Estimateurs non paramétriques de la fonction de répartition d'une variable censurée à droite sur petits domaines : approche basée sur un modèle

Sandrine Casanova & Eve Leconte
TSE-R
Université Toulouse 1 Capitole

10ème Colloque Francophone sur les Sondages, Lyon
25 octobre 2018

Position du problème

Estimation de la fonction de répartition (fdr) d'une variable d'intérêt censurée à droite sur petits domaines

$\hookrightarrow$ variable d'intérêt T = durée jusqu'à un événement d'intérêt

$\hookrightarrow$ durée censurée à droite si l'événement n'a pas eu lieu

$\hookrightarrow$ très peu d'articles dans le cadre des sondages

Position du problème

Estimation de la fonction de répartition (fdr) d'une variable d'intérêt censurée à droite sur petits domaines

$\hookrightarrow$ variable d'intérêt T = durée jusqu'à un événement d'intérêt

$\hookrightarrow$ durée censurée à droite si l'événement n'a pas eu lieu

$\hookrightarrow$ très peu d'articles dans le cadre des sondages

Exemple d'application

Enquête rétrospective sur l'insertion des jeunes filles diplômées du secondaire en Occitanie, 3 ans après la fin de leurs études

T : temps d'accès au premier emploi

$\hookrightarrow$ T censurée à droite pour celles qui n'ont pas trouvé d'emploi à la date de l'enquête

Domaine = niveau $\times$ type de formation

Plan

- Estimateurs de la fdr sur petits domaines en présence de censure
  1. Estimateurs utilisant uniquement les données du domaine
  2. Estimateur empruntant de la force aux autres domaines
- Simulations basées sur un modèle
- Exemple d'application
- Perspectives

Notations

Population U (taille N) partitionnée en m domaines U_i (taille N_i)
s échantillon de U de taille n
$s_i = s \cap U_i$ échantillon de taille n_i du domaine U_i.
t_{ij} valeur de T pour le $j$ème individu du domaine U_i : uniquement observée sur s_i et éventuellement censurée à droite par z_{ij}.
$\hookrightarrow$ sur s_i, on observe $y_{ij} = min(t_{ij}, z_{ij})$ et $\delta_{ij} = \mathbb{1}(t_{ij} \leq z_{ij})$

Notations

Population U (taille N) partitionnée en m domaines U_i (taille N_i)
s échantillon de U de taille n
$s_i = s \cap U_i$ échantillon de taille n_i du domaine U_i.
t_{ij} valeur de T pour le $j$ème individu du domaine U_i : uniquement observée sur s_i et éventuellement censurée à droite par z_{ij}.
$\hookrightarrow$ sur s_i, on observe $y_{ij} = min(t_{ij}, z_{ij})$ et $\delta_{ij} = \mathbb{1}(t_{ij} \leq z_{ij})$

Objectif : estimer la fdr F^i de T sur le domaine U_i

$$F^i(t) = \frac{1}{N_i} \sum_{j \in U_i} \mathbb{1}(t_{ij} \leq t)$$

- Information auxiliaire apportée par une covariable x continue connue sur tout U
- On suppose que le plan d'échantillonnage est non informatif
  $\hookrightarrow$ estimateurs basés sur un modèle
  $\hookrightarrow$ prédiction de $\mathbb{1}(t_{ij} \leq t)$ pour $j \in U_i \setminus s_i$

Estimateurs utilisant uniquement les données du domaine

- Approche directe :
  Cas censuré : t_{ij} non connu complètement sur s_i
  $\hookrightarrow$ on estime F^i par l'estimateur de Kaplan-Meier calculé avec les individus de s_i : $\widehat{F}^i_{\mathrm{KM}}$

Estimateurs utilisant uniquement les données du domaine

- Approche directe :
  Cas censuré : t_{ij} non connu complètement sur s_i
  $\hookrightarrow$ on estime F^i par l'estimateur de Kaplan-Meier calculé avec les individus de s_i : $\widehat{F}^i_{\text{KM}}$
- Approche indirecte : estimateur basé sur un modèle

Estimateurs de la fdr en population finie :

Chambers et Dunstan (1986)	$\rightarrow$	Dorfman et Hall (1993)	$\rightarrow$	Casanova et Leconte (2015)
paramétrique		non paramétrique		non paramétrique
sans censure		sans censure		avec censure

Estimateurs utilisant uniquement les données du domaine

- Approche directe :
  Cas censuré : t_{ij} non connu complètement sur s_i
  $\hookrightarrow$ on estime F^i par l'estimateur de Kaplan-Meier calculé avec les individus de s_i : $\widehat{F}^i_{\text{KM}}$
- Approche indirecte : estimateur basé sur un modèle

  Estimateurs de la fdr en population finie :

Chambers et Dunstan (1986)	$\rightarrow$	Dorfman et Hall (1993)	$\rightarrow$	Casanova et Leconte (2015)
paramétrique		non paramétrique		non paramétrique
sans censure		sans censure		avec censure

$\hookrightarrow$ Casanova et Leconte (2015) appliqué au domaine U_i

$$F^i(t) = \frac{n_i}{N_i} \underbrace{\left(\frac{1}{n_i} \sum_{j \in s_i} \mathbb{1}(t_{ij} \leq t) \right)}_{\text{estimé par } \widehat{F}^i_{\text{KM}}} + \frac{1}{N_i} \sum_{j \in U_i \setminus s_i} \underbrace{\mathbb{1}(t_{ij} \leq t)}_{\text{à prédire}}$$

Modèle non paramétrique de superpopulation sur les domaines U_i :

$$\xi : \ t_{ij} = m_i(x_{ij}) + \varepsilon_{ij}, \ i = 1, \ldots, m, \ j = 1, \ldots, N_i,$$

$m_i(x_{ij})$ médiane conditionnelle de T sachant $X = x_{ij}$
erreurs ε_{ij} variables i.i.d. de fdr G^i

Modèle non paramétrique de superpopulation sur les domaines U_i :

$$\xi : \ t_{ij} = m_i(x_{ij}) + \varepsilon_{ij}, \ i = 1, \ldots, m, \ j = 1, \ldots, N_i,$$

$m_i(x_{ij})$ médiane conditionnelle de T sachant $X = x_{ij}$
erreurs ε_{ij} variables i.i.d. de fdr G^i

On a $\mathbb{E}_\xi\left(\mathbb{1}(t_{ij} \leq t)\right) = G^i(t - m_i(x_{ij}))$

$\hookrightarrow$ $m_i(x_{ij})$ estimée par la solution de $\widehat{F}^i_{\text{SGKM}}(t \mid x_{ij}) = 0,5$
(deux paramètres de lissage h^i_X et h^i_T)

$\hookrightarrow$ fdr des erreurs G^i estimée par Kaplan-Meier calculé sur les résidus $y_{ij} - \hat{m}_i(x_{ij})$

$$\hookrightarrow \widehat{F}^i_{\text{M}}(t) = \frac{n_i}{N_i}\widehat{F}^i_{\text{KM}}(t) + \frac{1}{N_i}\sum_{j \in U_i \setminus s_i} \widehat{G}^i_{\text{KM}}(t - \widehat{m_i}(x_{ij}))$$

Estimateur empruntant de la force aux autres domaines

Estimateurs "petits domaines" de la fdr : approche (M)-quantiles

Tzavidis, Marchetti et Chambers (2010)	$\rightarrow$	Casanova (2012)	$\rightarrow$	Casanova et Leconte (2018)
paramétrique		non paramétrique		non paramétrique
sans censure		sans censure		avec censure

Modèle non paramétrique de superpopulation sur U :

$$\zeta \ : \ t_{ij} = m(q_i, x_{ij}) + \varepsilon_{ij}, \ i = 1, \ldots, m, \ j = 1, \ldots, N_i$$

q_i coefficient dans $(0, 1)$ caractérisant la position du domaine U_i,
$m(q_i, x_{ij})$ quantile conditionnel d'ordre q_i de T sachant $X = x_{ij}$
ε_{ij} variables i.i.d. (à i fixé) de fdr $G^{\prime i}$

$\hookrightarrow$ estimer $\mathbb{E}_\zeta\left(\mathbb{I}(t_{ij} \leq t)\right) = G'^{i}(t - m(q_i, x_{ij}))$

1. Estimation de q_i
   - Estimation des ordres-quantiles conditionnels des individus de s par $\hat{q}_{ij} = \hat{F}_{\text{SGKM}}(y_{ij} \mid x_{ij})$ à l'aide de tout l'échantillon (deux fenêtres h_X et h_T communes à tous les domaines)
   - Estimateur $\hat{q}_i$ de l'ordre q_i du domaine U_i : ordre médian des $\widehat{q_{ij}}$ des individus de s_i (par Kaplan-Meier)

$\hookrightarrow$ estimer $\mathbb{E}_\zeta\left(\mathbb{I}(t_{ij} \leq t)\right) = G'^{i}(t - m(q_i, x_{ij}))$

1. Estimation de q_i
   - Estimation des ordres-quantiles conditionnels des individus de s par $\hat{q}_{ij} = \hat{F}_{\text{SGKM}}(y_{ij} \mid x_{ij})$ à l'aide de tout l'échantillon (deux fenêtres h_X et h_T communes à tous les domaines)
   - Estimateur $\hat{q}_i$ de l'ordre q_i du domaine U_i : ordre médian des $\widehat{q_{ij}}$ des individus de s_i (par Kaplan-Meier)
2. Estimation du quantile conditionnel $m(q_i, x_{ij})$: solution de $\widehat{F}_{\text{SGKM}}(t \mid x_{ij}) = \hat{q}_i$

$\hookrightarrow$ estimer $\mathbb{E}_\zeta\left(\mathbb{I}(t_{ij} \leq t)\right) = G'^i(t - m(q_i, x_{ij}))$

1. Estimation de q_i
   - Estimation des ordres-quantiles conditionnels des individus de s par $\hat{q}_{ij} = \hat{F}_{\text{SGKM}}(y_{ij} \mid x_{ij})$ à l'aide de tout l'échantillon (deux fenêtres h_X et h_T communes à tous les domaines)
   - Estimateur $\hat{q}_i$ de l'ordre q_i du domaine U_i : ordre médian des $\widehat{q_{ij}}$ des individus de s_i (par Kaplan-Meier)
2. Estimation du quantile conditionnel $m(q_i, x_{ij})$: solution de $\widehat{F}_{\text{SGKM}}(t \mid x_{ij}) = \hat{q}_i$
3. Estimation de G'^i par Kaplan-Meier sur les résidus $y_{ij} - \hat{m}(\hat{q}_i, x_{ij})$ des individus de s_i

$$\hookrightarrow \widehat{F}^i_{\text{Q}}(t) = \frac{n_i}{N_i}\widehat{F}^i_{\text{KM}}(t) + \frac{1}{N_i}\sum_{j \in U_i \setminus s_i} \widehat{G}'^i_{\text{KM}}(t - \hat{m}(\hat{q}_i, x_{ij}))$$

Simulations basées sur un modèle

Description

- Génération de populations partitionnées en 10 domaines (de tailles fixes comprises entre 50 et 150) suivant le modèle log-linéaire de régression :
$$ln(t_{ij}) = 4 - 1,61 * x_{ij} + u_i + \varepsilon_{ij}$$
- $x_{ij} \sim U(1,4)$
- ε_{ij} terme d'erreur : distribution de valeur extrême $(Var(\varepsilon) = 1,645) \rightarrow t_{ij} \sim \mathcal{E}\text{xp}$
- Effet aléatoire du domaine U_i : $u_i \sim \mathcal{N}(0, \sigma^2)$
$\rightarrow \rho = \dfrac{\sigma^2}{\sigma^2 + Var(\varepsilon)}$ (10, 25 et 50 %)
- Délai de censure : $c_{ij} \sim U(0, c)$ (10, 25 et 50 % de censure)
- Données observées : $y_{ij} = min(t_{ij}, c_{ij})$ et $\delta_{ij} = \mathbb{1}(t_{ij} < c_{ij})$

- Taux de sondage : 1/20 (s_i de tailles 3, 5, 4, 4, 6, 5, 5, 4, 6, 5) et 1/10 (s_i de tailles 7, 9, 8, 7, 12, 9 , 9, 8, 12, 9)
- 1000 itérations
- Choix des paramètres de lissage de $\widehat{F}^i_{\mathrm{M}}$
  (h^i_X, h^i_T) choisi pour chaque domaine de façon à minimiser le critère $\mathrm{ASE}(\widehat{F}^i_{\mathrm{M}}) = \frac{1}{5}\sum_{k=1}^{5}\left(\widehat{F}^i_{\mathrm{M}}(tt_k) - F^i(tt_k)\right)^2$
  tt_k quantiles d'ordres 10 %, 25 %, 50 %, 75 % et 90 % de la loi de T.
- Choix des paramètres de lissage de $\widehat{F}^i_{\mathrm{Q}}$
  (h_X, h_T) tel que $\sum_{i=1}^{10}\mathrm{ASE}(\hat{F}^i_{\mathrm{Q}})$ minimum

Simulations basées sur un modèle

Résultats

Biais relatifs absolus moyennés (en %) sur les 10 domaines pour $\rho = 25\%$.

Taux de sondage : 1/20

Ordre quantile	$\tau = 10\%$ KM	M	Q	$\tau = 25\%$ KM	M	Q	$\tau = 50\%$ KM	M	Q
0.10	2.75	25.40	26.57	2.75	17.28	15.60	7.26	14.84	5.97
0.25	1.89	5.80	12.67	2.14	5.50	11.89	5.50	11.15	9.11
0.50	1.11	5.19	12.27	1.54	6.24	9.87	12.39	16.30	10.34
0.75	0.77	4.96	2.45	4.91	8.99	9.28	29.12	28.96	29.12
0.90	1.70	3.83	2.98	9.60	9.27	9.55	9.04	9.04	9.04

Taux de sondage : 1/10

Ordre quantile	$\tau = 10\%$ KM	M	Q	$\tau = 25\%$ KM	M	Q	$\tau = 50\%$ KM	M	Q
0.10	2.31	34.88	26.10	2.27	26.98	17.04	2.67	17.33	4.22
0.25	1.44	1.56	17.55	1.36	1.64	15.23	1.75	2.27	15.77
0.50	0.99	2.12	16.73	0.98	1.64	13.85	3.13	6.00	1.63
0.75	0.51	2.37	1.25	1.92	5.42	6.49	30.58	30.35	30.58
0.90	1.25	2.77	2.37	9.91	9.35	9.87	9.70	9.70	9.70

Racines carrées des erreurs quadratiques moyennes relatives (en %) moyennées sur les 10 domaines pour $\rho = 25\%$.

Taux de sondage : 1/20

Ordre	$\tau = 10\%$			$\tau = 25\%$			$\tau = 50\%$		
quantile	KM	M	Q	KM	M	Q	KM	M	Q
0.10	129.64	101.45	75.21	129.91	99.95	70.41	133.13	110.16	72.47
0.25	73.13	53.97	40.04	73.31	55.93	40.82	75.83	61.20	45.83
0.50	42.80	33.52	28.48	44.08	35.46	29.72	53.25	45.12	39.12
0.75	25.67	20.86	22.05	28.40	23.88	24.42	32.81	32.65	32.81
0.90	15.24	12.02	13.55	12.88	12.40	12.82	12.18	12.18	12.18

Taux de sondage : 1/10

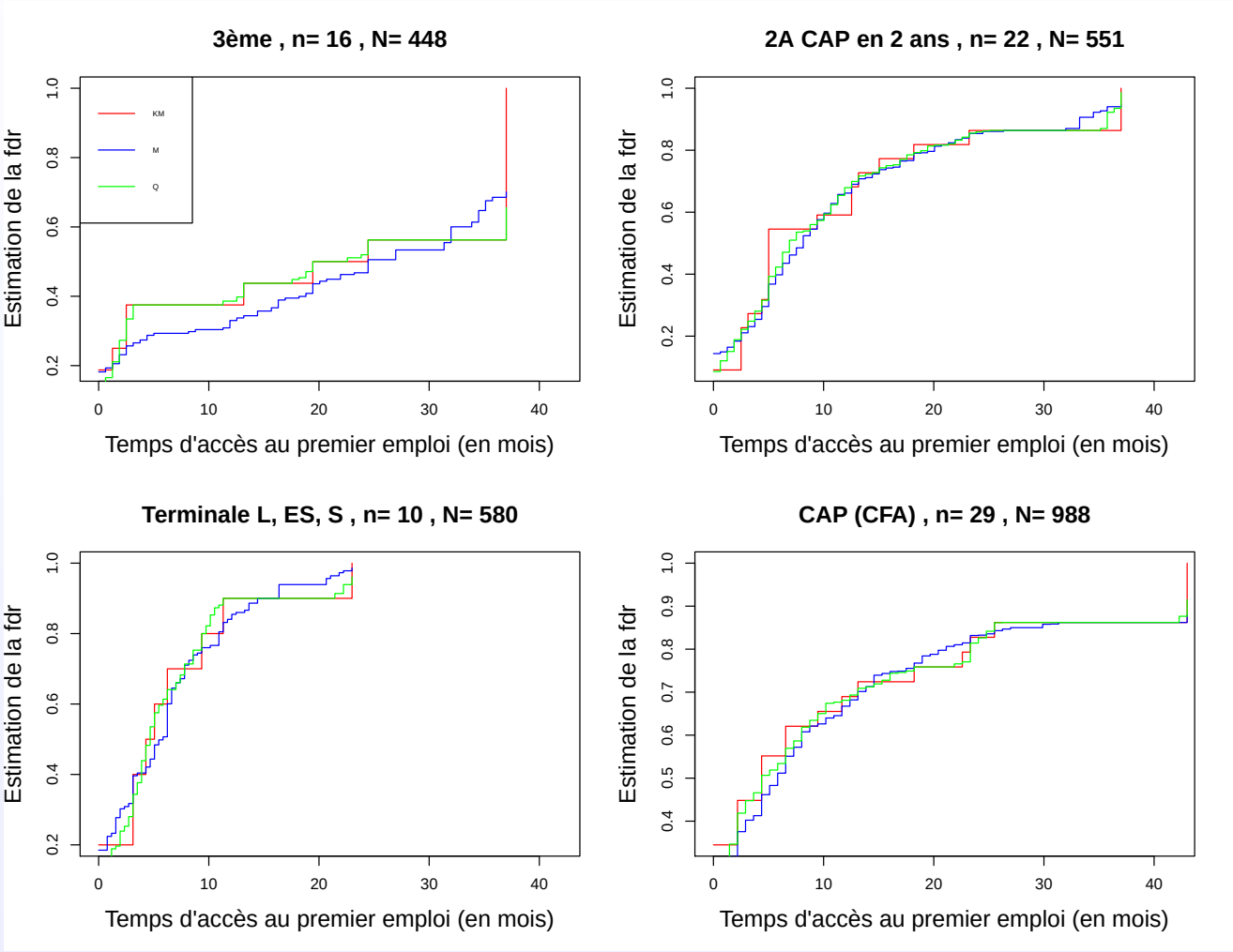
Ordre	$\tau = 10\%$			$\tau = 25\%$			$\tau = 50\%$		
quantile	KM	M	Q	KM	M	Q	KM	M	Q
0.10	90.70	65.95	61.33	90.84	60.93	57.38	91.60	58.40	55.55
0.25	52.23	29.10	34.18	52.39	29.47	33.96	54.01	33.70	36.90
0.50	30.05	22.35	25.02	30.83	23.58	24.53	37.12	29.87	27.71
0.75	17.88	15.37	15.99	21.06	17.57	18.72	34.45	34.22	34.45
0.90	11.52	9.49	10.50	13.25	12.42	13.20	12.90	12.90	12.90

Exemple d'application

- Données extraites de l'enquête "Génération 2010" du Centre d'Etudes et de REcherches sur les Qualifications (Céreq)
- Sous-population : jeunes filles diplômées sortant du secondaire en Occitanie en 2010 ($N = 10135$)
- Durée T : temps d'accès au premier emploi
  $\hookrightarrow$ censuré à droite pour celles qui n'ont pas trouvé d'emploi à la date de l'enquête (12,5% de censure)
- Enquête téléphonique rétrospective : jeunes filles interrogées 3 ans après la fin de leurs études

- Echantillon : obtenu par plan de sondage stratifié par région de formation et équilibré par type et spécialité de formation, tirage à probabilités inégales (algorithme du Cube)
  $\hookrightarrow n = 306$ jeunes filles
- Domaines : niveaux $\times$ types de formation
  $\hookrightarrow$ 34 domaines de tailles variant de 7 à 1480
  $\hookrightarrow$ tailles des échantillons s_i variant de 1 à 37
- Variable auxiliaire : taux de chômage de la zone d'emploi de l'établissement de fin d'études
- liaison significative avec T ($p = 0,014$)
  $\hookrightarrow$ 6,6 % de chance de plus de trouver un emploi si le taux de chômage local est plus faible de 1 %

Courbes de survie pour 4 domaines (niveau $\times$ formation)



Perspectives

- Diminution du biais à l'aide d'une approche "assistée par un modèle",
- Techniques de type bootstrap pour estimer le biais et la variance de l'erreur de prédiction et un IC de la fdr,
- Trouver d'autres exemples d'application.