Modelli di Durata : Introduzione

a.a. 2009/2010 - Quarto Periodo

Prof. Filippo DOMMA

Corso di Laurea Specialistica/Magistrale in Economia Applicata

Facoltà di Economia – UniCal

Richiami di Statistica

- Definizione di variabile casuale
- I momenti di una v.c.
- Le Famiglie di distribuzione
- L'Inferenza Statistica
- La Verosimiglianza
- Modello Lineare

Modello

Rappresentazione più o meno complessa della realtà

Ci permette di:

- Descrivere - Interpretare - Sintetizzare

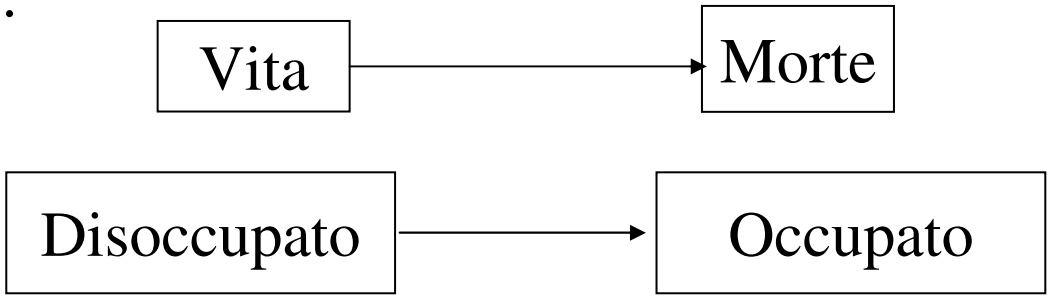
Fasi di costruzione di un modello:

-Specificazione
-Stima
-Valutazione (diagnostica)

L'obiettivo principale del corso consiste nella **specificazione** e nella **stima** di un modello che fornisca la rappresentazione della lunghezza dell'intervallo di tempo necessario affinché alcuni fenomeni economico sociali passano da uno stato ad un altro.

Gli stati vengono generalmente definiti dai valori assunti da una variabile qualitativa durante il periodo di osservazione, l'insieme di tutti i possibili valori di questa variabile viene chiamato **insieme** (o spazio) **degli stati**.

Esempio.



Definizione di Durata

I dati di durata fondamentalmente trattano con la misura della lunghezza di un intervallo di tempo tra due successive realizzazioni di un ben definito evento, spesso il cambio di stato di un individuo.

Definizione 1.

Per durata si intende il tempo necessario affinché l'evento di interesse si realizzi.

Evento di interesse: transizione (passaggio) da uno stato ad un altro

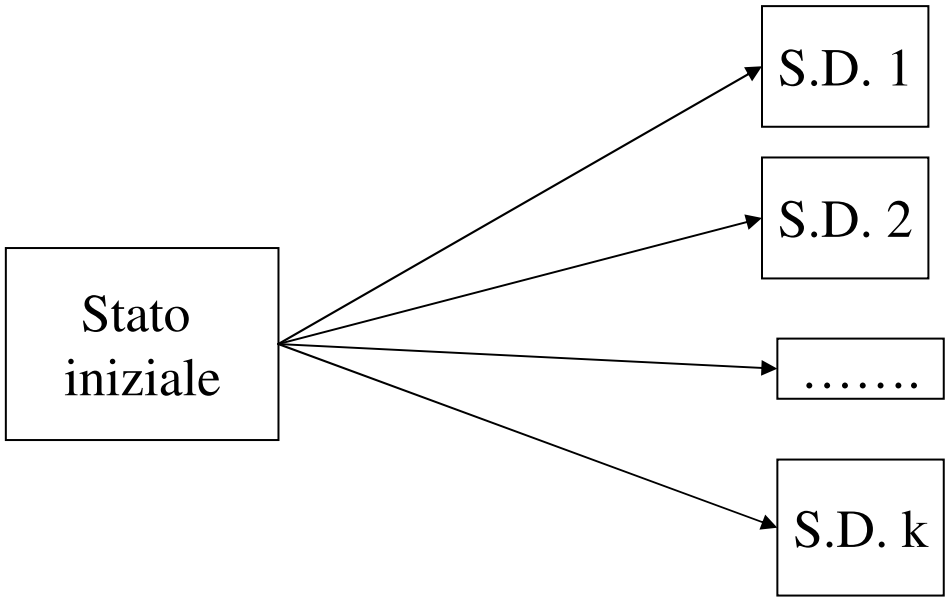
L'episodio (evento) si dice **non-ricorrente** (non ripetibile) quando può essere sperimentato al più una volta dal singolo individuo (*item*) della popolazione; al contrario, un evento si dice **ricorrente** (ripetibile) quando si può manifestare anche più di una volta nella vita di un individuo.

Una possibile Classificazione dei Modelli

1) **Episodio singolo**: un singolo episodio al termine del quale lo stato di destinazione o non viene osservato o è l'unico possibile.
Abbiamo, quindi, uno stato iniziale, un episodio ed un unico stato di destinazione

Se t_0 è il tempo nello stato iniziale e con t* indichiamo il tempo nello stato di destinazione allora la **durata** è $T=t^*-t_0$; se $t_0=0$ allora $T=t^*$.

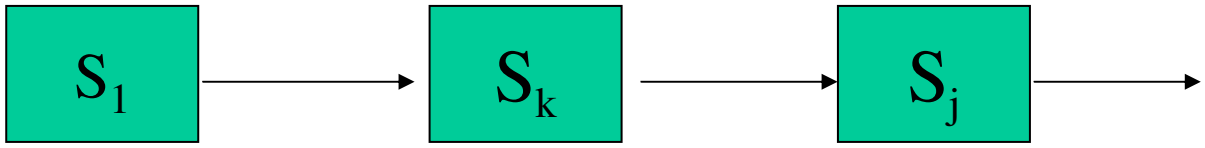
2) **multi-stati**: uno stato iniziale e più stati di destinazione (incompatibili ed esaustivi)



Processi a stati multipli. Lo stato iniziale occupato da un certo individuo può essere abbandonato per più di uno stato di destinazione (finale) esaustivi ed incompatibili, ognuno dei quali costituisce comunque uno stato assorbente. La durata per ogni individuo è unica, si ha cioè per ogni individuo una unica transizione, anche se, diversamente dal punto precedente lo stato finale occupato può essere diverso da individuo a individuo.

3) Eventi ripetibili a stati ed episodi multipli

- Ogni individuo può occupare una successione di stati. Ad ogni istante un individuo può occupare uno ed uno solo dei *k* stati possibili. Ogni volta che avviene una transizione dallo stato *k* si registra il nuovo stato *j* e il tempo di ingresso in quest'ultimo stato. La storia del processo è costituita dall'elenco degli stati successivamente occupati, dal momento di ingresso e di uscita di ognuno.



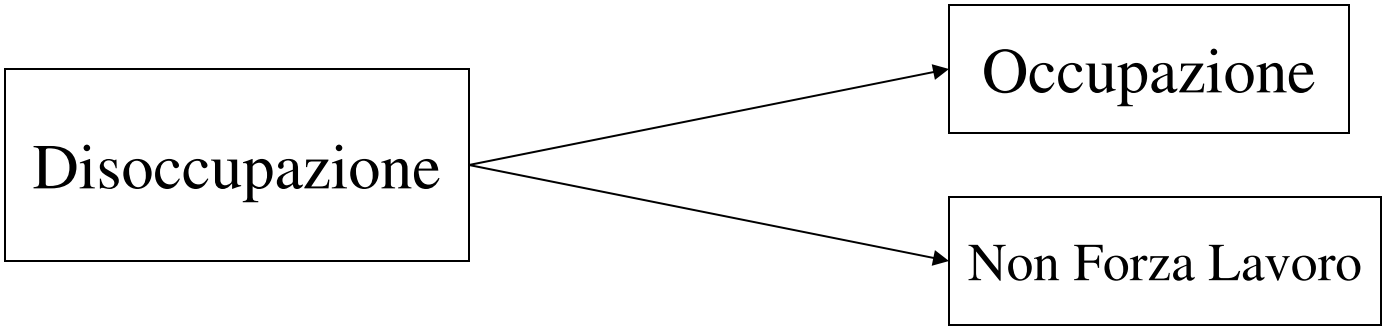
Esempio. Mobilità nel mercato del lavoro.

E' importante effettuare alcune **precisazioni** sulla **nozione di tempo** nei dati di durata poiché la sua **interpretazione non è univoca**. Infatti, basti pensare ad un esperimento clinico per rendersi conto che ci sono molti tempi da prendere in considerazione:
-Il tempo di calendario;
-Il tempo dalla nascita del paziente (età del paziente);
-Il tempo dalla diagnosi della malattia;
-Il tempo dal trattamento;
Il tempo "**più naturale**" è quello di calendario, poiché i dati vengo raccolti registrando le date in cui si verificano gli eventi di interesse. Tuttavia, il tempo di calendario solitamente è meno significativo rispetto ad altri tempi: ad esempio, in un'analisi demografica è più rilevante l'età anagrafica mentre in un esperimento clinico i tempi più importanti sono quelli della diagnosi e del trattamento. Dunque, la scelta del tempo su cui basare il modello dipende dal tipo di analisi che si vuole condurre. Comunque, è importante notare che certi tempi sono "soggettivi", in quanto l'origine dell'asse temporale è, di norma, diversa per ogni soggetto: ad esempio, il tempo della diagnosi inizia a decorrere dalla diagnosi della malattia che, di solito, avviene in una data diversa per ogni paziente. Quindi, se passiamo dal tempo di calendario (che viene usato per la raccolta dei dati) ad un tempo di tipo soggettivo effettuiamo un'operazione di riordinamento dei tempi.

I Dati di Durata: Esempi

Esempio 1. Mercato del Lavoro

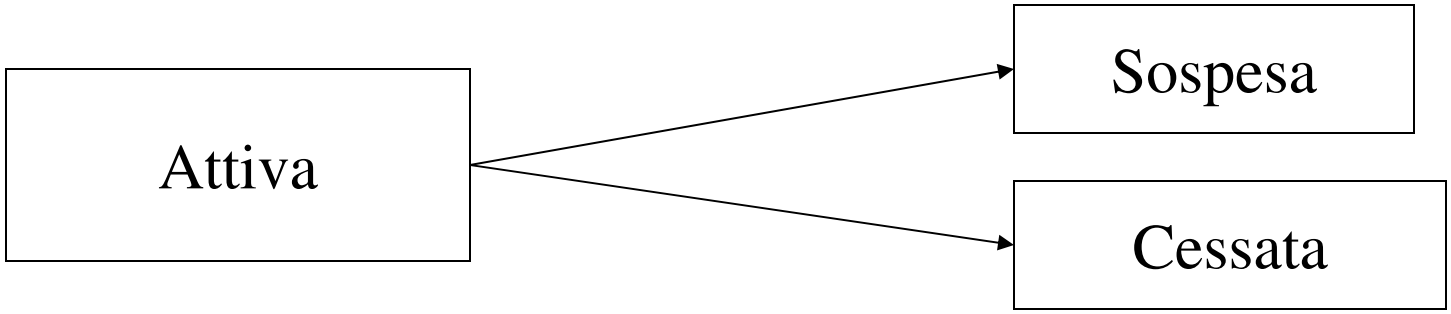
La popolazione degli individui con età > 15 anni viene suddivisa in:
-Forza Lavoro (individui Occupati + individui in cerca di occupazione)
-Non Forza Lavoro (persone che non sono alla ricerca di una occupazione).
Nel periodo (t_0,t_N), l'insieme degli stati è: {D, O, NFL}



Osservazione: la decisione di transitare da uno stato ad un altro e il tempo necessario per effettuare la transizione dipendono da un insieme di variabili legate al contesto economico (stato dell'economia, area geografica di appartenenza, …..) e un di variabili legate all'individuo (sesso, livello di istruzione, disponibilità alla mobilità, livello minimo del salario …..)

Esempio 2. Sopravvivenza delle Imprese

In un dato intervallo di tempo, consideriamo l'insieme delle imprese in una data area geografica (regione, macro regioni, nazione,…). Seguendo l'indagine dell'INPS, le imprese sono classificate in Attive, Sospese o Cessate. Di conseguenza l'insieme degli stati è: {A, S, C}



Osservazione: la sopravvivenza di una impresa dipende da una serie di variabili di contesto (area geografica, settore di attività, tassi di interesse, …..) e da un insieme di variabili legate all'impresa stessa (capacità manageriali, indicatori di bilancio, …….)

Esempio 2 (continua) Dati reali: fonte amministrativa – INPS

Settore di attività: 419 Ateco-ISTAT (Raccolta, depurazione acqua)
Provincia: Cosenza
Periodo di osservazione: dal 1/1/1989 al 31/12/1994
Registrazione mensile: Stato (Attiva, Sospesa, Cessata) e numero di lavoratori

Evento di interesse: Cessazione delle imprese
Durata: il tempo di transizione dallo stato di "Attività" a quello di "Cessazione"

Data di Creazione può essere diversa;
Nello "stato" attività si entra alla data di creazione (t_0).
Se la data di Cessazione (t*) è precedente al 31/12/1994 allora si realizza l'evento in t*
Se al 31/12/1994 l'impresa è attiva allora il dato è censurato in t**(= 31/12/1994)

CAP	Comune	Pr	aamm Creazione	aamm cessazione censura	Settore	s	L	s	L																										
87028	PRAIA A MARE	CS	5305	8903	419	1	6	1	6	3	0	3	0	3	0	3	0	3	0	3	0	3	0	3	0	3	0	3	0	3	0	3	0	3	0
87020	BELVEDERE MARITTIMO	CS	5405	9106	419	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	1	1	1	1	1	1	1
87100	COSENZA	CS	5503	**9412**	419	1	2	1	2	1	2	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
87068	ROSSANO	CS	5504	9301	419	1	5	1	5	1	5	1	5	1	4	1	4	1	4	1	4	1	4	1	4	1	3	1	4	1	4	1	3		
87030	COSENZA	CS	6108	**9412**	419	1	10	1	10	1	10	1	11	1	10	1	12	1	12	1	11	1	12	1	12	1	12	1	13	1	13				
87010	TERRANOVA DA SIBARI	CS	6305	**9412**	419	1	4	1	4	1	4	1	4	1	4	1	4	1	5	1	5	1	5	1	4	1	4	1	5	1	5				
87036	RENDE	CS	6307	**9412**	419	1	9	1	9	1	9	1	9	1	9	1	8	1	8	1	8	1	8	1	8	1	8	1	8						
87021	BELVEDERE MARITTIMO	CS	6308	8903	419	1	1	1	1	3	0	3	0	3	0	3	0	3	0	3	0	3	0	3	0	3	0	3	0	3	0				
87040	MARANO MARCHESATO	CS	6311	9011	419	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1				
87100	COSENZA	CS	6407	8912	419	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	3	0	3	0				
87022	CETRARO	CS	6508	**9412**	419	2	0	2	0	2	0	2	0	2	0	2	0	2	0	2	0	2	0	2	0	2	0	2	0	2	0				
87060	MANDATORICCIO	CS	6602	**9412**	419	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	2	1	2				
87027	PAOLA	CS	6604	**9412**	419	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1				
87040	MENDICINO	CS	6604	**9412**	419	1	7	1	7	1	7	1	7	1	7	1	8	1	9	1	9	1	9	1	9	1	8	1	8	1	8				
87040	ROSE	CS	6608	9208	419	1	4	1	3	1	3	1	5	1	4	1	4	1	4	1	5	1	5	1	5	1	5	1	5	1	4				
87040	COSENZA	CS	6701	**9412**	419	1	8	1	8	1	8	1	8	1	8	1	8	1	8	1	8	1	8	1	8	1	8	1	8	1	7				
87026	MORMANNO	CS	6701	**9412**	419	2	0	2	0	2	0	2	0	2	0	2	0	2	0	2	0	2	0	2	0	2	0	2	0	2	0				
87100	COSENZA	CS	6704	**9412**	419	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1				
87050	PANETTIERI	CS	6708	**9412**	419	1	2	1	2	1	1	1	1	1	1	1	2	1	2	1	2	1	1	1	1	1	1	1	1	1	1				
87040	CASTROLIBERO	CS	6710	**9412**	419	1	5	1	5	1	5	1	5	1	5	1	5	1	5	1	5	1	5	1	5	1	5	1	5	1	5				
87055	S.GIOVANNI IN FIORE	CS	6802	**9412**	419	1	5	1	5	1	5	1	4	1	6	1	5	1	5	1	7	1	7	1	7	1	6	1	6	1	6				
87029	SCALEA	CS	6803	**9412**	419	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	1	1	1	1	1	1	1				
87010	FRASCINETO	CS	6806	**9412**	419	1	3	1	3	1	2	1	2	1	2	1	3	1	3	1	3	1	4	1	4	1	4	1	4	1	4				

Figura 1. Inizio e fine attività (o data censurata) delle prime 16 imprese relative all'Esempio 2 (data di calendario)

N.B.: 1) La durata dell'impresa A è maggiore di quella dell'impresa B, anche se la data di cessazione di A è precedente a quella di B.

2) Quando parliamo "parità" tra due durate significa che due imprese hanno fatto registrare lo stesso tempo tra l'inizio dell'attività e quello di cessazione, ma non significa che sono cessate lo stesso giorno.

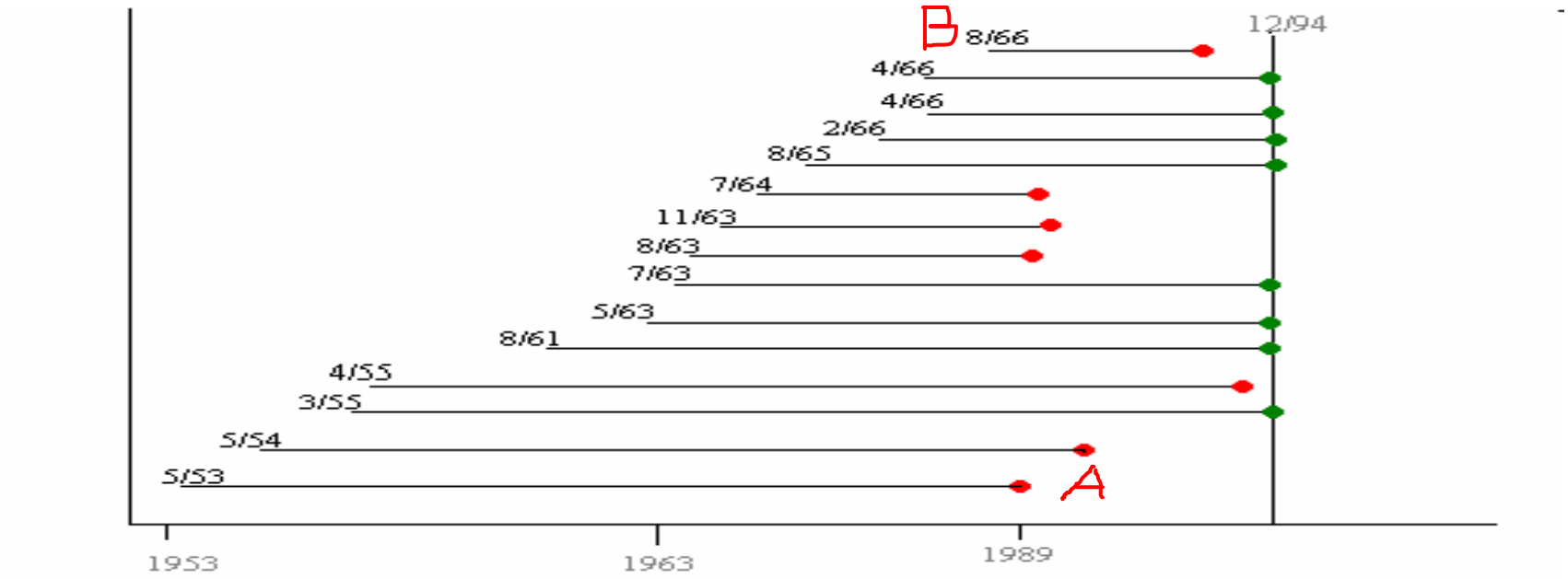
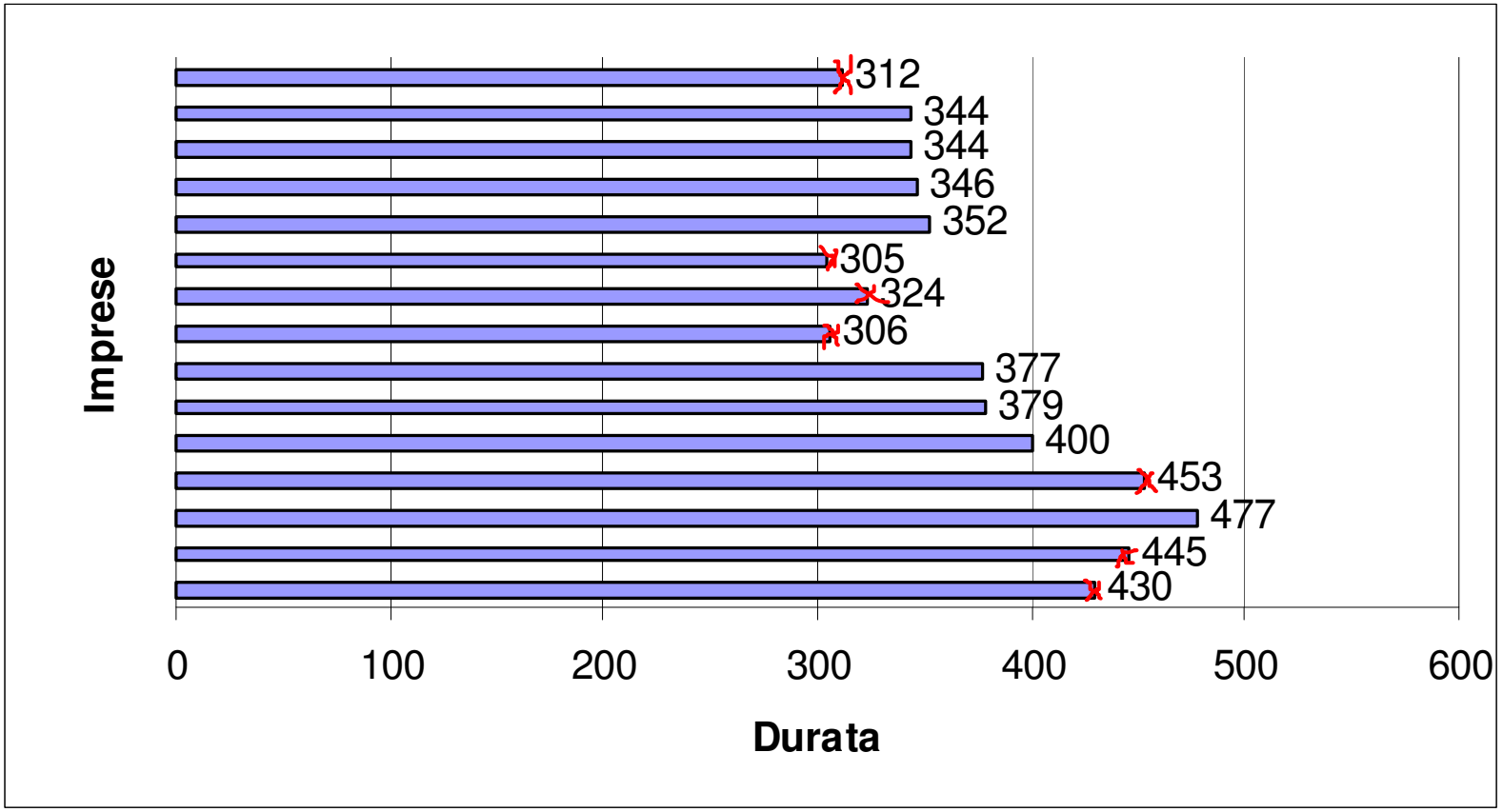


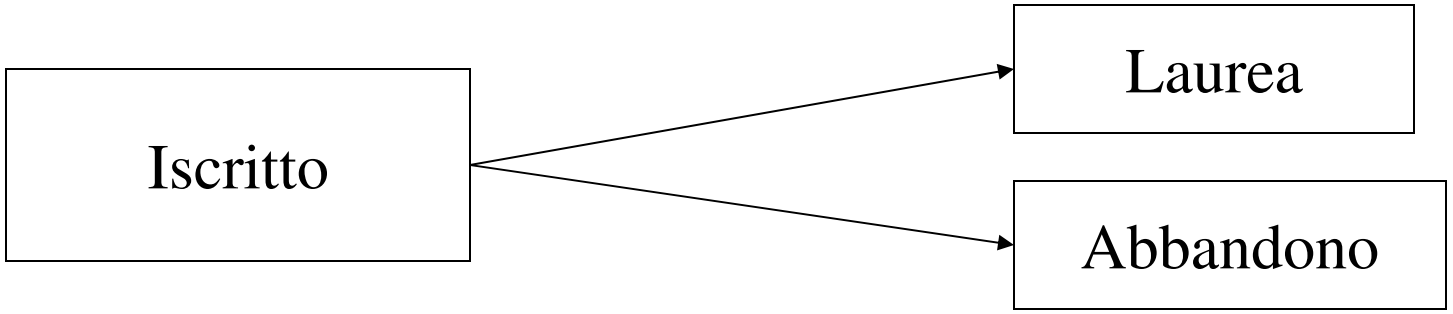
Figura 2. Grafico delle durate relative alle prime 16 imprese dell'Esempio 2.



x Complete

Esempio 3. La durata degli studi universitari

E' una situazione diversa rispetto agli esempi precedenti, perchè l'analisi si basa sulle coorti di immatricolati



Nella coorte i-esima gli individui entrano nello studio alla stessa data di calendario, t_0 è uguale per ogni immatricolato.

Alcuni Campi di Applicazione

-Mercato del lavoro (Durata della disoccupazione/occupazione);
-Sopravvivenza delle imprese;
-Comportamento dei consumatori;
-Assicurazioni ed Incidenti stradali;
-Analisi sulla famiglia: -permanenza dei figli; -1° figlio; -durata dei matrimoni;
-Criminologia e Sistema giudiziario: -ritorno nelle patrie galere; -durata dei processi;
-Durata dei Governi nella prima repubblica;
-Durata delle telefonate;
-Permanenza su una pagina web;
-Affidabilità dei materiali e dei sistemi;
-Ambito Biomedico e Demografico: -durata della vita delle persone, animali, piante,..
-Tempi di rigetto degli organi trapiantati;
-Durata della degenza ospedaliera;

Peculiarità dei dati di durata: la censura

La domanda che bisogna porsi a questo punto è la seguente:
Perché questo tipo di dati sono diversi da tutti gli altri ?
In altre parole, perché quest'area di ricerca necessita di una specifica metodologia statistica?

I motivi di fondo risiedono nel fatto che i dati di durata possono presentare informazioni parziali, rispetto all'evento di interesse, intrinseche all'indagine campionaria. Si pensi, ad esempio, ad una indagine sulla sopravvivenza delle imprese in una specifica area geografica, fissato l'intervallo di tempo entro cui effettuiamo la nostra indagine, diciamo 1980-2005, per un certo numero di aziende si avrà un'informazione completa in quanto realizzano l'evento di interesse (cessazione dell'attività entro il 2005), per altre si potrà dire solo che alla data di osservazione continuano a svolgere la loro attività (cioè si ha un'informazione parziale rispetto alla cessazione dell'attività ovverosia la durata di queste imprese va oltre la soglia limite di osservazione).
Nella letteratura statistica, questa aspetto particolare dei dati viene denominata **censura**, di conseguenza l'informazione parziale viene denominata osservazione censurata.
E' importante evidenziare che le tecniche statistiche sviluppate nel contesto dell'analisi di sopravvivenza, permettono di utilizzare nel processo inferenziale anche le suddette informazioni parziali così come vedremo nei paragrafi successivi.

Strumenti Metodologici

Sia T una v.c. continua e non-negativa con valori in [0,∞), interprete della **durata** di un generico evento (episodio), con **funzione di densità**, f(t;θ), parametrizzata da θ, cioè

$$T \sim f(t;\theta) \qquad \text{con} \qquad \theta \in \Theta \subset \Re^r$$

dove Θ è lo spazio parametrico. La **funzione di ripartizione**

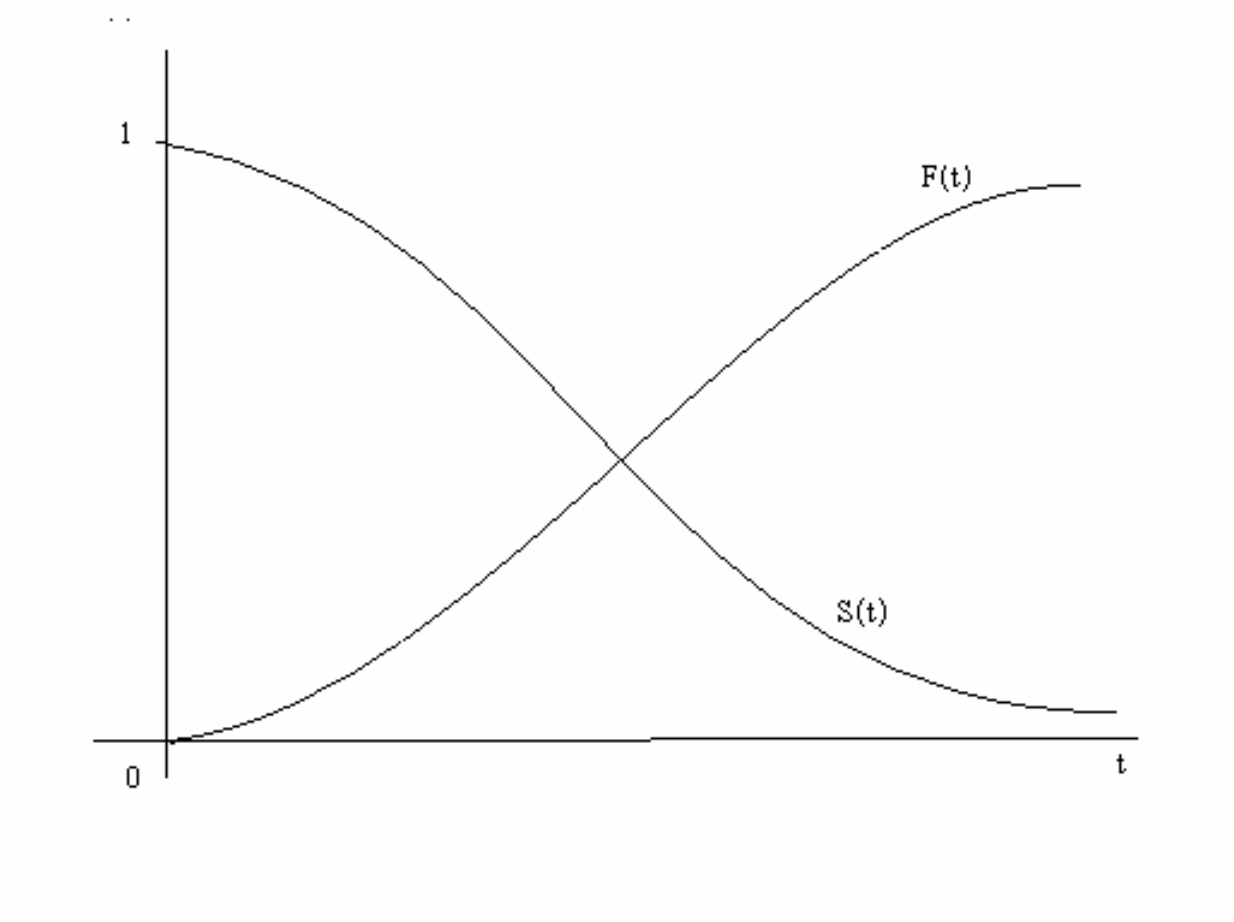
$$F(t;\theta) = P_r(T \leq t) = \int_0^t f(u;\theta)du$$

ci indica la probabilità che la durata di un episodio sia al più uguale a t. Di contro, la probabilità che la durata sia maggiore di t, detta **funzione di sopravvivenza**, è definita dalla seguente

$$S(t;\theta) = P_r(T \geq t) = 1 - F(t;\theta) = \int_t^\infty f(u;\theta)du$$

Si noti che, per ogni θ, S(t;θ) è monotona decrescente ed, inoltre, S(0;θ)=1 e S(+∞;θ)=0.

Grafico _._ Funzione di ripartizione e di sopravvivenza per T.



Un'importante funzione è la cosiddetta *hazard function*, definita da

$$h(t;\boldsymbol{\theta}) = \lim_{\Delta t \to 0} \frac{P_r\left[t \le T < t + \Delta t \,/\, T \ge t\right]}{\Delta t} = \frac{f(t;\boldsymbol{\theta})}{S(t;\boldsymbol{\theta})}$$

ci indica il tasso istantaneo che l'episodio termini nell'intervallo [t,t+Δt] dato che l'evento di interesse non si è verificato fino a t. Utilizzando la terminologia diffusa nei modelli di sopravvivenza h(t;θ) fornisce indicazioni sul tasso istantaneo di morte al tempo t, dato che l'individuo è sopravvissuto fino a t.

Osservazione: nel caso in cui la durata è ipotizzata continua, h(t; θ) non è una probabilità; infatti, si dimostra che

$$h(t;\boldsymbol{\theta}) \ge 0 \qquad \int_0^{\infty} h(t;\boldsymbol{\theta})\,dt = \infty$$

Si vedrà che nel caso discreto la *hazard function* è una probabilità condizionata.

In termini di mercato del lavoro la *hazard function* fornisce indicazioni circa il tasso istantaneo di uscita dallo stato di disoccupazione nell'intervallo $[t,t+\Delta t]$ per un individuo che risulta disoccupato fino a t.

Osservazione: la relazione che intercorre tra $h(t;\boldsymbol{\theta})$ e t è chiamata
dipendenza dalla durata

Se $\frac{\partial h(t;\boldsymbol{\theta})}{\partial t} > 0$ allora la dipendenza dalla durata è positiva
"ciò significa che la probabilità di uscita dallo stato di disoccupazione aumenta con l'allungarsi della permanenza nello stato stesso"

Se $\frac{\partial h(t;\boldsymbol{\theta})}{\partial t} < 0$ allora la dipendenza dalla durata è negativa
"l'uscita dallo stato di disoccupazione diventa meno probabile nel tempo"

Nel caso in cui il tempo è considerato una variabile casuale continua, si dimostra che tra le funzioni $f(t;\boldsymbol{\theta})$, $F(t;\boldsymbol{\theta})$, $S(t;\boldsymbol{\theta})$ e $h(t;\boldsymbol{\theta})$ esistono le seguenti relazioni matematicamente equivalenti

RE1) $$f(t;\boldsymbol{\theta}) = -\frac{\partial S(t;\boldsymbol{\theta})}{\partial t} = -S'(t;\boldsymbol{\theta})$$

RE2) $$h(t;\boldsymbol{\theta}) = \frac{f(t;\boldsymbol{\theta})}{S(t;\boldsymbol{\theta})} = -\frac{\partial \ln S(t;\boldsymbol{\theta})}{\partial t}$$

RE3) $$S(t;\boldsymbol{\theta}) = \exp\left(-\int_0^t h(u;\boldsymbol{\theta})du \right) = \exp\left(-H(t;\boldsymbol{\theta})\right)$$

dove $H(t;\boldsymbol{\theta}) = \int_0^t h(u;\boldsymbol{\theta})du$ è l'*hazard* cumulato.

RE4) $$f(t;\boldsymbol{\theta}) = h(t;\boldsymbol{\theta}) \times \exp\left(- \int_0^t h(u;\boldsymbol{\theta}) du \right)$$

data una *hazard function* è possibile risalire alla funzione di densità della v.c. T. Ciò evidenzia uno degli aspetti più importanti della funzione $h(t;\boldsymbol{\theta})$. Infatti, in questo ambito di studi, spesso si hanno delle informazioni qualitative sulla *hazard function* (del tipo: monotona decrescente, crescente; non-monotona, ecc.) del fenomeno in analisi, quindi ipotizzando una particolare forma funzionale per $h(t;\boldsymbol{\theta})$ si può determinare, tramite la relazione (RE4), la f.d. della v.c. T.

Vedremo più avanti alcuni metodi non-parametrici che permettono di costruire *hazard function* empirici (cioè direttamente dai dati osservati), dai quali è possibile estrarre informazioni sull'andamento teorico dell'*hazard function*.

Vita attesa residua

$$r(t;\boldsymbol{\theta}) = E\left[T - t / T \geq t\right] = \int_t^{\infty} (u - t)\frac{f(u;\boldsymbol{\theta})}{S(t;\boldsymbol{\theta})}du \qquad \text{per} \quad 0 \leq t < \infty$$

Con riferimento al mercato del lavoro $r(t;\boldsymbol{\theta})$ può essere interpretata come il tempo medio residuo prima di trovare una occupazione per un individuo che risulti essere disoccupato fino al tempo t.

Osservazione: se t=0 allora $r(t;\boldsymbol{\theta})=E(T)$

Si dimostra che:

RE5) $$h(t;\boldsymbol{\theta}) = \frac{r'(t;\boldsymbol{\theta}) + 1}{r(t;\boldsymbol{\theta})}$$

RE6) $$S(t;\boldsymbol{\theta}) = \exp\left(-\int_0^t h(u;\boldsymbol{\theta})du\right) = \exp\left(-\int_0^t \frac{r'(u;\boldsymbol{\theta})+1}{r(u;\boldsymbol{\theta})}du\right) = \frac{r(0;\boldsymbol{\theta})}{r(t;\boldsymbol{\theta})}\exp\left(-\int_0^t \frac{1}{r(u;\boldsymbol{\theta})}du\right)$$

Osservazione: $r(t;\boldsymbol{\theta})$ può essere utilizzate, ad esempio, per valutare l'impatto di un intervento pubblico sull'occupazione, quale la formazione professionale; infatti, l'intervento è positivo se riduce il tempo medio residuo di permanenza nello stato di disoccupazione degli individui sottoposti all'intervento rispetto a quelli non sottoposti.

Richiami di Inferenza

Nell'ambito dell'inferenza parametrica il metodo di stima usualmente utilizzato è quello di verosimiglianza, il motivo principale deriva dal fatto che gli stimatori di massima verosimiglianza, sotto opportune condizioni di regolarità, godono di proprietà ottimali (si rinvia alla letteratura specialistica per ulteriori approfondimenti). Supponiamo di estrarre un campione casuale indipendente ed identicamente distribuito (iid) dalla popolazione in esame di dimensione n, indicato con $(t_1, t_2,\ldots,t_n)$, la funzione di verosimiglianza risulta essere:

$$L(\boldsymbol{\theta}; t_1,..., t_n) = \prod_{i=1}^{n} f(t_i; \boldsymbol{\theta})$$

Al variare di $\boldsymbol{\theta}$ in $\boldsymbol{\Theta}$, $L(\boldsymbol{\theta};t_1,..,t_n)$ descrive la plausibilità (ovvero la verosimiglianza) che il campione osservato sia stato estratto dalla funzione di densità parametrizzata da $\boldsymbol{\theta}$. Evidentemente, il valore di $\boldsymbol{\theta}$ che massimizza $L(\boldsymbol{\theta};t_1,..,t_n)$ individuerà la funzione di densità che con maggiore verosimiglianza ha generato il campione. Usualmente si utilizza il logaritmo della f.v. ottenendo funzione di log-verosimiglianza

$$\ell(\boldsymbol{\theta}; t_1,...,t_n) = \ln L(\boldsymbol{\theta}; t_1,...,t_n) = \sum_{i=1}^{n} \ln f(t_i; \boldsymbol{\theta})$$

Lo stimatore di massima verosimiglianza $\hat{\boldsymbol{\theta}}$

di $\boldsymbol{\theta}$ può essere ottenuto risolvendo il sistema formato dalle derivate parziali della log-verosimiglianza rispetto alle singole componenti del vettore $\boldsymbol{\theta}$, cioè

$$\begin{cases} \dfrac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_1} = 0 \\ \dfrac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_2} = 0 \\ \ldots\ldots\ldots \\ \dfrac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_r} = 0 \end{cases}$$

Spesso per risolvere il sistema è necessario ricorrere a procedure iterative di analisi numerica, quali ad esempio il Newton-Raphson o il Fisher-scoring, i quali garantiscono, alla convergenza, una buona approssimazione della soluzione del sistema in esame.

Si dimostra che al divergere di n ad infinito

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)\xrightarrow{d}N_r\left(0,\boldsymbol{\Sigma}\right)$$

dove la matrice di varianze e covarianze asintotiche è pari all'inversa della matrice di informazione di Fisher, cioè

$$\boldsymbol{\Sigma}=\left[I(\boldsymbol{\theta})\right]^{-1}$$

L'elemento ij-esimo della matrice di informazione di Fisher è:

$$i_{ij}(\boldsymbol{\theta})=-E\left[\frac{\partial^2\ell(\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}\right]$$

per i,j=1,…r.

La verosimiglianza nel caso di dati censurati

Una caratteristica peculiare, fonte di ulteriori complicazioni nella fase di stima, dei dati relativi alla durata di un episodio è la eventualità di avere informazioni parziali sulla realizzazione dell'evento di interesse, tale situazione viene tecnicamente denominata **censura**.

In parole povere, si è in una situazione di dati censurati quando solo per alcuni individui si conosce il tempo di realizzazione dell'evento di interesse, per tutti gli altri si sa solo che esso eccede un certo valore.

Esempio.
Supponiamo di osservare l'attività (la "vita") di un gruppo di aziende in un intervallo di tempo (o,L]. Per le aziende che cesseranno la loro attività prima del limite L, la nostra informazione sarà completa in quanto l'evento di interesse (la cessazione - morte) si realizza prima del limite L. Per tutte le altre aziende siamo in possesso solo di una informazione parziale perché sappiamo solo che la loro attività va oltre il limite di tempo fissato, L (se cesseranno la loro attività succederà oltre L). In altre parole, una osservazione è censurata a destra in L se l'unica informazione che abbiamo è che la durata è maggiore o uguale ad L.
Analogamente, una osservazione è censurata a sinistra se l'inizio dell'episodio è minore o uguale ad L.

Per analizzare i dati censurati è necessario conoscere il meccanismo effettivo che genera la censura. Infatti, la censura può concretizzarsi per diverse ragioni, alcune delle quali verranno descritte di seguito.
L'obiettivo principale è quello di determinare la distribuzione campionaria e, quindi, la funzione di verosimiglianza nei diversi tipi di censura

Censura di I tipo

Alcuni esperimenti (osservazioni) vengono effettuati fino ad un fissato periodo di tempo, così la durata risulta nota solo per gli *items* che realizzano l'evento di interesse prima del prefissato valore.
Supponiamo, ad esempio, di osservare nel tempo l'attività di n aziende, indichiamo con T_i la durata della azienda i-esima e per ognuna di queste stabiliamo un tempo massimo di osservazione (tempo prefissato di osservazione), L_i.

Per la i-esima azienda, osserveremo la durata T_i se $T_i \leq L_i$

altrimenti la durata è censurata a destra in L_i.

I dati in questione possono essere rappresentati da n coppie di v.c. (t_i, δ_i)

dove

$$t_i = \min(T_i, L_i) \qquad \text{e} \qquad \delta_i = \begin{cases} 1 & \text{se} \quad T_i \leq L_i \\ 0 & \text{se} \quad T_i > L_i \end{cases}$$

Si dimostra che la funzione di densità congiunta di (t_i, δ_i) è data da:

$$\left[f(t_i;\theta)\right]^{\delta_i} \times \left[S(L_i;\theta)\right]^{1-\delta_i}$$

Infatti, la v.c. in esame ha una componente discreta ed una continua.
Per la parte <u>discreta</u> abbiamo:

$$P_r\{t_i = L_i\} = P_r\{\delta_i = 0\} = P_r\{T_i > L_i\} = S(L_i;\theta)$$

Per la parte <u>continua</u> si ha:

$$P_r\{t_i / \delta_i = 1\} = P_r\{t_i / T_i \le L_i\} = \frac{f(t_i;\theta)}{P_r\{T_i \le L_i\}} = \frac{f(t_i;\theta)}{1 - S(L_i;\theta)}$$

Da queste segue che

$$P_r\{t_i, \delta_i = 1\} = P_r\{t_i / \delta_i = 1\} \times P_r\{\delta_i = 1\} = P_r\{t_i / T_i \le L_i\} \times P_r\{T_i \le L_i\} = \frac{f(t_i;\theta)}{\{1 - S(L_i;\theta)\}} \times \{1 - S(L_i;\theta)\} = f(t_i;\theta)$$

$$P_r\{t_i, \delta_i = 0\} = P_r\{\delta_i = 0\} = P_r\{T_i > L_i\} = S(L_i;\theta)$$

Combinando queste ultime si ottiene la densità congiunta tra t_i e δ_i

La funzione di verosimiglianza delle n coppie di osservazioni risulta essere:

$$L(\theta;\mathbf{t},\delta)=\prod_{i=1}^{n}\left[f(t_i;\theta)\right]^{\delta_i}\times\left[S(L_i;\theta)\right]^{1-\delta_i}$$

Censura di II tipo

Questo schema di censura viene applicato nei casi in cui si seguono nel tempo, ad esempio, un insieme di pazienti oppure nelle applicazioni di tipo ingegneristico quando si valutano i tempi di rottura di un elemento meccanico. In tali situazioni, si registrano i tempi di rottura dei primi r elementi, senza arrivare alla rottura di tutti gli elementi sotto osservazione.
Il numero r di osservazioni è deciso prima della sperimentazione.
Formalmente, i dati consistono nelle r più piccole osservazioni

$$T_{(1)} \leq T_{(2)} \leq ... \leq T_{(r)}$$

da un campione casule iid di dimensione n $T_1,\ldots,T_n$, con r<n.
In altre parole, si hanno a disposizione le prime r statistiche d'ordine.
Con riferimento alla metodologia delle statistiche d'ordine (David, 1981) la distribuzione congiunta delle r più piccole osservazioni è data da:

$$\frac{n!}{(n-r)!} f\left(t_{(1)};\theta\right) \times .... \times f\left(t_{(r)};\theta\right) \left[S\left(t_{(r)};\theta\right)\right]^{n-r}$$