x^2 + \sigma_y^2 + 2\sigma_x\sigma_y\rho(x, y)$, $-1 \leq \rho(x, y) \leq +1$
  $\rho(x, y) = \frac{\overline{(x-\overline{x})(y-\overline{y})}}{\sigma_x\sigma_y} = \frac{\overline{xy}-\overline{x}\,\overline{y}}{\sigma_x\sigma_y}$–коэффициент корреляции x и y
  Следствия:
  - $D(x + \lambda) = D(x)$, где λ–числовая константа
  - $D(x + y) = D(x) + D(y)$ для независимых/некоррелированных x и y ($\rho(x, y) = 0$)

Среднее и дисперсия дискретных распределений I

Биноминальное распределение: $P(r|p,N) = C_N^r p^r q^{N-r}$, $q = 1-p$

$$\mu = E(r) = \sum_{r=0}^{N} r \frac{N!}{(N-r)!r!} p^r q^{N-r} = Np \sum_{r=1}^{N} \frac{(N-1)!}{(N-r)!(r-1)!} p^{r-1} q^{N-r} =$$

$$= (s = r-1,\ n = N-1) = Np \sum_{s=0}^{n} \frac{n!}{(n-s)!s!} p^s q^{n-s} = Np$$

$$E(r^2) = \sum_{r=0}^{N} r^2 P(r) = \sum_{r=2}^{N} r(r-1)P(r) + \sum_{r=1}^{N} rP(r) = N(N-1)p^2 + Np$$

$$\sigma^2 = E(r^2) - E^2(r) = Npq,\ \sigma = \sqrt{Np(1-p)},\ \gamma_1 = \frac{1-2p}{\sqrt{Np(1-p)}}$$

Распределение Пуассона: $P(r|\lambda) = (\lambda^r/r!)e^{-\lambda}$

$$\mu = E(r) = \sum_{r=0}^{\infty} r \frac{\lambda^r}{r!} e^{-\lambda} = \lambda \sum_{r=1}^{\infty} \frac{\lambda^{r-1}}{(r-1)!} e^{-\lambda} = (s = r-1) = \lambda \sum_{s=0}^{\infty} \frac{\lambda^s}{s!} e^{-\lambda} = \lambda$$

$$E(r^2) = \sum_{r=0}^{\infty} [r(r-1) + r] \frac{\lambda^r}{r!} e^{-\lambda} = \lambda^2 \sum_{r=2}^{\infty} \frac{\lambda^{r-2}}{(r-2)!} e^{-\lambda} + \lambda = \lambda^2 + \lambda$$

$$\sigma^2 = E(r^2) - E^2(r) = \lambda^2 + \lambda - \lambda^2 = \lambda,\ \sigma = \sqrt{\lambda},\ \gamma_1 = 1/\sqrt{\lambda}$$

Среднее и дисперсия дискретных распределений II

Пример 1: Эффективность регистрации продуктов реакции $e^+e^- \to \pi^+\pi^-$ в детекторе:

$$\varepsilon_{\text{det}} = \frac{\int\limits_{\text{det}} \varepsilon(\theta,\phi)\frac{d\sigma}{d\Omega}\,d\Omega}{\int\limits_{4\pi} \frac{d\sigma}{d\Omega}\,d\Omega},$$

где $\frac{d\sigma}{d\Omega}$ - дифференциальное сечение реакции $e^+e^- \to \pi^+\pi^-$, $\varepsilon(\theta,\phi)$ - эффективность реконструкции продуктов реакции (π^+ и π^-) если они уже попали в чувствительную область детектора.
Для определения ε_{det} проводится полное моделирование реакции $e^+e^- \to \pi^+\pi^-$. Сначала генерируется N_{gen} событий процесса методом Монте-Карло, потом моделируется прохождение частиц через вещество детектора (с помощью пакета Geant4) и отклик подсистем детектора, в конце накладываются условия отбора и подсчитываются зарегистрированные N_{det} событий. При этом величина N_{det} распределена биномиально, со средним $N_{\text{gen}}\varepsilon_0$, где ε_0 – истинная эффективность регистрации, и $\sigma(N_{\text{det}}) = \sqrt{N_{\text{gen}}\varepsilon_0(1-\varepsilon_0)}$. Тогда величина $\varepsilon_{\text{det}} = N_{\text{det}}/N_{\text{gen}}$, со средним ε_0 и стандартным отклонением $\sigma(\varepsilon_{\text{det}}) = \sqrt{\varepsilon_0(1-\varepsilon_0)/N_{\text{gen}}} \approx \sqrt{\varepsilon_{\text{det}}(1-\varepsilon_{\text{det}})/N_{\text{gen}}}$, называется расчётной эффективностью регистрации:

$$\varepsilon_{\text{det}} = \frac{N_{\text{det}}}{N_{\text{gen}}}, \quad \sigma(\varepsilon_{\text{det}}) = \frac{\sqrt{N_{\text{det}}\left(1-\frac{N_{\text{det}}}{N_{\text{gen}}}\right)}}{N_{\text{gen}}} = \frac{\sqrt{\varepsilon_{\text{det}}(1-\varepsilon_{\text{det}})}}{\sqrt{N_{\text{gen}}}}.$$

При этом погрешность расчётной эффективности регистрации уменьшается с ростом N_{gen} как $1/\sqrt{N_{\text{gen}}}$.

Среднее и дисперсия дискретных распределений III

Пример 2: Пусть в эксперименте на e^+e^- коллайдере был набран интеграл светимости $\mathcal{L}$, причём зарегистрировано N событий изучаемой реакции. Эффективность регистрации продуктов реакции – ε_{det}. Нужно определить полное сечение реакции σ_{tot}.

$$\overline{N} = \mathcal{L}\sigma_{\text{tot}}\varepsilon_{\text{det}},$$

$\mathcal{L} = \int_0^{T_{\text{exp}}} L(t)dt$, $L(t)$–(мгновенная) светимость коллайдера, которая определяется числом частиц в e^+e^- пучках, поперечными размерами пучков и частотой их столкновения. Число N зарегистрированных событий распределено по Пуассону, со средним $\overline{N}$ и стандартным отклонением $\sqrt{\overline{N}} \approx \sqrt{N}$, поэтому:

$$\sigma_{\text{tot}} = \frac{N}{\mathcal{L}\varepsilon_{\text{det}}}, \quad \Delta\sigma_{\text{tot}} = \frac{\sqrt{N}}{\mathcal{L}\varepsilon_{\text{det}}}$$

Отметим, что относительная погрешность сечения $(\Delta\sigma_{\text{tot}}/\sigma_{\text{tot}})$ падает с увеличением статистики (N) как $1/\sqrt{N}$.

Среднее и дисперсия непрерывных распределений I

Равномерное на отрезке $[a, b]$

$$\mu = \overline{x} = \int_a^b \frac{x dx}{b-a} = \frac{a+b}{2}, \quad \overline{x^2} = \int_a^b \frac{x^2 dx}{b-a} = \frac{1}{3}\frac{b^3-a^3}{b-a} = \frac{1}{3}(a^2+ab+b^2)$$

$$\sigma^2 = \overline{x^2} - \overline{x}^2 = \frac{(b-a)^2}{12}, \quad \sigma = \frac{b-a}{2\sqrt{3}}, \quad \gamma_1 = 0$$

Экспоненциальное, $x \in [0, +\infty)$

$$\mu = \int_0^\infty \frac{x}{\lambda} e^{-\frac{x}{\lambda}} dx = \lambda\Gamma(2) = \lambda, \quad \overline{x^2} = \int_0^\infty \frac{x^2}{\lambda} e^{-\frac{x}{\lambda}} dx = \lambda^2\Gamma(3) = 2\lambda^2$$

$$\sigma^2 = \overline{x^2} - \mu^2 = \lambda^2, \quad \sigma = \lambda, \quad \gamma_1 = \frac{\overline{(x-\lambda)^3}}{\lambda^3} = \frac{1}{\lambda^3}(\Gamma(4)\lambda^3 - 3\cdot 2\lambda^3 + 3\lambda^3 - \lambda^3) = 2$$

Нормальное, $x \in (-\infty, +\infty)$

$$\overline{x} = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} (x-\mu+\mu) e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \mu, \quad D = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} (x-\mu)^2 e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx =$$

$$\underset{y=\frac{x-\mu}{\sqrt{2}\sigma}}{=} \frac{4\sigma^2}{\sqrt{\pi}} \int_0^{+\infty} y^2 e^{-y^2} dy \underset{z=y^2}{=} \frac{2\sigma^2}{\sqrt{\pi}} \int_0^{+\infty} z^{1/2} e^{-z} dz = \frac{2\sigma^2}{\sqrt{\pi}}\Gamma\left(\frac{3}{2}\right) = \frac{\sigma^2}{\sqrt{\pi}}\Gamma\left(\frac{1}{2}\right) = \sigma^2, \quad \gamma_1 = 0$$

Распределение Коши, $x \in (-\infty, +\infty)$

Распределение Коши – частный случай важного в ФЭЧ распределения Брейта-Вигнера с параметрами x_0 (среднее) и Γ (ширина), описывающего сечение резонансного рождения частиц в конечном состоянии:

$$f(x|x_0,\Gamma) = \frac{1}{\pi}\cdot\frac{\Gamma/2}{(x-x_0)^2+\Gamma^2/4}$$

$$\mu = \frac{1}{\pi}\int\limits_{-\infty}^{+\infty}\frac{(x-x_0+x_0)\Gamma/2}{(x-x_0)^2+\Gamma^2/4}dx = x_0, \quad \text{V.p.}\frac{1}{\pi}\int\limits_{-\infty}^{+\infty}\frac{(x-x_0)\Gamma/2}{(x-x_0)^2+\Gamma^2/4}dx = 0$$

$$\sigma^2 = \frac{\Gamma/2}{\pi}\int\limits_{-\infty}^{+\infty}\frac{(x-x_0)^2}{(x-x_0)^2+\Gamma^2/4}dx = \frac{\Gamma/2}{\pi}\int\limits_{-\infty}^{+\infty}\left(1-\frac{\Gamma^2/4}{(x-x_0)^2+\Gamma^2/4}\right)dx \to \infty$$

т.е. дисперсия бесконечна.

Тем не менее, параметр Γ характеризует ширину распределения, представляя собой ширину распределения Брейта-Вигнера на полувысоте (FWHM).

Моменты и производящая функция моментов I

Моменты случайной величины

- Начальный момент n-ого порядка случайной величины x по определению равен $\mu'_n = E[x^n]$. Заметим, что $\mu'_0 = 1$, $\mu'_1 = \mu$.
- Центральный момент n-ого порядка случайной величины x по определению равен $\mu_n = E[(x-\mu)^n]$. Заметим, что $\mu_0 = 1$, $\mu_1 = 0$, $\mu_2 = \sigma^2$, $\mu_3 = \gamma_1\sigma^3$.
- Полный набор моментов $\{\mu'_n\}$ или $\{\mu_n\}$ полностью определяет распределение вероятности.

Производящая функция начальных моментов

Введём функцию, зависящую от всех моментов $\{\mu'_n\}$, которая сводит весь набор $\{\mu'_n\}$ к единственному выражению. Для этого введём вспомогательную переменную t. Производящая функция начальных моментов случайной величины x:

$$M'_x(t) = E(e^{xt}) = \begin{cases} \int\limits_{-\infty}^{+\infty} e^{xt} f(x)dx, \\ \sum\limits_{r=0}^{\infty} e^{rt} P_r. \end{cases}$$

Разлагая экспоненту в ряд,
$M'_x(t) = E[1 + xt + \frac{(xt)^2}{2!} + \frac{(xt)^3}{3!} + \ldots] = 1 + \mu'_1 t + \mu'_2 \frac{t^2}{2!} + \mu'_3 \frac{t^3}{3!} + \ldots$ видим, что μ'_n – коэффициент при $\frac{t^n}{n!}$. Таким образом:

$$\mu'_n = \int\limits_{-\infty}^{+\infty} x^n f(x)dx = \left.\frac{\partial^n M'_x(t)}{\partial t^n}\right|_{t=0}$$

Моменты и производящая функция моментов II

Производящая функция центральных моментов

$$M_x(t) = E[e^{(x-\mu)t}],\ \mu = \mu_1'.$$

Разлагая экспоненту в ряд,
$M_x(t) = E[1 + (x-\mu)t + (x-\mu)^2\frac{t^2}{2!} + (x-\mu)^3\frac{t^3}{3!} + ...] =$
$1 + 0 + \mu_2\frac{t^2}{2!} + \mu_3\frac{t^3}{3!} + ...$ видим, что μ_n – коэффициент при $\frac{t^n}{n!}$.
Заметим, что $M_x(t) = e^{-\mu t}M_x'(t)$.

Теорема 1: Если две производящие функции моментов (ПФМ) $M'_{x_1}(t)$ и $M'_{x_2}(t)$ одинаковы, то и исходные распределения вероятностей совпадают.

Теорема 2: ПФМ суммы независимых случайных величин равна произведению их производящих функций моментов.

Замечания:

- По заданному распределению вероятности легко найти $M_x(t)$, а значит и моменты μ_n. Часто это проще, чем вычислять моменты из самой ф.п.в.
- При нахождении распределения вероятности нескольких случайных величин иногда легче найти сперва ПФМ $M_x(t)$, а затем уже само распределение $f(x)$.

Моменты и производящая функция моментов III

Биномиальное распределение $P(r) = C_N^r p^r (1-p)^{N-r}$

$$M'(t) = \sum_{r=0}^{\infty} e^{rt} P(r) = \sum_{r=0}^{\infty} C_N^r (pe^t)^r (1-p)^{N-r} = (pe^t + 1 - p)^N$$

Распределение Пуассона $P(r) = (\mu^r / r!) e^{-\mu}$

$$M'(t) = \sum_{r=0}^{\infty} e^{rt} P(r) = e^{-\mu} \sum_{r=0}^{\infty} \frac{(\mu e^t)^r}{r!} = e^{\mu(e^t - 1)}$$

Равномерное распределение на отрезке $[a, b]$

$$M'(t) = \frac{1}{b-a} \int_a^b e^{xt} dx = \frac{e^{bt} - e^{at}}{bt - at}$$

Экспоненциальное распределение, $x \in [0, +\infty)$

$$M'(t) = \frac{1}{\lambda} \int_0^{\infty} e^{xt} e^{-\frac{x}{\lambda}} dx = \frac{1}{1 - \lambda t}, \ \lambda t < 1$$

Нормальное распределение

$$M'(t) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}} e^{xt} dx = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} e^{-\frac{(x - [\mu + t\sigma^2])^2}{2\sigma^2}} e^{\mu t + \sigma^2 \frac{t^2}{2}} dx = e^{\mu t + \sigma^2 \frac{t^2}{2}}$$

Моменты и производящая функция моментов IV

Пример: Сумма N независимых случайных величин $x = \sum_{i=1}^{N} x_i$, каждая из которых распределена по нормальному закону со средним μ_i и дисперсией σ_i^2 также распределена по нормальному закону со средним $\mu = \sum_{i=1}^{N} \mu_i$ и дисперсией $\sigma^2 = \sum_{i=1}^{N} \sigma_i^2$:

$$M'_x(t) = \prod_{i=1}^{N} M'_{x_i}(t) = \prod_{i=1}^{N} e^{\mu_i t + \sigma_i^2 \frac{t^2}{2}} = e^{t \sum \mu_i + \frac{t^2}{2} \sum \sigma_i^2}$$

Введение в математическую статистику

В теории вероятностей рассматриваются случайные величины с известными распределениями. Математическая статистика предоставляет методы, позволяющие по конечному числу экспериментов делать (с некоторой надёжностью) выводы о распределениях случайных величин, изучаемых в этих экспериментах.

Часто оказывается возможным сделать некоторые априорные (еще до проведения эксперимента) предположения об изучаемом распределении или о его свойствах (параметрах). В этом случае небходимо по экспериментальным данным подтвердить или опровергнуть эти предположения (гипотезы) или оценить численные значения параметров.

Пусть случайная величина имеет распределение F, которое частично или полностью нам неизвестно. Проведя эксперимент N раз в одинаковых условиях, получим набор значений. В серии уже проведенных экспериментов это набор фиксированных чисел, однако до проведения эксперимента мы не знаем, какие значения мы получим, и можно рассматривать их как случайные величины, имеющие одинаковое распределение.

Выборка, эмпирическая функция распределения

Выборкой $\{X_k\}_n = \{X_1, ..., X_n\}$ объёма n из распределения F называется набор из n независимых одинаково распределённых случайных величин, имеющих распределение F.

Элементы выборки, упорядоченные по возрастанию, составляют вариационный ряд, k-ый элемент которого обозначается как $X_{(k)}$ и называется k-ой порядковой статистикой.

Эмпирической функцией распределения, построенной по выборке $\{X_k\}_n$, называется случайная функция $F_n^*(y)$, при каждом $y \in R$ равная:

$F_n^*(y) =$ (число значений X_k таких, что $X_k < y)/n = \frac{1}{n} \sum_{k=1}^{n} I(X_k < y)$,

где величина $I(X_k < y)$ называется индикатором события $X_k < y$:

$$I(X_k < y) = \begin{cases} 1, & X_k < y \\ 0, & X_k \geq y \end{cases}$$

Свойства эмпирической функции распределения

Для любого $y \in R$ выполняется:

1. $E(F_n^*(y)) = F(y)$, т.е. $F_n^*(y)$ есть несмещённая оценка для $F(y)$.
2. Величина $nF_n^*(y)$ имеет биномиальное распределение с параметрами n и $p = F(y)$, поэтому $D(F_n^*(y)) = \frac{F(y)(1-F(y))}{n}$, т.е. отклонение эмпирической функции распределения от истинной уменьшается как $1/\sqrt{n}$ при увеличении n.
3. $\sqrt{n}(F_n^*(y) - F(y)) \underset{n\to\infty}{\to} N(0, \sigma^2)$, где $\sigma^2 = F(y)(1 - F(y))$ при $F(y) \neq 0, 1$, т.е. $F_n^*(y)$ есть асимптотически нормальная оценка для $F(y)$.

Гистограмма I

Для непрерывной величины гистограмма строится по сгруппированным данным (бинам): предполагаемую область значений (область выборочных данных) случайной величины делят на некоторое количество m интервалов (бинов) не обязательно равной длины $\Delta x_k,\ k = 1, ..., m$. Для каждого интервала производят подсчет количества значений n_k выборки, попавших в этот интервал, при этом очевидно $\sum_{k=1}^{m} n_k = N$. На каждом k-ом интервале строят прямоугольник, высота которого равна n_k. Если $f(x)$ – ф.п.в. данного распределения, то для его гистограммы имеем:

$$n_k = N \int_{\Delta x_k} f(x)dx \underset{f'\Delta x_k/f<<1}{\approx} Nf(x_k)\Delta x_k$$

В большинстве случаев все бины имеют одинаковую ширину $\Delta x_k = \Delta x,\ \forall k$, поэтому:

$$f(x_k) = \frac{n_k}{N\Delta x} = \text{const} \cdot n_k$$

Т.е. профиль гистограммы повторяет профиль ф.п.в. Таким образом, гистограмма, при надлежащем выборе ширины бина и достаточной статистике, позволяет визуализировать форму ф.п.в. $f(x)$. В случае, если $f(x)$ неизвестна, гистограмма позволяет по форме ф.п.в. сделать предположение о виде функции $f(x)$.

Гистограмма II

Мультиномиальное (полиномиальное) распределение – обобщение на случай $N > 1$ независимых испытаний случайного эксперимента с $k > 2$ возможными исходами. Для случайного вектора $\vec{n} = (n_1, ..., n_m)$ и набора вероятностей $\vec{p} = (p_1, ..., p_m)$:

$$P(\vec{n}|N, \vec{p}) = \frac{N!}{n_1!...n_m!} p_1^{n_1} p_2^{n_2} ... p_m^{n_m}, \quad \sum_{i=1}^{m} n_i = N, \quad \sum_{i=1}^{m} p_i = 1,$$

при этом $\overline{n_k} = E(n_k) = Np_k$, $\Delta n_k = \sigma_k = \sqrt{Np_k(1-p_k)}$.

Замечание 1: Очевидно, по построению гистограммы, набор чисел событий в бинах гистограммы распределён по мультиномиальному распределению, где $p_k = \int\limits_{\Delta x_k} f(x)dx \approx f(x_k)\Delta x_k$, причём $p_k \ll 1 \ \forall k$ и $\Delta n_k \approx \sqrt{Np_k} = \sqrt{\overline{n_k}} \approx \sqrt{n_k}$.

Замечание 2: Относительно выбора числа бинов нет строгих правил, однако стоит руководствоваться некоторыми "разумными" соображениями. При слишком грубом разбиении есть опасность потерять некоторые тонкие особенности в структуре распределения (например, узкий резонанс). С другой стороны, слишком мелкое разбиение приведёт к тому, что в каждый бин попадет 0 или 1 событие, и визуализация формы ф.п.в. станет невозможна. В качестве отправной точки можно взять $m = \sqrt{N}$, а потом подправить. Также ширину бина стоит выбирать достаточно круглым числом (0.1, 0.25, 0.5 и.т.п.; 0.10234789954 - плохая идея).

Выборочные характеристики и их свойства I

Статистикой называется любая (измеримая) функция от выборки случайной величины $\{X_k\}_n$: $Y = Y(X_1, ..., X_n)$. Будучи функцией от случайных величин, Y также является случайной величиной.

При определении оценки параметра θ в виде фиксированного численного значения θ^* следует помнить, что θ^* является величиной случайной, меняющейся в сериях повторных испытаний в соответствии с некоторым законом распределения $f(\theta^*|\theta)$. Поэтому задача оценки параметра включает не только определение конкретного "оптимального" значения параметра, но и определение ф.п.в. для этой случайной величины. При этом особой задачей статистики является построение такой оценки, которая бы включала максимум (всю) "информацию" о параметре, содержащуюся в исходных данных.
Качество оценки параметра определяется следующими характеристиками:
несмещённость, состоятельность, эффективность.

Выборочные характеристики и их свойства II

Несмещённость

Статистика $\theta^* = \theta^*(X_1, ..., X_n)$ называется несмещённой оценкой параметра θ, если $\forall \theta \in \Theta$ выполняется: $E(\theta^*) = \theta$.

Асимптотическая несмещённость

Статистика $\theta^* = \theta^*(X_1, ..., X_n)$ называется асимптотически несмещённой оценкой параметра θ, если $\forall \theta \in \Theta$ выполняется: $E(\theta^*) \underset{n \to \infty}{\to} \theta$.

Состоятельность

Статистика $\theta^* = \theta^*(X_1, ..., X_n)$ называется состоятельной оценкой параметра θ, если $\forall \theta \in \Theta$ имеет место сходимость по вероятности $\theta^* \underset{n \to \infty}{\to} \theta$.

Эффективность

Статистика $\theta^* = \theta^*(X_1, ..., X_n)$ называется эффективной оценкой параметра θ в некотором классе состоятельных оценок, если она не хуже в среднеквадратичном смысле $E((\theta^* - \theta)^2) \leq E((\theta_1^* - \theta)^2)$.

Асимптотическая нормальность

Статистика $\theta^* = \theta^*(X_1, ..., X_n)$ называется асимптотически нормальной оценкой параметра θ, если $\forall \theta \in \Theta$ имеет место сходимость:

$$\sqrt{n}(\theta^* - \theta) \to N(0, \sigma^2(\theta)),$$

где $\sigma^2(\theta)$ – дисперсия асимптотической нормальности.

Выборочные характеристики и их свойства III

Пример: Пусть $\{X_k\}_n$ – выборка случайной величины объёмом n, причём $\forall k,\ E(X_k) = \mu,\ D(X_k) = \sigma^2$. Введём выборочное среднее $\langle x \rangle$ и выборочную дисперсию s'^2:

$$\langle x \rangle = \frac{1}{n}\sum_{k=1}^{n} X_k, \quad s'^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \langle x \rangle)^2$$

$E(\langle x \rangle) = \frac{1}{n}\sum_{k=1}^{n} E(X_k) = \mu$, т.е. $\langle x \rangle$ является несмещённой оценкой μ.

$D(\langle x \rangle) = D(\frac{1}{n}\sum_{k=1}^{n} X_k) = \frac{1}{n^2}\sum_{k=1}^{n} D(X_k) = \frac{\sigma^2}{n},\ \sigma(\langle x \rangle) = \frac{\sigma}{\sqrt{n}}$

$E(s'^2) = E(\frac{1}{n}\sum_{k=1}^{n}(X_k - \mu - [\langle x \rangle - \mu])^2) =$

$\frac{1}{n}E(\sum_{k=1}^{n}((X_k - \mu)^2 - 2(X_k - \mu)(\langle x \rangle - \mu) + (\langle x \rangle - \mu)^2)) = \frac{n-1}{n}\sigma^2$

Т.о. выборочная дисперсия s'^2 является только асимптотически несмещённой оценкой истинной дисперсии σ^2.

Тогда $s^2 = s'^2 n/(n-1) = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \langle x \rangle)^2$ – несмещённая оценка дисперсии (далее будем пользоваться только s^2).