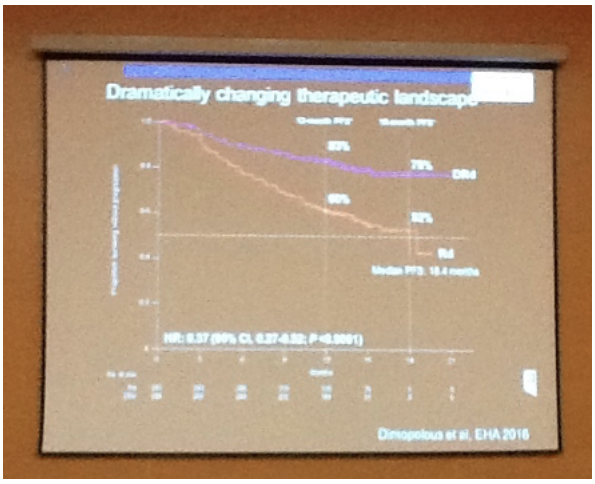
Università Magna Græcia di Catanzaro
Società dei Matematici e Naturalisti di Modena

Modelli di Sopravvivenza

www.dmi.units.it/borelli



La slide iniziale del Dr. Leif Bergsagel della Mayo Clinic
al congresso *Multiple Myeloma 2016* della U.M.G.

Autore
Massimo Borelli, Ph.D.

Contents

1 dati di tipo Time-to-Event **2**
1.1 la codifica dei dati 2
1.2 Rischio competitivo ed i modelli multistato 3

2 La cornice matematica **4**
2.1 Le definizioni 4
2.2 L'approccio non parametrico 6
2.3 L'approccio parametrico 6
2.3.1 La distribuzione esponenziale 6
2.3.2 La distribuzione di Weibull 7

3 Le curve di sopravvivenza con R **10**
3.1 L'approccio non parametrico 10
3.2 L'approccio parametrico 12

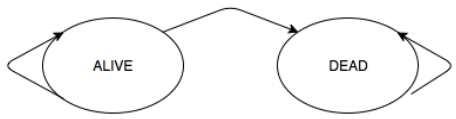
4 Sir David Cox: una leggenda della Statistica **14**
4.1 i modelli semiparametrici 14
4.2 Il log-rank test 15
4.3 La regressione di Cox 16
4.4 Interpretare un modello di Cox 16
4.5 Selezione del modello 18
4.6 Il modello minimale adeguato 19

5 Esercizi **20**

6 Modelli parametrici **21**
6.1 Strati 21
6.2 Modelli AFT - a rischio accelerato 22
6.3 Selezionare i dettagli del modello 24

7 Approfondimenti **25**
7.1 Rischio: *hazard* oppure *risk*? 25
7.2 La notazione log-lineare 25
7.3 Covariate che dipendono dal tempo 26

1 dati di tipo Time-to-Event

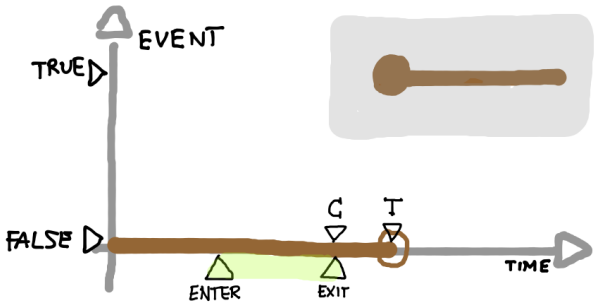


In questa figura vediamo un **modello di sopravvivenza** molto semplice, a due stati. Dal punto di vista matematico interesserebbe poter valutare le probabilità di transizione tra gli stati di questo grafico che prende il nome di **catena di Markov**, e che sono $p_{AA}, p_{AD}, p_{DD}, p_{DA}$. Diciamo che lo stato D, DEAD è assorbente perché avvenuta la transizione in quello stato non possiamo più ritornare nello stato antecedente: $p_{DD} = 1, p_{DA} = 0$.

1.1 la codifica dei dati

Nella **analisi di sopravvivenza** usualmente la variabile risposta che noi osserviamo e che caratterizza ciascun paziente è del tipo `(enter, exit, event)`, laddove la variabile logica `event` assume il valore `TRUE` qualora al tempo `exit` si è manifestato l'evento di interesse (DEAD); altrimenti, essa assume il valore `FALSE`.

Primo problema: per questo semplice modello morire a 17 anni o morire a 71 sarebbero due eventi equivalenti in questo contesto: non abbiamo un'evidenza specifica del 'tempo biologico'. Vedremo come migliorare questo fatto.

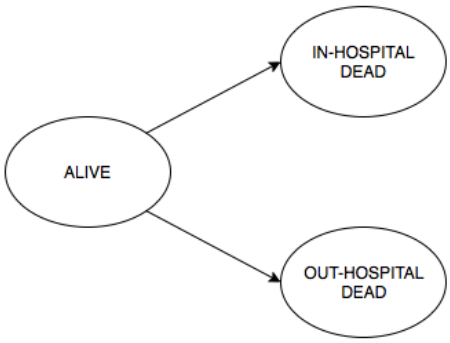


Secondo problema: abbiamo detto che nella tripla `(enter, exit, event)` possiamo avere `event = FALSE` qualora il tempo conclusivo della nostra osservazione (`exit`) anticipi quello dell'evento morte. Parliamo dunque di **eventi censurati a destra**, come vediamo nell'immagine qui sopra: il tempo T in cui accade `event = TRUE` segue l'evento C (abbiamo $T > C$) in cui registriamo

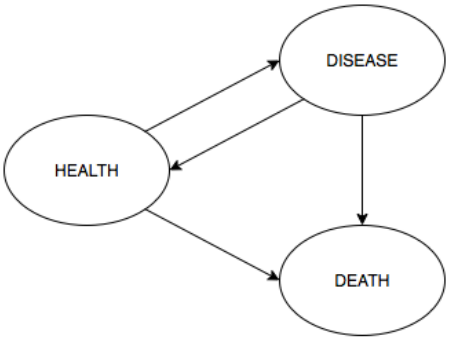
`exit`; nell'intervallo di osservazione `(enter, exit)` la **funzione indicatrice** di color marrone mostra che `event = FALSE`. Vedremo come gestire, in maniera abbastanza soddisfacente, questa difficoltà.

1.2 Rischio competitivo ed i modelli multistato

Modelli di sopravvivenza più avanzati sono ad esempio i modelli di **rischio competitivo**: ad esempio, il paziente può muovere in due diversi stati assorbenti, rappresentati dalla mortalità intra- ed extra-ospedaliera. Nel secondo stato assorbente vi possono essere dati censurati, mentre nel primo no – non è ragionevole supporre che la direzione ospedaliera ignori la data di decesso `exit` di un suo paziente.



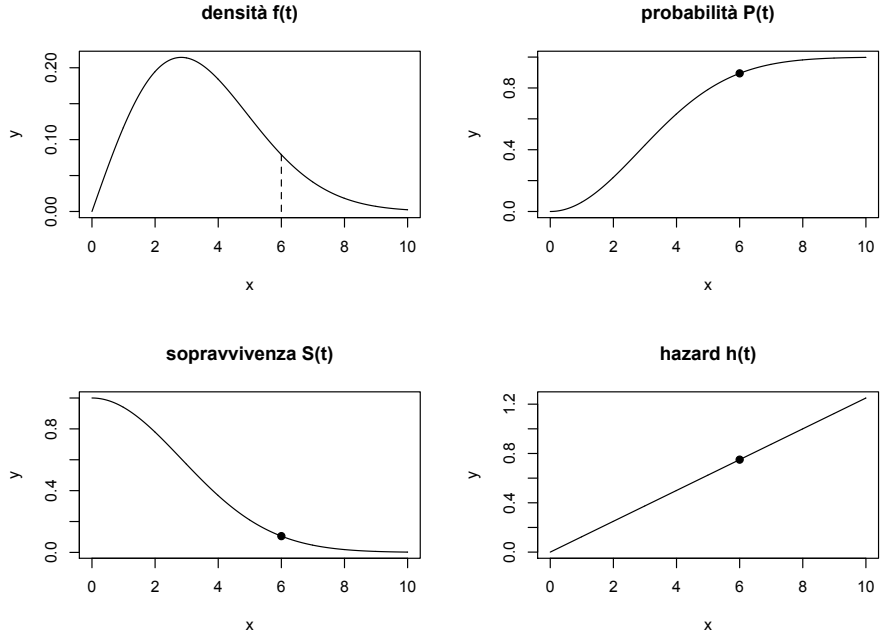
Invece un **modello multistato** ad esempio è il cosiddetto *illness-death model with recovery*: il paziente può entrare dello stato assorbente DEATH da uno qualsiasi dei due stati precedenti, nei quali è possibile ritransitare per un numero qualsiasi di volte.



2 La cornice matematica

2.1 Le definizioni

Matematicamente, ci sono di aiuto 5 funzioni che sono legate tra di loro e che descrivono in maniera equivalente il rischio di morte. Facciamo un esempio: mettiamo il caso che abbiamo seguito un certo numero di pazienti ed abbiamo visto che l'outcome sfavorevole sopraggiunge molto frequentemente circa dopo tre anni l'insorgenza della patologia, e dopo circa in dieci anni quasi nessuno sopravviva. Nella figura che segue, il pannello in alto a sinistra rappresenta la curva di **densità** $f(t)$ non è altro che una rappresentazione matematica dell'**istogramma** degli eventi. In quel disegno, il fatto che $f(6) = 0.09$ dice che circa il 9 per cento del nostro gruppo di pazienti è deceduto attorno al sesto anno dall'insorgenza della patologia.



Se ci chiediamo quale possa essere la percentuale di persone che hanno avuto outcome sfavorevole nei primi 6 anni, lo possiamo leggere in alto a destra nel grafico della **probabilità** $P(t)$: il pallino nero ci indica uno $0.90 = 90\%$ di persone. Questo significa, riguardando il pannello in alto a sinistra, che nell'intervallo da 0 a 6 della funzione di densità è racchiusa un'area pari allo $0.90 = 90\%$ dell'area totale di quell'istogramma. In termini di integrale:

$$P(6) = \int_0^6 f(t)dt = 0.90$$

Se invece ci chiediamo possa essere la percentuale di persone che sopravviveranno ai primi 6 anni, lo leggiamo dal grafico della **sopravvivenza** $S(t)$, che è esattamente il grafico della probabilità $P(t)$ 'capovolto'; infatti, il pallino nero vale qui circa $0.10 = 1 - 0.90$. In simboli matematici:

$$S(6) = 1 - P(6) = 1 - \int_0^6 f(t)dt = 1 - 0.90 = 0.10$$

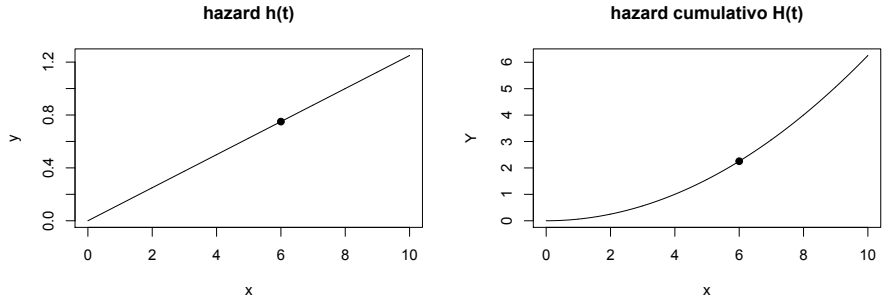
Per mezzo di queste tre funzioni, possiamo valutare l'**hazard** $h(t)$, che si può ottenere calcolando il rapporto $f(t)/S(t)$ e che indica la *probabilità istantanea* di morte al tempo t. Nel pannello in basso a destra, infatti, $h(6)$ vale circa 0.8; ed abbiamo già visto che $f(6)/S(6) = 0.08/0.10 = 0.8$. In questo particolare esempio l'hazard assomiglia ad una retta poiché abbiamo scelto come densità una particolare distribuzione di dati $f(t)$ (molto usata) che si chiama di Weibull.

Da ultimo possiamo considerare l'**hazard cumulativo** $H(t)$, che per l'hazard $h(t)$ gioca il medesimo ruolo della probabilità P rispetto alla densità f:

$$H(6) = \int_0^6 h(t)dt$$

Nel disegno sottostante, il pannello a sinistra mostra che $h(6) = 0.8$; pertanto, se calcoliamo l'area (= integrale) del triangolo di base 6 ed altezza 0.8 abbiamo:

$$\frac{6 \cdot h(6)}{2} = \frac{6 \cdot 0.8}{2} = 2.4 = H(6)$$



2.2 L'approccio non parametrico

I software statistici si occupano di calcolare quelli che vengono detti gli **stimatori non parametrici** della sopravvivenza $S(t)$, quando non conosciamo o non riusciamo a fornire un'espressione analitica della 'vera legge matematica' che descrive la densità $f(t)$:

- lo **stimatore di Nelson - Ålen**: $\hat{H} = \sum_{s \leq t} \hat{h}(s)$
- lo **stimatore di Kaplan - Meier**: $\hat{S} = \prod_{s<t}(1 - \hat{h}(s))$

Nel prossimo capitolo vedremo come fare per calcolare esplicitamente e visualizzare graficamente tutte queste quantità.

2.3 L'approccio parametrico

In certi casi abbiamo informazioni 'a priori' sui fenomeni su cui stiamo investigando che ci possono fornire la legge matematica della densità $f(t)$; ovvero, ci può interessare 'scoprire' se vi sia una particulare distribuzione statistica soggiacente al fenomeno (come ad esempio capita nei fenomeni distribuiti in maniera gaussiana, o in maniera log-normale).

2.3.1 La distribuzione esponenziale

La distribuzione esponenziale è la più semplice di tutte, ed è caratterizzata dal fatto che il tasso di rischio, `hazard rate`, $\lambda > 0$ è costante; il che è un vantaggio matematico, ma è uno svantaggio biomedico. Il vantaggio matematico è che i conti diventano particolamente facili:

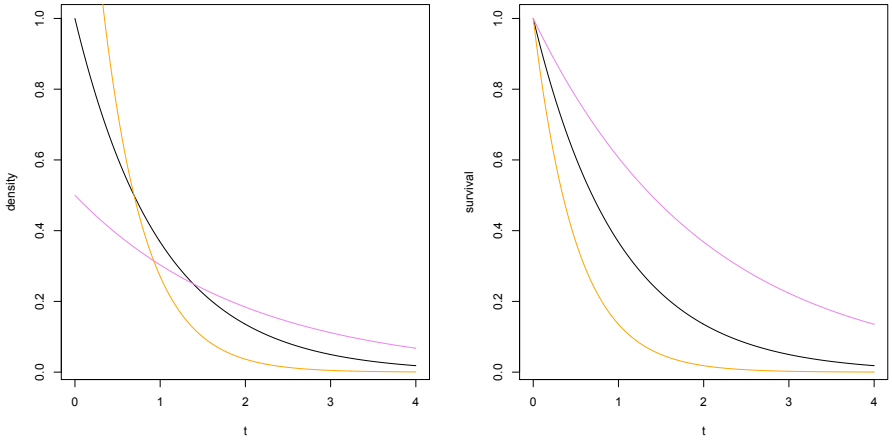
$$f(t; \lambda) = \lambda \exp(-\lambda t)$$
$$S(t; \lambda) = \exp(-\lambda t)$$
$$h(t; \lambda) = \lambda$$

Lo svantaggio biomedico sta nel fatto che, essendo il `rate` λ costante, un individuo giovane ed un individuo anziano hanno sempre la medesima probabilità di soccombere (ma c'è un trucco standard per bypassare questa limitazione: suddividere 'piece-wise' le età in classi di rischio; l'hazard rimane una funzione discontinua, ma f ed S diventano continue - anche se non derivabili).
Per disegnare la distribuzione esponenziale con R disponiamo dei comandi `dexp` e `pexp`. Ricordiamoci che essa è una legge che dipende da un solo parametro, come accade nella binomiale o nella Poisson.

```
t = seq(from = 0, to = 4, by = 0.01)
density1 = dexp(t, rate = 1)
density2 = dexp(t, rate = 2)
density05 = dexp(t, rate = 0.5)
survival1 = 1 - pexp(t, rate = 1)
survival2 = 1 - pexp(t, rate = 2)
```

```r
survival05 = 1 - pexp(t, rate = 0.5)
par(mfrow = c(1,2))
plot(t, density1, "l", ylab = "density")
lines(t, density2, col = "orange")
lines(t, density05, col = "violet")
plot(t, survival1, "l", ylab = "survival")
lines(t, survival2, col = "orange")
lines(t, survival05, col = "violet")
```

Listing 1: La distribuzione esponenziale con R.



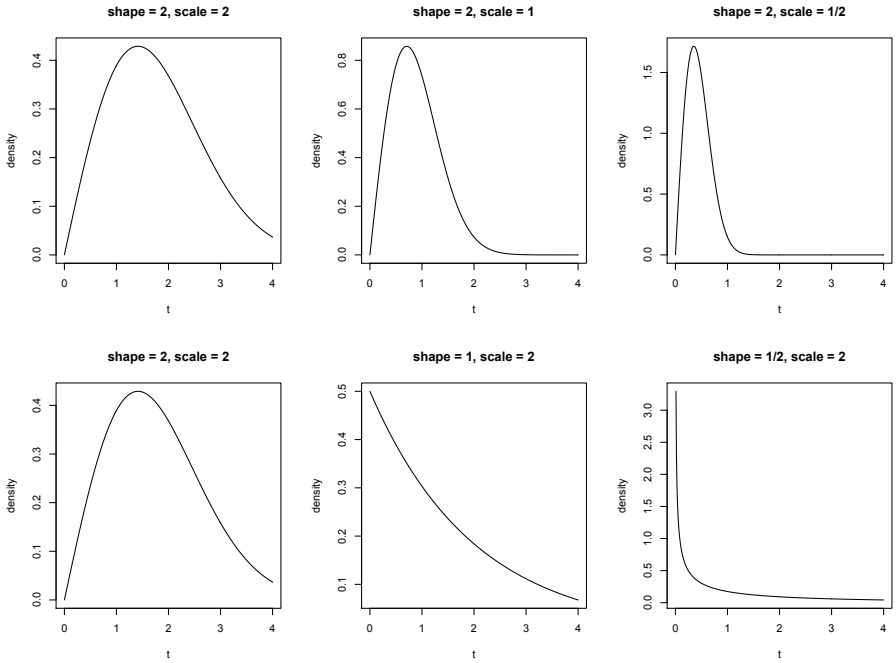
2.3.2 La distribuzione di Weibull

Negli anni '50 il matematico svedese Waloddi Weibull descrisse le interessanti proprietà della distribuzione che porta il suo nome e che possiamo rappresentare usando i comandi di R `dweibull` e `pweibull`. Si tratta di una distribuzione caratterizzata da due parametri (analogamente alla gaussiana, che ha un *parametro di centralità*, la media μ, e un *parametro di forma*, la deviazione standard σ):

- il *parametro di forma* (`shape`), $p > 0$
- il *parametro di scala* (`scale`), $\lambda > 0$

```r
t = seq(from = 0, to = 4, by = 0.01)
par(mfrow = c(2,3))
density22 = dweibull(t, shape = 2, scale = 2)
plot(t, density22, "l", ylab = "density", main = "shape = 2,
    scale = 2")
density21 = dweibull(t, shape = 2, scale = 1)
plot(t, density21, "l", ylab = "density", main = "shape = 2,
    scale = 1")
density205 = dweibull(t, shape = 2, scale = 1/2)
plot(t, density205, "l", ylab = "density", main = "shape = 2,
    scale = 1/2")
density22 = dweibull(t, shape = 2, scale = 2)
plot(t, density22, "l", ylab = "density", main = "shape = 2,
    scale = 2")
density12 = dweibull(t, shape = 1, scale = 2)
plot(t, density12, "l", ylab = "density", main = "shape = 1,
    scale = 2")
density052 = dweibull(t, shape = 1/2, scale = 2)
plot(t, density052, "l", ylab = "density", main = "shape = 1/
    2, scale = 2")
```

Listing 2: La distribuzione di Weibull con R.



Purtroppo, c'è una gran confusione in giro. Anche perché il parametro di scala $\lambda > 0$ della Weibull usa la stessa lettera greca λ in uso nella esponenziale (ma

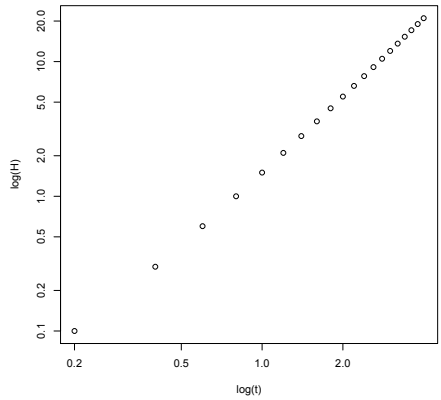
in quel caso λ non è la scala, è il `rate`; la scala, in quel caso, sarebbe $1/\lambda$ (che pasticcio!) In ogni caso, la funzione di rischio di una Weibull è:

$$h(t;\ p, \lambda) = \frac{p}{\lambda}\left(\frac{t}{\lambda}\right)^{(p-1)}$$

La distribuzione di Weibull, tra le altre belle proprietà matematiche, è caratterizzata dal fatto che - su scale logaritmiche - il rischio cumulativo $H(t)$ ed il tempo si comportano in modo lineare; ossia il grafico di $log(H(t))$ e $log(t)$ è una retta. Ecco quindi che se ci affidiamo ad un approccio non parametrico, e consideriamo lo stimatore $\hat{H}$ di Nelson - Ålen, possiamo visualmente apprezzare se il fenomeno che stiamo studiando è ben modellato da una distribuzione di Weibull (o esponenziale, che è una Weibull con $p = 1$).

```
t = seq(from = 0, to = 4, by = 0.2)
dens = dweibull(t, shape = 2, scale = 2)
Prob = pweibull(t, shape = 2, scale = 2)
Surv = 1 - Prob
haz = dens/Surv
Haz = cumsum(haz)
plot(t, Haz, log = "xy", xlab = "log(t)", ylab = "log(H)")
```

Listing 3: La distribuzione esponenziale con R.



3 Le curve di sopravvivenza con R

Vi sono vari pacchetti di R[9] che consentono di ottenere stime di base di sopravvivenza e delle loro relative rappresentazioni, come ad esempio `survival`[10] o `eha`[1]. Come esempio-guida scegliamo il dataset `addome` relativo a 75 pazienti oncologici di cui conosciamo l'`Eta` all'intervento in cui furono trattati con una determinata `Tecnica` chirurgica (`open` oppure `laparo`), di cui vediamo cinque record estratti casualmente:

	Mesi	Statoattuale	Tecnica	Eta
11	168	0	laparo	65
50	109	1	open	52
75	58	1	open	86
28	133	0	laparo	72
35	34	0	open	37

Ci interessa stimare la sopravvivenza espressa in `Mesi`, essendo la variabile logica `Statoattuale` l'indicatore dei dati censurati. Importiamo in R il dataset

```
library(survival)
library(eha)
url = "http://www.dmi.units.it/borelli/dataset/addome.csv"
addome = read.csv(url, header = TRUE)
# addome = read.csv("/Users/borelli/Documents/lezioni/2016
    catanzaro/sopravvivenza/Scripts/addome.csv", header = TRUE
    )
attach(addome)
names(addome)
head(addome)
str(addome)
addome[sample(1:75, 10),]
```

Listing 4: Importazione dalla rete in R del dataset `addome`.

Con il comando `str` vediamo che si tratta di un dataset con 75 righe (osservazioni) of 4 variabili.

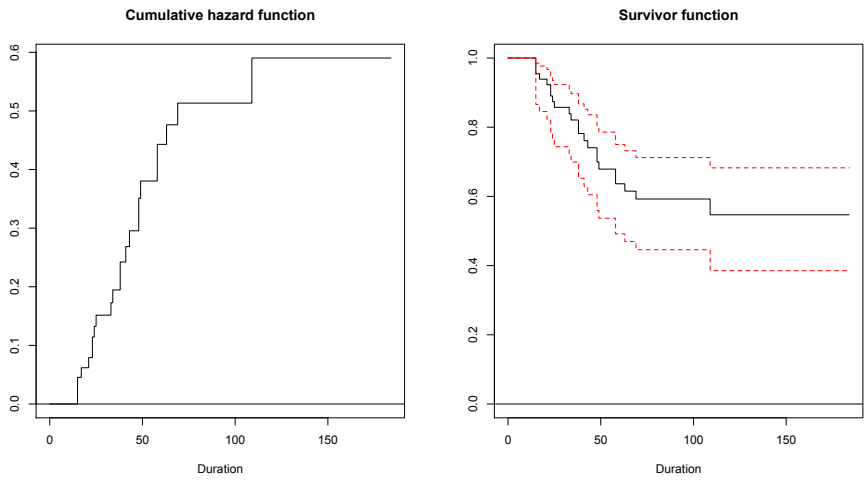
3.1 L'approccio non parametrico

Vogliamo innanzitutto visualizzare gli stimatori di Nelson - Ålen e di Kaplan - Meier del dataset `addome`:

```
par(mfrow = c(1,2))
with(addome, plot(Surv(Mesi, Statoattuale), fn = "cum"))
with(addome, plot(Surv(Mesi, Statoattuale), fn = "surv"))
```

Listing 5: Grafico degli stimatori di Nelson - Ålen e di Kaplan - Meier con R.

Se fossimo invece interessati a vedere numericamente l'evolversi della sopravvivenza, disponiamo del comando `survfit`:




```
km = survfit(Surv(Mesi, Statoattuale) ~ 1)
summary(km)
sort(Surv(Mesi, Statoattuale))
```

Listing 6: stimatori di Nelson - Ålen e di Kaplan - Meier con R.

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
15	66	3	0.955	0.0256	0.906	1.000
17	61	1	0.939	0.0296	0.883	0.999
21	58	1	0.923	0.0332	0.860	0.990
23	57	2	0.890	0.0392	0.817	0.971
24	54	1	0.874	0.0418	0.796	0.960
25	53	1	0.857	0.0441	0.775	0.948
33	47	1	0.839	0.0468	0.752	0.936
34	46	1	0.821	0.0492	0.730	0.923
38	42	2	0.782	0.0541	0.683	0.895
41	38	1	0.761	0.0564	0.658	0.880
43	37	1	0.741	0.0585	0.634	0.865
48	36	2	0.699	0.0621	0.588	0.832
49	34	1	0.679	0.0636	0.565	0.816
58	32	2	0.636	0.0663	0.519	0.781
63	30	1	0.615	0.0674	0.496	0.763
69	27	1	0.592	0.0687	0.472	0.744
109	13	1	0.547	0.0770	0.415	0.721

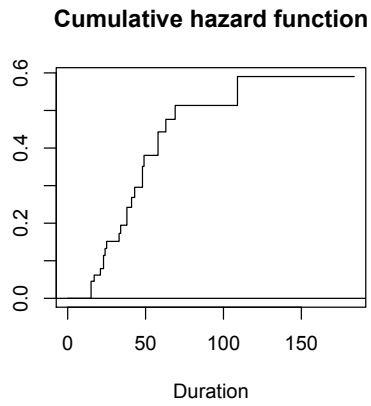
Per capire 'come scendono' i gradini del grafico di Kaplan - Meier, è bastato riordinare con il comando `sort` in maniera crescente i `Mesi` di sopravvivenza (tenendo conto dei dati censurati):

```
 [1]   1+   2+   3+   4+   4+   5+  10+  10+  14+  15+  15
[12]  15   15   16+  17   20+  20+  21   23+  23   23   24
[23]  25   28+  29+  30+  30+  30+  33   34+  34   35+  37+
[34]  38   38   39+  40+  41   43   48   48   49   52+  58
[45]  58   63   64+  65+  69   74+  77+  78+  82+  83+  87+
[56]  90+  93+  94+  97+ 102+ 107+ 108+ 109+ 109  113+ 114+
[67] 119+ 132+ 133+ 133+ 147+ 155+ 157+ 168+ 184+
```

3.2 L'approccio parametrico

Come annunciato nel precedente paragrafo, possiamo valutare 'ad occhio' se i nostri dati si conformano ad una distribuzione di Weibull se si dispongono approssimativamente in maniera rettilinea su scala logaritmica nel piano $(t, \hat{H}(t))$

```
with(addome, plot(Surv(Mesi, Statoattuale)), fn = "loglog" )
```



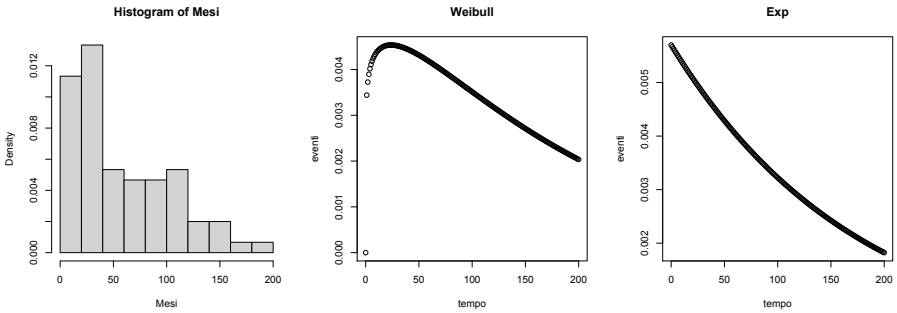
Con qualche riserva, possiamo provare ad ipotizzare che la Weibull sia una distribuzione aleatoria adeguata a rappresentare i dati. In tal caso, con il comando `phreg` del pacchetto `eha` otteniamo la stima dei parametri di forma p e di scala λ che meglio approssimano i nostri dati:

```
fitWeibull = phreg(Surv(Mesi,Statoattuale) ~ 1,data = addome)
fitWeibull
plot(fitWeibull)
```

Covariate	W.mean	Coef	Exp(Coef)	se(Coef)	Wald p
log(scale)		5.163		0.224	0.000
log(shape)		0.114		0.170	0.503

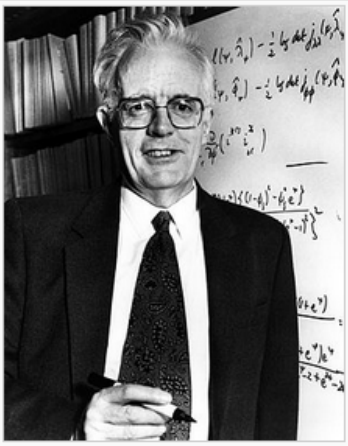
Vediamo che $\lambda = \exp(5.163) = 175$ e che $p = \exp(0.114) = 1.12$. Il valore di p è sospettosamente vicino ad uno (ed infatti l'ampiezza dell'errore standard 0.17 sopravanza la stima del coefficiente 0.11) e quindi potremmo assumere che la semplice distribuzione esponenziale di rate $1/\lambda = 1/175 = 0.006$ possa andarci bene. Ritorneremo su questo argomento in seguito.

```
par(mfrow = c(1,3))
hist(Mesi, freq = FALSE, col = "lightgrey")
tempo = 0:200
eventi = dweibull(tempo,  shape = 1.12 , scale = 174.68)
plot(tempo, eventi, main = "Weibull")
eventi = dexp(tempo, rate = 0.0057)
plot(tempo, eventi, main = "Exp")
```



4 Sir David Cox: una leggenda della Statistica

4.1 i modelli semiparametrici



David Cox (marzo 2006)

Sir David Cox, nella foto tratta da Wikipedia, compirà 92 anni il prossimo 15 luglio 2016. A lui dobbiamo l'intuizione di un modo, semplice nel suo concetto ma molto efficace, per stabilire se vi sia una 'differenza di rischio' tra diversi gruppi di una medesima popolazione – senza dover conoscere necessariamente il rischio di ciascuno dei gruppi. Pensiamo ai pazienti oncologici del dataset `addome`; le colleghe chirurghe mi riferiscono che capita piuttosto spesso in sala operatoria che con lesioni di minore entità si opti per una `Tecnica` di tipo laparoscopico, mentre in presenza di masse importanti l'approccio open sia quasi una scelta obbligata. E perciò non può stupire che la sopravvivenza del gruppo `laparo` differisca da quella del gruppo `open`.

Indichiamo con $h_L(t)$ la funzione di rischio del gruppo `laparo`, con $h_O(t)$ quella del gruppo `open`. Diremo che I due gruppi hanno un rischio proporzionale se esiste un numero costante $\psi > 0$, indipendente dal tempo t, per cui accada che:

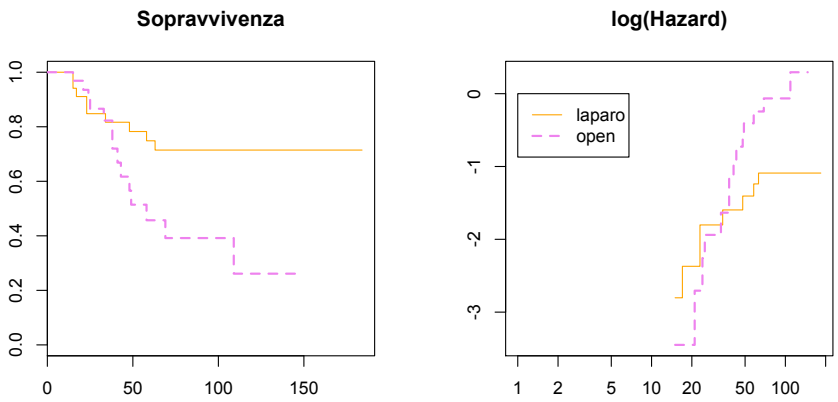
$$h_O(t) = \psi \cdot h_L(t)$$

Il primo trucco ingegnoso è quello di considerare il logaritmo di ψ, $\beta = \log(\psi)$ e scrivere:

$$h_O(t) = \exp(\beta) \cdot h_L(t)$$

Ora, non è affatto detto che in ogni caso i rischi di due gruppi siano proporzionali; è un ipotesi che va verificata controllando – almeno 'ad occhio' – il parallelismo degli stimatori di Nelson Ålen disegnati su scala log-log.

```
par(mfrow = c(1,2))
plot(survfit(Surv(Mesi,Statoattuale) ~ Tecnica), col=c("orange
    ", "violet"), lwd = c(1,2), lty = c(1,2), main = "
    Sopravvivenza")
plot(survfit(Surv(Mesi,Statoattuale) ~ Tecnica), col=c("orange
    ", "violet"), lwd = c(1,2), lty = c(1,2), fun="cloglog",
    main = "log(Hazard)")
legend(1, 0, levels(Tecnica), col=c("orange", "violet"), lwd =
     c(1,2), lty = c(1,2))
```



E qui, diremmo proprio che le cose non funzionano come vogliamo: non siamo in presenza di rischi proporzionali tra i due gruppi perché nel pannello di destra l'andamento è forse rettilineo (i.e. la distribuzione di Weibull potrebbe andar bene) ma le rette non sono parallele, e non possiamo dire che tra loro due ci sia una distanza costante $\beta = \log(\psi)$.

4.2 Il log-rank test

Proviamo ad eseguire un test (non parametrico, basato sui ranghi) per decidere se i due trattamenti si associno a sopravvivenze differenti (omettiamo i dettagli matematici; per approfondimenti si può vedere [1]). Il modo più semplice per eseguire il *log-rank test* è servirsi proprio della funzione 'Cox proportional hazards' `coxph` del pacchetto `survival`:

```
logrank = coxph(Surv(Mesi,Statoattuale) ~ Tecnica, data =
    addome)
summary (logrank)
```

L'output è molto ricco, ma se ci concentriamo sull'ultima riga, leggiamo:

```
Score (logrank) test = 4.65  on 1 df,   p=0.03101
```

possiamo pertanto decidere che nei due gruppi la sopravvivenza è diversa. Abbiamo ancora altre due informazioni immediatamente comprensibili:

```
              coef exp(coef)
Tecnicaopen 0.9108    2.4862
```

Tuttavia dubitiamo della bontà della stima del `coef` $\beta = 0.911$, che condurrebbe ad una costante di proporzionalità `exp(coef)` $\psi = 2.486$ stante il fatto che i rischi non ci sono apparsi proporzionali nel grafico della pagina precedente.

4.3 La regressione di Cox

Ritorniamo all'idea ingegnosa di scrivere la proporzionalità dei rischi in questo modo:

$$h_L(t) = \exp(\beta) \cdot h_O(t)$$

ed ancor meglio, definiamo per $x \in \{0, 1\}$ il rischio:

$$h(t; x) = \exp(\beta x) \cdot h_O(t)$$

Se $x = 0$ abbiamo

$$h(t; 0) = \exp(\beta \cdot 0) \cdot h_O(t) = 1 \cdot h_O(t) = h_O(t)$$

mentre se $x = 1$ abbiamo

$$h(t; 1) = \exp(\beta \cdot 1) \cdot h_O(t) = \exp(\beta) \cdot h_O(t) = h_L(t)$$

L'idea quindi di porre la $\mathbf{x}$ come termine esponenziale $\exp(\beta\mathbf{x})$ consente nella *regressione di Cox* di descrivere, in maniera moltiplicativa, la differenza di rischio tra due (o più) gruppi in un campione introducendo una (o più) covariate $\mathbf{x}$.

```r
modelloCox = coxph(Surv(Mesi,Statoattuale) ~ Tecnica + Eta,
    data = addome)
summary (modelloCox)
```

4.4 Interpretare un modello di Cox

Interpretiamo l'output:

```
               coef exp(coef) se(coef)     z Pr(>|z|)
Tecnicaopen 0.59613   1.81508  0.43818 1.360 0.173681
Eta         0.06617   1.06841  0.01734 3.815 0.000136 ***

            exp(coef) exp(-coef) lower .95 upper .95
Tecnicaopen     1.815     0.5509     0.769     4.284
Eta             1.068     0.9360     1.033     1.105
```

Innanzitutto seguiamo sempre l'**ordine alfabetico**, e quindi nella covariata `Tecnica` si assume che `laparo` è il riferimento ed `open` è il livello successivo. Pertanto:

$$h_O(t) = \psi \cdot h_L(t) = \exp(\beta) \cdot h_L(t)$$

$$h_O(t) = 1.815 \cdot h_L(t) = \exp(0.596) \cdot h_L(t)$$

Nota bene: non siamo in grado di dire precisamente quale sia il rischio ad un certo tempo t per un paziente del gruppo `laparo`, $h_L(t)$; ma sappiamo che un paziente con le medesime caratteristiche ma del gruppo `open` ha un rischio 1.815 volte superiore (ossia di circa l'82% in più). Si esprime questo fatto dicendo il modello di Cox e' di **tipo semiparametrico**.
Ci chiediamo: dunque la tecnica `open` è più rischiosa della tecnica `laparo`? No, non possiamo dirlo: l'incertezza della stima misurata dall'errore standard (`0.438`) è cospicua rispetto al valore β = `0.596`. Questo comporta che il p-value non sia significativo (`0.17`) e che con una fiducia del 95% si abbia che $\psi \in [0.769, 4.248]$: quando in un modello moltiplicativo il valore 1 cade nell'intervallo di fiducia allora non siamo asolutamente in grado di decidere se la covariata in questione (la `Tecnica`) sia protettiva ovvero aumenti il rischio di eventi infausti.

L'`Eta` invece appare essere un predittore significativo, lo deduciamo immediatamente dalle tre stelline, `***`: non stupisce infatti che la sopravvivenza delle persone rispetto alla patologia sia indissolubilmente legata alla loro anzianità. Si noti che noi non siamo in grado di dire quale sia il rischio $h(t; N)$ per un paziente di N anni di età, ma sappiamo che per un paziente con le medesime caratteristiche ma più anziano di un anno avremo che:

$$h(t; N + 1) = 1.068 \cdot h(t; N)$$

ossia un aumento di rischio di circa il 7% per ogni anno di età in più. Vale la pena osservare qui che differenza c'è tra `exp(coef)` = 1.068 ed `exp(-coef)` = 0.936: dire che chi è più vecchio di un anno ha un rischio che aumenta del 6.8% è come dire che chi è più giovane di un anno diminuisce il rischio del 6.4% (infatti $1 - 0.936 = 0.064$).

4.5 Selezione del modello

Sfruttando tutte le informazioni del dataset `addome`:

Mesi	Statoattuale	Tecnica	Eta
..	..	..	..

abbiamo considerato il modello additivo

`Surv(Mesi,Statoattuale)` $\sim$ `Tecnica + Eta`

Avremmo potuto anche studiare modelli più complessi di questo, come vedrete negli esercizi di fine capitolo. Ma sappiamo che l'obiettivo di un modello statistico è quello di essere **minimale ed adeguato**[3]. *Minimale*, nel senso che non deve includere covariate che non siano significative (vale infatti il principio di parsimonia di Occam); *adeguato* nel senso che non abbiamo dimenticato di introdurre nel modello informazioni che invece sarebbe stato opportuno considerare.

Esempio. Nel paragrafo 4.2 relativo al log-rank test, avevamo considerato il modello

`Surv(Mesi,Statoattuale)` $\sim$ `Tecnica`

il modello risultava 'significativo', certamente!, e sarebbe potuto essere un modello minimale; ma non era affatto adeguato poiché avremmo trascurato la covariata `Eta`, che è risultata essere invece un rilevante predittore.

Su tale questione, Göran Broström nel suo testo[1] cita alcuni consigli da tenere sempre presente:

- Disse lo statistico G. Box che non esiste il *vero* modello statistico, esiste solo quello *più utile*
- Potrebbe essere ancor più utile avere due modelli invece che uno
- Le covariate `importanti` debbono essere sempre presenti nel modello
- Si badi di evitare le procedure automatiche di selezione del modello
- Se in un modello vi sono termini di interazione tra due predittori, allora i predittori corrispondenti *devono* essere presenti nel modello (anche se non sono significativi)

Vale la pena di insistere su quella locuzione di '`importante`', che non vuol dire `significativo`. Ci sono ottimi paper a tale proposito, ma per iniziare consiglierei [5] e poi per approfondire [2]: sono libri forse un po' difficili da reperire in molte biblioteche di ateneo – ma per fortuna c'è `www.bookzz.org`.

Vediamo invece come il **test del rapporto di verosimiglianza** (Likelihood Ratio Test (LRT), si veda ad esempio l'Appendice di [4]) ci aiuti, partendo dal modello additivo che abbiamo studiato nel paragrafo precedente 4.4, a selezionare il modello minimale adeguato. Utilizziamo a tale scopo la funzione `drop1`:

```
modelloCox = coxph(Surv(Mesi,Statoattuale) ~ Tecnica + Eta,
    data = addome)
drop1 (modelloCox, test = "Chisq")
```

	Df	AIC	LRT	Pr(>Chi)
<none>		151.00		
Tecnica	1	150.91	1.90	0.1677
Eta	1	169.34	20.34	0.0000

Non c'è dubbio, il modello minimale adeguato ha come predittore l'età dei pazienti, e la tecnica non svolge un ruolo predittivo: infatti se tolgo ('drop') la covariata `Eta` il modello 4.2 differisce significativamente dal modello 4.4: è troppo 'povero', abbiamo 'perduto informazione'.

4.6 Il modello minimale adeguato

In definitiva, il modello che appare preferibile in quanto minimale ed adeguato è il modello 4.6:

```
minadeq = coxph(Surv(Mesi,Statoattuale) ~  Eta, data = addome)
summary(minadeq)
```

```
       coef exp(coef) se(coef)    z Pr(>|z|)
Eta 0.06949   1.07196  0.01733 4.01 6.07e-05 ***

    exp(coef) exp(-coef) lower .95 upper .95
Eta     1.072     0.9329     1.036     1.109
```

5 Esercizi

Il rischio aumenta in maniera proporzionale, o non proporzionale, all'età? Riprendete il modello 4.6 ed introducete come ulteriore covariata il termine di curvatura:

```
+ I(Eta^2)
```

1. il termine di curvatura è significativo? Se così fosse il rischio $h(t)$ non crescerebbe in maniera proporzionale all'età.
2. Nel modello con curvatura come cambia il rischio $h(t; N+1)$ per un paziente con $N+1$ anni di età rispetto ad un paziente con le medesime caratteristiche ma con N anni?

Tecnica chirurgica ed età interagiscono modificando il rischio? Studiate il modello 4.4 considerando un'interazione tra le due covariate, ossia con `*` invece che `+`:

1. il termine di interazione è significativo?
2. Nel modello con interazione come cambia il rischio $h(t; open, N+1)$ per un paziente con $N+1$ anni di età operato con la tecnica `open` rispetto ad un paziente con le medesime caratteristiche ma con N anni operato con la tecnica `laparo`?

```
curvatura = coxph(Surv(Mesi,Statoattuale) ~  Eta + I(Eta^2),
    data = addome)
interazione= coxph(Surv(Mesi,Statoattuale) ~  Eta * Tecnica,
    data = addome)
```

Prova di verifica. Pensando ad un vostro progetto di ricerca, proponete - o simulate - un dataset in cui un 'tempo di sopravvivenza' viene associato ad un paio di covariate possibili predittori. Effettuate l'analisi statistica proponendo il modello minimale adeguato e preparate una breve presentazione in cui discutete i risultati, illustrandoli con qualche grafico opportuno.

6 Modelli parametrici

Abbiamo finora imparato a modellare il rischio in maniera non parametrica, ed in maniera semiparametrica mediante la grande idea di Sir David Cox. Vediamo ora come potremmo stimare modelli di fenomeni dei quali siamo in grado di ipotizzare che posseggano una precisa struttura aleatoria sottostante, utilizzando la funzione `phreg` che abbiamo introdotto nel paragrafo 3.2 .

6.1 Strati

Abbiamo visto nella figura del paragrafo 4.1 che, distinguendo la sopravvivenza in base alla `Tecnica`, le 'rette' (sinonimo di distribuzione di Weibull) non sono parallele (sinonimo di mancanza di proporzionalità). Abbiamo anche visto nel paragrafo 4.5 relativo alla selezione del modello che la `Tecnica` non è un predittore. Tuttavia supponiamo (per finta!) che la covariata, quantunque non significativa, sia *importante* e che sia opportuno modellarla per considerazioni biologiche e mediche, non matematiche. In tal caso possiamo **stratificare** il rischio, che consente tecnicamente di creare delle *funzioni di verosimiglianza parziali* con valori comuni sulle covariate in modo da ottenere un'unica funzione di verosimiglianza moltiplicandole tra di loro.

```
modelloStrati = phreg(Surv(Mesi,Statoattuale) ~  Eta + strata(
    Tecnica), data = addome)
summary (modelloStrati)
```

Osservate l'interessante output, che dipana un dubbio che avevamo lasciato in sospeso nelle pagine precedenti: le metodiche `laparo` e `open` si comportano in maniera distinta.

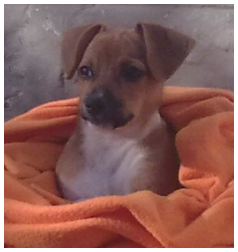
Covariate	W.mean	Coef	Exp(Coef)	se(Coef)	Wald p
Eta	54.394	0.073	1.076	0.018	0.000
log(scale):1		10.528		2.130	0.000
log(shape):1		-0.035		0.289	0.903
log(scale):2		7.289		0.919	0.000
log(shape):2		0.567		0.199	0.004

Mentre la tecnica 2 (in ordine alfabetico, `open`) ha una sua scala ed una sua forma, la tecnica 1 non ha una forma p statisticamente distinta dallo zero e, dunque, è sufficiente modellarla con una distribuzione esponenziale con *rate* $1/\lambda$ costante.

Esercizio. Provate a simulare una distribuzione casuale di `tempi` esponenziali e Weibull in accordo con i parametri appena determinati, di dimensioni analoghe a quelle dei dati di `addome`. Realizzate un grafico sul piano bidimensionale riportando su un asse i dati dei `Mesi` di sopravvivenza di `addome` e sull'altro i

tempi casuali, dopo averli preventivamente riordinati col comando `sort`. Che tipo di grafico vi aspettate di vedere? Riuscite a cogliere qualche analogia con un altro famoso grafico diagnostico?

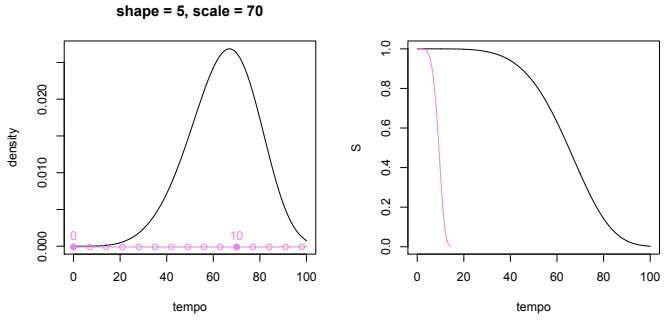
6.2 Modelli AFT - a rischio accelerato



E questa è l'immagine di Lilla Violetta de Glicinis, nobile meticcio. Al bar si sente spesso dire che un anno di vita dei cani equivalga a sette anni di noi umani. I libri di statistica, ad esempio [6], parlando di **accelerated failure time models** citano questa frase in maniera più precisa: la probabilità che un uomo sopravviva trascorsi t anni è uguale alla probabilità che un cane sopravviva trascorsi $t/7$ anni; in simboli:

$$S_{Massimo}(t) = S_{Lilla}(t/7)$$

Nella figura che segue abbiamo simulato una distribuzione di Weibull di forma $p = 5$ e scala $\lambda = 70$. Il tempo misurato in anni in colore nero è quello riferito al proprietario di Lilla, mentre la scala di misura a pallini viola misura il tempo in anni per la cagnolina. Vedete che per il proprietario il tempo modale vale circa 67 anni, mentre per Lilla il tempo modale è $67/7 = 9.6$ anni. La sopravvivenza del proprietario, in colore nero, raggiunge la metà (ovvero il tempo mediano di sopravvivenza) a circa 65.1 anni, mentre la Lilla ha un tempo mediano di sopravvivenza di 9.3 anni.



```r
tempo = seq(from = 0, to = 100, by = 0.1)
density = dweibull(tempo, shape = 5, scale = 70)
par(mfrow = c(1,2))
plot(tempo, density, "l", ylab = "density", main = "shape = 5,
     scale = 70")
lines(c(0,100), c(-0.0001,-0.0001), col = "violet")
lilla = seq(0, 100, by = 7)
points( lilla, rep(-0.0001, length(lilla)), col = "violet")
points( 70, -0.0001, col = "violet", bg = "violet", pch = 21)
points( 0, -0.0001, col = "violet", bg = "violet", pch = 21)
text( 70, -0.0001, "10", col = "violet", pos = 3)
text( 0, -0.0001, "0", col = "violet", pos = 3)
S = 1 - pweibull(tempo, shape = 5, scale = 70)
plot(tempo, S, "l")
points(tempo/7, S, col = "violet", "l")
```

Esercizio. Riguardando gli script qui sopra, sapreste dire con quali comandi di R sono riuscito a trovare i tempi modali ed i tempi mediani mio e di Lilla?

Anche i modelli a rischio accelerato sono di tipo **parametrico**: quindi si deve scegliere una distribuzione aleatoria dei dati – nelle scienze della vita la scelta ricade per default sulla variabile aleatoria di Weibull. I pacchetti `eha` e `survival` dispongono rispettivamente delle funzioni `aftreg` e `survreg` per determinare l'accelerazione del rischio. A nostro avviso, `aftreg` presenta un output più semplice da interpretare (la parametrizzazione di Therneau e Grambsch [10] differisce da quella scelta da Broström [1]). Vediamo un esempio:

```r
modelloAFT = aftreg(Surv(Mesi,Statoattuale) ~  Tecnica, data =
     addome)
summary (modelloAFT)
```

```
Covariate          W.mean      Coef Time-Accn  se(Coef)    Wald p
Tecnica
         laparo    0.637     0         1          (reference)
           open    0.363     0.910     2.484     0.347     0.009

Baseline parameters:
log(scale)                   5.514               0.303     0.000
log(shape)                   0.226               0.168     0.178
```

Siete ormai in grado di intuire quale sia il fattore di accelerazione dei rischio e quali siano i parametri della distribuzione di Weibull che modellano il fenomeno. Per **esercizio**, inoltre, provate a modellare la `Eta` invece che la `Tecnica`.

6.3 Selezionare i dettagli del modello

Abbiamo detto che di solito ci si affida spensieratamente alla distribuzione di Weibull come variabile aleatoria soggiacente i tempi di sopravvivenza. Ma vi possono essere altre possibilità:

- la distribuzione 'valore estremo', `dist = "ev"`
- la distribuzione di Gompertz, `dist = "gompertz"`
- la distribuzione con funzioni di rischio costanti a tratti, `dist = "pch"`
- la distribuzione di log-logistica, `dist = "loglogistic"`
- la distribuzione di log-normale, `dist = "lognormal"`

Il modo più scaltro per scegliere quale di queste distribuzioni sia migliore, non avendo elementi extra-statistici per decidere, possiamo ricorrere al valore massimo della log-verosimiglianza (`Max. log. likelihood`): scegliamo senza dubbio quello più vicino allo zero (quello 'più piccolo', escludendo il segno negativo).

Se invece siamo in dubbio se scegliere tra un modello a rischi proporzionali (`phreg`) ed uno a rischio accelerato (`aftreg`) conviene ricorrere ai criteri di informazione, come ad esempio AIC. Ora, siccome il numero di parametri equivale – a parità di covariate nel modello – si ritorna dunque a valutare la log-verosimiglianza, (`Max. log. likelihood`).

7 Approfondimenti

7.1 Rischio: *hazard* oppure *risk*?

Vogliamo mostrare con un esempio numerico che i due concetti che di solito vengono definiti in ambito epidemiologico e in quello dell'analisi di sopravvivenza in realtà coincidono. Supponiamo che in un gruppo A di 100 persone il 70% di esse sia stato colpito dall'evento sfavorevole 1, mentre in un gruppo B di altrettante persone solo il 20% ne sia stato colpito. Ecco il codice R per impostare la questione:

```
gruppo = factor(c(rep("A", 100), rep("B", 100)))
evento = c(rep(1, 70), rep(0, 30), rep(1, 20), rep(0, 80))
table(gruppo, evento)
(RR = 0.70/0.20)
```

	0	1
A	30	70
B	80	20

La statistica medica ci dice come calcolare il **relative risk**, rischio relativo:

$$RR = \frac{70\%}{20\%} = \frac{0.70}{0.20} = 3.5$$

Supponiamo di disporre di un'ulteriore informazione, ossia di conoscere il tempo in cui l'evento infausto si è manifestato; effettuiamo l'analisi di sopravvivenza con una regressione di Cox:

```
library(survival)
tempo = c(1:100, 1:100)
modello = coxph( Surv(tempo, evento) ~ gruppo)
(beta = -modello$coef)
(psi = exp(beta))
```

Eseguendo il codice osserverete dunque che le nozioni di rischio relativo RR e di **hazard ratio** coincidono: $\psi = \exp(\beta) = RR = 3.5$. Attenzione: c'è da ricordare che questo fatto dipende dalla scelta opportuna del `tempo`, che qui abbiamo costruito opportunamente ipotizzando che tutti gli eventi infausti accadono prima di quelli fausti. Se avessimo generato dei tempi casualmente la stima di ψ si sarebbe discostata, anche di parecchio, da quella del RR.

7.2 La notazione log-lineare

Ritorniamo all'esempio del cane Lilla, con il tempo di Massimo T e il tempo $T_{Lilla} = T/7$. Ricordando come funzionano i logaritmi, scriviamo:

$$\log(T) = \log(7 \cdot T_{Lilla}) = \log(7) + \log(T_{Lilla})$$

Quindi se chiamiamo $Y = \log(T)$, $\varepsilon = \log(T_{Lilla})$, $\log(7) = \beta\mathbf{x}$ ci riconduciamo alla forma

$$Y = \beta\mathbf{x} + \varepsilon$$

È questa la forma utilizzata nel pacchetto `survival`, laddove viene anche introdotto il parametro di scala σ che 'amplifica' i residui.

7.3 Covariate che dipendono dal tempo

I tubi endotracheali che vengono utilizzati nella ventilazione meccanica dei pazienti in terapia intensiva hanno una specie di palloncino ('cuffia') che viene fatto aderire alla trachea in modo che i fluidi non percolino dalle vie aeree superiori a quelle inferiori. Uno studio di Umberto Lucangelo [7] mostra che tubi di differenti qualità – e costo – possono avere medesimi risultati di contenimento di tale percolazione se si applica una lieve pressione residuale di fine espirio (PEEP). L'esperimento di Lucangelo mostra che, applicando – per sei ore – una leggera PEEP, tubi ottimali e scadenti si comportano allo stesso modo; togliendo la PEEP dopo la sesta ora il comportamento cambia considerevolmente e i tubi costosi prevalgono per qualità e sicurezza. In questi disegno sperimentale la PEEP è una **covariata che dipende dal tempo**: per questo tipo di studi il libro di Martinussen e Scheike [8] è ottimo (ma difficile) quanto il loro pacchetto `timereg`.

References

[1] Göran Broström. *Event History Analysis with R.* CRC Press, 2012.

[2] Kenneth P Burnham and David R Anderson. *Model selection and multimodel inference: a practical information-theoretic approach.* Springer Science & Business Media, 2003.

[3] Michael J Crawley. *Statistics: an introduction using R.* John Wiley & Sons, 2005.

[4] Julian J Faraway. *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models.* CRC press, 2005.

[5] Phillip I Good and James W Hardin. *Common errors in statistics (and how to avoid them).* John Wiley & Sons, 2012.

[6] David G Kleinbaum and Mitchel Klein. *Survival analysis: a self-learning text.* Springer Science & Business Media, 2006.

[7] Umberto Lucangelo, Walter A Zin, Vittorio Antonaglia, Lara Petrucci, Marino Viviani, Giovanni Buscema, Massimo Borelli, and Giorgio Berlot. Effect of positive expiratory pressure and type of tracheal cuff on the incidence of aspiration in mechanically ventilated patients in an intensive care unit. *Critical care medicine*, 36(2):409–413, 2008.

[8] Torben Martinussen and Thomas H Scheike. *Dynamic regression models for survival data*. Springer Science & Business Media, 2007.

[9] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[10] Terry M Therneau and Patricia M Grambsch. *Modeling survival data: extending the Cox model*. Springer Science & Business Media, 2000.

Acknoledgments. This text has been composed exploiting a Velimir Gayevskiy free template, developed within the Leslie Lamport free document typesetting system LaTeX.

Credits. These lecture notes are not original at all, following almost completely the nice and clear introductory textbook of Göran Broström.