Теория Вероятностей и Статистика

Оценивание характеристик распределения

Потанин Богдан Станиславович

доцент, научный сотрудник, кандидат экономических наук

2023–2024

- Часто, сделать заранее достаточно точное предположение о семействе распределений, из которого была получена выборка, весьма затруднительно.

- Часто, сделать заранее достаточно точное предположение о семействе распределений, из которого была получена выборка, весьма затруднительно.
- Тем не менее, некоторые отдельные характеристики (вероятности, медиана, математическое ожидание и т.д.) можно оценить, не накладывая существенных допущений о распределении и даже не учитывая его параметры.

- Часто, сделать заранее достаточно точное предположение о семействе распределений, из которого была получена выборка, весьма затруднительно.
- Тем не менее, некоторые отдельные характеристики (вероятности, медиана, математическое ожидание и т.д.) можно оценить, не накладывая существенных допущений о распределении и даже не учитывая его параметры.
- Рассмотрим методы, позволяющие оценивать различные характеристики распределения без предварительного оценивания параметров и без предположений о конкретном семействе распределений.

- Имеется выборка $X_1, ..., X_n$ из распределения $\Theta(\theta)$.

Оценивание вероятностей

Безусловные вероятности

- Имеется выборка $X_1, ..., X_n$ из распределения $\Theta(\theta)$.
- Необходимо оценить вероятность $P(X_1 \in A)$, где $A \subset R$.

Оценивание вероятностей

Безусловные вероятности

- Имеется выборка $X_1, ..., X_n$ из распределения $\Theta(\theta)$.
- Необходимо оценить вероятность $P(X_1 \in A)$, где $A \subset R$.
- Несмещенная и состоятельная оценка данной вероятности может быть получена как доля наблюдений в выборке, принадлежащих множеству A:

$$\hat{P}(X_1 \in A) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \in A) \sim \frac{1}{n}B(n, P(X_1 \in A))$$

$$\text{где: } I(X_i \in A) = \begin{cases} 1, \text{ если } X_i \in A \\ 0, \text{ в противном случае} \end{cases} \sim Ber(P(X_1 \in A))$$

- Имеется выборка $X_1, ..., X_n$ из распределения $\Theta(\theta)$.
- Необходимо оценить вероятность $P(X_1 \in A)$, где $A \subset R$.
- Несмещенная и состоятельная оценка данной вероятности может быть получена как доля наблюдений в выборке, принадлежащих множеству A:

$$\hat{P}(X_1 \in A) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \in A) \sim \frac{1}{n}B(n, P(X_1 \in A))$$

$$\text{где: } I(X_i \in A) = \begin{cases} 1, \text{ если } X_i \in A \\ 0, \text{ в противном случае} \end{cases} \sim Ber(P(X_1 \in A))$$

- Подставляя вместо наблюдений X_i их реализации x_i можно получить реализацию оценки соответствующей вероятности.

Оценивание вероятностей

Безусловные вероятности

- Имеется выборка $X_1, ..., X_n$ из распределения $\Theta(\theta)$.
- Необходимо оценить вероятность $P(X_1 \in A)$, где $A \subset R$.
- Несмещенная и состоятельная оценка данной вероятности может быть получена как доля наблюдений в выборке, принадлежащих множеству A:

$$\hat{P}(X_1 \in A) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \in A) \sim \frac{1}{n}B(n, P(X_1 \in A))$$

$$\text{где: } I(X_i \in A) = \begin{cases} 1, \text{ если } X_i \in A \\ 0, \text{ в противном случае} \end{cases} \sim Ber(P(X_1 \in A))$$

- Подставляя вместо наблюдений X_i их реализации x_i можно получить реализацию оценки соответствующей вероятности.

Доказательство: докажем несмещенность и состоятельность рассматриваемой оценки вероятности:

$$E\left(\hat{P}(X_1 \in A)\right) = \frac{1}{n}\sum_{i=1}^{n} E\left(I(X_i \in A)\right) = \frac{1}{n}\sum_{i=1}^{n} P(X_i \in A) = \frac{1}{n} \times nP(X_1 \in A) = P(X_1 \in A)$$

Оценивание вероятностей

Безусловные вероятности

- Имеется выборка $X_1, ..., X_n$ из распределения $\Theta(\theta)$.
- Необходимо оценить вероятность $P(X_1 \in A)$, где $A \subset R$.
- Несмещенная и состоятельная оценка данной вероятности может быть получена как доля наблюдений в выборке, принадлежащих множеству A:

$$\hat{P}(X_1 \in A) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \in A) \sim \frac{1}{n}B(n, P(X_1 \in A))$$

$$\text{где: } I(X_i \in A) = \begin{cases} 1, \text{ если } X_i \in A \\ 0, \text{ в противном случае} \end{cases} \sim Ber(P(X_1 \in A))$$

- Подставляя вместо наблюдений X_i их реализации x_i можно получить реализацию оценки соответствующей вероятности.

  Доказательство: докажем несмещенность и состоятельность рассматриваемой оценки вероятности:

$$E\left(\hat{P}(X_1 \in A)\right) = \frac{1}{n}\sum_{i=1}^{n} E\left(I(X_i \in A)\right) = \frac{1}{n}\sum_{i=1}^{n} P(X_i \in A) = \frac{1}{n} \times nP(X_1 \in A) = P(X_1 \in A)$$

$$\lim_{n\to\infty} Var\left(\hat{P}(X_1 \in A)\right) = \lim_{n\to\infty} \frac{1}{n^2}\sum_{i=1}^{n} Var\left(I(X_i \in A)\right) = \lim_{n\to\infty} \frac{n \times Var\left(I(X_1 \in A)\right)}{n^2} = 0$$

- Имеется выборка с реализацией $x = (5, 3, 1, -2, 3)$. Найдите реализации оценок вероятностей $P(X_1 = 3)$, $P(X_1 \geq 2.5)$, $P(X_1 = 2)$, $P(-10 \leq X_1 < 5)$, $P\left(X_1 \in \{2, -2, 1\}\right)$ и $P(X_1 \leq 100)$.

- Имеется выборка с реализацией $x = (5, 3, 1, -2, 3)$. Найдите реализации оценок вероятностей $P(X_1 = 3)$, $P(X_1 \geq 2.5)$, $P(X_1 = 2)$, $P(-10 \leq X_1 < 5)$, $P\left(X_1 \in \{2, -2, 1\}\right)$ и $P(X_1 \leq 100)$.
  Решение:
$$\hat{P}(X_1 = 3)(x) = \frac{1}{5}\left(I(x_1 = 3) + I(x_2 = 3) + I(x_3 = 3) + I(x_4 = 3) + I(x_5 = 3)\right) =$$

Оценивание вероятностей

Пример оценивания обычных вероятностей

- Имеется выборка с реализацией $x = (5, 3, 1, -2, 3)$. Найдите реализации оценок вероятностей $P(X_1 = 3)$, $P(X_1 \geq 2.5)$, $P(X_1 = 2)$, $P(-10 \leq X_1 < 5)$, $P\left(X_1 \in \{2, -2, 1\}\right)$ и $P(X_1 \leq 100)$.
  Решение:
$$\hat{P}(X_1 = 3)(x) = \frac{1}{5}\left(I(x_1 = 3) + I(x_2 = 3) + I(x_3 = 3) + I(x_4 = 3) + I(x_5 = 3)\right) =$$
$$= \frac{1}{5}(0 + 1 + 0 + 0 + 1) = \frac{2}{5} = 0.4$$

- Имеется выборка с реализацией $x = (5, 3, 1, -2, 3)$. Найдите реализации оценок вероятностей $P(X_1 = 3)$, $P(X_1 \geq 2.5)$, $P(X_1 = 2)$, $P(-10 \leq X_1 < 5)$, $P\left(X_1 \in \{2, -2, 1\}\right)$ и $P(X_1 \leq 100)$.
  Решение:
  $$\hat{P}(X_1 = 3)(x) = \frac{1}{5}\left(I(x_1 = 3) + I(x_2 = 3) + I(x_3 = 3) + I(x_4 = 3) + I(x_5 = 3)\right) =$$
  $$= \frac{1}{5}(0 + 1 + 0 + 0 + 1) = \frac{2}{5} = 0.4$$
  $$\hat{P}(X_1 \geq 2.5)(x) = \frac{1}{5}(1 + 1 + 0 + 0 + 1) = \frac{3}{5} = 0.6$$

- Имеется выборка с реализацией $x = (5, 3, 1, -2, 3)$. Найдите реализации оценок вероятностей $P(X_1 = 3)$, $P(X_1 \geq 2.5)$, $P(X_1 = 2)$, $P(-10 \leq X_1 < 5)$, $P\left(X_1 \in \{2, -2, 1\}\right)$ и $P(X_1 \leq 100)$.
  Решение:

$$\hat{P}(X_1 = 3)(x) = \frac{1}{5}\left(I(x_1 = 3) + I(x_2 = 3) + I(x_3 = 3) + I(x_4 = 3) + I(x_5 = 3)\right) =$$

$$= \frac{1}{5}\left(0 + 1 + 0 + 0 + 1\right) = \frac{2}{5} = 0.4$$

$$\hat{P}(X_1 \geq 2.5)(x) = \frac{1}{5}\left(1 + 1 + 0 + 0 + 1\right) = \frac{3}{5} = 0.6$$

$$\hat{P}(X_1 = 2)(x) = \frac{1}{5}\left(0 + 0 + 0 + 0 + 0\right) = 0$$

- Имеется выборка с реализацией $x = (5, 3, 1, -2, 3)$. Найдите реализации оценок вероятностей $P(X_1 = 3)$, $P(X_1 \geq 2.5)$, $P(X_1 = 2)$, $P(-10 \leq X_1 < 5)$, $P\left(X_1 \in \{2, -2, 1\}\right)$ и $P(X_1 \leq 100)$.
  Решение:

$$\hat{P}(X_1 = 3)(x) = \frac{1}{5}\left(I(x_1 = 3) + I(x_2 = 3) + I(x_3 = 3) + I(x_4 = 3) + I(x_5 = 3)\right) =$$

$$= \frac{1}{5}\left(0 + 1 + 0 + 0 + 1\right) = \frac{2}{5} = 0.4$$

$$\hat{P}(X_1 \geq 2.5)(x) = \frac{1}{5}\left(1 + 1 + 0 + 0 + 1\right) = \frac{3}{5} = 0.6$$

$$\hat{P}(X_1 = 2)(x) = \frac{1}{5}\left(0 + 0 + 0 + 0 + 0\right) = 0$$

$$\hat{P}(-10 \leq X_1 < 5)(x) = \frac{1}{5}\left(0 + 1 + 1 + 1 + 1\right) = \frac{4}{5} = 0.8$$

- Имеется выборка с реализацией $x = (5, 3, 1, -2, 3)$. Найдите реализации оценок вероятностей $P(X_1 = 3)$, $P(X_1 \geq 2.5)$, $P(X_1 = 2)$, $P(-10 \leq X_1 < 5)$, $P\left(X_1 \in \{2, -2, 1\}\right)$ и $P(X_1 \leq 100)$.
  Решение:

$$\hat{P}(X_1 = 3)(x) = \frac{1}{5}\left(I(x_1 = 3) + I(x_2 = 3) + I(x_3 = 3) + I(x_4 = 3) + I(x_5 = 3)\right) =$$

$$= \frac{1}{5}\left(0 + 1 + 0 + 0 + 1\right) = \frac{2}{5} = 0.4$$

$$\hat{P}(X_1 \geq 2.5)(x) = \frac{1}{5}\left(1 + 1 + 0 + 0 + 1\right) = \frac{3}{5} = 0.6$$

$$\hat{P}(X_1 = 2)(x) = \frac{1}{5}\left(0 + 0 + 0 + 0 + 0\right) = 0$$

$$\hat{P}(-10 \leq X_1 < 5)(x) = \frac{1}{5}\left(0 + 1 + 1 + 1 + 1\right) = \frac{4}{5} = 0.8$$

$$\hat{P}(X_1 \in \{2, -2, 1\})(x) = \frac{1}{5}\left(0 + 0 + 1 + 1 + 0\right) = \frac{2}{5} = 0.4$$

- Имеется выборка с реализацией $x = (5, 3, 1, -2, 3)$. Найдите реализации оценок вероятностей $P(X_1 = 3)$, $P(X_1 \geq 2.5)$, $P(X_1 = 2)$, $P(-10 \leq X_1 < 5)$, $P\left(X_1 \in \{2, -2, 1\}\right)$ и $P(X_1 \leq 100)$.
Решение:

$$\hat{P}(X_1 = 3)(x) = \frac{1}{5}\left(I(x_1 = 3) + I(x_2 = 3) + I(x_3 = 3) + I(x_4 = 3) + I(x_5 = 3)\right) =$$

$$= \frac{1}{5}(0 + 1 + 0 + 0 + 1) = \frac{2}{5} = 0.4$$

$$\hat{P}(X_1 \geq 2.5)(x) = \frac{1}{5}(1 + 1 + 0 + 0 + 1) = \frac{3}{5} = 0.6$$

$$\hat{P}(X_1 = 2)(x) = \frac{1}{5}(0 + 0 + 0 + 0 + 0) = 0$$

$$\hat{P}(-10 \leq X_1 < 5)(x) = \frac{1}{5}(0 + 1 + 1 + 1 + 1) = \frac{4}{5} = 0.8$$

$$\hat{P}(X_1 \in \{2, -2, 1\})(x) = \frac{1}{5}(0 + 0 + 1 + 1 + 0) = \frac{2}{5} = 0.4$$

$$\hat{P}(X_1 \leq 100)(x) = \frac{1}{5}(1 + 1 + 1 + 1 + 1) = \frac{5}{5} = 1$$

- Имеется выборка $X_1, ..., X_n$ из распределения $\Theta(\theta)$.

Оценивание вероятностей

Условные вероятности

- Имеется выборка $X_1, ..., X_n$ из распределения $\Theta(\theta)$.
- Необходимо оценить условную вероятность $P(X_1 \in A | X_1 \in B)$, где $A, B \subset R$ и $P(X_1 \in B) > 0$.

Оценивание вероятностей

Условные вероятности

- Имеется выборка $X_1, ..., X_n$ из распределения $\Theta(\theta)$.
- Необходимо оценить условную вероятность $P(X_1 \in A | X_1 \in B)$, где $A, B \subset R$ и $P(X_1 \in B) > 0$.
- Несмещенная и состоятельная оценка данной вероятности может быть получена как доля наблюдений в выборке, принадлежащих множеству $A \cap B$, среди наблюдений, принадлежащих B:

$$\hat{P}(X_1 \in A | X_1 \in B) = \frac{1}{m}\sum_{i=1}^{n} I(X_i \in A \cap B) = \frac{\hat{P}(X_1 \in A \cap B)}{\hat{P}(X_1 \in B)} \text{ , где } m = \sum_{i=1}^{n} I(X_i \in B)$$

- Имеется выборка $X_1, ..., X_n$ из распределения $\Theta(\theta)$.
- Необходимо оценить условную вероятность $P(X_1 \in A|X_1 \in B)$, где $A, B \subset R$ и $P(X_1 \in B) > 0$.
- Несмещенная и состоятельная оценка данной вероятности может быть получена как доля наблюдений в выборке, принадлежащих множеству $A \cap B$, среди наблюдений, принадлежащих B:

$$\hat{P}(X_1 \in A|X_1 \in B) = \frac{1}{m}\sum_{i=1}^{n} I(X_i \in A \cap B) = \frac{\hat{P}(X_1 \in A \cap B)}{\hat{P}(X_1 \in B)}\text{ , где } m = \sum_{i=1}^{n} I(X_i \in B)$$

- Подставляя вместо наблюдений X_i их реализации x_i можно получить реализацию оценки соответствующей условной вероятности.

- Имеется выборка $X_1, ..., X_n$ из распределения $\Theta(\theta)$.
- Необходимо оценить условную вероятность $P(X_1 \in A|X_1 \in B)$, где $A, B \subset R$ и $P(X_1 \in B) > 0$.
- Несмещенная и состоятельная оценка данной вероятности может быть получена как доля наблюдений в выборке, принадлежащих множеству $A \cap B$, среди наблюдений, принадлежащих B:

$$\hat{P}(X_1 \in A|X_1 \in B) = \frac{1}{m}\sum_{i=1}^{n} I(X_i \in A \cap B) = \frac{\hat{P}(X_1 \in A \cap B)}{\hat{P}(X_1 \in B)}\text{ , где } m = \sum_{i=1}^{n} I(X_i \in B)$$

- Подставляя вместо наблюдений X_i их реализации x_i можно получить реализацию оценки соответствующей условной вероятности.
  Доказательство: докажем состоятельность, используя теорему Слуцкого:

$$\begin{cases}\hat{P}(X_1 \in A \cap B) \xrightarrow{p} P(X_1 \in A \cap B)\\ \hat{P}(X_1 \in B) \xrightarrow{p} P(X_1 \in B)\end{cases} \implies \hat{P}(X_1 \in A|X_1 \in B) \xrightarrow{p} \frac{P(X_1 \in A \cap B)}{P(X_1 \in B)} = P(X_1 \in A|X_1 \in B)$$

- Имеется выборка $X_1, ..., X_n$ из распределения $\Theta(\theta)$.
- Необходимо оценить условную вероятность $P(X_1 \in A|X_1 \in B)$, где $A, B \subset R$ и $P(X_1 \in B) > 0$.
- Несмещенная и состоятельная оценка данной вероятности может быть получена как доля наблюдений в выборке, принадлежащих множеству $A \cap B$, среди наблюдений, принадлежащих B:

$$\hat{P}(X_1 \in A|X_1 \in B) = \frac{1}{m}\sum_{i=1}^{n} I(X_i \in A \cap B) = \frac{\hat{P}(X_1 \in A \cap B)}{\hat{P}(X_1 \in B)}\text{, где } m = \sum_{i=1}^{n} I(X_i \in B)$$

- Подставляя вместо наблюдений X_i их реализации x_i можно получить реализацию оценки соответствующей условной вероятности.
  Доказательство: докажем состоятельность, используя теорему Слуцкого:

$$\begin{cases}\hat{P}(X_1 \in A \cap B) \xrightarrow{p} P(X_1 \in A \cap B)\\ \hat{P}(X_1 \in B) \xrightarrow{p} P(X_1 \in B)\end{cases} \implies \hat{P}(X_1 \in A|X_1 \in B) \xrightarrow{p} \frac{P(X_1 \in A \cap B)}{P(X_1 \in B)} = P(X_1 \in A|X_1 \in B)$$

Пример: имеется реализация выборки $x = (5, 3, 1, -2, 3)$, найдем реализацию оценки $P(X_1 \leq 3|X_1 > 0)$.

$$m = I(X_1 > 0) + ... + I(X_5 > 0) = 1 + 1 + 1 + 0 + 1 = 4$$

- Имеется выборка $X_1, ..., X_n$ из распределения $\Theta(\theta)$.
- Необходимо оценить условную вероятность $P(X_1 \in A|X_1 \in B)$, где $A, B \subset R$ и $P(X_1 \in B) > 0$.
- Несмещенная и состоятельная оценка данной вероятности может быть получена как доля наблюдений в выборке, принадлежащих множеству $A \cap B$, среди наблюдений, принадлежащих B:

$$\hat{P}(X_1 \in A|X_1 \in B) = \frac{1}{m}\sum_{i=1}^{n} I(X_i \in A \cap B) = \frac{\hat{P}(X_1 \in A \cap B)}{\hat{P}(X_1 \in B)}\text{, где } m = \sum_{i=1}^{n} I(X_i \in B)$$

- Подставляя вместо наблюдений X_i их реализации x_i можно получить реализацию оценки соответствующей условной вероятности.

  Доказательство: докажем состоятельность, используя теорему Слуцкого:

$$\begin{cases}\hat{P}(X_1 \in A \cap B) \xrightarrow{p} P(X_1 \in A \cap B)\\ \hat{P}(X_1 \in B) \xrightarrow{p} P(X_1 \in B)\end{cases} \implies \hat{P}(X_1 \in A|X_1 \in B) \xrightarrow{p} \frac{P(X_1 \in A \cap B)}{P(X_1 \in B)} = P(X_1 \in A|X_1 \in B)$$

Пример: имеется реализация выборки $x = (5, 3, 1, -2, 3)$, найдем реализацию оценки $P(X_1 \leq 3|X_1 > 0)$.

$$m = I(X_1 > 0) + ... + I(X_5 > 0) = 1 + 1 + 1 + 0 + 1 = 4$$

$$\hat{P}(X_1 \leq 3|X_1 > 0)(x) = \frac{1}{4}\left(I(0 < x_1 \leq 3) + ... + I(0 < x_5 \leq 3)\right) = \frac{1}{4}(0 + 1 + 1 + 0 + 1) = \frac{3}{4}$$

- **Выборочная (эмпирическая) функция распределения** определяется как:

$$\hat{F}_n(t) = \hat{P}(X_1 \leq t) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq t) \sim \frac{1}{n}B(n, P(X_1 \leq t))$$

- **Выборочная (эмпирическая) функция распределения** определяется как:

$$\hat{F}_n(t) = \hat{P}(X_1 \leq t) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq t) \sim \frac{1}{n}B(n, P(X_1 \leq t))$$

- Выборочная функция распределения в точке t является несмещенной и состоятельной оценкой функции распределения в этой же точке, то есть $\hat{F}_n(t) \xrightarrow{p} F_{X_1}(t), \forall t \in R$.

- **Выборочная (эмпирическая) функция распределения** определяется как:

$$\hat{F}_n(t) = \hat{P}(X_1 \leq t) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq t) \sim \frac{1}{n}B(n, P(X_1 \leq t))$$

- Выборочная функция распределения в точке t является несмещенной и состоятельной оценкой функции распределения в этой же точке, то есть $\hat{F}_n(t) \xrightarrow{p} F_{X_1}(t), \forall t \in R$.
  Пример: Найдем реализацию выборочной функции распределения для выборки с реализацией $x = (5, 3, 1, -2, 3)$.

- **Выборочная (эмпирическая) функция распределения** определяется как:

$$\hat{F}_n(t) = \hat{P}(X_1 \leq t) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq t) \sim \frac{1}{n}B(n, P(X_1 \leq t))$$

- Выборочная функция распределения в точке t является несмещенной и состоятельной оценкой функции распределения в этой же точке, то есть $\hat{F}_n(t) \xrightarrow{p} F_{X_1}(t), \forall t \in R$.
  Пример: Найдем реализацию выборочной функции распределения для выборки с реализацией $x = (5, 3, 1, -2, 3)$.

$$\hat{F}_n(t)|(X = x) = \left\{ \begin{array}{l} \\ \\ \\ \end{array} \right.$$

- **Выборочная (эмпирическая) функция распределения** определяется как:

$$\hat{F}_n(t) = \hat{P}(X_1 \leq t) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq t) \sim \frac{1}{n}B(n, P(X_1 \leq t))$$

- Выборочная функция распределения в точке t является несмещенной и состоятельной оценкой функции распределения в этой же точке, то есть $\hat{F}_n(t) \xrightarrow{p} F_{X_1}(t), \forall t \in R$.
Пример: Найдем реализацию выборочной функции распределения для выборки с реализацией $x = (5, 3, 1, -2, 3)$.

$$\hat{F}_n(t)|(X = x) = \begin{cases} (0+0+0+0+0)/5 = 0, \text{ если } t < -2 \\ \\ \\ \\ \end{cases}$$

- **Выборочная (эмпирическая) функция распределения** определяется как:

$$\hat{F}_n(t) = \hat{P}(X_1 \leq t) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq t) \sim \frac{1}{n}B(n, P(X_1 \leq t))$$

- Выборочная функция распределения в точке t является несмещенной и состоятельной оценкой функции распределения в этой же точке, то есть $\hat{F}_n(t) \xrightarrow{p} F_{X_1}(t), \forall t \in R$.
  Пример: Найдем реализацию выборочной функции распределения для выборки с реализацией $x = (5, 3, 1, -2, 3)$.

$$\hat{F}_n(t)|(X = x) = \begin{cases} (0+0+0+0+0)/5 = 0, \text{ если } t < -2 \\ (0+0+0+1+0)/5 = 0.2, \text{ если } -2 \leq t < 1 \\ \\ \\ \end{cases}$$

- **Выборочная (эмпирическая) функция распределения** определяется как:

$$\hat{F}_n(t) = \hat{P}(X_1 \leq t) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq t) \sim \frac{1}{n}B(n, P(X_1 \leq t))$$

- Выборочная функция распределения в точке t является несмещенной и состоятельной оценкой функции распределения в этой же точке, то есть $\hat{F}_n(t) \xrightarrow{p} F_{X_1}(t), \forall t \in R$.
Пример: Найдем реализацию выборочной функции распределения для выборки с реализацией $x = (5, 3, 1, -2, 3)$.

$$\hat{F}_n(t)|(X = x) = \begin{cases} (0+0+0+0+0)/5 = 0, \text{ если } t < -2 \\ (0+0+0+1+0)/5 = 0.2, \text{ если } -2 \leq t < 1 \\ (0+0+1+1+0)/5 = 0.4, \text{ если } 1 \leq t < 3 \\ \\ \\ \end{cases}$$

- **Выборочная (эмпирическая) функция распределения** определяется как:

$$\hat{F}_n(t) = \hat{P}(X_1 \leq t) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq t) \sim \frac{1}{n}B(n, P(X_1 \leq t))$$

- Выборочная функция распределения в точке t является несмещенной и состоятельной оценкой функции распределения в этой же точке, то есть $\hat{F}_n(t) \xrightarrow{p} F_{X_1}(t), \forall t \in R$.
Пример: Найдем реализацию выборочной функции распределения для выборки с реализацией $x = (5, 3, 1, -2, 3)$.

$$\hat{F}_n(t)|(X = x) = \begin{cases} (0+0+0+0+0)/5 = 0, \text{ если } t < -2 \\ (0+0+0+1+0)/5 = 0.2, \text{ если } -2 \leq t < 1 \\ (0+0+1+1+0)/5 = 0.4, \text{ если } 1 \leq t < 3 \\ (0+1+1+1+1)/5 = 0.8, \text{ если } 3 \leq t < 5 \\ \end{cases}$$

- **Выборочная (эмпирическая) функция распределения** определяется как:

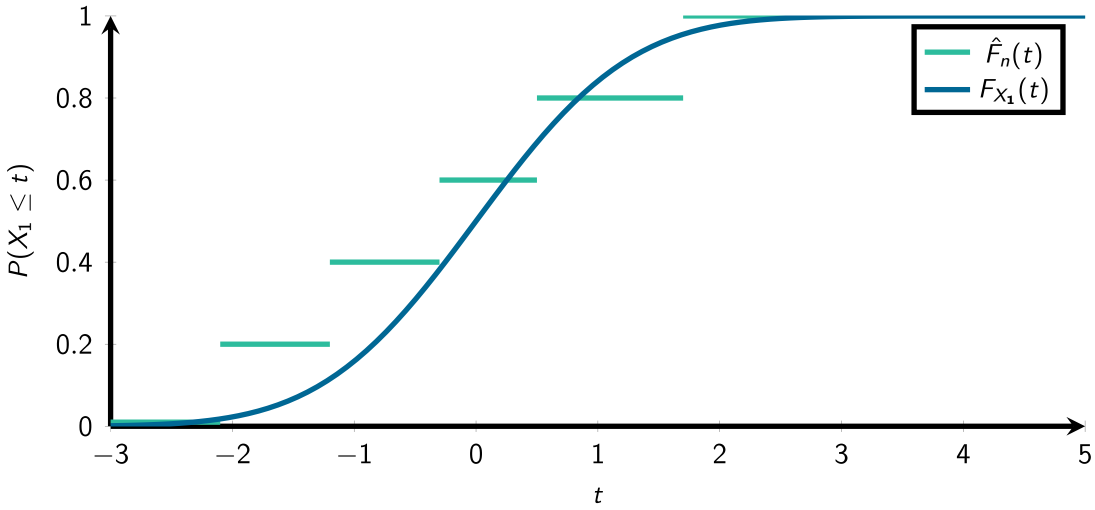
$$\hat{F}_n(t) = \hat{P}(X_1 \leq t) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq t) \sim \frac{1}{n}B(n, P(X_1 \leq t))$$

- Выборочная функция распределения в точке t является несмещенной и состоятельной оценкой функции распределения в этой же точке, то есть $\hat{F}_n(t) \xrightarrow{p} F_{X_1}(t), \forall t \in R$.
  Пример: Найдем реализацию выборочной функции распределения для выборки с реализацией $x = (5, 3, 1, -2, 3)$.

$$\hat{F}_n(t)|(X = x) = \begin{cases} (0+0+0+0+0)/5 = 0, \text{ если } t < -2 \\ (0+0+0+1+0)/5 = 0.2, \text{ если } -2 \leq t < 1 \\ (0+0+1+1+0)/5 = 0.4, \text{ если } 1 \leq t < 3 \\ (0+1+1+1+1)/5 = 0.8, \text{ если } 3 \leq t < 5 \\ (1+1+1+1+1)/5 = 1, \text{ если } t \geq 5 \end{cases}$$

Оценивание вероятностей

График эмпирической функции распределения на фоне теоретической функции распределения



- Погрешность при использовании выборочной функции распределения вместо истинной можно записать как $|\hat{F}_n(t) - F_{X_1}(t)|$.

Оценивание вероятностей

Верхняя граница погрешности

- Погрешность при использовании выборочной функции распределения вместо истинной можно записать как $|\hat{F}_n(t) - F_{X_1}(t)|$.
- При помощи неравенства Чебышева можно найти верхнюю границу для вероятности того, что соответствующая погрешность превысит определенное значение:

$$P(|\hat{F}_n(t) - F_{X_1}(t)| > \varepsilon) \leq \frac{F_{X_1}(t)(1 - F_{X_1}(t))}{n\varepsilon^2} \leq \frac{1}{4n\varepsilon^2}$$

- Погрешность при использовании выборочной функции распределения вместо истинной можно записать как $|\hat{F}_n(t) - F_{X_1}(t)|$.
- При помощи неравенства Чебышева можно найти верхнюю границу для вероятности того, что соответствующая погрешность превысит определенное значение:

$$P(|\hat{F}_n(t) - F_{X_1}(t)| > \varepsilon) \leq \frac{F_{X_1}(t)(1 - F_{X_1}(t))}{n\varepsilon^2} \leq \frac{1}{4n\varepsilon^2}$$

- Чем больше объем выборки n, тем меньше максимально возможная вероятность получить погрешность, превышающую ε.

Оценивание вероятностей

Верхняя граница погрешности

- Погрешность при использовании выборочной функции распределения вместо истинной можно записать как $|\hat{F}_n(t) - F_{X_1}(t)|$.
- При помощи неравенства Чебышева можно найти верхнюю границу для вероятности того, что соответствующая погрешность превысит определенное значение:

$$P(|\hat{F}_n(t) - F_{X_1}(t)| > \varepsilon) \leq \frac{F_{X_1}(t)(1 - F_{X_1}(t))}{n\varepsilon^2} \leq \frac{1}{4n\varepsilon^2}$$

- Чем больше объем выборки n, тем меньше максимально возможная вероятность получить погрешность, превышающую ε.
- Например, если объем выборки составляет $n = 1000000$ миллион наблюдений, то вероятность того, что погрешность превысит $\varepsilon = 0.01$ окажется меньше, чем:

$$\frac{1}{4 \times 1000000 \times 0.01^2} = 0.0025$$

- Выборочное среднее $\overline{X}_n$ является несмещенной и состоятельной оценкой математического ожидания $E(X_1)$.

- Выборочное среднее $\overline{X}_n$ является несмещенной и состоятельной оценкой математического ожидания $E(X_1)$.
  Доказательство:

$$E(\overline{X}_n) = E(X_1)$$

- Выборочное среднее $\overline{X}_n$ является несмещенной и состоятельной оценкой математического ожидания $E(X_1)$.
Доказательство:

$$E(\overline{X}_n) = E(X_1)$$

$$\lim_{n\to\infty} Var(\overline{X}_n) = \lim_{n\to\infty} Var(X_1)/n = 0$$

- Выборочное среднее $\overline{X}_n$ является несмещенной и состоятельной оценкой математического ожидания $E(X_1)$.
  Доказательство:
  $$E(\overline{X}_n) = E(X_1)$$
  $$\lim_{n\to\infty} Var(\overline{X}_n) = \lim_{n\to\infty} Var(X_1)/n = 0$$
- По аналогии несмещенные и состоятельные оценки начальных моментов $E(X_1^k)$, где $k \in N$, могут быть получены с использованием выборочных начальных моментов соответствующего порядка:
  $$\hat{m}_k = \overline{X^k}_n = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

- Выборочное среднее $\overline{X}_n$ является несмещенной и состоятельной оценкой математического ожидания $E(X_1)$.
Доказательство:

$$E(\overline{X}_n) = E(X_1)$$

$$\lim_{n\to\infty} Var(\overline{X}_n) = \lim_{n\to\infty} Var(X_1)/n = 0$$

- По аналогии несмещенные и состоятельные оценки начальных моментов $E(X_1^k)$, где $k \in N$, могут быть получены с использованием выборочных начальных моментов соответствующего порядка:

$$\hat{m}_k = \overline{X^k}_n = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

Пример: рассчитаем реализацию третьего начального выборочного момента по выборке с реализациями $x = (5, 2, 5)$:

$$\hat{m}_k(x) = \frac{1}{3}\left(5^3 + 2^3 + 5^3\right) = 86$$

- **Выборочная дисперсия** рассчитывается как:

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

- **Выборочная дисперсия** рассчитывается как:

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

- Она является смещенной, но состоятельной оценкой дисперсии $Var(X_1)$.

- **Выборочная дисперсия** рассчитывается как:

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

- Она является смещенной, но состоятельной оценкой дисперсии $Var(X_1)$.
  Доказательство: покажем, что оценка является смещенной:

- **Выборочная дисперсия** рассчитывается как:

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

- Она является смещенной, но состоятельной оценкой дисперсии $Var(X_1)$.
  Доказательство: покажем, что оценка является смещенной:

$$E(S_n^2) = \frac{1}{n}\sum_{i=1}^{n}E((X_i - \overline{X}_n)^2) = \frac{1}{n} \times nE((X_1 - \overline{X}_n)^2) = E((X_1 - \overline{X}_n)^2) =$$

Выборочные характеристики

Выборочная дисперсия

- **Выборочная дисперсия** рассчитывается как:

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

- Она является смещенной, но состоятельной оценкой дисперсии $Var(X_1)$.
  Доказательство: покажем, что оценка является смещенной:

$$E(S_n^2) = \frac{1}{n}\sum_{i=1}^{n}E((X_i - \overline{X}_n)^2) = \frac{1}{n} \times nE((X_1 - \overline{X}_n)^2) = E((X_1 - \overline{X}_n)^2) =$$

$$= Var(X_1 - \overline{X}_n) + E(X_1 - \overline{X}_n)^2 = Var(X_1) + Var(\overline{X}_n) - 2Cov(X_1, \overline{X}_n) + 0^2 =$$

- **Выборочная дисперсия** рассчитывается как:

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

- Она является смещенной, но состоятельной оценкой дисперсии $Var(X_1)$.
  Доказательство: покажем, что оценка является смещенной:

$$E(S_n^2) = \frac{1}{n}\sum_{i=1}^{n}E((X_i - \overline{X}_n)^2) = \frac{1}{n} \times nE((X_1 - \overline{X}_n)^2) = E((X_1 - \overline{X}_n)^2) =$$

$$= Var(X_1 - \overline{X}_n) + E(X_1 - \overline{X}_n)^2 = Var(X_1) + Var(\overline{X}_n) - 2Cov(X_1, \overline{X}_n) + 0^2 =$$

$$= Var(X_1) + \frac{Var(X_1)}{n} - 2Cov\left(X_1, \frac{1}{n}X_1\right) = Var(X_1) \times \left(1 + \frac{1}{n} - 2\frac{1}{n}\right) = \frac{n-1}{n}Var(X_1)$$

- **Выборочная дисперсия** рассчитывается как:

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

- Она является смещенной, но состоятельной оценкой дисперсии $Var(X_1)$.
  Доказательство: покажем, что оценка является смещенной:

$$E(S_n^2) = \frac{1}{n}\sum_{i=1}^{n}E((X_i - \overline{X}_n)^2) = \frac{1}{n} \times nE((X_1 - \overline{X}_n)^2) = E((X_1 - \overline{X}_n)^2) =$$

$$= Var(X_1 - \overline{X}_n) + E(X_1 - \overline{X}_n)^2 = Var(X_1) + Var(\overline{X}_n) - 2Cov(X_1, \overline{X}_n) + 0^2 =$$

$$= Var(X_1) + \frac{Var(X_1)}{n} - 2Cov\left(X_1, \frac{1}{n}X_1\right) = Var(X_1) \times \left(1 + \frac{1}{n} - 2\frac{1}{n}\right) = \frac{n-1}{n}Var(X_1)$$

- Состоятельная и несмещенная оценка дисперсии может быть получена с использование **исправленной (скорректированной) выборочной дисперсии**:

$$\hat{\sigma}^2 = \frac{n}{n-1}S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

- Обозначим через $X_{(i)}$ наблюдение, которое является i-м по величине в выборке. Оно называется i-**й порядковой статистикой**. Статистики $X_{(1)} = \min(X_1, ..., X_n)$ и $X_{(n)} = \max(X_1, ..., X_n)$ именуются **минимальной и максимальной порядковыми статистиками** соответственно.

Выборочные характеристики

Вариационный ряд

- Обозначим через $X_{(i)}$ наблюдение, которое является i-м по величине в выборке. Оно называется i-**й порядковой статистикой**. Статистики $X_{(1)} = \min(X_1, ..., X_n)$ и $X_{(n)} = \max(X_1, ..., X_n)$ именуются **минимальной и максимальной порядковыми статистиками** соответственно.
- Последовательность $X_{(1)}, ..., X_{(n)}$ **именуется вариационным рядом**.

- Обозначим через $X_{(i)}$ наблюдение, которое является i-м по величине в выборке. Оно называется **i-й порядковой статистикой**. Статистики $X_{(1)} = \min(X_1, ..., X_n)$ и $X_{(n)} = \max(X_1, ..., X_n)$ именуются **минимальной и максимальной порядковыми статистиками** соответственно.
- Последовательность $X_{(1)}, ..., X_{(n)}$ **именуется вариационным рядом**.
- Распределение максимальной экстремальной статистики определяется как:

$$F_{X_{(n)}}(x) = F_{\max(X_1,...,X_n)}(x) = P(\max(X_1, ..., X_n) \leq x) = P(X_1 \leq x, ..., X_n \leq x) =$$

- Обозначим через $X_{(i)}$ наблюдение, которое является i-м по величине в выборке. Оно называется **i-й порядковой статистикой**. Статистики $X_{(1)} = \min(X_1, ..., X_n)$ и $X_{(n)} = \max(X_1, ..., X_n)$ именуются **минимальной и максимальной порядковыми статистиками** соответственно.
- Последовательность $X_{(1)}, ..., X_{(n)}$ **именуется вариационным рядом**.
- Распределение максимальной экстремальной статистики определяется как:

$$F_{X_{(n)}}(x) = F_{\max(X_1,...,X_n)}(x) = P(\max(X_1, ..., X_n) \leq x) = P(X_1 \leq x, ..., X_n \leq x) =$$
$$= P(X_1 \leq x) \times ... \times P(X_n \leq x) = F_{X_1}(x) \times ... \times F_{X_n}(x) = \left(F_{X_1}(x)\right)^n$$

- Обозначим через $X_{(i)}$ наблюдение, которое является i-м по величине в выборке. Оно называется **i-й порядковой статистикой**. Статистики $X_{(1)} = \min(X_1, ..., X_n)$ и $X_{(n)} = \max(X_1, ..., X_n)$ именуются **минимальной и максимальной порядковыми статистиками** соответственно.
- Последовательность $X_{(1)}, ..., X_{(n)}$ **именуется вариационным рядом**.
- Распределение максимальной экстремальной статистики определяется как:

$$F_{X_{(n)}}(x) = F_{\max(X_1,...,X_n)}(x) = P(\max(X_1, ..., X_n) \leq x) = P(X_1 \leq x, ..., X_n \leq x) =$$
$$= P(X_1 \leq x) \times ... \times P(X_n \leq x) = F_{X_1}(x) \times ... \times F_{X_n}(x) = \left(F_{X_1}(x)\right)^n$$

- Распределение минимальной экстремальной статистики определяется как:

$$F_{X_{(1)}}(x) = F_{\min(X_1,...,X_n)}(x) = P(\min(X_1, ..., X_n) \leq x) = 1 - P(\min(X_1, ..., X_n) > x) =$$

- Обозначим через $X_{(i)}$ наблюдение, которое является i-м по величине в выборке. Оно называется i-**й порядковой статистикой**. Статистики $X_{(1)} = \min(X_1, ..., X_n)$ и $X_{(n)} = \max(X_1, ..., X_n)$ именуются **минимальной и максимальной порядковыми статистиками** соответственно.
- Последовательность $X_{(1)}, ..., X_{(n)}$ **именуется вариационным рядом**.
- Распределение максимальной экстремальной статистики определяется как:

$$F_{X_{(n)}}(x) = F_{\max(X_1,...,X_n)}(x) = P(\max(X_1, ..., X_n) \leq x) = P(X_1 \leq x, ..., X_n \leq x) =$$
$$= P(X_1 \leq x) \times ... \times P(X_n \leq x) = F_{X_1}(x) \times ... \times F_{X_n}(x) = \left(F_{X_1}(x)\right)^n$$

- Распределение минимальной экстремальной статистики определяется как:

$$F_{X_{(1)}}(x) = F_{\min(X_1,...,X_n)}(x) = P(\min(X_1, ..., X_n) \leq x) = 1 - P(\min(X_1, ..., X_n) > x) =$$
$$= 1 - P(X_1 > x) \times ... \times P(X_n > x) = 1 - \left(1 - F_{X_1}(x)\right) \times ... \times (1 - F_{X_n}(x)) = 1 - \left(1 - F_{X_1}(x)\right)^n$$

Выборочные характеристики

Вариационный ряд

- Обозначим через $X_{(i)}$ наблюдение, которое является i-м по величине в выборке. Оно называется **i-й порядковой статистикой**. Статистики $X_{(1)} = \min(X_1, ..., X_n)$ и $X_{(n)} = \max(X_1, ..., X_n)$ именуются **минимальной и максимальной порядковыми статистиками** соответственно.
- Последовательность $X_{(1)}, ..., X_{(n)}$ **именуется вариационным рядом**.
- Распределение максимальной экстремальной статистики определяется как:

$$F_{X_{(n)}}(x) = F_{\max(X_1,...,X_n)}(x) = P(\max(X_1, ..., X_n) \leq x) = P(X_1 \leq x, ..., X_n \leq x) =$$
$$= P(X_1 \leq x) \times ... \times P(X_n \leq x) = F_{X_1}(x) \times ... \times F_{X_n}(x) = \left(F_{X_1}(x)\right)^n$$

- Распределение минимальной экстремальной статистики определяется как:

$$F_{X_{(1)}}(x) = F_{\min(X_1,...,X_n)}(x) = P(\min(X_1, ..., X_n) \leq x) = 1 - P(\min(X_1, ..., X_n) > x) =$$
$$= 1 - P(X_1 > x) \times ... \times P(X_n > x) = 1 - \left(1 - F_{X_1}(x)\right) \times ... \times (1 - F_{X_n}(x)) = 1 - \left(1 - F_{X_1}(x)\right)^n$$

- Распределение i-й порядковой статистики определяется как:

$$F_{X_{(i)}}(x) = \sum_{k=i}^{n} C_n^k \left(1 - F_{X_1}(x)\right)^{n-k} \left(F_{X_1}(x)\right)^k$$

Выборочные характеристики
Вариационный ряд

- Обозначим через $X_{(i)}$ наблюдение, которое является i-м по величине в выборке. Оно называется **i-й порядковой статистикой**. Статистики $X_{(1)} = \min(X_1, ..., X_n)$ и $X_{(n)} = \max(X_1, ..., X_n)$ именуются **минимальной и максимальной порядковыми статистиками** соответственно.
- Последовательность $X_{(1)}, ..., X_{(n)}$ **именуется вариационным рядом**.
- Распределение максимальной экстремальной статистики определяется как:

$$F_{X_{(n)}}(x) = F_{\max(X_1,...,X_n)}(x) = P(\max(X_1, ..., X_n) \leq x) = P(X_1 \leq x, ..., X_n \leq x) =$$
$$= P(X_1 \leq x) \times ... \times P(X_n \leq x) = F_{X_1}(x) \times ... \times F_{X_n}(x) = \left(F_{X_1}(x)\right)^n$$

- Распределение минимальной экстремальной статистики определяется как:

$$F_{X_{(1)}}(x) = F_{\min(X_1,...,X_n)}(x) = P(\min(X_1, ..., X_n) \leq x) = 1 - P(\min(X_1, ..., X_n) > x) =$$
$$= 1 - P(X_1 > x) \times ... \times P(X_n > x) = 1 - \left(1 - F_{X_1}(x)\right) \times ... \times \left(1 - F_{X_n}(x)\right) = 1 - \left(1 - F_{X_1}(x)\right)^n$$

- Распределение i-й порядковой статистики определяется как:

$$F_{X_{(i)}}(x) = \sum_{k=i}^{n} C_n^k \left(1 - F_{X_1}(x)\right)^{n-k} \left(F_{X_1}(x)\right)^k$$

Пример: найдем вероятность того, что в выборке из $n = 5$ экспоненциальных случайных величин с параметром $\lambda = 0.2$ наибольшее значение не превысит 10:

$$F_{X_{(5)}}(10) = (1 - e^{-0.2\times10})^5 \approx 0.483$$

- Гистограмма является оценкой функции плотности.

Оценивание функции плотности

Гистограмма

- Гистограмма является оценкой функции плотности.
- Разобьем выборку на m интервалов равной длины h, где интервал обозначим как b_k.

Оценивание функции плотности
Гистограмма

- Гистограмма является оценкой функции плотности.
- Разобьем выборку на m интервалов равной длины h, где интервал обозначим как b_k.
- Гистограмма предполагает следующую (как правило смещенную) оценку функции плотности в точке $t \in b_k$:

$$\hat{f}_{X_i}(t) = \frac{1}{nh}\sum_{i=1}^{n} I(X_i \in b_k)$$

Оценивание функции плотности

Гистограмма

- Гистограмма является оценкой функции плотности.
- Разобьем выборку на m интервалов равной длины h, где интервал обозначим как b_k.
- Гистограмма предполагает следующую (как правило смещенную) оценку функции плотности в точке $t \in b_k$:

$$\hat{f}_{X_i}(t) = \frac{1}{nh}\sum_{i=1}^{n} I(X_i \in b_k)$$

- Эффективность оценки зависит от параметра h. Существуют различные подходы к подбору оптимального значения данного параметра, например, по правилу Райса (Rice rule) $h = 2n^{1.5}$.

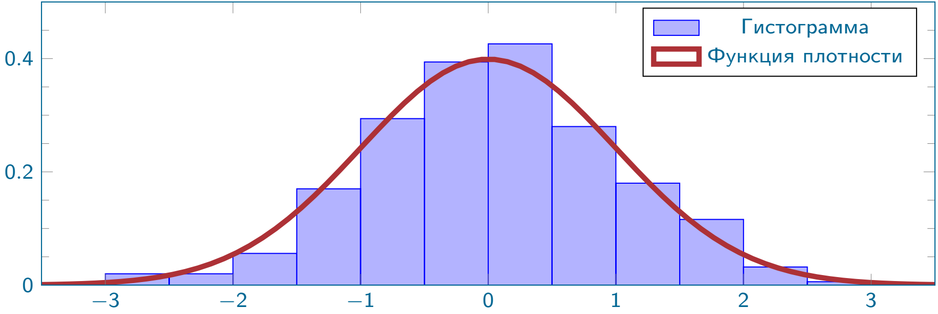
Оценивание функции плотности

Гистограмма

- Гистограмма является оценкой функции плотности.
- Разобьем выборку на m интервалов равной длины h, где интервал обозначим как b_k.
- Гистограмма предполагает следующую (как правило смещенную) оценку функции плотности в точке $t \in b_k$:

$$\hat{f}_{X_i}(t) = \frac{1}{nh}\sum_{i=1}^{n} I(X_i \in b_k)$$

- Эффективность оценки зависит от параметра h. Существуют различные подходы к подбору оптимального значения данного параметра, например, по правилу Райса (Rice rule) $h = 2n^{1.5}$.



- Функция плотности при помощи ядер оценивается следующим образом:

$$\hat{f}_{X_1}(t) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{t - x_i}{h}\right)$$

Оценивание функции плотности

Ядерное оценивание

- Функция плотности при помощи ядер оценивается следующим образом:

$$\hat{f}_{X_1}(t) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{t - x_i}{h}\right)$$

- Функция K является неотрицательной и именуется **ядром**. В качестве нее часто используют функцию плотности стандартного нормального распределения $K = \phi(t)$ (гауссовское ядро).

Оценивание функции плотности

Ядерное оценивание

- Функция плотности при помощи ядер оценивается следующим образом:

$$\hat{f}_{X_1}(t) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{t - x_i}{h}\right)$$

- Функция K является неотрицательной и именуется **ядром**. В качестве нее часто используют функцию плотности стандартного нормального распределения $K = \phi(t)$ (гауссовское ядро).
- Параметр h именуется **шириной окна** и его подбирают из соображения максимизации эффективности.

- Функция плотности при помощи ядер оценивается следующим образом:

$$\hat{f}_{X_1}(t) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{t - x_i}{h}\right)$$

- Функция K является неотрицательной и именуется **ядром**. В качестве нее часто используют функцию плотности стандартного нормального распределения $K = \phi(t)$ (гауссовское ядро).
- Параметр h именуется **шириной окна** и его подбирают из соображения максимизации эффективности.