Теорія страхового та фінансового ризику

Ямненко Р.Є

КНУ імені Тараса Шевченка

20 березня 2017 р.

Зміст

Питання статистичних висновків

1.1. Розглянемо питання статистичних висновків. При досить м'яких умовах на розподіл T, ми можемо одержати всю інформацію оцінивши $F(t), S(t), f(t)$ або μ_t для всіх $t \geq 0$.

1.2. Найпростішим експериментом є спостереження великої кількості новонароджених. Пропорція живих у віці $t > 0$ дає оцінку $S(t)$. Оцінка буде східчастою функцією, і чим більшою за об'ємом буде вибірка тим ближче до гладкої функції буде оцінка. Для застосувань оцінку потрібно згладити додатково. Нам не потрібно припускати, що T належить деякому параметричному сімейству. Це *непараметричний підхід* в оцінюванні. Зрозуміло, що так ми одержимо емпіричну функцію розподілу T.

1.3.

Ясно, що існують певні практичні труднощі:

(a) Навіть, якщо знайдено задовільну групу осіб, експеримент повинен тривати біля 100 років до завершення.

(b) План спостережень вимагає спостерігати кожну смерть у всіх осіб у вибірці. На практиці з тих чи інших причин спостереження втрачаються і виключення їх з аналізу може привести до зміщення результату. У статистиці цю проблему називають *цензуруванням*. Все, що ми знаємо про деяких осіб це те, що вони померли після певного віку.

1.4. У медицинській статистиці, де тривалість життя часто коротка, непараметричні оцінки дуже важливі.

У цьому розділі ми покажемо, як експеримент, описаний вище, може бути виправлений, щоб допускати цензурування. Інакше ми будемо повинні використовувати інший план спостережень і базувати наші висновки на даних зібраних протягом короткого часу: 3 або 4 роки.

Наслідком цензурування є те, що ми більше не спостерігаємо ту ж саму сукупність осіб протягом їх спільного терміну життя, тому ми не можемо робити вибірку із того самого розподілу. Можливо розумно розширити припущення у моделі так, щоб смертність осіб, що народились у році y моделювалась випадковою величиною T^y.

На практиці звичайно ділять дослідження до одного року віку. Ми повернемся до подібних досліджень у Лекції 7.

1.5. Спостереження осіб між цілими значеннями віку x і $x+1$, а також обмеження періоду спостережень також є цензуруванням. Цензурування може відбутися як у непередбачені моменти часу, наприклад припинення дії страхового полісу, так і у визначений час, наприклад при досягненні віку $x+1$ або при закінчені терміну досліджень.

Механізм цензурування

2.1. Цензурування — це основна характерна ознака даних про виживання (у дійсності аналіз виживання треба визначити, як аналіз цензурованих даних). Механізм, який приводить до цензурування, грає важливу роль у статистичних висновках.

Наведемо деякі із найбільш загальновживаних типів цензурування, які не є взаємовиключними:

(a) **Цензурування справа**. Дані є цензурованими справа, якщо механізм цензурування припиняє спостереження над процесом. Прикладом є закінчення досліджень у фіксовану дату.

(b) **Цезурування зліва.** Дані є цензурованими зліва, якщо механізм цензурування не дає можливості взнати, коли процес попав у стан, за яким ми хочемо спостерігати. Прикладом є регулярний медичний огляд. Виявлення стану пацієнта (хвороба) при огляді дає нам інформацію лише про те, що початок хвороби припадає на період після останнього огляду. Час, що пройшов з початку хвороби, цензурований зліва.

(c) **Інтервальне цензурування.** Дані є інтервально цензуровані, якщо план спостереження дає можливість сказати, що подія, яка нас цікавить, відбулась впродовж деякого інтервалу часу. Прикладом є актуарні дослідження, коли ми можемо знати лише календарний рік смерті.

(d) **Випадкове цензурування.** Якщо цензурування випадкове, тоді момент часу C_i, в який спостереження за тривалістю i-того життя цензурується, є випадковою величиною. Спостереження будуть цензуровані, якщо $C_i < T_i$, де T_i – це випадкова тривалість життя i-ї особи. Прикладом випадкового цензурування є випадок, коли особа випадає із спостереження з причин інших ніж смерть і момент виходу із спостереження не відомий заздалегідь. Випадкове цензурування є окремим випадком цензурування справа.

(e) **Неінформативне цензурування.** Цензурування є неінформативним, якщо воно не дає інформації про тривалість життя $\{T_i\}$. При випадковому цензуруванні незалежність кожної пари

$$T_i, C_i$$

є достатньою умовою того, що цензурування є неінформативним. Інформативне цензурування більш складне для аналізу, зокрема тому, що функція правдоподібності не може бути факторизована (розкладена на множники).

(f) **Цензурування типу I.** Якщо моменти цензурування $\{C_i\}$ відомі заздалегідь (вироджений випадок випадкового цензурування), тоді має місце механізм цензурування I-го типу.

(g) **Цензурування типу II.** Якщо обстеження продовжуються доти, поки не наступить визначена кількість смертей, тоді говорять, що має місце цензурування типу II. Це може спростити аналіз, тому, що кількість подій, що нас цікавить є невипадковою.

2.2. Зрозуміло, що план спостережень можливо вводить цензурування певного виду, і при аналізі це треба брати до уваги.

Цензурування також може залежати від того, що спостереження велися до фіксованої дати, наприклад, якщо є достатньо вагомі свідчення, накопичені протягом курсу експериментального лікування, то дослідження можна припинити достроково, щоб при позитивних результатах цей курс лікування впровадити для всіх пацієнтів, а при негативних результатах цей курс лікування відмінити.

Оцінка Каплана-Мейера (оцінка добуток-границя)

3.1. У цьому підрозділі ми розробимо емпіричну функцію розподілу, яка допускає цензурування.

3.2. Будемо розглядати тривалість життя як функцію часу t, без врахування початкового віку x. Одержані результати можна буде застосовувати для новонароджених, для тих хто досяг віку x, або для тих осіб, які мали спільні властивості в момент часу $t = 0$, наприклад мали однаковий діагноз медичного стану.

Медичні дослідження часто грунтуються на моменті постановки діагнозу, або на моменті початку лікування, і якщо вік пацієнта враховується при аналізі, то як пояснювальна змінна у регресійній моделі.

3.3. Припустимо, що ми спостерігаємо за популяцією з N осіб при наявності неінформативного цензурування, і припустимо, що ми виявили m смертей. Нехай

$$t_1 < t_2 < \ldots < t_k$$

— це впорядковані моменти часу, в які спостерігалась смерть.

Ми не припускаємо, що $k = m$, тому в один і той же момент часу можна спостерігати більше ніж одну смерть. Припустимо, що d_j смертей спостерігали в момент t_j $(1 \leq j \leq k)$, тому

$$d_1 + \ldots + d_k = m.$$

Спостереження інших $N - m$ осіб цензуровано (тобто ми не будемо намагатись продовжувати спостереження за цими особами у нашому дослідженні). Припустимо, що c_j осіб цензуровано протягом проміжку часу між t_j і t_{j+1} $(0 \leq l \leq k)$, де $t_0 = 0,\ t_{k+1} = +\infty$, щоб допускати цензурування спостережень після останнього моменту спостереження смерті t_k. Тому

$$c_0 + c_1 + \ldots + c_k = N - m.$$

Отже c_j – це число осіб, які випали із спостереження між t_j і t_{j+1} з причин інших ніж смерть. Будемо вважати, що всі c_j цензуровані спостереження попадають у відкритий інтервал (t_j, t_{j+1}). Якщо особа цензурується у той самий час, що і спостереження смерті іншої особи, то вважаємо, що смерть настала раніше.

Нехай $t_{j1}, t_{j2}, \ldots, t_{jc_j}$ — це моменти часу (не обов'язково різні) із інтервалу (t_j, t_{j+1}), у які спостереження були цензуровані. Зручно позначити через n_j число осіб, які вижили у групі ризику до моменту t_j^-, тобто якраз до j-го моменту спостереження смерті.

Для одержання функції правдоподібності цих спостережень і не роблячи ніяких припущень стосовно виду функції розподілу $F(t)$, будемо вважати, що:

(a) ймовірність того, що смерть трапиться у момент часу t_j дорівнює

$$F(t_j) - F(t_j^-);$$

(b) ймовірність того, що особа доживе до менту її цензурування t_{jl} дорівнює

$$1 - F(t_{jl})$$

при неінформативному цензуруванні.

Смерті і цензурування незалежні, тому загальна функція правдоподібності має вигляд

$$L = \prod_{j=1}^{k} (F(t_j) - F(t_j^-))^{d_j} \prod_{j=0}^{k} \prod_{l=1}^{c_j} (1 - F(t_{jl})).$$

Будемо шукати функцію $F(t)$, яка максимізує L, при єдиній умові, що вона є функцією розподілу. Оскільки функція розподілу є неспадна, а $t_{j1}, \ldots, t_{jc_j} \in (t_j, t_{j+1})$, то кожен множник $1 - F(t_{jl})$ максимізується, якщо

$$F(t_{jl}) = F(t_j) \quad \forall t_{jl}.$$

Таким чином,

$$F(t) = F(t_j), \ \forall t \in [t_j, t_{j+1}).$$

При цьому для всіх j:

$$F(t_j) > F(t_j^-),$$

бо інакше $L = 0$. Таким чином будь-яка оцінка максимальної правдоподібності для $F(t)$ буде східчастою функцією із стрибками у моменти спостереження смерті.

3.4.

Тепер зручно поширити на дискретні розподіли означення сили смертності (або ризику), яке ми дали в Розділі 5 для неперервних розподілів.

Припустимо, що $F(t)$ — це дискретний розподіл із стрибками в точках $t_1, \ldots, t_k$. Покладемо

$$\lambda_j = P(T = t_j \mid T \geq t_j), \qquad 1 \leq j \leq k.$$

Цю величину називають дискретною функцією ризику. Символ λ_j використовується, щоб не плутати із звичайною силою смертності.

Оскільки

$$t_0 = 0,\ F(0) = 0,\ d_0 = 0,\ t_{k+1} = +\infty$$

і

$$\lambda_j = \frac{P(T = t_j)}{P(T \geq t_j)} = \frac{F(t_j) - F(t_j^-)}{1 - F(t_j^-)},$$

то L можна переписати у вигляді

$$L = \prod_{j=1}^{k} \left(\frac{F(t_j) - F(t_j^-)}{1 - F(t_j^-)} \right)^{d_j} \prod_{j=0}^{k} (1 - F(t_j^-))^{d_j} \prod_{l=1}^{c_j} (1 - F(t_{jl})) =$$

$$= \prod_{j=1}^{k} (\lambda_j)^{d_j} \prod_{j=0}^{k} (1 - F(t_j^-))^{d_j} (1 - F(t_j))^{c_j}.$$

Випадкова величина T набуває значення $t_1, t_2, \ldots, t_k$, тому

$$P(T > t_j) = P(T \geq t_{j+1}),\ j = 0, 1, \ldots, k$$

$$P(T \geq t_0) = P(T \geq t_1) = 1$$

і

$$\prod_{j=0}^{k} (1 - F(t_j^-))^{d_j} (1 - F(t_j))^{c_j} = \prod_{j=0}^{k} P^{d_j}(T \geq t_j) P^{c_j}(T > t_j) =$$

$$= \prod_{j=1}^{k} P^{d_j}(T \geq t_j) P^{c_j}(T \geq t_{j+1}) =$$

$$= P^{c_1}(T \geq t_2) \prod_{j=2}^{k} P^{d_j}(T \geq t_j) P^{c_j}(T \geq t_{j+1}).$$

Із означення маємо

$$1-\lambda_j = P(T > t_j \mid T \geq t_j) = \frac{P(T > t_j)}{P(T \geq t_j)} = \frac{P(T \geq t_{j+1})}{P(T \geq t_j)}.$$

Тому

$$(1-\lambda_1)(1-\lambda_2)\cdots(1-\lambda_j) = \frac{P(T \geq t_2)}{P(T \geq t_1)} \cdot \frac{P(T \geq t_3)}{P(T \geq t_2)} \cdots \frac{P(T \geq t_{j+1})}{P(T \geq t_j)}$$

$$= P(T \geq t_{j+1}) = P(T > t_j).$$

Оскільки $F(t) = F(t_j) = P(T \leq t_j)$ при $t \in [t_j, t_{j+1})$, то

$$F(t) = 1 - P(T > t) = 1 - \prod_{t_j \leq t} (1-\lambda_j). \qquad (*)$$

Далі

$$P^{d_j}(T \geq t_j) \cdot P^{c_j}(T \geq t_{j+1}) =$$

$$= (1-\lambda_1)^{d_j+c_j} \cdots (1-\lambda_{j-1})^{d_j+c_j}(1-\lambda_j)^{c_j}.$$

Із означення маємо

$$n_{j+1} = n_j - d_j - c_j,\ j = 1, 2, \ldots, k.$$

Тому

$$P^{d_j}(T \geq t_j) \cdot P^{c_j}(T \geq t_{j+1}) =$$

$$= (1 - \lambda_1)^{n_j - n_{j+1}} \cdots (1 - \lambda_{j-1})^{n_j - n_{j+1}} (1 - \lambda_j)^{c_j}, \quad j = 2, 3, \ldots, k.$$

Оскільки $n_{k+1} = 0$, то

$$P^{c_1}(T \geq t_2) \prod_{j=2}^{k} P^{d_j}(T \geq t_j) \cdot P^{c_j}(T \geq t_{j+1}) = \prod_{j=1}^{k} (1 - \lambda_j)^{n_j - d_j}.$$

Тому

$$L = \prod_{j=1}^{k} (\lambda_j)^{d_j} (1 - \lambda_j)^{n_j - d_j}.$$

3.5. Розглянувши $\ln L$ і дослідивши цю функцію на максимум одержимо оцінку максимальної правдоподібності

$$\hat{\lambda}_j = \frac{d_j}{n_j}, \qquad 1 \leq j \leq k.$$

А використавши властивість інваріантності оцінок максимальної правдоподібності одержимо із (*) оцінку

$$\hat{F}(t) = 1 - \prod_{t_j \leq t} (1 - \hat{\lambda}_j).$$

3.6.

Одержана оцінка — це *оцінка Каплана-Мейера* або *оцінка добуток-границя*. Її можна розглядати з різних точок зору:

(a) Вивчаючи ймовірність смерті протягом малих вікових інтервалів, ми можемо вибирати розбиття часової осі так як нам хочется. Зручний вибір — мати дуже малі часові інтервали, які містять кожен момент t_j (достатньо малі, щоб виключити моменти цензурування t_{jl}) і довші часові інтервали, які містять тільки цензуровані спостереження.

Єдина інформація, яку ми одержимо із останніх інтервалів, це те, що протягом них не було смертей. Тому можна вважати функцію $F(t)$ сталою на цих інтервалах. В той же час малі інтервали дають біноміальну оцінку тривалостей життя, які спостерігались.

(b) Як альтернативний підхід, ми можемо вибирати все більш тонке розбиття осі часу і оцінювати $1 - F(t)$ як добуток ймовірностей виживання для кожного інтервалу.

Тоді із даного означення дискретної сили смертності, ми одержимо оцінку Каплана-Мейера, якщо крок розбиття спрямувати до нуля. Це є причиною назви "оцінка добуток-границя яка іноді використовується.

3.7. Тільки ті особи із групи ризику, для яких проводилось спостереження за тривалістю життя $\{t_j\}$ дають внесок у оцінку. Тому не обов'язково починати спостереження всіх осіб в один і той же час або вік.

Оцінка буде мати місце і для даних, зрізаних зліва, за умови, що зрізання є неінформативним, в тому сенсі, що початок спостереження у визначений час або вік не залежить від залишку майбутнього життя.

Порівняння розподілів тривалості життя

4.1. Оцінка Каплана-Мейера часто використовується для порівняння розподілів тривалості двох або більше популяцій, наприклад при порівнянні методів лікування. Тому важливо знати її статистичні властивості.

Має місце наближена формула (формула Грінвуда) для дисперсії $\tilde{F}(t)$:

$$\mathrm{Var}[\tilde{F}(t)] \approx (1 - \hat{F}(t))^2 \sum_{t_j \le t} \frac{d_j}{n_j(n_j - d_j)}$$

прийнятна для більшості t, але яка може применшувати дисперсію у хвості розподілу.

Оцінка Нельсона-Аалена

5.1. Альтернативний непараметричний підхід — це оцінка *інтегрального ризику*

$$\Lambda = \int_0^t \mu_s \, ds + \sum_{t_j \leq t} \lambda_j,$$

де інтеграл відповідає неперервній частині розподілу, а сума — дискретній частині.

Оскільки методологія була розроблена статистиками, то широко використовується термін "інтегральний ризик в той час як термін "інтегральна сила смертності" майже не використовується.

5.2. Оцінка Нельсона-Аалена інтегрального ризику має вигляд

$$\hat{\Lambda}_t = \sum_{t_j \le t} \frac{d_j}{n_j}.$$

5.3. Оцінку Каплана-Мейера можна наблизити через $\hat{\Lambda}_t$:

$$\hat{F}_t = 1 - \prod_{t_j \le t} \left(1 - \frac{d_j}{n_j}\right) \approx 1 - \exp\left\{-\sum_{t_j \le t} \frac{d_j}{n_j}\right\} = 1 - \exp\{-\hat{\Lambda}_t\}.$$

5.4.

У відповідності до формули Грінвуда для дисперсії оцінки Каплана-Мейера, має місце формула для дисперсії оцінки Нельсона-Аалена:

$$\mathrm{Var}[\tilde{\Lambda}_t] \approx \sum_{t_j \le t} \frac{d_j(n_j - d_j)}{n_j^3}.$$