Ανάλυση επιβίωσης

Kaplan Meier καμπύλες και Logrank test

Γεωργία Σαλαντή
Εργαστήριο Υγιεινής και Επιδημιολογίας
Πανεπιστήμιο Ιωαννίνων

Πρόγραμμα

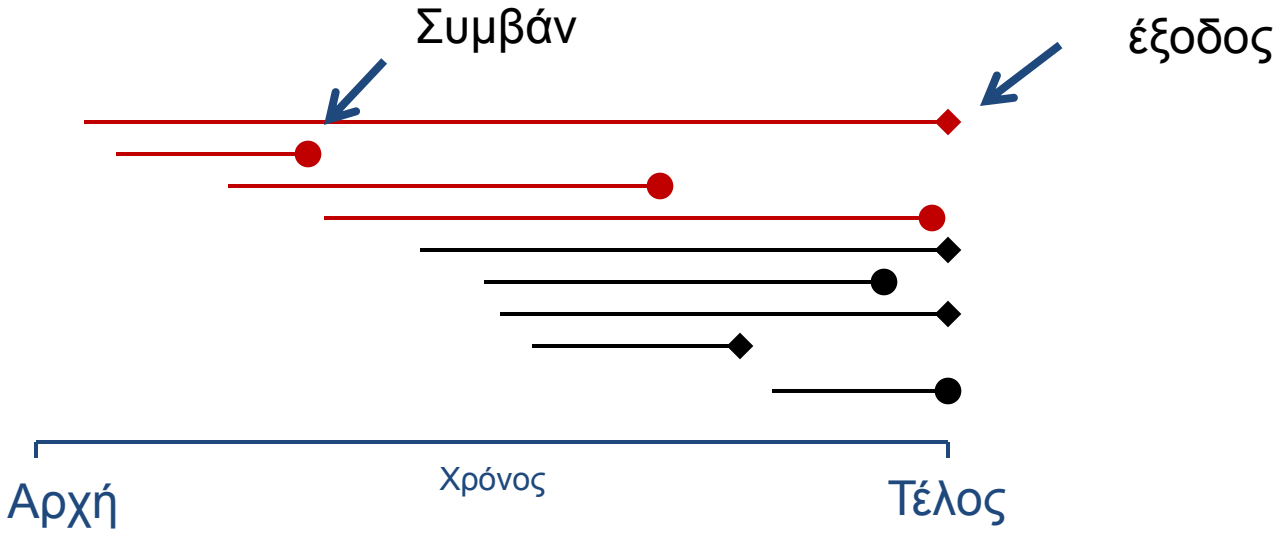
Μη παραμετρική ανάλυση

- Kaplan-Meier καμπύλες επιβίωσης
- Logrank τεστ

Μοντέλα παλινδρόμησης

- Παλινδρόμηση Cox

Δεδομένα του τύπου *'χρόνος για το συμβάν'*



$$RR = \frac{\frac{1}{4}}{\frac{3}{5}} = 0.41$$

Το RR επιβίωσης είναι παραπλανητικό διότι δεν παίρνει υπόψη του τη διάσταση του χρόνου

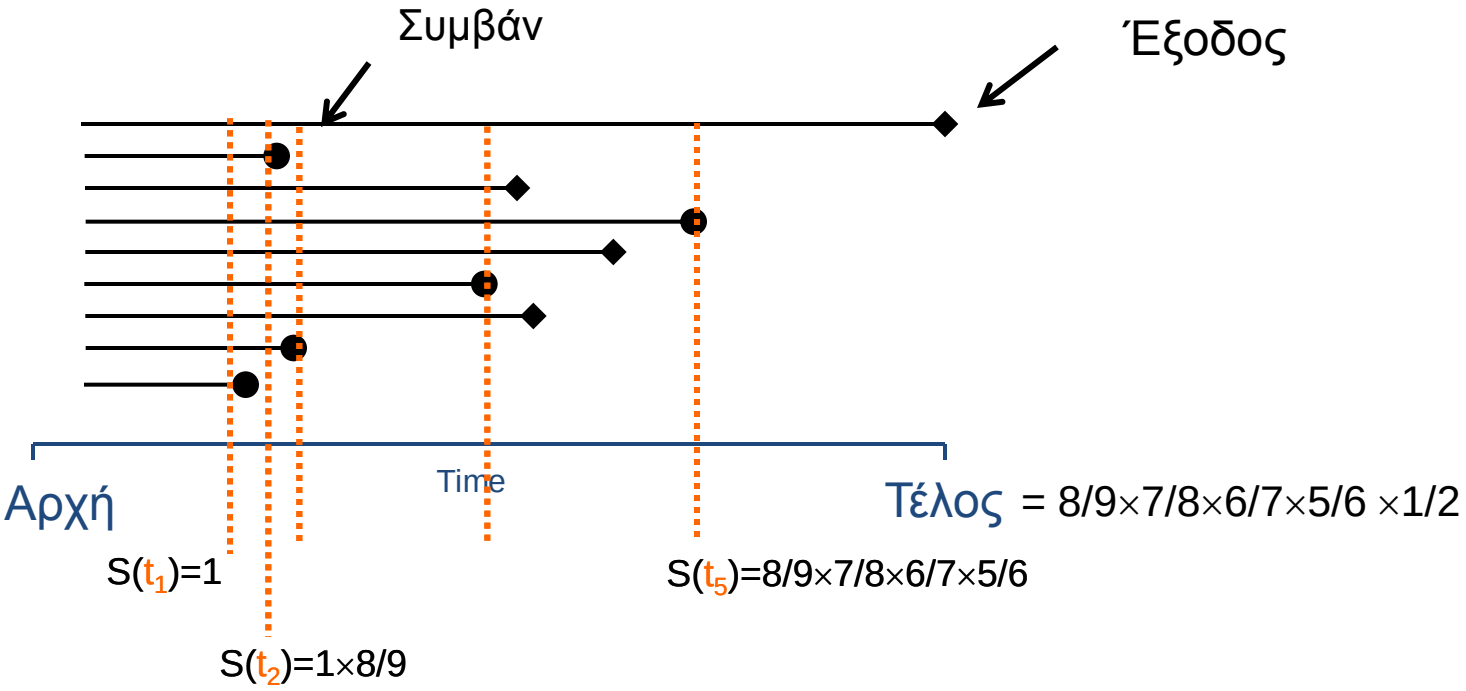
Διαφορές από τους μέσους όρους των χρόνων επιβίωσης δεν παίρνουν υπόψη αυτούς που έφυγαν από την μελέτη

Σύγκριση δυο ομάδων ως προς την επιβίωση

- Χρησιμοποιώντας τις συναρτήσεις επιβίωσης
  - Kaplan-Meier καμπύλες
  - Το Logrank τεστ
- Χρησιμοποιώντας την hazard function (συνάρτηση στιγμιαίου κινδύνου)
  - Cox μοντέλο

Σύγκριση με καμπύλες επιβίωσης

Καμπύλη επιβίωσης S(t)



Kaplan Meier Εκτιμητής επιβίωσης

Η συνάρτηση επιβίωσης αναφέρεται στην πιθανότητα κάποιος να επιβιώσει μέχρι το χρονικό σημείο t

Εκτιμητής Kaplan-Meier

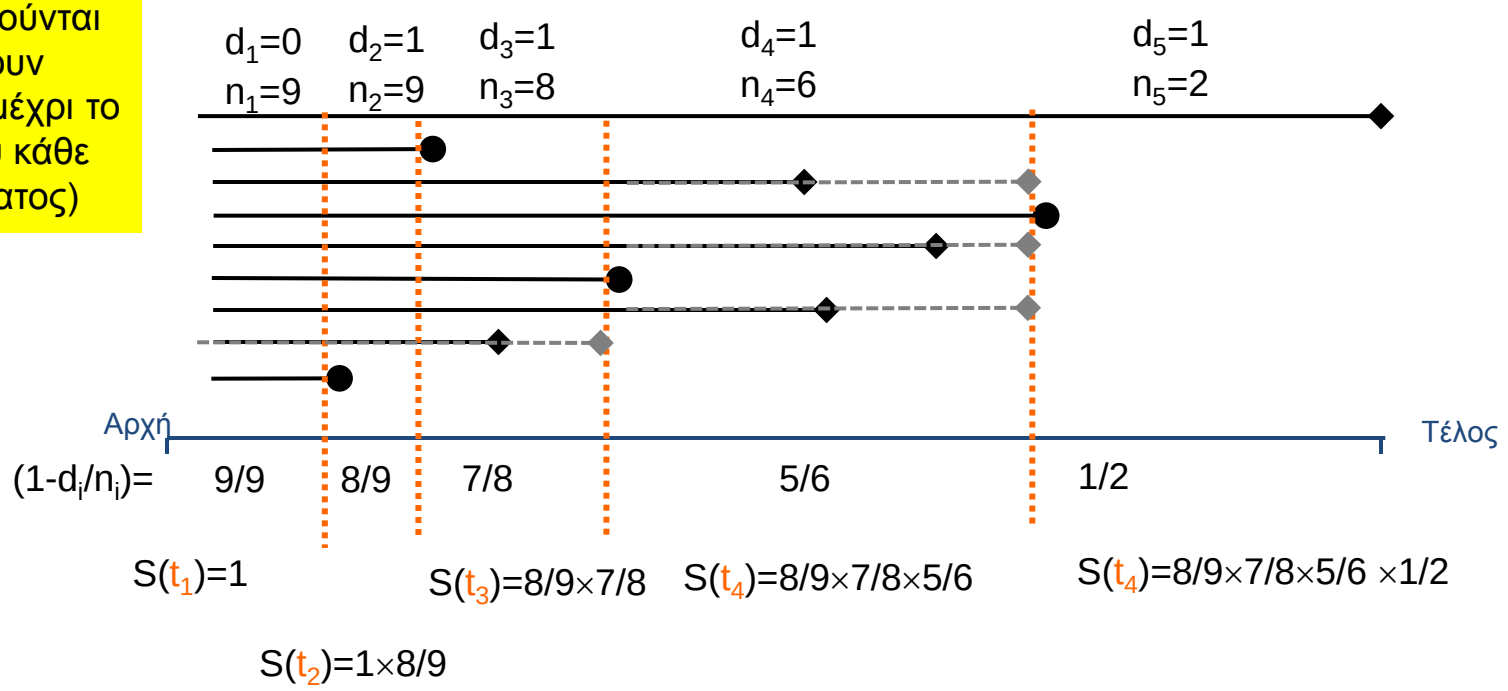
- Χωρίζουμε τον χρόνο σε διαστήματα t_i που ορίζονται από τα γεγονότα (π.χ. θανάτους *i=1,2,…)*
- Στην αρχή κάθε διαστήματος (**πριν** το συμβάν) υπάρχουν n_i άτομα ‘σε κίνδυνο’
- Ο εκτιμητής Kaplan-Meier είναι

$$S(t) = \prod_{t_i < t} \left(1 - \frac{d_i}{n_i} \right)$$

Τα συμβάντα λαμβάνουν χώρα στην αρχή κάθε διαστήματος και οι έξοδοι λαμβάνουν χώρα στο τέλος κάθε διαστήματος (και θεωρούνται ότι έχουν επιβιώσει μέχρι το τέλος του κάθε διαστήματος)

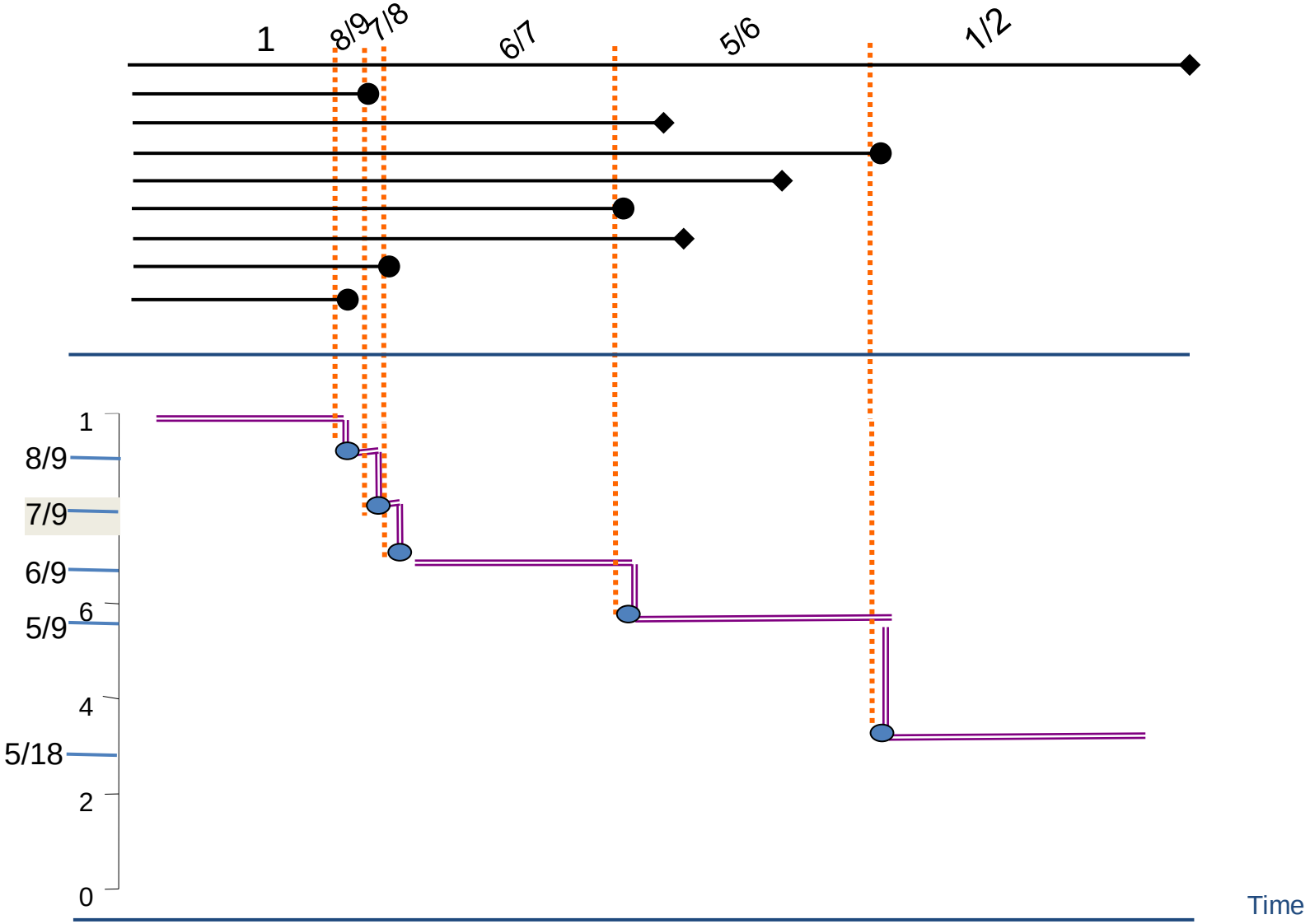
Kaplan Meier Εκτιμητής επιβίωσης

Καμπύλη επιβίωσης S(t)



Η συνάρτηση επιβίωσης αναφέρεται στην πιθανότητα κάποιος να επιβιώσει μέχρι το χρονικό σημείο t

Καμπύλη επιβίωσης S(t)



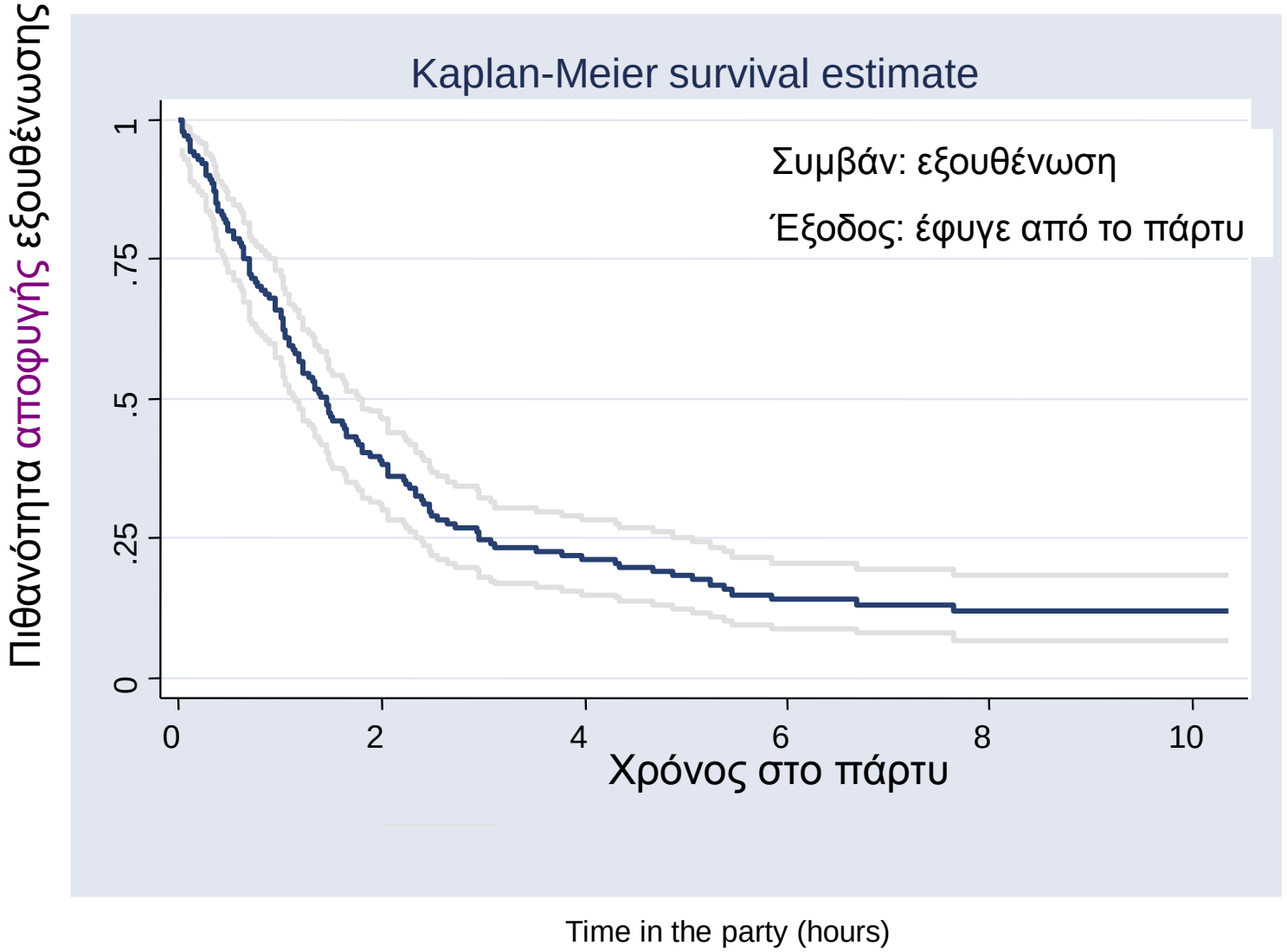
Διασπορά S(t)

Όπως όλα τα μεγέθη, η καμπύλη επιβίωσης έχει αβεβαιότητα

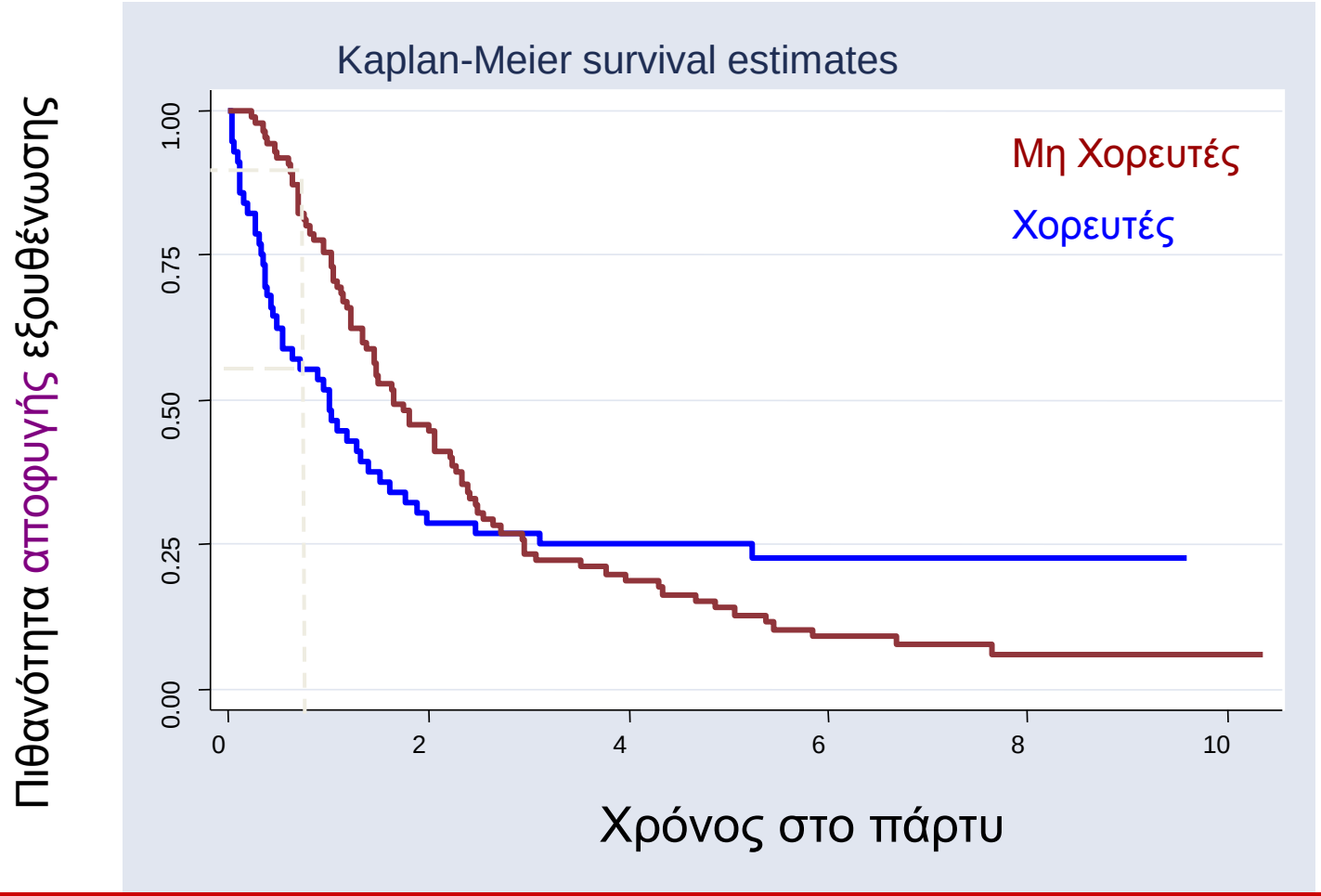
$$var(S(t)) = (S(t))^2 \sum_{t_i < t} \left(\frac{d_i}{n_i (n_i - d_i)} \right)$$

προς το τέλος, οι καμπύλες τείνουν να γίνονται πιο αβέβαιες, με μεγαλύτερα διαστήματα εμπιστοσύνης

Π.χ.: Χρόνος μέχρι εξουθένωσης σε πάρτυ



Π.χ: Σύγκριση καμπύλων επιβίωσης



Πως θα συγκρίνουμε τις δύο καμπύλες συνολικά;

Logrank τεστ

- Έλεγχος για την υπόθεση H_0

  'Οι καμπύλες επιβίωσης για τις δύο ομάδες είναι ίδιες'

- Είναι μη παραμετρικό τεστ
  - Καμία παραδοχή για την μορφή των καμπύλων επιβίωσης

- Συνυπολογίζει τις εξόδους
- Μέγιστη ισχύ όταν οι καμπύλες δεν διασταυρώνονται

Logrank τέστ

- Σε κάθε χρονικό συμβάν *t* κάνουμε υπολογισμούς
  - Υπολογίζουμε τα προσδοκώμενα συμβάντα υποθέτοντας ίση πιθανότητα και στις δύο ομάδες
    e.g. $O_{t,Dancing}$, $E_{t,Dancing}$, $O_{t,no\ Dancing}$, $E_{t,no\ Dancing}$
  - Υπολογίζουμε ένα άλλο μέγεθος V_t (πιο περίπλοκο…)

- Αθροίζουμε τους υπολογισμούς από όλα τα χρονικά σημεία

$$\text{Logrank} = \frac{\left(\sum_t O_{t,Dancers} - \sum_t E_{t,Dancers}\right)^2}{\sum_t V_t} \sim \chi_{1df}$$

Logrank τεστ: υπολογισμοί

Χρόνος	O_D	N_D	O_{ND}	N_{ND}	$E_{t,D}$	V_t
1h	1	20	0	15	1×20/35	0.24
1.2h	1	19	1	15	2×19/34	0.48
….	…	…	…	…		
10.5h	…	…	…	…		
Άθροισμα	*O*				*E*	*V*

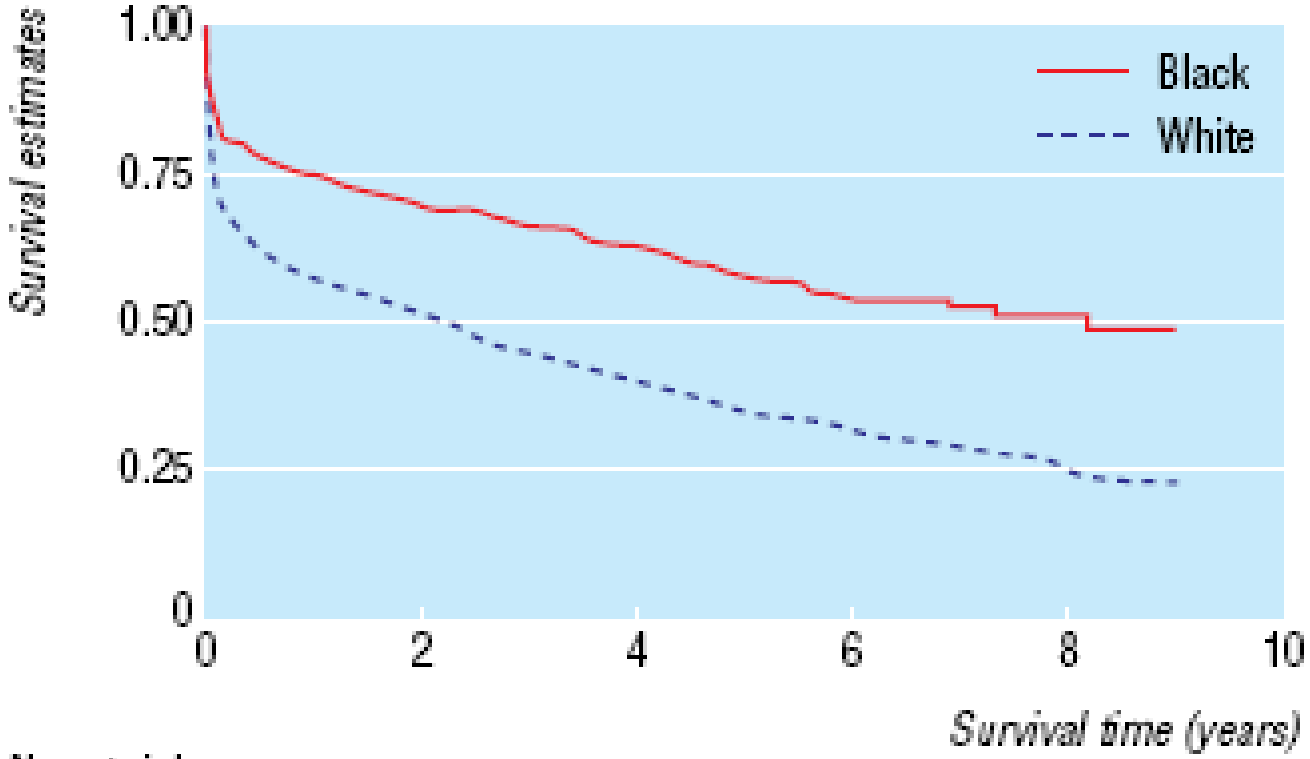
$$E_{t,D} = \frac{(O_D + O_{ND}) \times N_D}{(N_D + N_{ND})}$$

$$V_t = \frac{N_D \times N_{ND} \times (O_D + O_{ND}) \times (N_D + N_{ND} - (O_D + O_{ND}))}{(N_D + N_{ND})^2 \times (N_D + N_{ND} - 1)}$$

$$\text{Logrank} = \frac{(O - E)^2}{V}$$

Survival differences after stroke in a multiethnic population: follow-up study with the south London stroke register

Charles D A Wolfe, Nigel C Smeeton, Catherine Coshall, Kate Tilling, Anthony G Rudd

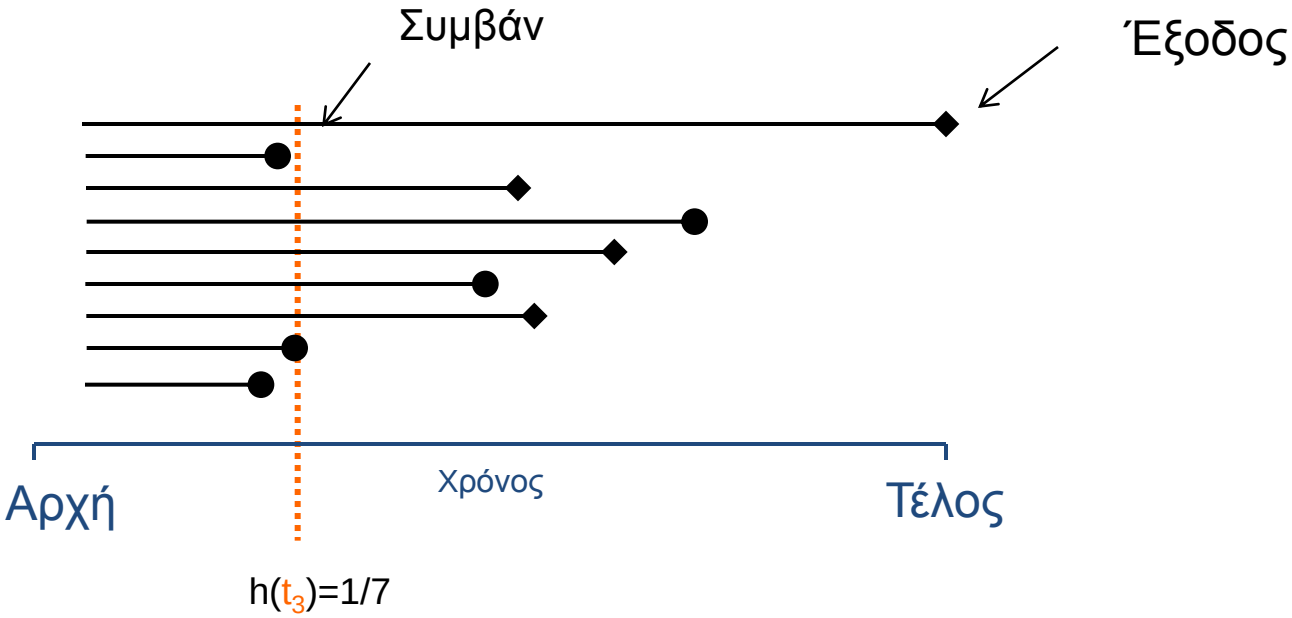


Nos at risk

White	1721	802	482	231	58
Black	414	253	160	81	23

The Kaplan-Meier survival curve showed a clear difference between the two groups (figure), with white people having poorer survival (log rank test $P < 0.001$). After adjustment for age

Συνάρτηση στιγμιαίου κινδύνου
Hazard function



$$h(t) = \frac{\text{events at t}}{\text{indiv at risk at t}} = \frac{d_t}{r_t}$$

Όταν $t \to 0$

Η hazard function περιγράφει την στιγμιαία πιθανότητα ενός γεγονότος στον χρόνο t μεταξύ των ατόμων που επιβίωσαν μέχρι t.

Σύγκριση των στιγμιαίων κινδύνων

$$HR = \frac{\text{Στιγμιαίος κίνδυνος (hazard) στους χορευτές}}{\text{Στιγμιαίος κίνδυνος στους μη χορευτές}}$$

$$SE(\ln HR) = \frac{1}{\sqrt{V}}$$

V είναι η διασπορά του Logrank τεστ

- HR=1 or lnHR=0 ο στιγμιαίος κίνδυνος στις δύο ομάδες είναι ίδιος
- 95% CI : $\ln HR \pm \frac{1.96}{\sqrt{V}}$
- π.χ. HR=2

"Ένας χορευτής έχει διπλό στιγμιαίο κίνδυνο να εξαντληθεί σε σχέση με έναν μη χορευτή'

Αυτός ο υπολογισμός παίρνει υπόψη του αυτούς που έφυγαν από το πάρτυ (και δεν ξέρουμε τι απέγιναν!)

Σύγκριση επιβίωσης με μοντέλα για hazard functions

Το μοντέλο Cox

- Μοντελοποιούμε την hazard function παίρνοντας υπόψη και μεταβλητές
- Συγκρίνουμε δύο ομάδες:

Ομάδα 1 (Όχι χορευτές)

$h(t) = h_0(t)$

Ομάδα 2 (χορευτές)

$h(t) = h_0(t) \times e^{\beta}$

$HR = e^{\beta}$

Baseline hazard (στιγμιαίος κίνδυνος στην ομάδα αναφοράς)

$$h(t) = h_0(t) \times e^{\beta \times D}$$

$D = 1$ Χορευτής

$D = 0$ Όχι χορευτής

Cox μοντέλο

Μοντέλο Cox : Γενίκευση

$$h(t) = h_0(t) \times e^{\beta_1 \times X_1 + \beta_2 \times X_2 + \ldots + \beta_n \times X_n}$$

$X_1, X_2, \ldots, X_n$ μεταβλητές (συνεχείς ή διχότομες)

Διχότομες μεταβλητές: HR συγκρίνει στιγμιαίους κινδύνους μεταξύ ατόμων που, π.χ. έχουν εκτεθεί στους παράγοντες X_1 and X_2 με αυτούς που έχουν μόνο έκθεση στον X_3

$$HR = \frac{h_0 \times e^{\beta_1 + \beta_2}}{h_0 \times e^{\beta_3}} = e^{\beta_1 + \beta_2 - \beta_3}$$

Συνεχείς μεταβλητές: HR συγκρίνει στιγμιαίους κινδύνους μεταξύ ατόμων που έχουν τιμές u_1 για την έκθεση X_1 με αυτούς που έχουν τιμή u_1 για τον ίδιο παράγοντα

$$HR = \frac{h_0 \times e^{\beta_1 u_1}}{h_0 \times e^{\beta_1 u_2}} = e^{\beta_1 (u_1 - u_2)}$$

Το μοντέλο Cox : Ερμηνεία των συντελεστών

Διχότομη μεταβλητή: Χ είναι ‘χορός’

Ο στιγμιαίος κίνδυνος για έναν χορευτή είναι e^{β} φορές τον στιγμιαίο κίνδυνο για έναν μη χορευτή

Συνεχής μεταβλητή : Χ είναι ‘ηλικία σε χρόνια’

Ο στιγμιαίος κίνδυνος θα είναι e^{β} φορές για ένα άτομο ένα χρόνοι μεγαλύτερο από κάποιο άλλο

Κατηγορική μεταβλητή : Χ είναι ’μη καπνιστής’ (0), ‘πρώην καπνιστής’ (1) ή ‘καπνιστής’ (2)

Πρέπει να φτιάξουμε ‘ψευτομεταβλητές’

V_1 για σύγκριση 1 με 0

V_2 για σύγκριση 2 με 0

Έπειτα ερμηνεύουμε όπως για τις διχότομες μεταβλητές

Το μοντέλο Cox : Φανταστικό παράδειγμα

$$h(t) = h_0(t) \times e^{\beta_1 \times \text{Χορός} + \beta_2 \times \text{Κάπνισμα} + \beta_3 \times \text{Ηλικία}}$$

Coefficient	Value	SE	HR	95% Conf limits		P-value
β_1	0.41	0.020	1.50	1.44	1.56	< 0.01
β_2	0.01	0.010	1.01	0.99	1.03	0.32
β_3	0.10	0.003	1.10	1.09	1.11	< 0.01

$e^{0.41}=1.5$ $e^{0.01}=1$ $e^{0.1}=1.1$

Υπόθεση ανάλογων στιγμιαίων κινδύνων (Proportional hazards assumption)

Το Cox μοντέλο υποθέτει

1. Ο λόγος στιγμιαίων κινδύνων δεν αλλάζει με το χρόνο
   - Η επίδραση των μεταβλητών του μοντέλου δεν αλλάζουν με το χρόνο = Δεν υπάρχει αλληλεπίδραση χρόνου-μεταβλητής
   - Το αντίθετο: $\beta(t)$ ο συντελεστής είναι συνάρτηση του χρόνου
2. Οι μεταβλητές μπαίνουν στο μοντέλο με γραμμικό τρόπο
   - Το αντίθετο: $\beta(X)$ ο συντελεστής είναι συνάρτηση της μεταβλητής

Επιλογή των μεταβλητών

- *'Στατιστικά σημαντικές'* μεταβλητές είναι αυτές που δίνουν στατ. σημαντικό HR (CI δεν περιέχει 1) ή β (CI δεν περιέχει 0)
- Κρατάμε μέσα στο μοντέλο στατ. σημαντικές μεταβλητές ή κλινικά σημαντικούς προγνωστικούς παράγοντες
- Αποφασίζουμε για το καλύτερο μοντέλο όπως στην λογαριθμιστική παλινδρόμηση:

με βάση την πιθανοφάνεια

(όσο μεγαλύτερη τόσο καλύτερο)

Με το Likelihood Ratio test (λόγοι πιθανοφανειών)

Επανάληψη περί Likelihood, Deviance και προσαρμογή του μοντέλου

- Likelihood
  - Όσο μεγαλύτερη τόσο καλύτερη η προσαρμογή
- Deviance $D = -2 \log(\text{Likelihood})$
  - Όσο μικρότερη τόσο καλύτερη η προσαρμογή

- Όσο πιο πολλές μεταβλητές στο μοντέλο, τόσο πιο καλή φαίνεται να είναι η προσαρμογή του μοντέλου

LRT: γενίκευση

- LRT= D(μοντέλο χωρίς τις p μεταβλητές)- D(μοντέλο με τις p μεταβλητές)

$LRT \sim X^2$ με p βαθμούς ελευθερίας

  - P-value<0.05 : Το μοντέλο με τις επιπλέον μεταβλητές είναι στατιστικά καλύτερο από το μοντέλο χωρίς τις μεταβλητές $\Rightarrow$ κρατάμε τις μεταβλητές ως σημαντικές
  - P-value>0.05 : Το μοντέλο με τις επιπλέον μεταβλητές δεν είναι στατιστικά καλύτερο από το μοντέλο χωρίς τις μεταβλητές $\Rightarrow$πετάμε τις μεταβλητές ως ασήμαντες
- Προσοχή ! Τέτοιοι 'απλοϊκοί' κανόνες δεν είναι πάντα χρήσιμοι – να κοιτάμε και την κλινική πλευρά του θέματος

Επιλογή μοντέλων: φανταστικό παράδειγμα

$$h(t) = h_0(t) \times e^{\beta_1 \times \text{Dancing} + \beta_2 \times \text{Smoking} + \beta_3 \times \text{Age}}$$

$$h(t) = h_0(t) \times e^{\beta_1 \times \text{Dancing} + \beta_2 \times \text{Smoking}}$$

$$h(t) = h_0(t) \times e^{\beta_1 \times \text{Dancing}}$$

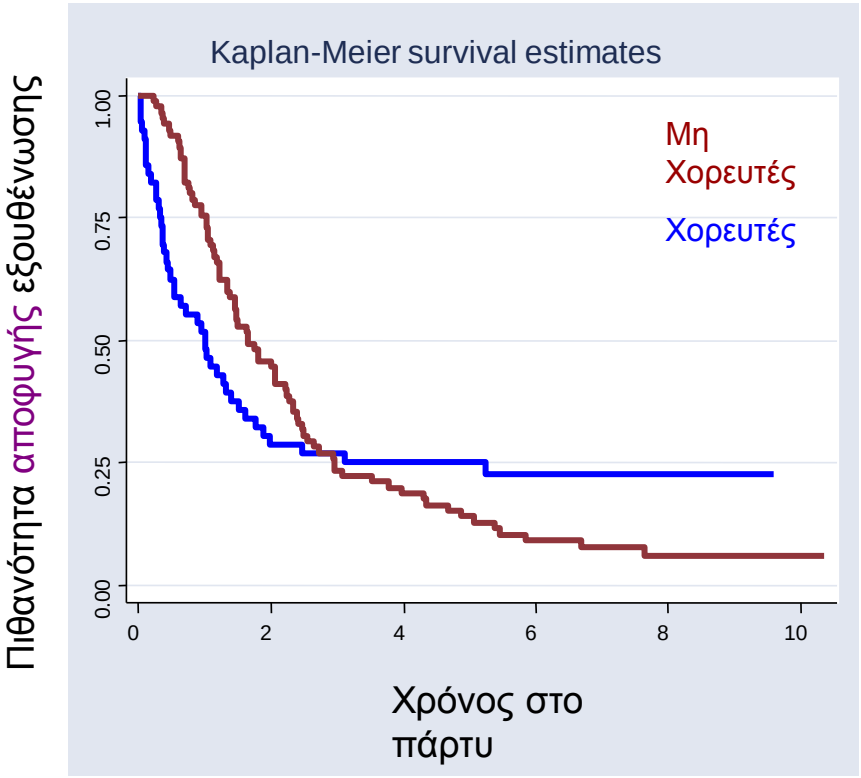
Σύγκρινε για να αποφανθείς για την σημαντικότητα της ηλικίας (df=1)

Σύγκρινε για να αποφανθείς για την σημαντικότητα του καπνίσματος (df=1)

για
ς και
df=2)

Μοντέλο	Deviance	Σύγκριση	X^2	P-value (1 df)
Dancing Smoking Age	$D_1 = 238.5$			
Dancing Smoking	$D_2 = 239.8$	$D_2 - D_1$	1.3	0.25
Dancing	$D_3 = 244.1$	$D_3 - D_2$	4.3	0.04

Υπόθεση ανάλογων στιγμιαίων κινδύνων



- $h(t)=h0(t)\times e^{\beta\times \text{Χορός}}$
- Έστω $\beta=0.4$
- $e^{\beta} = 1.5 = HR$
- Ένας χορευτής έχει 1.5 φορές τον στιγμιαίο κίνδυνο να εξουθενωθεί σε σχέση με έναν μη χορευτή
- Αυτό είναι μάλλον παραπλανητικό! (δες τις Kaplan-Meier καμπύλες)
- Αυτός είναι ένας μάλλον πολύ ‘στο περίπου’ τρόπος για να ελέγξουμε την υπόθεση αναλογίας – υπάρχουν τεστ και καλύτερες μέθοδοι