Ανάλυση Επιβίωσης

Επικ. Καθ. Σ. Ζημερας
Τμήμα Στατιστικής και Αναλογιστικών – Χρηματοοικονομικών Μαθηματικών
Πανεπιστήμιο Αιγαίου
Σάμος
2020

Συναρτήσεις επιβίωσης

Έστω ότι έχουμε μια ομάδα $\boldsymbol{t}_i$ που για κάθε i αναπαριστούν τους χρόνους επιβίωσης στην κάθε περίπτωση από το δείγμα που έχουμε συλλέξει.

Αν θεωρήσουμε μια τυχαία και συνεχής μεταβλητή $\boldsymbol{T \epsilon R^+}$ έτσι ώστε κάθε $\boldsymbol{t}_i$ να είναι μια συγκεκριμένη τιμή που αυτή παίρνει τότε:

Η **συνάρτηση πυκνότητας πιθανότητας θανάτου** (σ.π.π. ή probability density function), ορίζεται ως :

$$f(t) = \lim_{\Delta t \to 0^+} \frac{Pr(t < T < t + \Delta t)}{\Delta t}$$

Συναρτήσεις επιβίωσης

- Ειδικότερα για συνεχείς τ.μ. ορίζεται ως

$$f(t)=\frac{dF(T)}{dt}=\lim_{\Delta t\to 0}\frac{p(t\le T\le t+\Delta t)}{\Delta t}=\lim_{\Delta t\to 0}\frac{F(t+\Delta t)-F(t)}{\Delta t},t\ge 0$$

- Για διακριτές τ.μ. ορίζεται ως

$$f(t)=P(T=t)=\begin{cases} f_j,\ t=a_j,\ j=1,2,...,n \\ 0,\ t\neq a_j,\ j=1,2,...,n \end{cases}$$

Η ποσότητα f(t)Δt, για Δt μικρό, εκφράζει την πιθανότητα αποτυχίας στο διάστημα [t,t+Δt).

Η σ.π.π. έχει τις εξής ιδιότητες : $f(t)>0$

$$\sum_{i=0}^{\infty} f(t)dt=1$$

Συναρτήσεις επιβίωσης

- Είναι γνωστό ότι το γινόμενο f(t)dt=dF(t) δηλώνει την απειροστή πιθανότητα που στην περίπτωση της ανάλυσης επιβίωσης δίνει την «αποτυχία »,δηλαδή εδώ ο θάνατος να συμβεί στο απειροστό διάστημα [t,t+Δt).
- Η σχέση

$$f(t) = \frac{dF(T)}{dt}$$

μπορεί να θεωρηθεί ότι αποτελεί το στιγμιαίο ρυθμό θανάτων. Το διάγραμμα συνάρτησης πυκνότητας της διάρκειας ζωής f(t) ονομάζετε καμπύλη θανάτου.

Συναρτήσεις επιβίωσης

- Επίσης, η πιθανότητα ότι ένα νεογέννητο να αποβιώσει μεταξύ ηλικιών t_1 και t_2 είναι

$$P\left(t_1 \le T \le t_2\right) = \int_{t_1}^{t_2} f\left(u\right)du = F\left(t_2\right) - F\left(t_1\right)$$

- Η δεσμευμένη πιθανότητα είναι ότι ένα νεογέννητο θα αποβιώσει μεταξύ των ηλικιών t_1 και t_2 δεδομένου της επιβίωσης στην ηλικία t_1 είναι

$$P\left(t_1 \le T \le t_2 \,|\, T \ge t_1\right) = \frac{F\left(t_2\right) - F\left(t_1\right)}{1 - F\left(t_1\right)}$$

Συναρτήσεις επιβίωσης

- **Συνάρτηση κατανομής διάρκειας ζωής** ορίζεται ως

$$F(T) = P(T \leq t)$$

για κάθε 0≤t≤T, όπου δηλώνει την πιθανότητα ένα νεογέννητο να αποβιώσει έως την ηλικία T.

Συναρτήσεις επιβίωσης

Συνάρτηση επιβίωσης

Η **συνάρτηση επιβίωσης (survival function)**, συμβολίζεται με S(t) και εκφράζει την πιθανότητα ο χρόνος επιβίωσης T να είναι μεγαλύτερος της χρονικής στιγμής t :

$$S(t) = P(T > t) = \int_t^{\times} f(u)du = 1 - F(t) \qquad \text{για } \textbf{συνεχείς τ.μ.}$$

$$S(t) = \sum_{u \geq t} f(u) = \sum_{a_j \geq t} f(a_j) = \sum_{a_j \geq t} f_j \qquad \text{για } \textbf{διακριτές τ.μ..}$$

Η S(t) ισοδύναμα εκφράζει την πιθανότητα επιβίωσης μέχρι τη χρονική στιγμή t, όπου $t \in [0, \infty)$ και $S(t) \in [0,1]$

Μέσω της σχέσης της με την α.σ.κ. προκύπτει ότι η συνάρτηση επιβίωσης είναι μία φθίνουσα και συνεχής συνάρτηση του t για την οποία ισχύει : $S(0) = 1$, $\lim_{t \to +\infty} S(t) = 0$ και

$$S(t_a) \geq S(t_b) \Leftrightarrow t_a \leq t_b$$

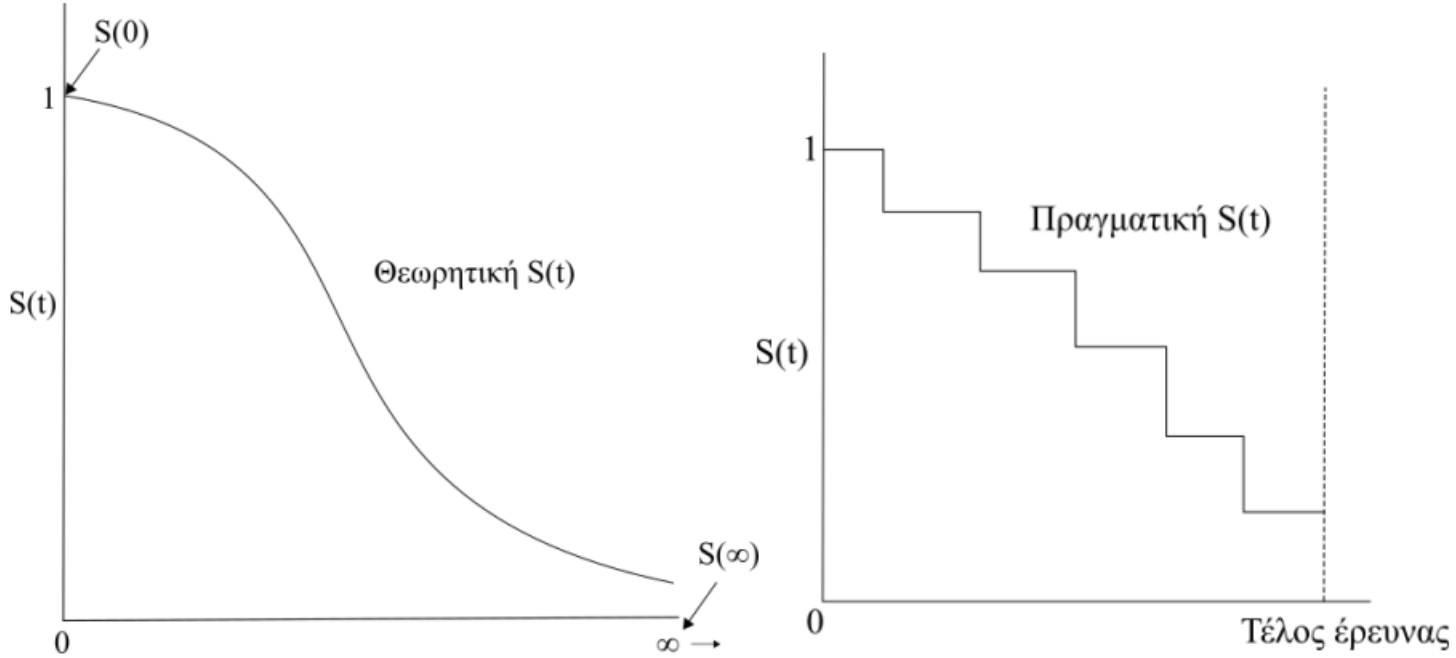
Συναρτήσεις επιβίωσης

Η γραφική παράσταση της S(t) συναρτήσει του t είναι γνωστή ως **καμπύλη επιβίωσης (survival curve)** και είναι ιδιαίτερα σημαντική στην ανάλυση δεδομένων χρόνου επιβίωσης, αφού ο υπολογισμός των πιθανοτήτων επιβίωσης για διάφορες τιμές του t μας παρέχει σημαντική πληροφόρηση για τα δεδομένα μας. Η S(t) μπορεί να εκτιμηθεί με παραμετρικούς ή μη παραμετρικούς τρόπους.

Στην θεωρία, η γραφική παράσταση της συνάρτησης επιβίωσης παρίσταται από μία λεία καμπύλη. Πρακτικά όμως, χρησιμοποιώντας δεδομένα, η μορφή της είναι μία **βηματική συνάρτηση (step function).**

Επιπροσθέτως, λόγω του ότι η περίοδος μελέτης δεν είναι ποτέ απεριόριστη σε διάρκεια και είναι πιθανόν να υπάρχουν ανταγωνιστικοί κίνδυνοι για αποτυχία, μπορεί να υπάρχουν άτομα στα οποία δεν συμβαίνει το γεγονός που μελετάται.

Συναρτήσεις επιβίωσης



Στο **γράφημα 1** παρίσταται η θεωρητική καμπύλη της συνάρτησης επιβίωσης, ενώ στο **γράφημα 2** η πραγματική βηματική συνάρτηση της συνάρτησης επιβίωσης.

Συναρτήσεις κινδύνου

Συνάρτηση κινδύνου

Η **συνάρτηση κινδύνου (hazard function)** συμβολίζεται με h(t) και εκφράζει την πιθανότητα αποβίωσης ή αλλιώς το στιγμιαίο ρυθμό αποβίωσης τη χρονική στιγμή t, δεδομένου ότι το άτομο έχει επιβιώσει μέχρι τη χρονική στιγμή t :

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t | T \ge t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{P([t \le T < t + \Delta t] \square [T \ge t])}{P(T \ge t)}$$

$$= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{P(t \le T < t + \Delta t)}{P(T \ge t)} = \frac{f(t)}{S(t)}$$

στην περίπτωση συνεχούς τ.μ., για την οποία ισχύει ότι : $h(t) \ge 0$, για κάθε $t \ge 0$ και $\int_0^{+\infty} h(t)\,dt \to +\infty$

Συναρτήσεις κινδύνου

Στην περίπτωση διακριτής τ.μ. (έστω ότι η Τ λαμβάνει τις τιμές $\alpha_1, \alpha_2, ..., \alpha_n$) η συνάρτηση κινδύνου δίνεται από την σχέση :

$$h(a_j) \,..\, h_j = P(T = a_j \,|\, T \geq a_j) = \frac{P(T = a_j)}{P(T \geq a_j)} = \frac{f(a_j)}{S(a_j)} = \frac{f(t)}{\sum_{k:\, a_k \geq a_j} f(a_k)}$$

προκύπτει η εξίσου σημαντική σχέση : $f(t) = h(t) \times S(t)$

Μέσω της μορφής της γραφικής παράστασης της συνάρτησης κινδύνου, μπορούμε να αναγνωρίσουμε συγκεκριμένα μοντέλα επιβίωσης. Για παράδειγμα, εάν πρόκειται για μία σταθερή συνάρτηση κινδύνου τότε το μοντέλο επιβίωσης είναι το εκθετικό.

Συναρτήσεις κινδύνου

Αφού $F_T(t) = \int_0^t f_T(x)dx \ , t \geq 0 \Leftrightarrow f_T(t) = F'_T(t) = -S'(t), \ t \geq 0$

$$h_T(t) = \frac{-S'(t)}{S(t)} \Leftrightarrow$$

$$-h_T(t) = \left[\ln\left(S(t)\right)\right]' \Leftrightarrow$$

$$-\int_0^t h_T(x)dx = \ln\left[S(t)\right] - \ln\left[S(0)\right] \Leftrightarrow$$

$$S(t) = \exp\left[-\int_0^t h_T(x)dx\right], t \geq 0$$

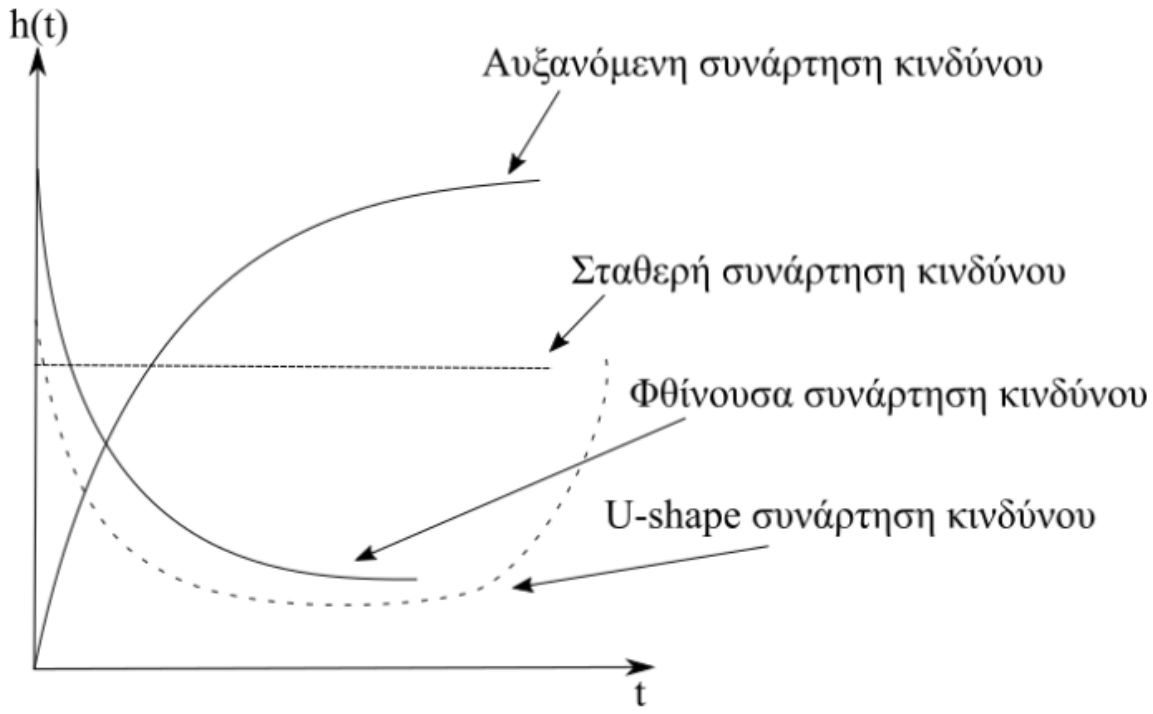
Συναρτήσεις κινδύνου

Προφανώς η συνάρτηση πυκνότητας πιθανότητας $f_T(t) = h_T(t) \cdot S(t)$ της συνεχούς μεταβλητής T γίνεται:

$$f_T(t) = h_T(t) \exp\left[-\int_0^t h_T(x) dx \right], t \geq 0,$$

δηλαδή ορίζεται μονοσήμαντα από την συνάρτηση κινδύνου $h_T(t)$.

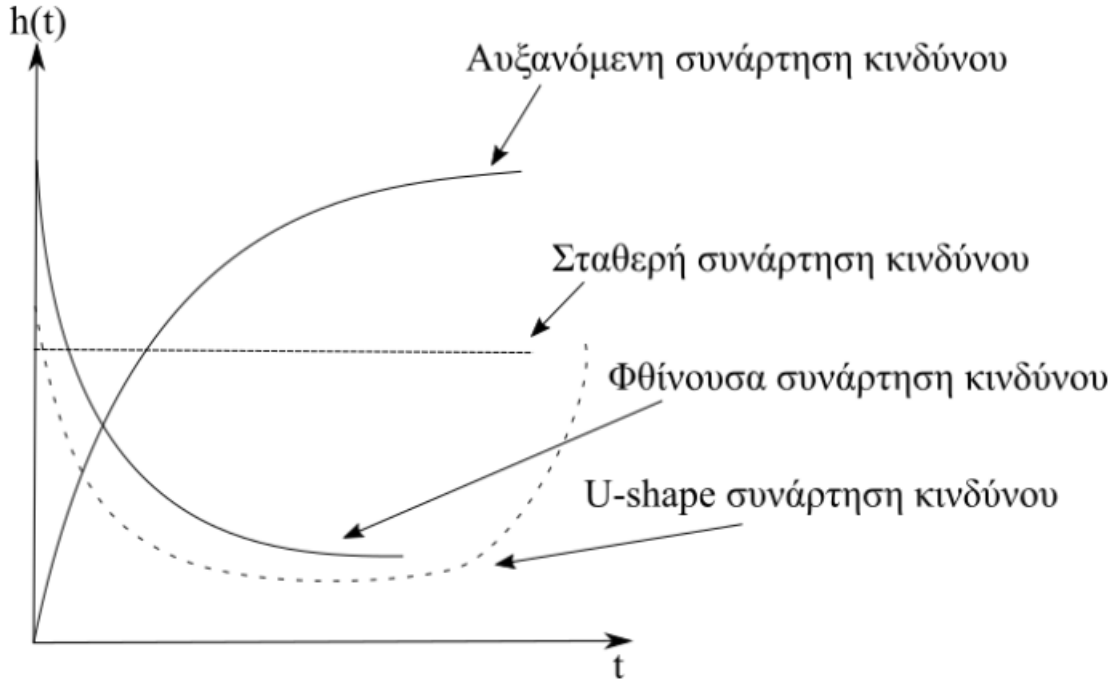
Συναρτήσεις κινδύνου



- Σταθερή συνάρτηση κινδύνου (π.χ. διαδικασία Poisson, αποτελεί μία μη ρεαλιστική περίπτωση) : $h(t) = \lambda$

- Αυξανόμενη συνάρτηση κινδύνου (π.χ. ηλικία συνταξιοδότησης) : $h(t_2) \geq h(t_1)$ αν $t_2 \geq t_1$

Συναρτήσεις κινδύνου



- Φθίνουσα συνάρτηση κινδύνου (π.χ. μετά από μία εγχείρηση) : $h(t_2) \leq h(t_1)$ αν $t_2 \geq t_1$

- U-shape συνάρτηση κινδύνου (ή αλλιώς bathtub) : Περιγράφει την ανθρώπινη θνησιμότητα από τη γέννηση μέχρι και την ηλικία θανάτου.

Συναρτήσεις κινδύνου

Παράδειγμα

Έστω ότι η συνάρτηση βαθμού κινδύνου του χρόνου λειτουργίας ενός είδους μηχανών είναι η $h_T(t)= \frac{4}{t}$, $t \geq 1$. Τότε η συνάρτηση επιβίωσης θα είναι

$$S(t) = \exp\left[-\int_1^t \frac{4}{t} dx \right] = \exp\left[-4\left(\ln t - \ln 1\right) \right] = \exp\left(-4 \ln t\right) = t^{-4} = \frac{1}{t^4}, \text{ με } t \geq 1.$$

Συναρτήσεις κινδύνου

- Χρόνος διακριτός

Στην περίπτωση της διακριτής κατανομής η συνάρτηση βαθμού κινδύνου θα είναι η $h_T(t) = \frac{P(T=t)}{P(T \geq t)} = \frac{P(T=t)}{S(t)}$, $t = 0, 1, 2, \ldots$

Ισχύει τότε:

$$h_T(t) = \frac{P(T \geq t) - P(T \geq t+1)}{P(T \geq t)} \Leftrightarrow$$

$$h_T(t)P(T \geq t) = P(T \geq t) - P(T \geq t+1) \Leftrightarrow$$
$$P(T \geq t+1) - [1 - h_T(t)]P(T \geq t) = 0$$

Η παραπάνω εξίσωση διαφορών έχει μοναδική λύση την

$$P(T \geq t) = P(T \geq 0)\prod_{i=0}^{t-1}(1 - h_T(i)), \text{ για } t = 0, 1, 2, \ldots$$

Αθροιστική συνάρτηση κινδύνου

Η **αθροιστική συνάρτηση κινδύνου (cumulative hazard function)** συμβολίζεται με H(t) και ορίζεται ως :

$$H(t) = \int_0^t h(u)du$$

για συνεχείς τ.μ.,η οποία είναι μία αύξουσα συνάρτηση του t, με $H(0) = 0$ και $\lim_{t \to +\infty} H(t) = +\infty$

ενώ σε διακριτές τ.μ. :

$$H(t) = \sum_{k:a_k<t} h_k$$

Σχέσεις συναρτήσεων

Σχέση μεταξύ συνάρτησης επιβίωσης, συνάρτησης κινδύνου και αθροιστικής συνάρτησης κινδύνου

Για μία αριστερά συνεχή συνάρτηση επιβίωσης S(t), είναι εύκολο να δειχθεί ότι :

$$f(t) = -S(t) \quad \text{ή} \quad S(t) = -f(t)$$

Χρησιμοποιώντας τα παραπάνω μπορεί να δειχθεί :

$$-\frac{d}{dt}(\log S(t)) = -\frac{1}{S(t)}S(t) = -\frac{-f(t)}{S(t)} = \frac{f(t)}{S(t)}.$$

Επομένως ένας εναλλακτικός τρόπος να γραφεί η h(t) : $h(t) = -\frac{d}{dt}(\log S(t))$

Σχέσεις συναρτήσεων

Στην συνεχή περίπτωση ισχύει :

$$H(t)=\int_0^t h(u)\,du=\int_0^t \frac{f(u)}{S(u)}\,du=\int_0^t -\frac{d}{du}\log S(u)\,du=-\log S(t)+\log S(0)\Rightarrow S(t)=e^{-H(t)}$$

Ενώ στη διακριτή περίπτωση :

Υποθέτουμε ότι $a_j < t \le a_{j+1}$. Τότε,

$$S(t)=P(T\ge a_1, T\ge a_2,\ldots, T\ge a_{j+1})$$

$$=P(T\ge a_1)P(T\ge a_2 \mid T\ge a_1)\ldots P(T\ge a_{j+1}\mid T\ge a_j)$$

$$=(1-h_1)\ldots(1-h_j)$$

Είναι εμφανές ότι η h(.) ορίζεται αν και μόνο αν έχει οριστεί η f(.) ή η S(.) και το αντίστροφο.

Μέσοι χρόνοι

Μέσος χρόνος ζωής

Ως **μέσος χρόνος ζωής (mean survival),** ορίζεται η συνάρτηση :

$$\mu = \int_0^{\times} uf(u)du \quad \textbf{για συνεχείς τ.μ. T}$$

$$\mu = \sum_{j=1}^{n} a_j f_j \quad \textbf{για διακριτές τ.μ. T}$$

Μπορεί να αποδειχθεί ότι η διακύμανση (variance) της τ.μ. T σχετίζεται με την συνάρτηση επιβίωσης μέσω της σχέσης :

$$Var(T) = 2\int_0^{\infty} tS(t)dt - \left[\int_0^{\infty} S(t)dt\right]^2$$

Διάμεσος χρόνου ζωής

Έστω τ ο μικρότερος χρόνος τ.ω. $S(\tau_{0.5}) = 0.5$ **(1.17).** Ο τ ονομάζεται **διάμεσος χρόνου ζωής (median survival).** Παρομοίως οποιοδήποτε άλλο εκατοστημόριο μπορεί να ορισθεί.

Το 25-οστό εκατοστημόριο δίνεται από τον τύπο $\hat{S}(\tau_{0.25}) \leq 0.75$ όπου $\tau_{0.25}$ ο μικρότερος χρόνος ώστε να ισχύει η σχέση ενώ $\hat{S}(\tau_{0.75}) \leq 0.25$ το 75-οστό εκατοστημόριο.

Στην πράξη η διάμεσος μπορεί να μην αντιστοιχεί ακριβώς σε κάποιον χρόνο αποτυχίας. Σε αυτή την περίπτωση η εκτιμώμενη διάμεσος είναι ο μικρότερος χρόνος τ.ω. $\hat{S}(\tau_{0.5}) \leq 0.5$.

Λόγω του ότι η κατανομή μίας τ.μ. συνήθως δεν είναι συμμετρική (π.χ. εκθετική), συχνά χρησιμοποιούμε την διάμεσο χρόνου ζωής. Η διάμεσος χρόνο ζωής συνήθως εκτιμάται καλύτερα, ειδικά μη παραμετρικά, από τον μέσο χρόνο ζωής. Σε περιπτώσεις λογοκριμένων δεδομένων επιβίωσης, η διάμεσος είναι καταλληλότερη από τον μέσο.

Μέση υπολοιπόμενη ζωή

Ως **μέση υπολοιπόμενη ζωή (mean residual life),** ορίζεται η συνάρτηση

$$mrl(t) = E(T - t \mid T > t),\ t \geq 0$$

εκφράζει την αναμενόμενη ζωή μίας παρατήρησης

που έχει ήδη ηλικία t, δηλαδή έχει επιβιώσει ως τη χρονική στιγμή t.

Ισχύει ότι $\mu = mrl(0)$.

Για μία συνεχή τ.μ. Τ ισχύει :

$$mrl(t) = \frac{1}{S(t)} \int_t^{\infty} S(s)\, ds$$

Μέση υπολοιπόμενη ζωή

Μπορεί να αποδειχθεί ότι $\mu = E(T) = \int_0^{\times} S(s)ds$

$$\mu = E(T) = \int_0^{\infty} tf(t)dt = \int_0^{\infty} t[-S'(t)]dt = [-tS(t)]_0^{\infty} - \int_0^{\infty} t'(-S(t))dt$$

$$= -\lim_{s\to\infty} s \cdot S(s) + \int_0^{\infty} S(t)dt = \int_0^{\infty} S(t)dt$$

$$-\lim_{s\to\infty} s \cdot S(s) = 0$$

Μέση υπολοιπόμενη ζωή

Στην περίπτωση διακριτής τ.μ. Τ ισχύει :

$$mrl(t) = \frac{1}{S(t)} \sum_{t=0}^{\infty} S(t),\ t = 0,1,2,...$$

Για παράδειγμα, στην εκθετική κατανομή η mrl είναι η :

$$mrl(t) = \frac{1}{S(t)} \int_{t}^{\infty} S(x)dx = \frac{1}{e^{-\lambda t}} \int_{t}^{\infty} e^{-\lambda x} dx = -\frac{1}{\lambda e^{-\lambda t}} (0 - e^{-\lambda t}) = \frac{1}{\lambda},\ \lambda > 0.$$

Στάδια ανάλυσης επιβίωσης

Στάδιο 1 : Συλλογή των δεδομένων επιβίωσης και περαιτέρω διαχωρισμός τους σε κατηγορίες ως προς την λογοκρισία τους, εάν αυτό κρίνεται απαραίτητο.

Στάδιο 2 : Προκαταρκτική ανάλυση των δεδομένων που έχουν συλλεχθεί, όπως ο υπολογισμός διαφόρων στατιστικών μέτρων (μέση τιμή, διακύμανση, διάμεσος, ελάχιστη και μέγιστη τιμή).

Στάδιο 3 : Επιλογή κατάλληλου μοντέλου για την περιγραφή των δεδομένων. Συχνά υπάρχουν περισσότερες από μία κατανομές που περιγράφουν επαρκώς τα δεδομένα, έτσι ο ερευνητής έχει την δυνατότητα να επιλέξει όποιο πληρεί τα κριτήριά του.

Στάδια ανάλυσης επιβίωσης

Στάδιο 4 : Εκτίμηση των παραμέτρων του μοντέλου. Η ακρίβεια της εκτίμησης εξαρτάται από το πλήθος των δεδομένων καθώς και την μέθοδο που χρησιμοποιείται για την εκτίμηση.

Στάδιο 5 : Προσαρμογή του μοντέλου που έχει επιλεγεί στα δεδομένα. Σημαντικός είναι ο έλεγχος εγκυρότητας του μοντέλου, καθώς αυτό μετά την προσαρμογή του μπορεί να κριθεί κατάλληλο ή μη επαρκές.