# Анализ кривых дожития в условиях цензурирования с приложением в нейрохирургии

Матвеева Юлия Алексеевна, гр. 522

Санкт-Петербургский государственный университет
Математико-механический факультет
Кафедра статистического моделирования

Научный руководитель: к.ф.-м.н., доц. Алексеева Н.П.
Рецензент: к.ф.-м.н. Коробейников А.И.

Санкт-Петербург
2011г.

- $\tilde{Z}$, $\tilde{n}$ — независимые случайные величины.
- Носитель $\tilde{n}$ содержится в множестве натуральных чисел.
- $\{n_k\}_{k=1}^m$ — реализации повторных независимых копий $\{\tilde{n}_k\}_{k=1}^m$ величины $\tilde{n}$.
- $\{Z_j\}$ — реализации повторных независимых копий $\{\tilde{Z}_j\}$ величины $\tilde{Z}$.

Выборка $\{Z_j\}$ разбита на группы в соответствии с реализациями $\{n_k\}_{k=1}^m$:

$$\overbrace{Z_1, \ldots, Z_{n_1}}, \quad \overbrace{Z_{n_1+1}, \ldots, Z_{n_1+n_2}}, \quad \overbrace{Z_{n_1+n_2+1}, \ldots, Z_{n_1+n_2+n_3}},$$

$$\ldots, \overbrace{Z_{n_1+\cdots+n_{m-1}+1}, \ldots, Z_{n_1+\cdots+n_m}},$$

- $\tilde{\mathcal{M}}(n) = \sup\{x : \hat{F}_n(x) \leqslant \frac{1}{2}\}$ — выборочная медиана, построенная по выборке объема $n$ из величины $\tilde{Z}$.
- $\mathcal{M}_k$ — реализация выборочной медианы $k$-й группы.

**Медианное цензурирование:**

Величины $\{Z_j\}$ не наблюдаются, а наблюдаются лишь наборы

$$(\mathcal{M}_k, n_k)\,.$$

**Медианное цензурирование:**

Величины $\{Z_j\}$ не наблюдаются, а наблюдаются лишь наборы

$$(\mathcal{M}_k, n_k)\,.$$

**Задача:**

Построение и анализ (непараметрической) оценки функции распределения

$$F_0(x) = \mathsf{P}(\tilde{Z} \leqslant x)$$

в условиях медианного цензурирования.

**Метод:**

Метод максимального правдоподобия.

- **Правое цензурирование:** Наблюдается величина $\tilde{Y} = \left[\tilde{C} = \min\{\tilde{Z}, \tilde{T}\}, \tilde{\delta} = 1_{\{\tilde{Z} \leqslant \tilde{T}\}}\right]$, где $\tilde{Z}$ и $\tilde{T}$ независимы.
- **Интервальное цензурирование типа $K$:**

$$K \geq 1, \qquad \tilde{T} = \{0 =: \tilde{T}_0 < \tilde{T}_1 < \cdots < \tilde{T}_K < \tilde{T}_{K+1} := \infty\},$$
$$\tilde{Z} \quad \text{и} \quad \tilde{T} \quad \text{независимы},$$
$$\tilde{\delta}_k = 1_{\{\tilde{Z} \in (\tilde{T}_{k-1}, \tilde{T}_k)\}}, k = 1, \ldots, K+1.$$

  Наблюдается $(\tilde{T}, \tilde{\delta})$.
- **Интервальное цензурирование смешанного типа:** Определяется как смесь (по $K$) интервальных цензурирований типа $K$. $\tilde{K} \geq 1$ — целочисленная случайная величина.

$$\tilde{T} = \{0 =: \tilde{T}_{(0,K)} < \tilde{T}_{(1,K)} < \cdots < \tilde{T}_{(K,K)} < \tilde{T}_{(K+1,K)} := \infty\}, K = 1, 2, 3, \ldots,$$
$$\tilde{Z} \quad \text{и} \quad (\tilde{T}, \tilde{K}) \quad \text{независимы},$$
$$\tilde{\delta}_{(k,K)} = 1_{\{\tilde{Z} \in (\tilde{T}_{(k-1,K)}, \tilde{T}_{(k,K)})\}}, k = 1, \ldots, K+1, K = 1, 2, 3, \ldots.$$

  Наблюдается $(\tilde{K}, \tilde{\delta}_{(\cdot,\tilde{K})}, \tilde{T}_{(\cdot,\tilde{K})})$.

J. Sun. *The Statistical Analysis of Interval-censored Failure Time Data*, Springer, 2006.

Задача непараметрического оценивания функции распределения, вообще говоря, бесконечномерна.

**Структура работы:**

1. Сведение задачи к конечномерной и анализ ее свойств (существование решения, выпуклость).
2. Явное аналитическое решение задачи в частном случае и анализ асимптотических свойств соответствующей оценки.
3. Анализ асимптотических свойств оценки в общем случае (состоятельность, несмещенность, асимптотическая нормальность).
4. Аппробация разработанного метода на реальных данных.

## Лемма

*Введем порядковые статистики* $\tilde{Z}_{[1]} \leqslant \tilde{Z}_{[2]} \leqslant \ldots \leqslant \tilde{Z}_{[n]}$. *Тогда для выборочной медианы, соответствующей выборке* $\{\tilde{Z}_j\}_{j=1}^n$, *верно следующее:*

$$\tilde{\mathcal{M}}(n) = \tilde{Z}_{[l]}, \quad l = \left\lceil \frac{n+1}{2} \right\rceil.$$

## Лемма

*Функция распределения* $G_n(x)$ *выборочной медианы* $\tilde{\mathcal{M}}(n)$ *имеет вид:*

$$G_n(x) = \mathsf{P}\left(\tilde{\mathcal{M}}(n) \leqslant x\right) = \Psi_n[F_0(x)],$$

*где* $\Psi_n(\cdot)$ *— регуляризованная неполная бета-функция* $B(a_n, b_n)$ *с параметрами* $a_n = \left\lceil \frac{n+1}{2} \right\rceil$ *и* $b_n = \left\lfloor \frac{n+1}{2} \right\rfloor$*:*

$$\Psi_n(y) = \varkappa_n \int\limits_0^y t^{\left\lceil \frac{n+1}{2} \right\rceil - 1} (1-t)^{\left\lfloor \frac{n+1}{2} \right\rfloor - 1} \mathrm{d}t, \; y \in [0,1],$$

$$\varkappa_n = \frac{\Gamma(n+1)}{\Gamma(\left\lceil \frac{n+1}{2} \right\rceil)\Gamma(\left\lfloor \frac{n+1}{2} \right\rfloor)}.$$

$$L_N(F \mid \{(\mathcal{M}_k, n_k)\}_{k=1}^m) = \prod_{k=1}^m \mathsf{P}_{\tilde{\mathcal{M}}(n_k)}(\mathcal{M}_k)\, \mathsf{P}(\tilde{n} = n_k).$$

$$\mathsf{P}_{\tilde{\mathcal{M}}(n_k)}(\mathcal{M}_k) = \mathsf{P}(\tilde{\mathcal{M}}(n_k) = \mathcal{M}_k). \tag{1}$$

$$\mathsf{P}_{\tilde{\mathcal{M}}(n_k)}(\mathcal{M}_k) = p_{\tilde{\mathcal{M}}(n_k)}(\mathcal{M}_k). \tag{2}$$

$$L_N(F \mid \{(\mathcal{M}_k, n_k)\}_{k=1}^m) = \prod_{k=1}^{m} \mathsf{P}_{\tilde{\mathcal{M}}(n_k)}(\mathcal{M}_k)\, \mathsf{P}(\tilde{n} = n_k).$$

$$\mathsf{P}_{\tilde{\mathcal{M}}(n_k)}(\mathcal{M}_k) = \mathsf{P}(\tilde{\mathcal{M}}(n_k) = \mathcal{M}_k). \tag{1}$$

$$\mathsf{P}_{\tilde{\mathcal{M}}(n_k)}(\mathcal{M}_k) = p_{\tilde{\mathcal{M}}(n_k)}(\mathcal{M}_k). \tag{2}$$

- $\mathfrak{D}$ — множество распределений, имеющих атомы в точках наблюдения,
- $\mathfrak{C}$ — множество соответствующих им функций распределения.

Задача приобретает вид нахождения:

$$\widehat{F}_m = \operatorname*{argmax}_{F \in \mathfrak{C}} l_m(F),$$

$$l_m(F) = \sum_{k=1}^{m} \log\Big(\Psi_{n_k}\left[F(\mathcal{M}_k)\right] - \Psi_{n_k}\left[F(\mathcal{M}_k - 0)\right]\Big).$$

## Утверждение

1. На множестве $\mathfrak{D}$ распределений, имеющих атомы в наблюдаемых медианах, максимум функции правдоподобия достигается.
2. Максимайзеры функции правдоподобия сосредоточены на конечном множестве наблюдаемых точек $\{\mathcal{M}_1, \ldots, \mathcal{M}_m\}$ и присваивают каждой из них положительную вероятность.

### Утверждение

1. На множестве $\mathfrak{D}$ распределений, имеющих атомы в наблюдаемых медианах, максимум функции правдоподобия достигается.
2. Максимайзеры функции правдоподобия сосредоточены на конечном множестве наблюдаемых точек $\{\mathcal{M}_1, \ldots, \mathcal{M}_m\}$ и присваивают каждой из них положительную вероятность.

Упорядочим точки наблюдения $\mathcal{M}_1 < \mathcal{M}_2 < \ldots < \mathcal{M}_s$ и перейдем к новым переменным:

$$x_\nu = F(\mathcal{M}_\nu),\ \nu = 1, \ldots, s,$$

$$\Xi_m := \sum_{\nu=1}^{s} \sum_{k:\mathcal{M}_k = \mathcal{M}_\nu} \log\left(\Psi_{n_k}[x_\nu] - \Psi_{n_k}[x_{\nu-1}]\right) \to \max,$$

$$x_0 := 0 < x_1 < \ldots < x_s = 1.$$

## Утверждение

1. На множестве $\mathfrak{D}$ распределений, имеющих атомы в наблюдаемых медианах, максимум функции правдоподобия достигается.
2. Максимайзеры функции правдоподобия сосредоточены на конечном множестве наблюдаемых точек $\{\mathcal{M}_1, \ldots, \mathcal{M}_m\}$ и присваивают каждой из них положительную вероятность.

Упорядочим точки наблюдения $\mathcal{M}_1 < \mathcal{M}_2 < \ldots < \mathcal{M}_s$ и перейдем к новым переменным:

$$x_\nu = F(\mathcal{M}_\nu), \ \nu = 1, \ldots, s,$$

$$\Xi_m := \sum_{\nu=1}^{s} \sum_{k:\mathcal{M}_k=\mathcal{M}_\nu} \log\left(\Psi_{n_k}[x_\nu] - \Psi_{n_k}[x_{\nu-1}]\right) \to \max,$$

$$x_0 := 0 < x_1 < \ldots < x_s = 1.$$

## Утверждение

Функция $\Xi_m(\cdot)$ является строго выпуклой (вниз).

### Утверждение

В случае групп одинакового объема $n$ максимум функции правдоподобия достигается на единственной функции распределения, которая определяется формулой

$$\widehat{F}_m(x) = \Psi_n^{-1}(\widehat{G}_m(x)).$$

Здесь $\widehat{G}_m(x)$ эмпирическая функция распределения медианы $\tilde{\mathcal{M}}(n)$, построенная по наблюдениям $\mathcal{M}_1, \ldots, \mathcal{M}_m$.

Факторы, облегчающие задачу нахождения оценки $\widehat{F}_m$:

- Строгая монотонность функции $\Psi_n(\cdot)$ на интервале $[0, 1]$ поиска решений.
- Промежутки постоянства функции распределения $\widehat{G}_m(x)$ совпадают с промежутками постоянства функции распределения $\widehat{F}_m(x)$.
- Наличие затабулированных как прямых, так и обратных значений функции $\Psi_n(\cdot) = B(\lceil \frac{n+1}{2} \rceil, \lfloor \frac{n+1}{2} \rfloor)$.

## Теорема

*Оценка* $\widehat{F}_m(\cdot) = \Psi_n^{-1}(\widehat{G}_m(\cdot))$ *как оценка функции распределения* $F_0(\cdot)$ *исходной случайной величины* $\tilde{Z}$ *обладает следующими свойствами:*

- **равномерная состоятельность** :

$$\sup_{x\in\mathbb{R}} |\widehat{F}_m(x) - F_0(x)| \xrightarrow{a.s.} 0 \quad \text{при} \quad m \to \infty.$$

- **равномерная асимптотическая несмещенность** :

$$\boldsymbol{E} \sup_{x\in\mathbb{R}} |\widehat{F}_m(x) - F_0(x)| \xrightarrow{m\to\infty} 0.$$

- *В точках* $x$, *где* $0 < F_0(x) < 1$, *оценка* $\widehat{F}_m(x) = \Psi_n^{-1}\left[\widehat{G}_m(x)\right]$ *является* **асимптотически нормальной** :

$$\mathcal{L}(\sqrt{m}(\widehat{F}_m(x) - F_0(x)) \Rightarrow \mathrm{N}(0, D),$$

$$\text{где} \quad D = \frac{\Psi_n\left[F_0(x)\right]\left(1 - \Psi_n\left[F_0(x)\right]\right)}{\left(\Psi_n'\left[F_0(x)\right]\right)^2}.$$

- $\{\tilde{Z}_j\}$ — ненаблюдаемые независимые копии случайной величины $\tilde{Z}$ с конечным носителем,
- $\tilde{n}_k$ — независимые от $\{\tilde{Z}_j\}$ и между собой копии случайной величины $\tilde{n}$ с конечным носителем,
- известны пары реализаций $\{\, (\mathcal{M}_k, n_k) \,\}_{k=1}^{m}$.
- $\widehat{F}_m(\cdot)$ — оценка максимального правдоподобия функции распределения $F_0(\cdot)$ случайной величины $\tilde{Z}$.

**Равномерная сильная состоятельность:**

### Теорема

В перечисленных условиях $\widehat{F}_m(\cdot)$ является равномерно сильно состоятельной оценкой функции распределения $F_0(\cdot)$ случайной величины $\tilde{Z}$:

$$\sup_{x\in\mathbb{R}} |\widehat{F}_m(x) - F_0(x)| \xrightarrow{\text{a.s.}} 0 \quad \text{при} \quad m \to \infty.$$

Техника доказательства: Schick, A., Yu, Q., 2000.

**Равномерная асимптотическая несмещенность:**

### Следствие

*Оценка* $\widehat{F}_m(x)$ *является равномерно асимптотически несмещенной:*

$$\boldsymbol{E} \sup_{x\in\mathbb{R}} |\widehat{F}_m(x) - F_0(x)| \xrightarrow{m\to\infty} 0.$$

$\mathcal{Z} = \{z_1 < z_2 < \ldots < z_{s-1} < z_s\}$ — конечный носитель величины $\tilde{Z}$.

$$\mathbf{x} = \begin{pmatrix} F(z_1) \\ \vdots \\ F(z_{s-1}) \end{pmatrix} =: \mathbf{F}(\mathbf{z}), \qquad \mathbf{x}_0 = \mathbf{F}_0(\mathbf{z}), \qquad \hat{\mathbf{x}}_m = \widehat{\mathbf{F}}_m(\mathbf{z}).$$

$$\boxed{0 =: x_0 < x_1 < \cdots < x_{s-1} < x_s := 1.}$$

### Утверждение

*Введем матрицу вторых производных*

$$\Sigma_m(\mathbf{x}) = \Sigma_m(F) = \left(\frac{\partial^2 \Xi_m}{\partial x_i \partial x_j}(\mathbf{x})\right)_{i,j=1}^{s-1}.$$

*Информационная матрица Фишера*

$$J_0 = \boldsymbol{E}\left[\frac{1}{m}\nabla\Xi_m(\mathbf{x}_0)\left(\nabla\Xi_m(\mathbf{x}_0)\right)^T\right] = -\boldsymbol{E}\left[\frac{1}{m}\Sigma_m(\mathbf{x}_0)\right]$$

*является положительно определенной.*

**Асимптотическая нормальность:**

### Теорема

1. *Оценка* $\widehat{F}_m(x)$ *является асимптотически нормальной:*

$$\mathcal{L}\left(\sqrt{m}\,\Delta(\widehat{F}_m, F_0)\right) \Rightarrow N(0, J_0^{-1}),$$

$$\text{где} \quad \Delta(\widehat{F}_m, F_0) = \begin{pmatrix} \widehat{F}_m(z_1) - F_0(z_1) \\ \vdots \\ \widehat{F}_m(z_{s-1}) - F_0(z_{s-1}) \end{pmatrix}.$$

2. $-\frac{1}{m}\Sigma_m(\widehat{F}_m)$ *является сильно состоятельной оценкой* $J_0$*:*

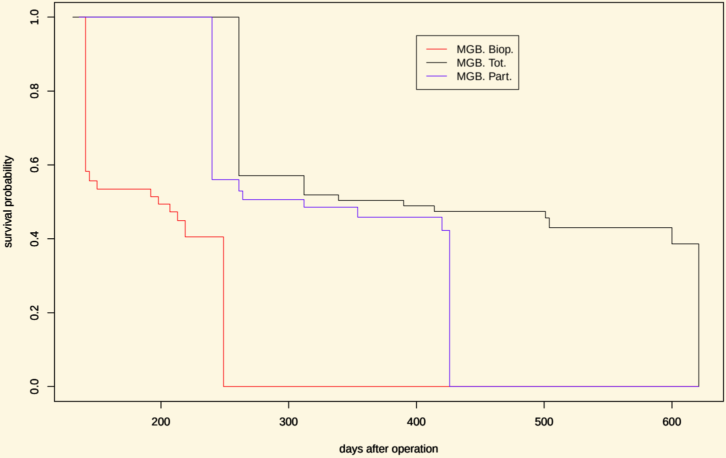
$$-\frac{1}{m}\Sigma_m(\hat{\mathbf{x}}_m) \xrightarrow{a.s.} -\boldsymbol{E}\left[\frac{1}{m}\Sigma_m(\mathbf{x}_0)\right] = J_0 \quad \text{при} \quad m \to \infty.$$

**Данные:**

Военной академией им. Кирова были предоставлены данные о времени жизни пациентов с различными типами опухолей головного мозга, подвергнутых различным способам хирургического лечения. Данные были медианно-цензурированными, то есть имели вид набора пар

$$\{(\mathcal{M}_k, n_k)\}.$$

- язык программирования и среда разработки: **R** ;
- нахождение значений бета-функции: функция **Rbeta** из библиотеки **zipfR** упомянутой среды;
- решение оптимизационной задачи нахождения оценки максимального правдоподобия: функция **constrOptim** из базовой библиотеки упомянутой среды
- использован алгоритм оптимизации:
  - метод деформируемых многогранников (Nelder-Mead).



MGB. Biop.
MGB. Tot.
MGB. Part.
survival probability
0.0
0.2
0.4
0.6
0.8
1.0
200
300
400
500
600
days after operation