제 7장: 기타 표집

7.1 포획-재포획 추정

1. 문제

   - 집단의 크기 N에 대한 추정
   - closed population 가정
   - homogeneous capture probability 가정
   - capture-recapture의 절차는 다음의 두 단계 절차이다

     ① n_1개의 개체를 포획하고 tag를 붙인 후 방류한다

     ② 일정시간이 지난후 n_2개의 개체를 포획하고 다시 포획된 (꼬리표가 있는) 개체의 수를 센다. 이를 m이라 하자.
   - 그러면 직관적으로 $N : n_1 = n_2 : m$의 비례관계로부터 $\hat{N} = \frac{n_1 n_2}{m}$의 추정량을 계산할 수 있다.

2. 비추정량

   - 앞에서 소개한 $\hat{N} = \frac{n_1 n_2}{m}$은 특별한 상황에서 모합에 대한 비추정량으로 표현될 수 있고 비추정량의 성질을 이용하여 $\hat{N}$의 통계적 성질을 계산할 수 있다.
   - $\hat{N}$을 비추정량으로 재표현 하기위하여 다음의 용어를 정의한다.

$$y_i = 1, \quad i = 1, 2, \cdots, N$$

$$x_i = \begin{cases} 1 & i\text{번째 물고기에 꼬리표가 달린 경우} \\ 0 & \text{그렇지 않은 경우} \end{cases}$$

   - 우리는 $\tau_x = \sum_{i=1}^{N} x_i = n_1$임을 알고 있고 따라서 이를 보조 변수로 이용하여

$$\tau_y = \sum_{i=1}^{N} y_i = N$$

을 추정하고자 한다.

- $\sum_{j=1}^{n_2} Y_j = n_2$이고 $\sum_{j=1}^{n_2} X_j = m$ 이므로
  모 비 $\beta = \frac{\tau_y}{\tau_x}$의 추정량은

$$b = \frac{\sum_{j=1}^{n_2} Y_j}{\sum_{j=1}^{n_2} X_j} = \frac{n_2}{m}$$

  이고, 따라서

$$\hat{\tau}_y = \hat{N} = b\tau_x = \frac{n_2}{m} n_1 = \frac{n_1 n_2}{m}$$

  이다.

- 추정량의 분산과 분산추정량:

$$\begin{aligned} \hat{\mathrm{Var}}(\hat{\tau}_y) &= \hat{\mathrm{Var}}(\hat{N}) \\ &= \left(\tau_x\right)^2 \cdot \hat{\mathrm{Var}}(b) \\ &\approx \left(\tau_x\right)^2 \cdot \frac{N - n_2}{N} \frac{s_e^2}{n_2 \bar{X}^2}. \end{aligned}$$

  RECALL: 4장의 비 추정량에서

$$\mathrm{Var}(b) \approx \frac{1}{\mu_x^2} \frac{N - n}{N - 1} \frac{1}{n} \sigma_e^2,$$

  이고 σ_e^2는 잔차 $y_i - \beta x_i$의 분산이다.

  앞의 RECALL에 기반하여, 여기서 $\tau_x = n$, $\bar{X} = \frac{m}{n_2}$,

$$\begin{aligned} s_e^2 &= \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (Y_j - bX_j)^2 \\ &= \frac{1}{n_2 - 1} \left\{ m \left(1 - \frac{n_2}{m}(1)\right)^2 + (n_2 - m) \left(1 - \frac{n_2}{m}(0)\right)^2 \right\} \\ &= \frac{n_2(n_2 - m)}{m(n_2 - 1)}. \end{aligned}$$

따라서

$$\begin{aligned}\hat{\mathrm{Var}}(\hat{N}) &= n_1^2 \frac{1}{n_2(\frac{m}{n_2})^2} \frac{n_2(n_2-m)}{m(n_2-1)} \\ &= \frac{n_1^2 n_2(n_2-m)}{m^3}.\end{aligned}$$

7.1.1 Direct sampling vs inverse sampling

- Inverse sampling에서는 capture-recapture의 recapture procedure에서

  "m마리의 꼬리표 달린 물고기들이 잡힐 때까지 물고기를 임의표집"

  한다.

- 편의상 N과 n이 충분히 크다 하면 비복원추출은 복원추출로 근사 할 수 있으므로 re-capture 절차에서 복원추출을 가정한다.

- Recpature절차는

  $$n_2 \sim \text{Negative Binomial}(m, p = \frac{n_1}{N})$$

  의 분포를 따르게 되고, pmf는

  $$\binom{n_2-1}{m-1} p^m (1-p)^{n_2-m}$$

  이 된다.

- 또 $\mathbb{E}\big(n_2\big) = \frac{m}{p}$ 이므로

  $$\mathbb{E}\left(\frac{n_2}{m}\right) = \frac{1}{p} = \frac{N}{N_1}$$

  이고

  $$\mathbb{E}\left(\frac{n_1 n_2}{m}\right) = N$$

  이 된다.

- 따라서 추정량은 $\hat{N} = \frac{n_1 n_2}{m}$으로 앞의 추정량과 동일하게 표현되고 inverse sampling 가정하에 **비편향추정량**이다.

- 추정량의 분산은:

$$\begin{aligned}\mathrm{Var}(\hat{N}) &= \mathrm{Var}\left(\frac{n_1 n_2}{m}\right) \\ &= \frac{n_1^2}{m^2}\cdot\frac{m(1-p)}{p^2} \\ &= \frac{n_1^2}{m^2}\cdot\left\{m\frac{\frac{N-n_1}{N}}{(\frac{n_1}{N})^2}\right\} \\ &= \frac{N(N-n_1)}{m}\end{aligned}$$

- 분산의 추정량은:

$$\begin{aligned}\hat{\mathrm{Var}}(\hat{N}) &= \frac{1}{m}\cdot\frac{n_1 n_2}{m}\cdot\left(\frac{n_1 n_2}{m}-n_1\right) \\ &= \frac{1}{m^3}n_1^2 n_2(n_2-m).\end{aligned}$$

분산 추정량이 "direct sampling 기반" 비추정량의 분산추정량과 동일함을 알 수 있다.

7.2 사각표집

- 모집단의 크기를 측정하는게 목적이고
  단위면적당 개체들의 수 – 즉, 밀도(density) 또는 rate(λ) –를 측정하여
  [밀도]×[총단위(총면적)]로 개체들의 총수를 추정한다.

- 전체 면적을 구획으로 분할하게 되고 분할된 구획을 사각(quadrat) 이라 부른다.

  총면적 A에서 관심있는 개체의 수 N을 추정하고자 관심지역을 면적 a의 $\mathbb{K}$개의 사각으로 분할한다. $A = Ka$,

- 표집: 전체 K개의 사각중 k개의 사각을 단순임의표집하여 개체 수를 센다.

$$Y_i = \text{번째 사각에서 관측된 개체의 수}, \quad i = 1, 2, \cdots, k,$$

라 하면

$$Y_i \sim \text{Poisson}(\lambda a), \quad \lambda = \text{단위 면적당 개수}$$

분포를 따른다.

- 추정:

$$\hat{\lambda} = \frac{1}{a}\bar{Y} = \frac{1}{a}\frac{1}{k}\sum_{i=1}^{k} Y_i$$

이고, 전체 개체 수 "N"의 추정량은

$$\hat{N} = \hat{\lambda}A$$

이다.

- 이제 $\hat{N}$의 통계적 성질을 살펴보자.

$\hat{N}$은 불편 추정량이다.

다음으로 추정량의 분산과 분산의 추정량을 살펴보면, 먼저 추정량의 분산은

$$\begin{aligned} \text{Var}(\hat{\lambda}) &= \text{Var}\left(\frac{1}{ka}\sum_{i=1}^{k} Y_i\right) \\ &= \frac{1}{k^2a^2} \cdot \left(k\lambda a\right) = \frac{1}{ka}\lambda \end{aligned}$$

이므로, 이를 이용하여 분산과 분산의 추정량을 계산하면

$$\hat{\text{Var}}(\hat{\lambda}) = \frac{1}{ka} \cdot \hat{\lambda} = \frac{1}{ka} \cdot \left(\frac{1}{a}\bar{Y}\right) = \frac{1}{ka^2}\bar{Y}$$

이다.

- 분산에 대하여 다른 방식으로

$$\text{Var}(\hat{\lambda}) = \frac{1}{ka^2}\text{Var}(Y_i) = \frac{1}{ka^2}\sigma_y^2, \quad \hat{\text{Var}}(\hat{\lambda}) = \frac{1}{ka^2}s_y^2,$$

여기서 $s_y^2 = \frac{1}{k-1}\sum_{i=1}^{k}(Y_i - \bar{Y})^2$을 생각 할 수도 있다.

7.2.1 Stocked quadrat sampling

- 개체수를 정확하게 세는 작업이 어려운 경우 숫자를 세는 대신 존재여부만을 기록한다. 이러한 표집방법을 stocked-quadrat-sampling(적재-사각-표집)이라 한다.
- $Z_i = \begin{cases} 1 & i\text{번째 사각에서 개체가 있는 경우} \\ 0 & \text{그렇지 않은 경우} \end{cases}$
- 개체의 수 Y_i에 대하여 포아송모형을 가정하면
  $Z_i = I(Y_i \geq 1)$이고

$$\begin{aligned} P(Z_i = 1) &= P(Y_i \geq 1) = 1 - P(Y_i = 0) = 1 - e^{-\lambda a} \\ P(Z_i = 0) &= e^{-\lambda a}. \end{aligned}$$

- 따라서 λ의 추정량을 계산하면
  $Y_i = 0$(또는 $Z_i = 0$)인 사각의 개수를 q라 하면

$$\lambda = -\frac{1}{a} \log P(Z_i = 0)$$

  이므로

$$\hat{\lambda} = \frac{1}{a} \log(\frac{q}{k})$$

  가 된다. 여기서

$$q \sim \text{Binomial}(k, e^{-\lambda a})$$

  의 분포를 따른다.

- $\hat{\lambda}$의 통계적 성질:

  편의(bias)를 가진 추정량이다.

추정량의 분산을 계산하면

$$
\begin{aligned}
\mathrm{Var}(\hat{\lambda}) &= \frac{1}{a^2}\ \mathrm{Var}\left(\log\left(\frac{q}{k}\right)\right) \\
&\left\{\begin{aligned}
&\log\left(\frac{q}{k} - e^{-\lambda a} + e^{-\lambda a}\right) \\
&= \log\left(e^{-\lambda a}\right)\cdot \log\left\{1 + \left(\frac{q}{k}e^{\lambda a} - 1\right)\right\} \\
&\approx (-\lambda a)\cdot\left\{\frac{q}{k}e^{\lambda a} - 1\right\}
\end{aligned}\right. \\
&= \frac{1}{a^2}\cdot\lambda^2 a^2\ \mathrm{Var}\left\{\frac{q}{k}e^{\lambda a} - 1\right\} \\
&= \lambda^2\ \frac{1}{k^2}\ e^{2\lambda a}\ k\, e^{-\lambda a}\left(1 - e^{-\lambda a}\right) \\
&= \frac{\lambda^2}{k}\, e^{\lambda a}\left(1 - e^{-\lambda a}\right) = \frac{\lambda^2}{k}\left(e^{\lambda a} - 1\right)
\end{aligned}
$$

따라서

$$\hat{\mathrm{Var}}(\hat{\lambda}) = \frac{\hat{\lambda}^2}{k}\left(e^{\hat{\lambda}a} - 1\right)$$

이다.

7.3 임의화 반응 (randomized response)

민감한 질문(sensitive question)을 해야하는 경우
전혀 무관한 무해한 질문(innocuous question)을 동시에 하여
응답자로 하여금 임의선택된 질문에 대하여 답하도록 하는 절차이다.

예를 들어

(질문 1) 뇌물을 받은 적이 있습니까 ? ① 예 ② 아니오

(질문 2) 뇌물을 받은 적이 없습니까 ? ① 예 ② 아니오

카드 7장에 (질문 1)을 다른 카드 3장에 (질문 2)를 적어 통에 넣고 임의로 1장의 카드를 선택하여 예, 아니오의 답만 기록한다. ($\pi = 0.7$)

이 때

$$p = \text{Prob}\big(\text{예}\big) = p_s\pi + (1 - p_s)(1 - \pi)$$

이고

$$p_s = \frac{p - (1 - \pi)}{2\pi - 1} \quad (\pi \neq \frac{1}{2})$$

이다.

앞의 실험을 n명에게 수행하여 m명이 예라고 대답한 경우

$$\hat{p} = \frac{m}{n}$$

이고 여기서

$$m \sim \text{Binomial}(n, p)$$

의 분포를 따른다.

- 따라서 추정량은

$$\hat{p_s} = \frac{\hat{p} - (1 - \pi)}{2\pi - 1}$$

이고

- 통계적 성질:

$$\begin{aligned}
\mathbb{E}(\hat{p_s}) &= p_s \\
\text{Var}(\hat{p_s}) &= \frac{1}{(2\pi - 1)^2}\text{Var}(\hat{p}) = \frac{1}{(2\pi - 1)^2}\frac{1}{n}p(1 - p) \\
\hat{\text{Var}}(\hat{p_s}) &= \frac{1}{(2\pi - 1)^2}\frac{1}{n}\hat{p}(1 - \hat{p})
\end{aligned}$$

이다.

다른 예제로 교재의 p244 (예 7.3)을 참조하기 바란다.