Einführung in die Überlebensanalyse, Teil 2

Patrizia Fedi

02.05.06

1 Zensierung (Censoring)

1.1 Rechtsseitige Zensierung

1.1.1 Typ I Zensierung

Oftmals wird aus Zeit- oder Kostengründen eine Studie beendet, bevor das interessierende Ereignis bei allen Teilnehmern eingetroffen ist. Bei diesem Vorgang spricht man von *Typ I Zensierung.*
Sei C_r definiert als der fest gewählte Zeitpunkt des Abbruchs der Studie. Die Lebensdauer eines Teilnehmers ist zufällig und wird mit X bezeichnet. Die exakte Lebensdauer eines Patienten ist also genau dann bekannt, wenn $X < C_r$ gilt.
Ist $X \geq C_r$, dann wird die Zeit dieses Patienten gleich C_r gesetzt.

Sei x eine Realisierung von X. Üblicherweise wird x durch ein Paar (t, δ) repräsentiert. Dabei ist $t = \min(x, C_r)$ und

$$\delta = \begin{cases} 1, \text{ falls die exakte Lebensdauer bekannt ist,} \\ 0, \text{ falls die Zeit zensuriert ist,} \end{cases}$$

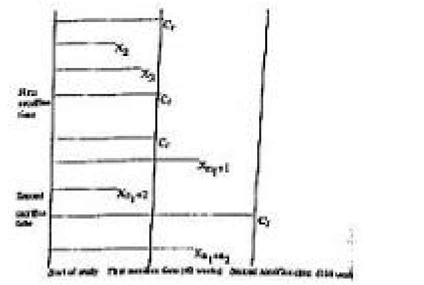
also $\delta = 1_{\{x < C_r\}}$.

Solche Paare bilden die Daten eines rechtsseitig zensurierten Experimentes.

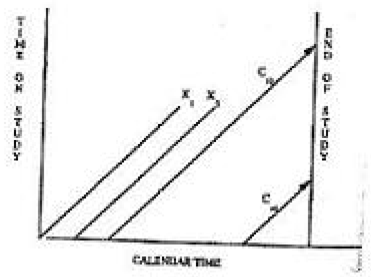
Es gibt zwei Arten von Typ I Zensierung.

1. *Schrittweise Typ I Zensierung* : Diese Form wird bei Tieren angewendet. Es werden unterschiedliche, aber feste Zensierungs-Zeiten eingesetzt. Dies soll z.B. Auskunft geben über den natürlichen Verlauf von nicht tödlichen Krankheiten.

2. *Verallgemeinerte Typ I Zensierung* : Hierbei geht es um Studien mit festgelegter Endzeit, bei denen die Individuen jedoch zu unterschiedlichen Zeiten beitreten. Der Zensierungs-Zeitpunkt wird bei Eintritt bekannt. Diese Form von Zensierung wird üblicherweise durch ein Lexis-Diagramm präsentiert.



Typ I Zensierung mit zwei verschiedenen Zensierungszeiten



Lexis-Diagramm für verallgemeinerte Typ I Zensierung

1.1.2 Typ II Zensierung

Bei Zensierung vom Typ II läuft die Studie bis zum Ausscheiden der ersten r von n Teilnehmer. Dabei ist $r < n$ zu Beginn festgelegt. Die Startzeit ist bei allen Teilnehmern dieselbe. Da es sehr lange dauern kann, bis das gewünschte Ereignis bei jedem Teilnehmenden eintritt und dieser ausscheidet, können durch dieses Verfahren Zeit und Kosten gespart werden.

Bemerkung. *Die Anzahl Ausscheidungen r und die Anzahl zensurierter Beobachtungen $n - r$ sind fest vorgegeben. Die r-te beobachtete Lebensdauer $X_{(r)} =: C_r$, d.h. die Dauer des Experiments, ist jedoch zufällig.*

Müssen Teilnehmer aufgrund von Ereignissen entlassen werden, die nichts der Studie zu tun haben, so spricht man von *Zufalls-Zensierung* . Diese tritt häufig in Kombination mit Zensierung vom Typ I auf.

1.2 Linksseitige Zensierung und Intervall-Zensierung

Bei einigen Studienteilnehmern kann es vorkommen, dass das interessierende Ereignis bereits eingetroffen ist, bevor diese Person in der Studie beobachtet wurde. Der genaue Ereignis-Zeitpunkt ist jedoch nicht bekannt.

Der linksseitige Zensierungs-Zeitpunkt C_l ist Startzeit der Beobachtungsphase, also fest gewählt. Die Lebensdauer X ist zufällig und genau dann bekannt, wenn $X > C_l$ gilt.

Die Daten einer Stichprobe werden wiederum durch Paare (t, ϵ) repräsentiert. Für eine Realisierung x von X ist dabei $t = \max(x, C_l)$ und $\epsilon = 1_{\{x > C_r\}}$.

Oft tritt in einem Experiment sowohl links-, als auch rechtsseitige Zensierung auf. Die Lebensdaten heissen dann *doppelt zensuriert* .
Die Daten werden auch hier durch ein Paar (t, μ) dargestellt. Es ist
$t = \max[\min(x, C_r), C_l]$ und

$$\mu = \begin{cases} 1, \text{ falls t der Todeszeitpunkt ist,} \\ 0, \text{ falls die Beobachtung rechtsseitig zensuriert ist,} \\ -1, \text{ falls die Beobachtung linksseitig zensuriert ist.} \end{cases}$$

Bei *Intervall-Zensierung* ist nur bekannt, dass das Ereignis innerhalb eines Intervalls $(L, R]$ aufgetreten ist.

Bemerkung. *Jede Kombination von Rechts-, Links- und Intervall-Zensierung kann in einer Studie vorkommen.*

Bemerkung. *Intervall-Zensierung ist eine Verallgemeinerung von links- und rechtsseitiger Zensierung.*

Beispiel. Bei einer Studie zum Gebrauch von Marihuana wurde nach dem Zeitpunkt des ersten Konsums gefragt. Neben der exakten Angabe des Zeitpunktes gab es folgende Antwortmöglichkeiten:

- 'Ich habe bereits konsumiert, kann aber nicht mehr sagen wann.'
  $\rightarrow$ Dies entspricht einem linksseitig zensurierten Ereigniszeitpunkt.

- 'Ich habe bisher noch kein Marihuana konsumiert.'
  $\rightarrow$ Diese Antwort ist eine rechtsseitig zensurierte Beobachtung.

Insgesamt erhalten wir ein doppelt zensuriertes Stichproben-Schema.

Beispiel. Wird bei einem Arztbesuch eine Krankheit festgestellt, so ist nur bekannt, dass der Ausbruch der Krankheit irgendwo zwischen den beiden Arztbesuchen lag. Die Beobachtung ist Intervall-zensuriert.

2 Stutzen (Truncation)

Beim *Stutzen* wird eine Bedingung definiert, die gewisse Individuen von den Untersuchenden abschirmt. Patienten werden nur unter bestimmten Umständen in die Studie aufgenommen. Erfüllen sie die Bedingung nicht, werden sie nicht beachtet.

Sei X der zufällige Zeitpunkt des Eintreffens des interessierenden Ereignisses beim Patienten. Die Entscheidung, wer in die Studie gelangt, werde zum Zeitpunkt Y gefällt ($\rightarrow$ das Stutzen findet zum fest gewählten Zeitpunkt Y statt).

Bei *linksseitig gestutzten Stichproben* werden nur Patienten mit $X > Y$ miteinbezogen.

Bemerkung. *Im Gegensatz zur linksseitigen Zensierung, bei der wir eine Teilinformation über die Teilnehmer erhalten, werden beim linksseitigen Stutzen die Individuen gar nicht erst in die Studie aufgenommen.*

Bei *rechtsseitig gestutzten Stichproben* werden nur Patienten berücksichtigt, die ein bestimmtes Ereignis bereits hinter sich haben ($X < Y$).

Beispiel. Bei einer AIDS-Studie werden nur Patienten aufgenommen, bei denen die Krankheit ausgebrochen ist ($\rightarrow$ rechtsseitiges Stutzen). Die Anzahl der HIV-infizierten Personen ist unbekannt, da diese Personen nicht erfasst werden.
Wird jedoch getestet, wie lange sich der Ausbruch von AIDS durch ein Medikament herauszögern lässt, werden die HIV-infizierten Personen berücksichtigt. Wer bereits AIDS hat, kann an der Studie nicht teilnehmen ($\rightarrow$ linksseitiges Stutzen).

3 Konstruktion der Likelihood

Bemerkung. *Wir gehen im Folgenden davon aus, dass die Lebensdauer X und die Zensierungs-Dauer C_r unabhängig sind.*

Für die Likelihood der Daten benutzen wir folgende Komponenten.

- bei exakter Lebensdauer x: $f(x)$
  ($x \in D$:=Menge der exakt bekannten Todeszeitpunkte)
- bei rechtsseitig zensurierten Beobachtungen x: $S(C_r)$
  ($x \in R$:=Menge der rechtsseitig zensurierten Beobachtungen)
- bei linksseitig zensurierten Beobachtungen x: $1 - S(C_l)$
  ($x \in L$:=Menge der linksseitig zensurierten Beobachtungen)
- bei linksseitig gestutzten Beobachtungen x: $f(x)/S(Y)$
- bei rechtsseitig gestutzten Beobachtungen x: $f(Y)/[1 - S(Y)]$
- bei Intervall-zensurierten Beobachtungen x: $[S(L) - S(R)]$
  ($x \in I$:=Menge der Intervall-zensurierten Beobachtungen)

Die Likelihood-Funktion wird nun gebildet, indem die Komponenten zusammengefügt werden.

$$L(x_1, \ldots, x_n) = \textstyle\prod_{i=1}^n L(x_i)$$

$$= \prod_{i:x_i \in D} f(x_i) \prod_{i:x_i \in R} S(C_r) \prod_{i:x_i \in L} [1 - S(C_l)] \prod_{i:x_i \in I} [S(L) - S(R)] \qquad (1)$$

Im Fall von Typ I Zensierung mit Zensierungs-Zeitpunkt C_r gilt für eine Realisierung x von X:

$$L(x) = \begin{cases} f(x), \text{ falls } x < C_r, \\ S(C_r), \text{ falls } x \geq C_r. \end{cases} \qquad (2)$$

Mit $t = \min(x, C_r)$ und $\delta = 1_{\{x < C_r\}}$ erhalten wir aus (2):

$$L(t) = \big(f(t)\big)^{\delta} \big(S(t)\big)^{1-\delta}.$$

Bei einer n-Zufallsstichprobe mit Daten (t_i, δ_i) können wir die Likelihood demnach als

$$L(t_1, \ldots, t_n) = \prod_{i=1}^{n} L(t_i) = \prod_{i=1}^{n} \big(f(t_i)\big)^{\delta_i} \big(S(t_i)\big)^{1-\delta_i} \tag{3}$$

schreiben, was dasselbe ist wie (1).

Beispiel. Seien $X_1, \ldots, X_n$ i.i.d., $X_i \sim NE(\lambda)$. Dann gilt $f(x_i) = \lambda e^{-\lambda x_i}$ und $S(x_i) = \mathbb{P}[X_i > x_i] = e^{-\lambda x_i}$, $(i = 1, \ldots, n)$.
Für eine n-Zufallsstichprobe erhalten wir mit (3) also folgende Likelihood:

$$L(t_1, \ldots, t_n) = \prod_{i=1}^{n} \big(\lambda e^{-\lambda t_i}\big)^{\delta_i} \big(e^{-\lambda t_i}\big)^{1-\delta_i} = \prod_{i=1}^{n} \lambda^{\delta_i} e^{-\lambda t_i} = \lambda^{a} e^{-\lambda s},$$

wobei $a := \sum_{i=1}^{n} \delta_i$ und $s := \sum_{i=1}^{n} t_i$.
Dabei kann man a als die Anzahl der ausgeschiedenen Patienten und s als die gesamte Testzeit betrachten.