JOHNS HOPKINS
BLOOMBERG SCHOOL of PUBLIC HEALTH

# Lecture 12

## Finish case-study of log-linear regression applied to binned survival data
## Continuous time survival analysis

# Review of Lecture 12

- The data contains information about *time to death* for inpatients hospitalized for a severe mental disorder. Survival time from hospitalization is in years.

- Patients are censored: i.e. we don't get to follow patients long enough to see when the event occurs for all patients.

- In the data, "censor" is 1 if censored; 0 if the patient died; "age" of hospitalization for mental disorder is in years; "male" is 1 for males and 0 for females.

```
##    survive censor age male event
## 1        1      0  58    0     1
## 2        1      0  51    0     1
## 3        2      0  55    0     1
## 4       11      0  48    0     1
## 5       14      0  47    0     1
## 6       22      0  28    0     1
## 7       24      0  45    0     1
## 8       26      0  43    0     1
## 9       31      1  31    0     0
## 10      32      0  25    0     1
## 11      35      1  35    0     0
## 12      35      1  33    0     0
## 13      36      1  25    0     0
## 14      37      1  30    0     0
## 15      40      0  36    0     1
```

# Review of Lecture 12

- We “binned” the information about survival into 10-year increments of follow-up

```
##    survive censor age male event
## 1        1      0  58    0     1
## 2        1      0  51    0     1
## 3        2      0  55    0     1
## 4       11      0  48    0     1
## 5       14      0  47    0     1
## 6       22      0  28    0     1
## 7       24      0  45    0     1
## 8       26      0  43    0     1
## 9       31      1  31    0     0
## 10      32      0  25    0     1
## 11      35      1  35    0     0
## 12      35      1  33    0     0
## 13      36      1  25    0     0
## 14      37      1  30    0     0
## 15      40      0  36    0     1
```

```
##   Cutoff male pyears  n event  rate midp
## 1   0-10    0    124 15     3 0.024    5
## 2  11-20    0    105 12     2 0.019   15
## 3  21-30    0     82 10     3 0.037   25
## 4  31-40    0     36  7     2 0.056   35
## 5   0-10    1    110 11     0 0.000    5
## 6  11-20    1    110 11     0 0.000   15
## 7  21-30    1     95 11     3 0.032   25
## 8  31-40    1     25  6     1 0.040   35
```

# Review of Lecture 12

- Incidence: risk per unit time of the event occurring among those that enter the interval

- Hazard: the limit of the incidence rate as the interval width goes to zero
  - Crude estimate: number of events divided by the person-time experienced in the interval

- Want a smooth estimate of incidence/hazard using a log linear model: $\lambda_i = exp(X_i^{'}\beta)$

- Assume the number of events per interval $Y_i$ ~ Poisson($PT_i\lambda_i$)

$$E(Y_i) = \lambda_i PT_i = exp(log(PT_i) + X_i^{'}\beta)$$

# Review of Lecture 12

- Model A:

$$\text{Model A: } E(Y_i) = \lambda_i PT_i = exp(log(PT_i) + \beta_0)$$

```
fitA = glm(event~1,offset=log(pyears),data=binned,family="poisson")
summary(fitA)
```

```
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.8933     0.2673  -14.57   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
lincom(fitA,"(Intercept)",eform=TRUE)
```

```
##               Estimate     2.5 %     97.5 %    Chisq   Pr(>Chisq)
## (Intercept) 0.02037846 0.0120692 0.03440838 212.2069 4.533119e-48
```

# Review of Lecture 12

- Model fitted values, i.e. Expected deaths per interval time

```
##   Cutoff male pyears  n event  rate midp  expected
## 1   0-10    0    124 15     3 0.024    5 2.5269287
## 2  11-20    0    105 12     2 0.019   15 2.1397380
## 3  21-30    0     82 10     3 0.037   25 1.6710335
## 4  31-40    0     36  7     2 0.056   35 0.7336245
## 5   0-10    1    110 11     0 0.000    5 2.2416303
## 6  11-20    1    110 11     0 0.000   15 2.2416303
## 7  21-30    1     95 11     3 0.032   25 1.9359534
## 8  31-40    1     25  6     1 0.040   35 0.5094614
```

# Log-linear model, Model B

- Model B

$$\text{Model B: } E(Y_i) = \lambda_i PT_i = exp(log(PT_i) + \beta_0 + \beta_1 male_i)$$

```r
fitB = glm(event~1+male,offset=log(pyears),data=binned,family="quasipoisson")
summary(fitB)
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3.5467     0.3864  -9.180 9.42e-05 ***
## male         -0.8959     0.7228  -1.239    0.261
...
```

# Log-linear model; Model B

```
lincom(fitB,c("(Intercept)","(Intercept)+male","male"),eform=TRUE)
```

```
##                    Estimate   2.5 %       97.5 %     Chisq    Pr(>Chisq)
## (Intercept)        0.02881844 0.01351409  0.06145458 84.26334 4.330714e-20
## (Intercept)+male   0.01176471 0.003552796 0.03895757 52.88403 3.538347e-13
## male               0.4082353  0.09899778  1.683432   1.536181 0.2151872
```

# Log-linear model; Models C and D

- We would expect the hazard of death to depend on how long one has been in the hospital since we do not live forever. Models C and D estimate the relative risk of death for men as compared to women, controlling for a time-varying baseline hazard

- In survival analysis, we refer to the "baseline hazard" as the hazard function when setting exposure variables to 0.

$$E(Y_i) = \lambda_i PT_i = exp(log(PT_i) + f(time_i) + X_i^{'}\beta)$$

$$\text{Model C: } E(Y_i) = \lambda_i PT_i = exp(log(PT_i) + \beta_0 + \beta_1 midp_i + \beta_2 male_i)$$

$$\text{Model D: } E(Y_i) = \lambda_i PT_i = exp(log(PT_i) + \beta_0 + \beta_1 I(midp_i = 15) + \beta_2 I(midp_i = 25) + \beta_3 I(midp_i = 35) + \beta_4 male_i)$$

# Log-linear model; Models C and D

```
fitC = glm(event~1+male+midp,offset=log(pyears),data=binned,family="quasipoisson")
summary(fitC)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.55525    0.60351  -7.548 0.000647 ***
## male        -0.88461    0.54687  -1.618 0.166674
## midp         0.05391    0.02444   2.206 0.078504 .
```

```
lincom(fitC,"male",eform=TRUE)
```

```
##       Estimate    2.5 %   97.5 %    Chisq Pr(>Chisq)
## male 0.4128736 0.141358 1.205907 2.616609  0.1057502
```

# Log-linear model; Models C and D

```
fitD = glm(event~1+male+as.factor(midp),offset=log(pyears),data=binned,family="quasipoisson")
summary(fitD)
```

```
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -4.0321     0.6024  -6.694   0.0068 **
## male                -0.8909     0.5975  -1.491   0.2327
## as.factor(midp)15   -0.2863     0.9187  -0.312   0.7757
## as.factor(midp)25    1.0282     0.7122   1.444   0.2445
## as.factor(midp)35    1.2965     0.8219   1.577   0.2128
```

```
lincom(fitD,"male",eform=TRUE)
```

```
##        Estimate     2.5 %   97.5 %    Chisq Pr(>Chisq)
## male 0.4102721 0.1271973 1.323323 2.223383  0.1359349
```

# Log-linear models: Model E

- Finally, we look for evidence that the relative rate for men as compared to women changes over the duration of follow-up
  - I.e. the proportional hazards assumption is inadequate for our data.

- Model E: we center the midpoint variable at 20 years duration so that the male coefficient has a more reasonable interpretation and include interaction between male and years of hospitalization

$$\text{Model E: } E(Y_i) = \lambda_i PT_i = exp(log(PT_i) + \beta_0 + \beta_1(midp_i - 20) + \beta_2 male_i + \beta_3(midp_i - 20)male_i)$$

```r
binned$midc = binned$midp - 20
fitE = glm(event~1+male*midc,offset=log(pyears),data=binned,family="quasipoisson")
summary(fitE)
```

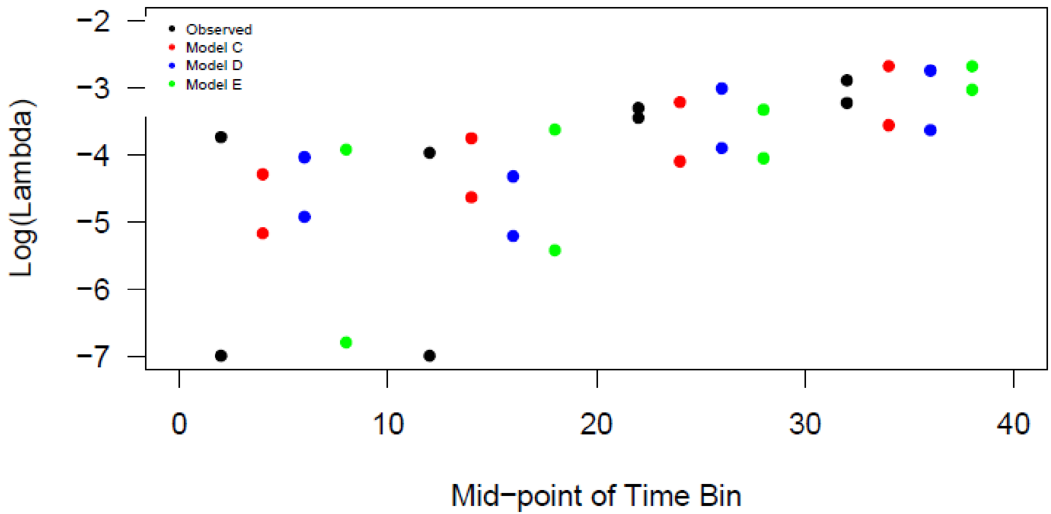
```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.46957    0.24520 -14.150 0.000145 ***
## male        -1.26510    0.58774  -2.152 0.097711 .
## midc         0.02981    0.02344   1.272 0.272360
## male:midc    0.10782    0.05454   1.977 0.119211
```

# Log-linear models: Model E

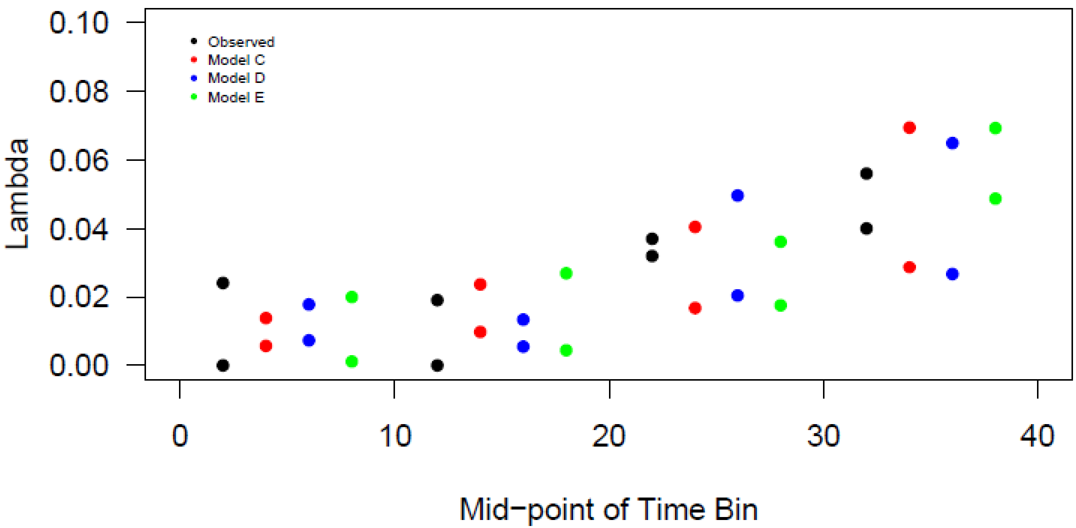
```
lincom(fitE,c("male-15*male:midc","male-5*male:midc","male+5*male:midc","male+15*male:midc"),eform=TRUE
```

```
##                    Estimate   2.5 %       97.5 %    Chisq     Pr(>Chisq)
## male-15*male:midc 0.05600358 0.004905048 0.639423  5.38189   0.02034682
## male-5*male:midc  0.1646096  0.03624414  0.7476055 5.460166  0.0194548
## male+5*male:midc  0.4838319  0.1839557   1.272553  2.165207  0.1411656
## male+15*male:midc 1.422112   0.3629106   5.572732  0.2553863 0.6133077
```

# Model comparison

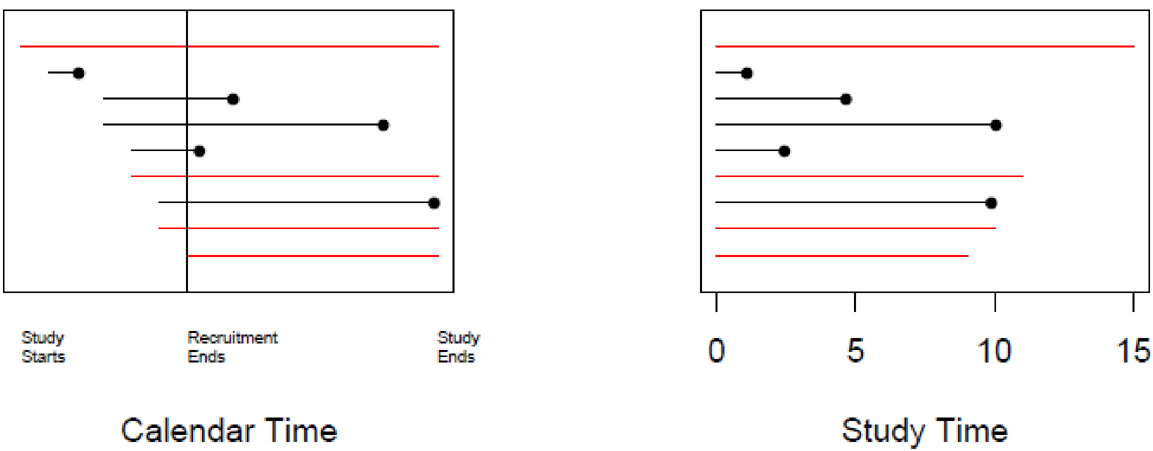


# Model comparison



# Continuous time survival analysis

- Binning survival times is convenient when working from administrative data or data where you do not have access to individual level data
- Most natural to treat time as continuous
- Review definition of censoring



# Continuous time survival analysis

- Absent censoring, the survival outcome $Y_i$, is the time from start of an at risk period to when the event of interest occurs.

- In the presence of censoring, we get to see $\delta_i = 1$ if the event occurs and 0 if the even is censored
  - $T_i = \min(D_i, C_i)$ where $D_i$ is the time when the event occurs and $C_i$ is the time of censoring
  - Data for patient i is $(T_i, \delta_i)$

- Goals:
  - Determine if the survival experience differs across exposure groups
  - Predict survival experience

# Key survival analysis definitions

Let $T$ be a time to event random variable, $T \geq 0$.

Then we will define a series of quantities that can be used to describe the distribution of $T$.

- Cumulative Distribution Function: $F(t) = Pr(T \leq t)$
- Survival Function: $S(t) = Pr(T > t) = 1 - F(t)$
- Density function: $f(t) = \frac{d}{dt}F(T)$

- Hazard function: $h(t) = \lim_{dt \to 0} \frac{Pr(t < T \leq t + dt | T > t)}{dt}$

# Key survival analysis definitions

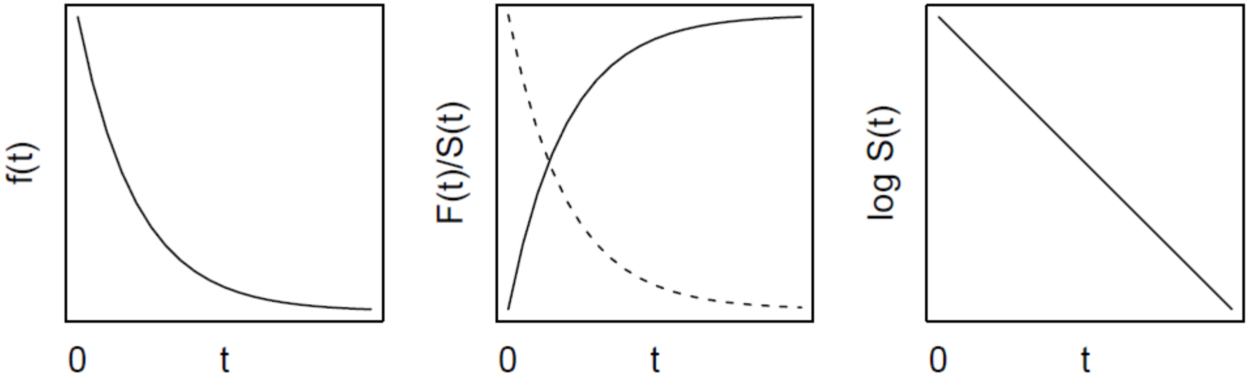
$$
\begin{aligned}
h(t) &= \lim_{dt \to 0} \frac{Pr(t < T \leq t + dt | T > t)}{dt} \\
&= \lim_{dt \to 0} \frac{Pr(t < T \leq t + dt \text{ and } T > t)}{Pr(T > t)dt} \\
&= \lim_{dt \to 0} \frac{Pr(t < T \leq t + dt \text{ and } T > t)}{dtS(t)} \\
&= \frac{f(t)}{S(t)} \\
&= \frac{f(t)}{1 - F(t)} \\
&= \frac{dF(t)}{dt} / [1 - F(t)] \\
&= -\frac{d}{dt}[1 - F(t)]/[1 - F(t)] \\
&= -\frac{d}{dt} S(t)/S(t) \\
&= -\frac{d}{dt} log_e S(t)
\end{aligned}
$$

- Cumulative hazard function: $H(t) = \int_0^t h(u)du = log_e S(t)$. This implies: $S(t) = e^{-\int_0^t h(u)du} = e^{-H(t)}$

# Common Parametric Models; Exponential

Assume $T \sim Exponential(\lambda)$ then

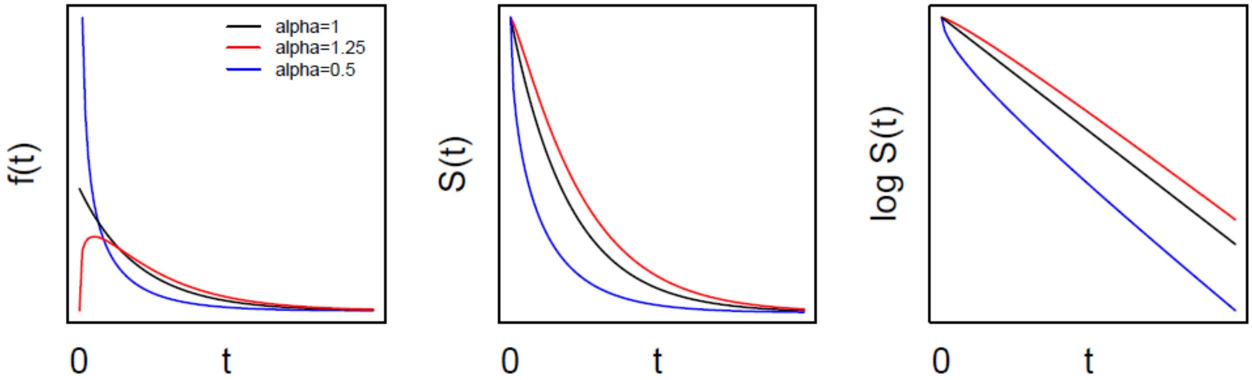
- $F(t) = 1 - e^{-\lambda t}$, $S(t) = e^{-\lambda t}$
- $f(t) = \frac{d}{dt}(1 - e^{-\lambda t}) = \lambda e^{-\lambda t}$
- $E(T) = 1/\lambda$, $Var(T) = 1/\lambda^2$
- $h(t) = f(t)/S(t) = \lambda e^{-\lambda t}/e^{-\lambda t} = \lambda$, i.e. a constant hazard model



# Common Parametric Models: Gamma

Assume $T \sim Gamma(\alpha, \lambda)$, then

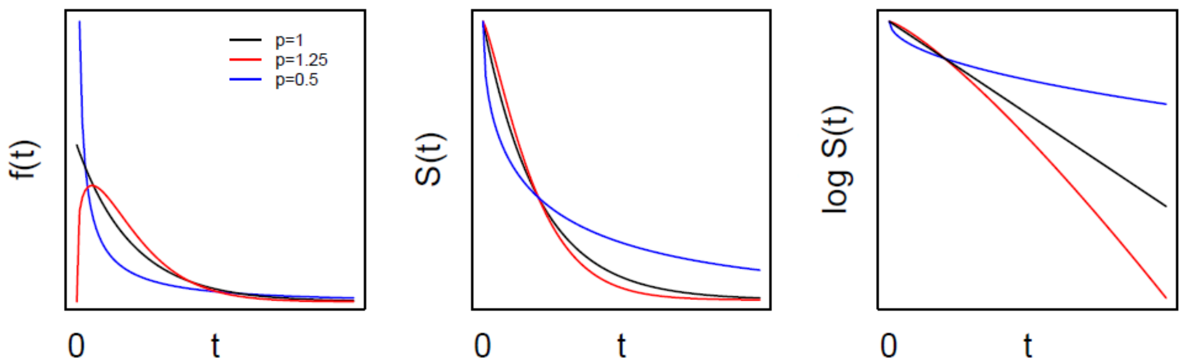
- $f(t) = \frac{\lambda^{\alpha} t^{\alpha-1} e^{-\lambda t}}{\Gamma(\alpha)}$, $t > 0$, $\Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1} e^{-t} dt$
- $F(t)$, $S(t)$ and $h(t)$ have to be solved by numerical integration; there are no closed form solutions.



# Common Parametric Models: Weibull

Assume $T \sim Weibull(\lambda, p)$, then

- $f(t) = p\lambda t^{p-1} e^{-(\lambda t)^p}$
- $F(t) = 1 - e^{-(\lambda t)^p}$, $S(t) = e^{-(\lambda t)^p}$
- $h(t) = p\lambda^p t^{p-1}$
- When $p = 1$, $Weibull(\lambda, 1) = Exponential(\lambda)$.



# Analysis of survival analysis outcomes in continuous time

- Estimating S(t) via Kaplan-Meier survival function estimate (now)
- Testing whether S1(t) = S2(t), via the log-rank test (Lab 7)
- Regression of survival outcomes on exposures via Cox Proportional Hazards regression models (Lecture 14)

# Kaplan-Meier estimate of the survival function

The Kaplan-Meier estimate of the survival function $S(t)$ is also known as the **Product-limit** estimator.

This estimator for the survival function assumes that:

- censoring is unrelated to prognosis, i.e. event process and censoring process are independent
- the survival probabilities are the same for subjects recruited early and late in the study
- the events happened at the times specified

To construct the Kaplan-Meier estimator, you need to order the unique event times and compute:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Event times: | $t_1$ | $<$ | $t_2$ | $<$ | ... | $<$ | $t_J$ |
| No. at risk: | $N_1$ | $>$ | $N_2$ | $>$ | ... | $>$ | $N_J$ |
| No. of events: | $y_1$ | | $y_2$ | | ... | | $y_J$ |

The estimate of $S(t)$ is 1 if $t < t_1$ and

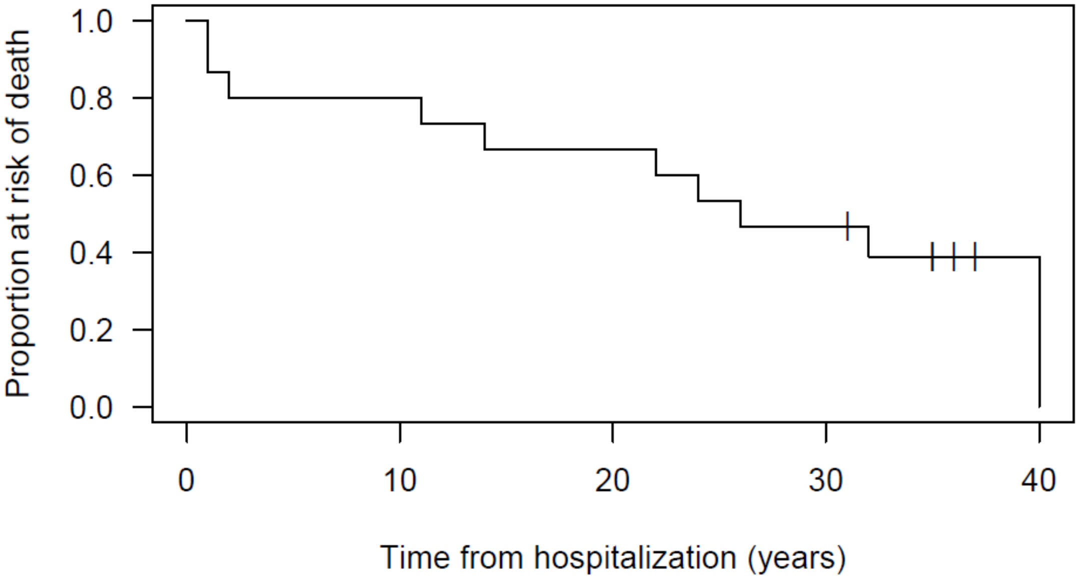
$$\hat{S}(t) = \prod_{j:t_j \leq t} \left( \frac{N_j - y_j}{N_j} \right)$$

# Kaplan-Meier estimate of the survival function

- Using the data for inpatients hospitalized for a severe mental disorder, we will be computing the Kaplan-Meier estimate of the survival function for the female patients. Survival time from hospitalization is in years.

| | 1 | 1 | 2 | 11 | 14 | 22 | 24 | 26 | 31+ | 32 | 35+ | 35+ | 36+ | 37+ | 40 | Ni | yi | (Ni-yi)/Ni | S(t) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 15 | 2 | 0.867 | 0.867 |
| 2 | | | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 13 | 1 | 0.923 | 0.800 |
| 3 | | | | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 12 | 0 | 1.000 | 0.800 |
| 4 | | | | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 12 | 0 | 1.000 | 0.800 |
| 5 | | | | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 12 | 0 | 1.000 | 0.800 |
| 6 | | | | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 12 | 0 | 1.000 | 0.800 |
| 7 | | | | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 12 | 0 | 1.000 | 0.800 |
| 8 | | | | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 12 | 0 | 1.000 | 0.800 |
| 9 | | | | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 12 | 0 | 1.000 | 0.800 |
| 10 | | | | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 12 | 0 | 1.000 | 0.800 |
| 11 | | | | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 12 | 1 | 0.917 | 0.733 |
| 12 | | | | | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 11 | 0 | 1.000 | 0.733 |
| 13 | | | | | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 11 | 0 | 1.000 | 0.733 |
| 14 | | | | | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 11 | 1 | 0.909 | 0.667 |

# Kaplan-Meier estimate of survival function



# Greenwood's formula for variance of S(t)

An estimate of the variance of $\hat{S}(t)$ based on Greenwood's formula (application of Delta method) is:

$$\hat{V}ar(\hat{S}(t)) = \hat{S}(t)^2 \sum_{j:t_j \le t} \frac{y_j}{N_j(N_j - y_j)}$$

A 95% confidence interval for $S(t)$ can be derived as:

$$\hat{S}(t) \pm 1.96\sqrt{\hat{V}ar(\hat{S}(t))}$$

with imposing the constraint that the confidence interval lies in $[0, 1]$, i.e. if the bounds of the confidence interval go outside $[0, 1]$, set the values to 0 or 1, respectively. This is unappealing in many respects!

# Variance of S(t) estimate based on complementary log-log

An alternative to Greenwoods formula for the variance, a variance estimate can be derived based on the complementary Log-Log transformation.

Let $v(t) = log[-logS(t)]$. Note that $S(t) \in [0,1]$ and $v(t) \in [-\infty, \infty]$.

$$\hat{V}ar(\hat{v}(t)) = \sum_{j:t_j \leq t} \frac{y_j}{N_j(N_j - y_j)} \left[ \sum_{j:t_j \leq t} log\left(\frac{N_j - y_j}{N_j}\right) \right]^{-2}$$

The 95% confidence interval for $v(t)$ is given by:

$$\hat{v}(t) \pm 1.96\sqrt{\hat{V}ar(\hat{v}(t))}$$

where we can define the upper and lower bound as $\hat{v}_L(t)$ and $\hat{v}_U(t)$.

NOTE: $S(t) = exp(-exp(v(t)))$, so the 95% confidence interval for $S(t)$ is:

$$[exp(-exp(\hat{v}_U(t))), exp(-exp(\hat{v}_L(t)))]$$

# Example calculations: Greenwood's formula

Compute the 95% confidence interval for $S(2)$:

1. Using Greenwood's formula:

$$
\begin{aligned}
\hat{V}ar(\hat{S}(2)) &= \hat{S}(2)^2 \sum_{j:t_j \le 2} \frac{y_j}{N_j(N_j - y_j)} \\
&= \hat{S}(2)^2 \left[ \frac{y_1}{N_1(N_1 - y_1)} + \frac{y_2}{N_2(N_2 - y_2)} \right] \\
&= 0.8^2 \left[ \frac{2}{15 \times (15-2)} + \frac{1}{13 \times (13-1)} \right] \\
&= 0.0107
\end{aligned}
$$

95% CI for $S(2)$: $0.8 \pm 1.96 * \sqrt{0.0107} \rightarrow (0.598, 1.003)$

# Example calculations: Complementary log-log

2. Using the Complementary Log-Log transformation

$$
\begin{aligned}
\hat{v}(2) &= log(-log(\hat{S}(2))) \\
&= log(-log(0.8)) \\
&= -1.50
\end{aligned}
$$

$$
\begin{aligned}
\hat{V}ar(\hat{v}(2)) &= \sum_{j:t_j \le 2} \frac{y_j}{N_j(N_j - y_j)} \left[ \sum_{j:t_j \le 2} log\left(\frac{N_j - y_j}{N_j}\right) \right]^{-2} \\
&= \left[ \frac{y_1}{N_1(N_1-y_1)} + \frac{y_2}{N_2(N_2-y_2)} \right] \left[ log\left(\frac{N_1-y_1}{N_1}\right) + log\left(\frac{N_2-y_2}{N_2}\right) \right]^{-2} \\
&= \left[ \frac{2}{15\times 13} + \frac{1}{13\times 12} \right] [log(13/15) + log(12/13)]^{-2} \\
&= 0.335
\end{aligned}
$$

95% CI for $v(2)$ is: $\hat{v}(2) \pm 1.96\sqrt{\hat{V}ar(\hat{v}(2))}$ is $-1.50 \pm 1.96\sqrt{0.335}$ is $(-2.63, -0.36)$.

95% CI for $S(2)$ is: $(exp(-exp(-0.36)), exp(-exp(-2.63)))$ is $(0.50, 0.93)$.

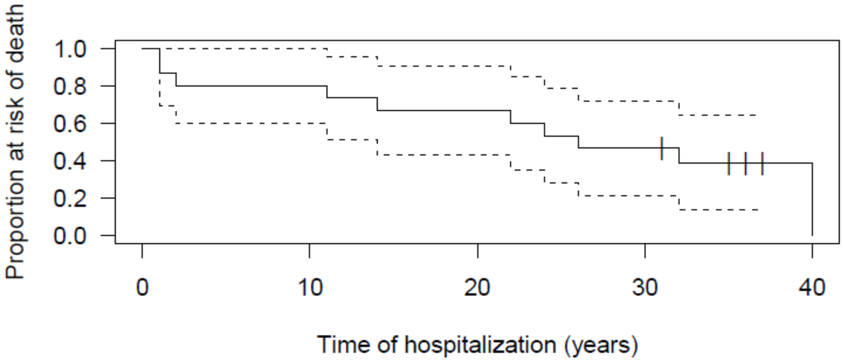
# R implementation

```
library(survival)
St.green = survfit(Surv(survive,event) ~ 1, data = d.female,
                   type = "kaplan-meier",
                   conf.type = "plain")
St.cll = survfit(Surv(survive,event) ~ 1, data = d.female,
                 type = "kaplan-meier",
                 conf.type = "log-log")
summary(St.green)
```

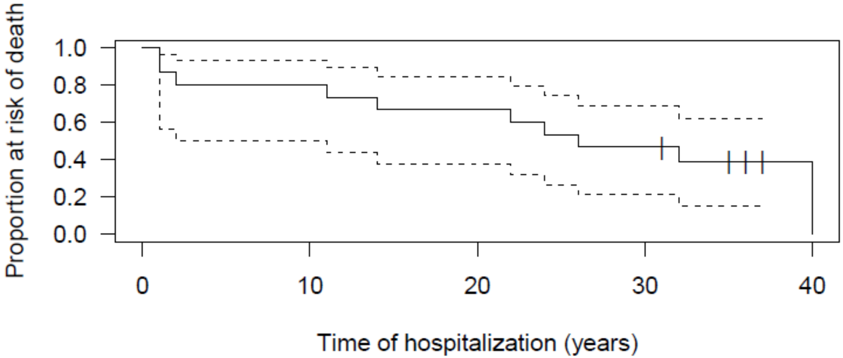
```
## Call: survfit(formula = Surv(survive, event) ~ 1, data = d.female,
##     type = "kaplan-meier", conf.type = "plain")
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     1     15       2    0.867  0.0878        0.695        1.000
##     2     13       1    0.800  0.1033        0.598        1.000
##    11     12       1    0.733  0.1142        0.510        0.957
##    14     11       1    0.667  0.1217        0.428        0.905
##    22     10       1    0.600  0.1265        0.352        0.848
##    24      9       1    0.533  0.1288        0.281        0.786
##    26      8       1    0.467  0.1288        0.214        0.719
##    32      6       1    0.389  0.1287        0.137        0.641
##    40      1       1    0.000     NaN          NaN          NaN
```

# R implementation

Confidence intervals: Greenwoods formula




Confidence intervals: Log( − Log S(t))



# Where to next....

- Lab: log-rank test comparing two survival functions
- Thursday: Cox proportional hazards model