Непараметрическое оценивание

Малов Сергей Васильевич

Санкт-Петербургский государственный электротехнический университет

19/26 сентября 2020 г.

План

1. Выборочное распределение
2. Выборочные характеристики
3. Непараметрическая оценка плотности распределения

Эмпирическое распределение

Пусть $X_1, X_2, \ldots, X_n$ – выборка из распределения $\mathbb{P}_\theta$, $\theta \in \Theta$.

- Истинное значение $\mathbb{P}_\theta$ будем называть теоретическим распределением.
- Никаких априорных предположений о виде теоретического распределения не делают
- Распределение характеризуется функцией распределения
$$F(x) = P(X_1 < x),\ x \in \mathbb{R}$$
- В качестве параметра можно использовать теоретическую функцию распределения $\theta \equiv F$
- Параметризация: Θ – совокупность всех функций распределения случайной величины

Эмпирическое распределение

По исходной выборке построим дискретное распределение $\mathbb{P}_n$, имеющее атомы $1/n$ в точках $X_1, X_2, \ldots, X_n$

- данное распределение называется эмпирическим
- соответствующая функция распределения F_n называется эмпирической
- эмпирическая функция распределения имеет вид

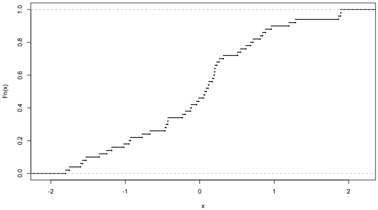
$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i < x\}},$$

где

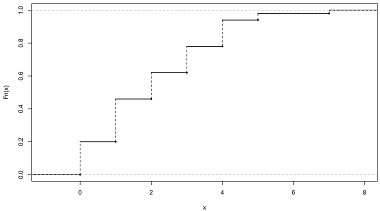
$$\mathbb{1}_A = \begin{cases} 1, & \text{если } A \text{ выполнено;} \\ 0, & \text{в остальных случаях.} \end{cases}$$

Эмпирическое распределение

Примеры построения графиков выборочных функций распределения:



- Нормальное распределение $\mathcal{N}(0,1)$
- Размер выборки n=50



- Распределение Пуассона $\lambda = 2$
- Размер выборки n=50

Порядковые статистики

Порядковые статистики и ранги:

- Упорядоченный набор наблюдений $X_{(1)} \leq \ldots \leq X_{(n)}$ – вариационный ряд
  - Элементы вариационного ряда – порядковые статистики
  - Переход от исходного набора наблюдений к порядковым статистикам необратим
- Для восстановления исходного набора наблюдений требуются ранги $R_1, \ldots, R_n$
  - Ранг i-го наблюдения R_i – его номер в вариационном ряду
  - При наличии повторных наблюдений ранги определены неоднозначно
- В выборочной модели
  - набор порядковых статстик – достаточная статистика
  - ранги не несут информации о распределении выборки

Эмпирическое распределение

Пусть $X^*_{(1)} < \ldots < X^*_{(k)}$ – элементы вариационного ряда без повторений

- Таблица эмпирического распределения:

Значение	$X^*_{(1)}$	$\ldots$	$X^*_{(k)}$
Вероятность	$\frac{\#\{i : X_i = X^*_{(1)}\}}{n}$	$\ldots$	$\frac{\#\{i : X_i = X^*_{(n)}\}}{n}$

- Эмпирические распределения, построенные по исходной выборке и по соответствующему вариационному ряду, совпадают
- Эмпирическая функция распределения является статистикой, ее используют в качестве оценки для теоретической функции распределения

Теорема (Гливенко–Кантелли)

Пусть $\mathrm{X} = (X_1, X_2, \ldots, X_n)$ – выборка из распределения P_θ, $\theta \in \Theta$; $F(x)$ и $F_n(x)$ – теоретическая и эмпирическая функции распределения соответственно. Тогда с вероятностью 1

$$\lim_{n\to\infty} D_n(\mathrm{X}) = \lim_{n\to\infty} \sup_{-\infty<x<\infty} |F_n(x) - F(x)| \xrightarrow[n\to\infty]{} 0.$$

Идея доказательства. Поточечная сходимость получается непосредственно из усиленного закона больших чисел в схеме Бернулли с вероятностями успеха $F(x)$, а равномерная получается, поскольку значения на $\pm\infty$ у функций $F(x)$ и $F_n(x)$ совпадают и конечны. ■

- Эмпирическая функция распределения – статистика, использующаяся в качестве оценки теоретической функции распределения
- С ростом размера выборки эмпирическая функция распределения приближается к теоретической (состоятельность)

Теорема Гливенко–Кантелли

Теорема (Гливенко–Кантелли)

Пусть $\mathrm{X} = (X_1, X_2, \ldots, X_n)$ – выборка из распределения P_θ, $\theta \in \Theta$; $F(x)$ и $F_n(x)$ – теоретическая и эмпирическая функции распределения соответственно. Тогда с вероятностью 1

$$\lim_{n\to\infty} D_n(\mathrm{X}) = \lim_{n\to\infty} \sup_{-\infty<x<\infty} |F_n(x) - F(x)| \xrightarrow[n\to\infty]{} 0.$$

Идея доказательства. Поточечная сходимость получается непосредственно из усиленного закона больших чисел в схеме Бернулли с вероятностями успеха $F(x)$, а равномерная получается, поскольку значения на $\pm\infty$ у функций $F(x)$ и $F_n(x)$ совпадают и конечны. ■

- Эмпирическая функция распределения – статистика, использующаяся в качестве оценки теоретической функции распределения
- С ростом размера выборки эмпирическая функция распределения приближается к теоретической (состоятельность)

Теорема (Гливенко–Кантелли)

Пусть $\mathrm{X} = (X_1, X_2, \ldots, X_n)$ – выборка из распределения P_θ, $\theta \in \Theta$; $F(x)$ и $F_n(x)$ – теоретическая и эмпирическая функции распределения соответственно. Тогда с вероятностью 1

$$\lim_{n\to\infty} D_n(\mathrm{X}) = \lim_{n\to\infty} \sup_{-\infty<x<\infty} |F_n(x) - F(x)| \xrightarrow[n\to\infty]{} 0.$$

Идея доказательства. Поточечная сходимость получается непосредственно из усиленного закона больших чисел в схеме Бернулли с вероятностями успеха $F(x)$, а равномерная получается, поскольку значения на $\pm\infty$ у функций $F(x)$ и $F_n(x)$ совпадают и конечны. ■

- Эмпирическая функция распределения – статистика, использующаяся в качестве оценки теоретической функции распределения
- С ростом размера выборки эмпирическая функция распределения приближается к теоретической (состоятельность)

Теорема Колмогорова

Отклонение теоретической функции распределения от эмпирической

$$D_n(\mathrm{X}) = \sup_{-\infty<x<\infty} |F_n(x) - F(x)|$$

называется статистикой Колмогорова

Теорема (Колмогоров)

Если $F(x)$ непрерывна, то для любого положительного значения t

$$\lim_{n\to\infty} \mathbb{P}(\sqrt{n}D_n \leq t) = K(t) = \sum_{j=-\infty}^{\infty}(-1)^j e^{-2j^2t^2}.$$

- Доказательство основано на преобразовании Смирнова
- Существует точное распределение статистики Колмогорова при фиксированном n
- Функция $K(t)$ называется функцией распределения Колмогорова.
- На практике данная теорема дает хорошее приближение уже при $n \geq 20$.

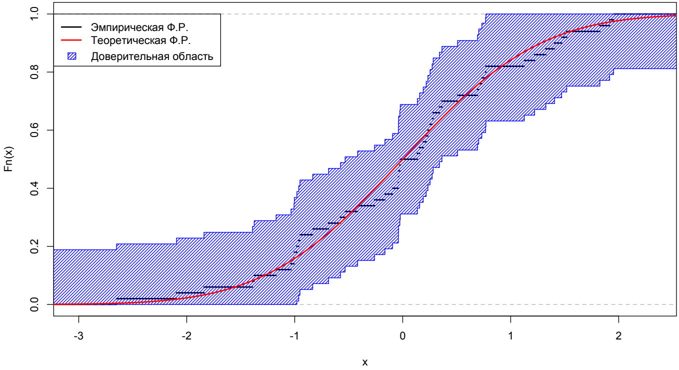
Теорема Колмогорова

Построение доверительной области с использованием Теоремы Колмогорова.

- Задаемся малым числом $\alpha > 0$
- Находим такое z_α, что $K(z_\alpha) = 1 - \alpha$
- Если выборка имеет объем n, то $\mathbb{P}(\sqrt{n}D_n \leq z_\alpha) \geq 1 - \alpha$ означает, что неравенство
$$F_n(x) - z_\alpha/\sqrt{n} < F(x) < F_n(x) + z_\alpha/\sqrt{n}$$
выполняется одновременно для всех x с вероятностью не менее $1 - \alpha$
- Совокупность всех функций распределения F, удовлетворяющих условию
$$F_n(x) - z_\alpha/\sqrt{n} < F(x) < F_n(x) + z_\alpha/\sqrt{n}$$
при всех $x \in \mathbb{R}$, является доверительной уровня доверия $1 - \alpha$
- Аналогично, доверительная область может быть построена исходя из точного распределения D_n

Теорема Колмогорова

Доверительная область, полученная с использованием точного распределения статистики Колмогорова



- Теоретическое распределение – нормальное $\mathcal{N}(0,1)$
- Размер выборки $n = 50$

План

1 Выборочное распределение

2 Выборочные характеристики

3 Непараметрическая оценка плотности распределения

Выборочные числовые характеристики

Пусть $\alpha : \mathcal{C} \to \mathbb{R}$ – числовая характеристика, $\mathcal{C}$ – класс распределений

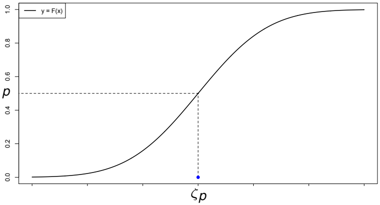
- $\alpha(F)$ – теоретическое значение числовой характеристики, $F \in \mathcal{C}$
- статистика $\alpha(F_n)$ (если $F_n \in \mathcal{C}$) – выборочный аналог числовой характеристики

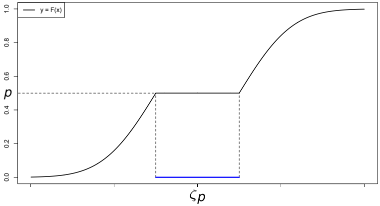
Некоторые примеры:

Название	Теоретическая характеристика	Выборочная характеристика
Мат. ожидание	$E_F X = \int x\, dF(x)$	$\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$
k-й момент	$E_F X^k = \int x^k\, dF(x)$	$\overline{X^k} = \frac{1}{n}\sum_{i=1}^n X_i^k$
g-момент	$E_F g(X) = \int g(x)\, dF(x)$	$\overline{g(X)} = \frac{1}{n}\sum_{i=1}^n g(X_i)$
Дисперсия	$D_F X = \int x^2 dF(x) - (E_F X)^2 = \int (x - E_F X)^2 dF(x)$	$s^2 = \overline{X^2} - \overline{X}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2$
p-квантиль	$\zeta_p : \mathbb{P}(\xi \le \zeta_p) \ge p$, $\mathbb{P}(\xi \ge \zeta_p) \ge 1 - p$	$Z_{n,p} = \begin{cases} X_{([np]+1)}, & \text{если } np \notin \mathbb{Z}; \\ \left[X_{([np]+1)}, X_{([np]+1)}\right], & np \in \mathbb{Z}. \end{cases}$

Квантили, графическое представление

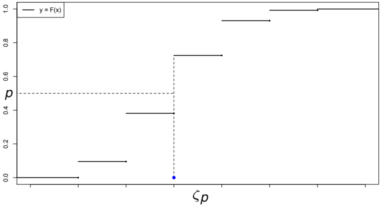
Абсолютно непрерывные распределения

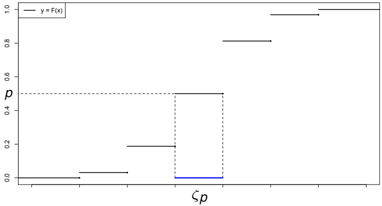




- ζ_p определена однозначно
- ζ_p – любая точка интервала

Дискретные распределения





- ζ_p определена однозначно
- ζ_p – любая точка интервала

Рассмотрим числовые характеристики двух типов

- I тип: числовые характеристики, представимые в виде
$$\alpha(F) = H(\mathbb{E}_F g_1(X), \ldots, \mathbb{E}_F g_k(X))$$
  - H – непрерывная функция k аргументов
  - $F \in \mathcal{C} = \{F : \mathbb{E}_F |g_i(X)| < \infty, i = 1, \ldots, k\}$.
- II тип: $\alpha(F)$ непрерывный функционал в равномерной метрике
  - Для любой последовательности функций распределения $\{F_n^*\}_{n=1}^{\infty}$: $F_n^* \in \mathcal{C}$,
$$\sup_x |F_n^*(x) - F(x)| \xrightarrow[n\to\infty]{} 0 \quad \Rightarrow \quad \alpha(F_n) \xrightarrow[n\to\infty]{} \alpha(F)$$

Теорема (состоятельность)

Пусть $X_1, X_2, \ldots, X_n$ – выборка из распределения F. Если при каждом $\mathbb{P}_\theta \in \mathcal{C}$ числовая характеристика $\alpha(F)$ первого или второго типа существует, то с вероятностью 1 с ростом размера выборки
$$\alpha(F_n) \xrightarrow[n\to\infty]{} \alpha(F).$$

Асимптотические свойства выборочных характеристик

Идея доказательства

Для статистик I типа выражение $\frac{1}{n}\sum_{i=1}^{n} g_j(X_i)$ представляет собой сумму независимых одинаково распределенных случайных величин с математическим ожиданием $\mathbb{E}_F g_j(X) = \int_{-\infty}^{\infty} g_j(x)\, dF(x)$, $j = 1, \ldots, k$. Используя усиленный закон больших чисел, и непрерывность функции H получаем требуемое утверждение.

Для статистик II типа утверждение теоремы следует непосредственно из теоремы Гливенко – Кантелли. ■

- Числовые характеристики I типа
  - Математическое ожидание
  - Дисперсия
  - Моменты (центральные, абсолютные, g-моменты)
  - Асимметрия, эксцесс
- Числовые характеристики II типа:
  - ζ_p (квантиль порядка p), если $\mathcal{C}$ – класс абсолютно непрерывных распределений: $F(\zeta_p)$ строго возрастает в окрестности ζ_p
    - $p = 1/2$ – медиана
    - $p = 1/4$ и $p = 3/4$ – квартили
  - точная верхняя (нижняя) грань носителя $\sup\{\mathrm{supp}(X)\}$ ($\inf\{\mathrm{supp}(X)\}$), если $\mathcal{C}$ – класс абсолютно непрерывных распределений

Идея доказательства

Для статистик I типа выражение $\frac{1}{n}\sum_{i=1}^{n} g_j(X_i)$ представляет собой сумму независимых одинаково распределенных случайных величин с математическим ожиданием $\mathbb{E}_F g_j(X) = \int_{-\infty}^{\infty} g_j(x)\,dF(x)$, $j = 1, \ldots, k$. Используя усиленный закон больших чисел, и непрерывность функции H получаем требуемое утверждение.

Для статистик II типа утверждение теоремы следует непосредственно из теоремы Гливенко – Кантелли. ■

- Числовые характеристики I типа
  - Математическое ожидание
  - Дисперсия
  - Моменты (центральные, абсолютные, g-моменты)
  - Асимметрия, эксцесс
- Числовые характеристики II типа:
  - ζ_p (квантиль порядка p), если $\mathcal{C}$ – класс абсолютно непрерывных распределений: $F(\zeta_p)$ строго возрастает в окрестности ζ_p
    - $p = 1/2$ – медиана
    - $p = 1/4$ и $p = 3/4$ – квартили
  - точная верхняя (нижняя) грань носителя $\sup\{\mathrm{supp}(X)\}$ ($\inf\{\mathrm{supp}(X)\}$), если $\mathcal{C}$ – класс абсолютно непрерывных распределений

Теорема (Асимптотическая нормальность)

Пусть $X_1, X_2, \ldots, X_n$ – выборка из распределения F; α – числовая характеритика I типа:

(i) $\mathbb{E}_F|g_i(X)|^2 < \infty$, $i = 1, \ldots, n$

(ii) H – непрерывно дифференцируемая функция k переменных с ненулевым дифференциалом в точке $(\mathbb{E}_F g_1(X), \ldots, \mathbb{E}_F g_k(X))$

Тогда
$$\sqrt{n}(\alpha(F_n) - \alpha(F)) \Rightarrow \mathcal{N}(0, \sigma_\alpha^2) \qquad (*)$$

Идея доказательства

Отметим, что $\xi_1, \ldots, \xi_n$ – независимые и одинаково распределенные случайные векторы, $\xi_j = (\xi_{1j}, \ldots, \xi_{kj})$, $\xi_{ij} = g_i(X_j)$. Используя ЦПТ в векторной форме получаем, что
$$\sqrt{n}\left(\tfrac{1}{n}\sum_{j=1}^n \xi_j - \mathbb{E}_F\xi_1\right) \Rightarrow \mathcal{N}(0, \Sigma),$$
где Σ – матрица ковариации ξ_1. По формуле Тейлора
$$\sqrt{n}(\alpha(F_n) - \alpha(F)) = \sum_{i=1}^k \frac{\partial}{\partial x_i} H(x_1^*, \ldots, x_k^*)\sqrt{n}(\mathbb{E}_{F_n}\xi_i - \mathbb{E}_F\xi_i)$$
$(x_1^*, \ldots, x_k^*)$ – точка на отрезке, соединяющим $(\xi_1, \ldots, \xi_n)$ и $\mathbb{E}_F(\xi_1, \ldots, \xi_n)$. Далее, с учетом теоремы о состоятельности и непрерывности H получаем сходимость (*), где $\sigma_\alpha^2 = \nabla H(\mathbb{E}_F\xi_{11}, \ldots, \mathbb{E}_F\xi_{k1})' \, \Sigma \, \nabla H(\mathbb{E}_F\xi_{11}, \ldots, \mathbb{E}_F\xi_{k1})$. ∎

Теорема (Асимптотическая нормальность)

Пусть $X_1, X_2, \ldots, X_n$ – выборка из распределения F; α – числовая характеритика I типа:

(i) $\mathbb{E}_F|g_i(X)|^2 < \infty$, $i = 1, \ldots, n$

(ii) H – непрерывно дифференцируемая функция k переменных с ненулевым дифференциалом в точке $(\mathbb{E}_F g_1(X), \ldots, \mathbb{E}_F g_k(X))$

Тогда

$$\sqrt{n}(\alpha(F_n) - \alpha(F)) \Rightarrow \mathcal{N}(0, \sigma_\alpha^2) \qquad (*)$$

Идея доказательства

Отметим, что $\xi_1, \ldots, \xi_n$ – независимые и одинаково распределенные случайные векторы, $\xi_j = (\xi_{1j}, \ldots, \xi_{kj})$, $\xi_{ij} = g_i(X_j)$. Используя ЦПТ в векторной форме получаем, что

$$\sqrt{n}\left(\frac{1}{n}\sum_{j=1}^n \xi_j - \mathbb{E}_F\xi_1\right) \Rightarrow \mathcal{N}(0, \Sigma),$$

где Σ – матрица ковариации ξ_1. По формуле Тейлора

$$\sqrt{n}(\alpha(F_n) - \alpha(F)) = \sum_{i=1}^k \frac{\partial}{\partial x_i} H(x_1^*, \ldots, x_k^*)\sqrt{n}(\mathbb{E}_{F_n}\xi_i - \mathbb{E}_F\xi_i)$$

$(x_1^*, \ldots, x_k^*)$ – точка на отрезке, соединяющим $(\xi_1, \ldots, \xi_n)$ и $\mathbb{E}_F(\xi_1, \ldots, \xi_n)$. Далее, с учетом теоремы о состоятельности и непрерывности H получаем сходимость (*), где $\sigma_\alpha^2 = \nabla H(\mathbb{E}_F\xi_{11}, \ldots, \mathbb{E}_F\xi_{k1})' \, \Sigma \, \nabla H(\mathbb{E}_F\xi_{11}, \ldots, \mathbb{E}_F\xi_{k1})$. ∎

Теорема

Пусть $X_1, X_2, \ldots, X_n$ – выборка из абсолютно непрерывного распределения F с плотностью распределения f, дифференцируемой в точке ζ_p: $f'(\zeta_p) > 0$;

$$\eta_n = \sqrt{\frac{n}{p(1-p)}}\, f(\zeta_p)(Z_{n,p} - \zeta_p),$$

$p \in (0,1)$. Тогда при $n \to \infty$ имеет место сходимость по распределению $\eta_n \Rightarrow N(0,1)$, т. е.

$$\mathbb{P}_F(\eta_n < x) \underset{n\to\infty}{\longrightarrow} \Phi(x), \quad x \in \mathbb{R},$$

где

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp(-t^2/2)\, dt$$

есть функция стандартного нормального распределения $N(0,1)$.

Доказательство

Пусть $k = np$, если np – целое, и $k = [np] + 1$ в остальных случаях. Рассмотрим случайные величины

$$Y_i = \sqrt{\frac{n}{pq}}\, f(\zeta_p)(X_i - \zeta_p), \qquad J_i(t) = 1_{\{Y_i < t\}},$$

$i = 1, 2, \ldots, n$, где $1_A = 1$, если событие A произошло, и $1_A = 0$ в противном случае. Очевидно, что $\eta_n = Y_{(k)}$. При этом $\mathbb{P}_F(Y_{(k)} < t) = \mathbb{P}_F(\sum_{i=1}^n J_i(t) \geq k)$. Отметим, что $J_1(t), \ldots, J_n(t)$ – испытания Бернулли с вероятностью успеха

$$\mathbb{P}_F(J_i(t) = 1) = \mathbb{P}_F(Y_i < t) = \mathbb{P}_F\left(X_i < \zeta_p + \sqrt{\frac{pq}{n}}\frac{t}{f(\zeta_p)}\right) = F\left(\zeta_p + \sqrt{\frac{pq}{n}}\frac{t}{f(\zeta_p)}\right).$$

Используя разложение Тейлора в окрестности ζ_p получаем

$$a = \mathbb{E}_F J_i = \mathbb{P}_F(J_i(t) = 1) = F(\zeta_p) + f(\zeta_p)\frac{t\sqrt{pq}}{\sqrt{n}f(\zeta_p)} + o(1/\sqrt{n}) = F(\zeta_p) + t\sqrt{pq/n} + o(1/\sqrt{n}),$$

$$\mathbb{D}_F J_i = a(1-a).$$

Рассмотрим

$$\mathbb{P}_F\left(\sum_{i=1}^n J_i \geq k\right) = \mathbb{P}_F\left(\frac{\sum_{i=1}^n J_i - na}{\sqrt{na(1-a)}} \geq \frac{k - na}{\sqrt{na(1-a)}}\right).$$

Остается отметить, что

$$\frac{k - np - t\sqrt{np(1-p)} + o(\sqrt{n})}{\sqrt{np(1-p)(1+o(1))}} \xrightarrow[n\to\infty]{} -t.$$

Используя центральную предельную теорему и симметричность стандартного нормального распределения, получаем:

$$\mathbb{P}_F\left(\sum_{i=1}^n J_i \geq k\right) \xrightarrow[n\to\infty]{} 1 - \Phi(-t) = \Phi(t). \qquad \blacksquare$$

План

1. Выборочное распределение

2. Выборочные характеристики

3. Непараметрическая оценка плотности распределения

Гистограмма: дискретное распределение выборки

- Эмпирическое распределение может использоваться для оценивания дискретной плотности распределения.
  - введем частоты $\nu(x)/n$, где $\nu(x)$ – число наблюдений, имеющих значение x.
  - с ростом n, согласно закону больших чисел, частоты будут сходиться к теоретическим значениям $q_\theta(x) = \mathbb{P}_\theta(X_1 = x)$.
- Приближая распределения дискретными и используя частоты в качестве оценок вероятностей, мы получаем наглядный способ представления данных
- Для наглядного представления данных часто используют значения $\nu(x)$ без нормировки на число наблюдений

Гистограмма частот

Гистограмма: абсолютно непрерывное распределение выборки

- вещественную прямую разбивают на интервалы $\{I_i\}_{i \in \mathbb{N}}$ (их счетное число) одинаковой длины $h > 0$
- вычислим выборочные аналоги вероятностей попадания в соответствующие интервалы ν_i/n, где ν_i – число наблюдений попавших в i-й интервал
- Функция $H(x) = \nu_j/(nh)$, $x \in I_j$, $j \in \mathbb{N}$ называется гистограммой
  - площадь подграфика $H(x; h)$ равна единице при любом $h > 0$
  - $\nu_i/n \xrightarrow[n\to\infty]{} \mathbb{P}_\theta(X_1 \in I_i)$ по вероятности (и с вероятностью 1) при каждом фиксированном θ
  - если плотность непрерывна, то $H(x; h), x \in I_j$ является оценкой некоторого среднего значения плотности на интервале I_j

Оценивание плотности распределения

Пусть $X_1, X_2, \dots, X_n$ – выборка из абсолютно непрерывного распределения с плотностью распределения p

- Уменьшая подходящим образом $h = h_n$ ($nh_n \xrightarrow[n\to\infty]{} \infty$, $h_n \to 0$), получаем оценку теоретической плотности распределения
- При наличии производной p' в окрестности точки x граница среднеквадратической ошибки (риска) локальной оценки $p(x)$ с помощью гистограммы задается соотношением

$$\mathbb{E}_F(H(x;h) - p(x))^2 \le d^2h^2 + p_{\max}/(nh) + p_{\max}^2/n$$

  - $d = \sup_{x\in I_x}(|p'(x)|)$; $p_{\max} = \sup_{x\in I_x}(p(x))$
  - I_x – интервал разбиения, содержащий точку x
- Оптимизация правой части неравенства дает

$$h_{\mathrm{opt}} = \left(p_{\max}/(nd^2)\right)^{1/3}$$

- При больших n разумно использовать шаг $h_n = c\,n^{-1/3}$
  - $c > 0$ – некоторая константа

Использование свойств непрерывности и гладкости плотности p

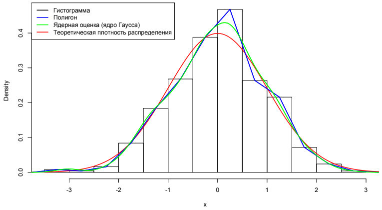
- Если функция плотности непрерывная, то ломанными ее можно приблизить лучше, чем ступенчатыми функциями
  - полигон частот – кусочно-линейная непрерывная функция, совпадающая с гистограммой в середине каждого интервала
  - площадь подграфика полигона частот тоже равна единице
  - для оценки непрерывных плотностей целесообразно использовать полигон частот вместо гистограммы
- Современные методы оценивания гладкой плотности распределения основаны на построении ядерных оценок

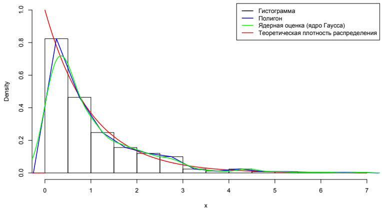
$$f_n(x) = (nh_n)^{-1} \sum_{i=1}^{n} K\Big((x - X_i)/h_n\Big)$$

  - K – неотрицательная функция ограниченной вариации («ядро»)
  - обычно K симметрична отностительно нуля

Оценивание плотности распределения

Примеры построения графиков оценок плотности распределения:





- Нормальное распределение $\mathcal{N}(0,1)$
- Размер выборки n=500
- Плотность распределения гладкая

- Экспоненциальное распределение
- Размер выборки n=500
- Плотность расперделения имеет разрыв в нуле