

생존분석의 추정과 비교 : 보충자료

이용희

November 30, 2018

Contents

1 생존함수와 위험함수	2
1.1 예제: 지수분포	2
1.2 예제: 와이블분포	2
2 생명표방법을 이용한 생존함수 추정	3
3 누적한계추정법에 의한 생존함수 추정	4
4 비모수적 방법을 이용한 생존함수의 비교	5
4.1 2×2 분할표에서 두 비율의 비교	5
4.2 코크란-맨텔-헨젤 검정	6
4.3 비모수적 방법을 이용한 생존함수의 비교	6

1 생존함수와 위험함수

1.1 예제: 지수분포

만약 생존시간 T 가 지수분포(Exponential distribution)을 따른다고 하자.

$$f(t; \lambda) = \lambda e^{-\lambda t}$$

위에서 λ 는 흔히 사망률, 실패율(failure rate)라고 부르며 $E(T) = 1/\lambda$ 이다.

생존함수(survaival function)와 위험함수(hazard function)은 다음과 같이 주어진다.

만약 생존시간이 지수분포를 따른다면 위험함수는 상수 λ 이므로

$$\begin{aligned} S(t) &= P(T > t) = 1 - P(T \leq t) \\ &= 1 - \int_0^t \lambda e^{-\lambda t} \\ &= 1 - (1 - e^{-\lambda t}) \\ &= e^{-\lambda t} \\ h(t) &= \frac{f(t)}{S(t)} \\ &= \lambda e^{-\lambda t} / e^{-\lambda t} \\ &= \lambda \end{aligned}$$

1.2 예제: 와이블분포

만약 생존시간이 와이블 분포(Weibull distribution)를 따른다면 확률밀도함수는 다음과 같다.

$$f(t) = \frac{\lambda t^{\lambda-1}}{\theta^\lambda} \exp \left[- \left(\frac{t}{\theta} \right)^\lambda \right], \quad \theta, \lambda > 0, t \geq 0.$$

평균은 $E(T) = \theta \Gamma(1 + 1/\lambda)$ 이다. 여기서 $\lambda = 1$ 이면 지수분포가 된다.

이때 생존 함수는 다음과 같다.

$$\begin{aligned} S(t) &= 1 - \int_0^t \frac{\lambda t^{\lambda-1}}{\theta^\lambda} \exp \left[- \left(\frac{t}{\theta} \right)^\lambda \right] dt \\ &= \exp \left[- \left(\frac{t}{\theta} \right)^\lambda \right]. \end{aligned}$$

또한 위험함수는 다음과 같다.

$$\begin{aligned}
 h(t) &= \frac{f(t)}{S(t)} \\
 &= \frac{\frac{\lambda t^{\lambda-1}}{\theta^\lambda} \exp \left[- \left(\frac{t}{\theta} \right)^\lambda \right]}{\exp \left[- \left(\frac{t}{\theta} \right)^\lambda \right]} \\
 &= \left(\frac{\lambda}{\theta^\lambda} \right) t^{\lambda-1}.
 \end{aligned}$$

위에서 볼 수 있듯이 $\lambda = 1$ 이면 지수분포가 되고 위험함수는 상수이다. 만약에 $\lambda > 1$ 이면 위험함수는 시간에 따라 증가하며 $\lambda < 1$ 이면 위험함수는 감소한다.

2 생명표방법을 이용한 생존함수 추정

시간 $(0, \infty)$ 구간을 다음과 같은 경계선을 이용하여 $k + 1$ 개의 구간으로 나누었다고 가정하자. ($t_0 = 0$)

$$(t_0, t_1] \quad (t_1, t_2] \quad \dots \quad (t_{k-1}, t_k] \quad (t_k, \infty)$$

생존함수는 다음과 같은 조건부 확률의 축차식으로 구할 수 있다.

$$\begin{aligned}
 S(t_i) &= p(T > t_i) \\
 &= P(T > t_i | T > t_{i-1}) P(T > t_{i-1}) \\
 &= P(T > t_i | T > t_{i-1}) P(T > t_{i-1} | T > t_{i-2}) P(T > t_{i-2}) \\
 &= \dots \\
 &= P(T > t_i | T > t_{i-1}) P(T > t_{i-1} | T > t_{i-2}) P(T > t_{i-2}) \dots P(T > t_2 | T > t_1) P(T > t_1)
 \end{aligned}$$

생명표방법을 이용한 생존함수 추정은 표를 이용하면 편리하다. 다음은 표를 생명표방법을 이용하여 생존함수를 추정하는 예를 보여준다.

구간	위험그룹인원수	사망자 수	중도절단 수	유효인원수	사망율	생존율	생존함수 추정
I_i	n_i	d_i	c_i	n'_i	$\frac{d_i}{n'_i}$	$1 - \frac{d_i}{n'_i}$	$\hat{S}(t_i)$
0-1	126	47	19	116.5	0.40	0.60	0.60
1-2	60	5	17	51.5	0.10	0.90	0.54
2-3	38	2	15	30.5	0.07	0.93	0.50
3-4	21	2	9	16.5	0.12	0.88	0.44
4-5	10	0	6	7.0	0.00	1.00	0.44

위에서 유효인원수는 다음과 같이 계산한다.

$$n'_i = n_i - \frac{c_i}{2}$$

또한 생존함수의 추정식은는 조건부 확률의 축차식을 이용하여 다음과 같이 계산한다.

$$\hat{S}(t_i) = \prod_{k=1}^i \hat{p}_k = \prod_{k=1}^i \left(1 - \frac{d_i}{n'_i}\right)$$

3 누적한계추정법에 의한 생존함수 추정

표본으로 추출한 생존시간들을 순서대로 나열한 다음 누적한계추정법은 생존함수를 다음의 식으로 추정한다.

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_k}{n_k}\right)^{\delta_i} = \prod_{t_i \leq t} \left(\frac{n_k - d_k}{n_k}\right)^{\delta_i}$$

다음 주어진 자료로부터 생존함수를 추정하는 방법을 알아보자. +는 중도절단된 자료를 표시한다.

9, 13, 13+, 18, ,23 ,28+, ,31, ,34, ,45+, ,48, ,161+

누적한계추정법은 생존함수를 다음과 같이 축차적으로 구할 수 있다.

$$\begin{aligned}\hat{S}(0) &= 1 \\ \hat{S}(9) &= \hat{S}(0) \times \frac{10}{11} = 0.91 \\ \hat{S}(13) &= \hat{S}(9) \times \frac{9}{10} = 0.82 \\ \hat{S}(18) &= \hat{S}(13) \times \frac{7}{8} = 0.72 \\ \hat{S}(23) &= \hat{S}(18) \times \frac{6}{7} = 0.61 \\ \hat{S}(31) &= \hat{S}(23) \times \frac{4}{5} = 0.49 \\ \hat{S}(34) &= \hat{S}(31) \times \frac{3}{4} = 0.37 \\ \hat{S}(48) &= \hat{S}(34) \times \frac{1}{2} = 0.18\end{aligned}$$

4 비모수적 방법을 이용한 생존함수의 비교

4.1 2×2 분할표에서 두 비율의 비교

다음과 같은 2×2 분할표에서 두 비율을 비교한다고 가정하자.

처리/반응여부	반응	반응 안함	합계
1	a	b	n_1
2	c	d	n_2
합계	m_1	m_2	n

처리 1과 2가 서로 독립인 집단에 적용되었다면 두 집단의 반응 비율이 같다는 가설 $H_0 : p_1 = p_2$ 를 다음과 같은 통계량으로 검정할 수 있다.

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}}$$

위에서

$$\hat{p}_1 = \frac{a}{n_1} \quad \hat{p}_2 = \frac{c}{n_2} \quad \hat{p} = \frac{m_1}{n}$$

n 이 충분히 크면 귀무가설 하에서 통계량 z 는 정규분포를 따르므로 $|z| > z_{\alpha/2}$ 이면 귀무가설을 기각한다.

카이제곱 분포의 정의에 의하여 통계량 z 의 제곱 $\chi^2 = z^2$ 은 자유도가 1인 카이제곱분포를 따른다.

이때 통계량 z 의 제곱은 2×2 분할표에서 동일성 검정에 대한 카이제곱 통계량(교과서 46 페이지)과 동일하다.

$$\chi^2 = z^2 = \frac{n(ad - bc)^2}{n_1 n_2 m_1 m_2}$$

만약에 귀무가설이 참이고 주변합 n_1, n_2, m_1, m_2 가 고정되어 있다고 가정하면 첫번째 행, 첫번째 열의 값 $A = a$ 의 분포는 다음과 같은 초기하분포(Hypergeometric distribution)를 따른다.

$$P(A = a) = \frac{\binom{n_1}{a} \binom{n_2}{m_1 - a}}{\binom{n}{m_1}}$$

위의 초기하분포에서 $A = a$ 의 기대값과 분산은 다음과 같다.

$$E_0(A) = \frac{n_1 m_1}{n}, \quad Var_0(A) = \frac{n_1 n_2 m_1 m_2}{n^2 (n - 1)}$$

따라서 다음이 성립하며

$$ad - bc = n[a - E_0(A)], \quad n_1 n_2 m_1 m_2 = n^2(n-1)Var_0(A)$$

이제 2×2 분할표에서 동일성 검정에 대한 카이제곱 통계량은 초기하분포와 다음과 같은 관계를 가진다.

$$\chi^2 = z^2 = \frac{n(ad - bc)^2}{n_1 n_2 m_1 m_2} = \frac{n}{n-1} \left[\frac{a - E_0(A)}{\sqrt{Var_0(A)}} \right]^2$$

4.2 코크란-멘텔-헨젤 검정

교과서 53-55 페이지 참조

전체 자료가 다음과 같이 K 개의 2×2 분할표로 나누어 진다면

처리/반응여부	반응	반응 안함	합계
1	n_{k11}	n_{k12}	n_{k1+}
2	n_{k21}	n_{k22}	n_{k2+}
합계	n_{k+1}	n_{k+2}	n_k

전체 자료에 대하여 처리의 반응률에 대한 다음과 같은 가설검정을

$$H_0 : p_{k1} = p_{k2}, \quad k = 1, 2, \dots, K \quad \text{vs.} \quad H_0 : \text{not } H_0$$

다음과 같은 코크란-멘텔-헨젤 검정 통계량(CMH 검정 통계량)으로 검정할 수 있다.

$$Q_{CMH} = \frac{[\sum_{k=1}^K n_{k11} - \sum_{k=1}^K E(n_{k11}|H_0)]^2}{\sum_{k=1}^K Var(n_{k11}|H_0)}$$

위의 통계량에서 각 2×2 분할표에 대한 기대값과 분산은 다음과 같이 계산할 수 있다.

$$E(n_{k11}|H_0) = \frac{n_{k1+}n_{k+1}}{n_k} \quad Var(n_{k11}|H_0) = \frac{n_{k1+}n_{k2+}n_{k+1}n_{k+2}}{n_k^2(n_k - 1)}$$

CMH 검정 통계량 Q_{CMH} 은 귀무가설 H_0 가 옳은 경우에 자유도 1을 가지는 카이제곱 분포를 따른다.

4.3 비모수적 방법을 이용한 생존함수의 비교

두 개의 독립 집단에 대하여 다음과 같이 생존시간을 관측하였다고 하자.

group 1 $(X_{11}, \delta_{11}), (X_{12}, \delta_{12}), \dots, (X_{1n_1}, \delta_{1n_1})$

group 2 $(X_{21}, \delta_{21}), (X_{22}, \delta_{22}), \dots, (X_{2n_2}, \delta_{2n_2})$

두 개의 집단에 대한 생존함수가 동일하다는 다음 가설을 고려하자.

$$H_0 : S_1 = S_2 \quad H_1 : S_1 \neq S_2$$

위의 가설은 두 집단의 생존시간을 모두 합쳐서 순서대로 나열하고 중도절단이 없는 자료들에서 다음과 같은 2×2 분할표를 작성한 다음 CMH-검정 통계량을 이용하여 검정할 수 있다.

처리/반응여부	사망	생존	합계
1	a	b	n_1
2	c	d	n_2
합계	m_1	m_2	n

다음과 같은 예제 자료를 고려해보자

[group 1] 3, 5, 7, 9+, 18

[group 2] 12, 19, 20, 20+, 33+

두 표본을 합쳐서 순서대로 놓으면 다음과 같다.

3, 5, 7, 9+ 12, 18, 19, 20, 20+, 33+

이제 중도절단이 없는 자료들(3,5,7,12,18,19,20)에 대하여 각각 2×2 분할표를 작성하고 CMH-검정 통계량을 계산할 수 있다. 각 분할표와 관련 통계량을 다음과 같은 표로 정리할 수 있다.

이제 다음과 같이 CMH 통계량을 계산할 수 있다.

$$CMH \chi^2 = \frac{[\sum(a - E_0(A))]^2}{\sum[m_1 m_2 / (n - 1)][n_1 n_2 / n^2]}$$

여기서

X	n	m_1	n_1	a	$E_0(A)$	$a - E_0(A)$	$m_1 m_2 / (n - 1)$	$n_1 n_2 / n^2$
3	10	1	5	1	0.50	0.50	1	0.2500
5	9	1	4	1	0.44	0.56	1	0.2469
7	8	1	3	1	0.38	0.62	1	0.2344
12	6	1	1	0	0.17	-0.17	1	0.1389
18	5	1	1	1	0.20	0.80	1	0.1600
19	4	1	0	0	0	0	1	0
20	3	1	0	0	0	0	1	0

$$\begin{aligned}\sum(a - E_0(A)) &= 0.50 + 0.56 + 0.62 - 0.17 + 0.80 \\ &= 2.31\end{aligned}$$

$$\begin{aligned}\sum[m_1 m_2 / (n - 1)][n_1 n_2 / n^2] &= (1)(0.2500) + (1)(0.2469) + (1)(0.2344) \\ &\quad + (1)(0.1389) + (1)(0.1600) + (1)(0) + (1)(0) \\ &= 1.0302\end{aligned}$$

따라서

$$CMH \chi^2 = \frac{(2.31)^2}{1.0302} = 5.1796$$

유의수준 $\alpha = 0.05$ 에서 $\chi^2(1, 0.95) = 3.84159 < 5.1796$ 이므로 H_0 를 기각한다. 즉 두 집단의 생존 함수는 같지 않다.