Modeling

Survival Analysis

Dr Vedaste Ndahindwa

University of Rwanda
School of Public Health

Contents

Background

- In logistic regression, we were interested in studying how risk factors were associated with presence or absence of disease for example.
- Or we may have study dropout, and therefore subjects who we are not sure if they had disease or not. In these cases, logistic regression is not appropriate.
- Survival analysis is used to analyze data in which the time until the event is of interest. The response is often referred to as a failure time, survival time, or event time.

Examples

- Time until tumor recurrence.
- Time until cardiovascular death after some treatment intervention
- Time until AIDS for HIV patients
- Time until a machine part fails

The survival time response

- Usually continuous
- May be incompletely determined for some subjects
- For some subjects we may know that their survival time was at least equal to some time t. Whereas, for other subjects, we will know their exact time of event.
- Incompletely observed responses are censored
- Is always ≥ 0.

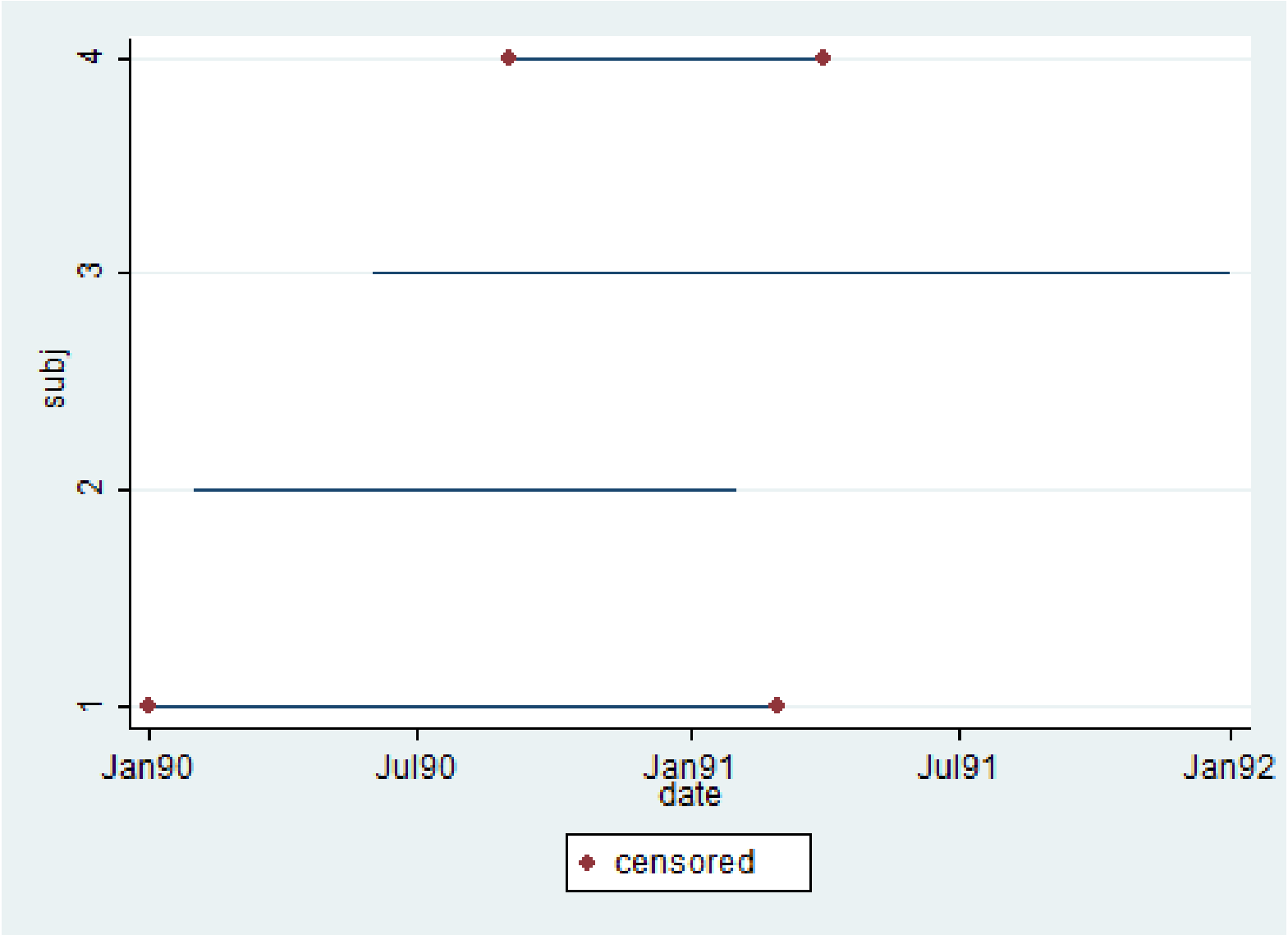
Censoring defined

Definition: Censoring occurs when cases are lost

What are the types:

- **Left censoring**: When the patient experiences the event in question before the beginning of the study observation period.
- **Interval censoring**: When the patient is followed for awhile and then goes on a trip for awhile and then returns to continue being studied.
- **Right censoring**:
  - A patient is lost to follow-up within the study period.
  - Experiences the event after the observation period

Censoring



Analysis issues

- If there is no censoring, standard regression procedures could be used.
- However, these may be inadequate because
  - Time to event is restricted to be positive and has a skewed distribution..
  - The probability of surviving past a certain point in time may be of more interest than the expected time of event
  - The hazard function, used for regression in survival analysis, can lend more insight into the failure mechanism than linear regression.

Terminology and notation

- T denotes the response variable, $T \geq 0$
- The survival function is

$$S(t) = Pr(T > t) = 1 - F(t)$$

- The survival function gives the probability that a subject will survive past time t.
- As t ranges from 0 to ∞, the survival function has the following properties
  - It is non-increasing
  - At time $t = 0, S(t) = 1$. In other words, the probability of surviving past time 0 is 1
  - At time $t = \infty, S(t) = S(\infty) = 0$. As time goes to infinity, the survival curve goes to 0.

Survival data

How do we record and represent survival data with censoring?

- T_i denotes the response for the i^{th} subject
- Let C_i denote the censoring time for the i^{th} subject
- Let δ_i denote the event indicator
  - 1 if the event was observed ($Ti \leq C_i$)
  - 0 if the response was censored ($Ti > C_i$)
- The observed response is $Y_i = min(T_i, C_i)$.

Survival Analysis Preprocessing

The stset command

- This command identifies the survival time variable as well as the censoring variable.
- It sets up stata variables that indicate the entry, exit, and censoring time

```
stset studytime, failure(died)
```

Summary description of survival data

`stdes`

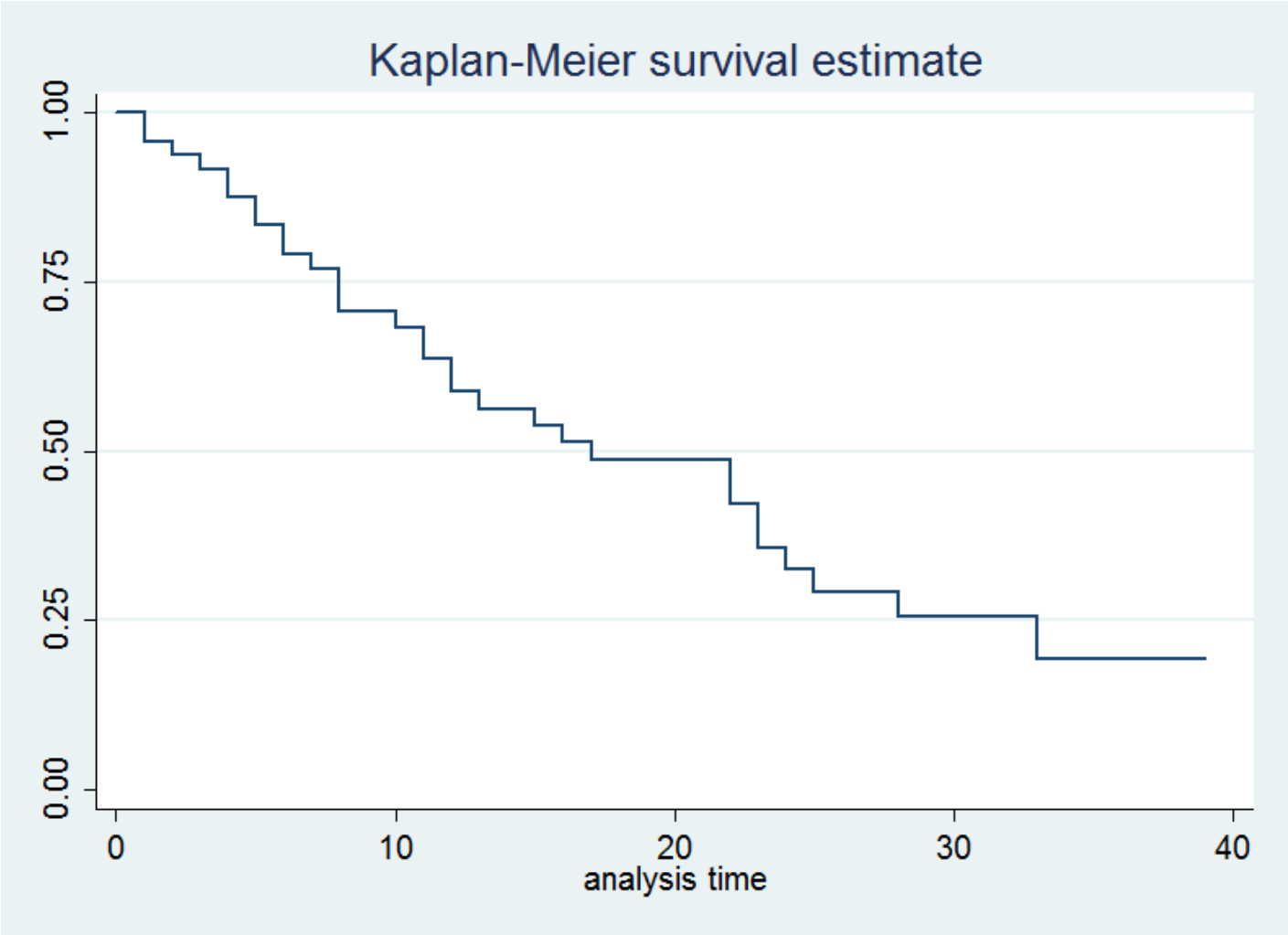
- This command describes summary information about the data set.
- It provides summary statistics about the number of subjects, records, time at risk, failure events, etc

Stata output

```
. stdes

        failure _d:  died
  analysis time _t:  studytime

                                |-------------- per subject --------------|
Category                  total        mean         min      median         max
------------------------------------------------------------------------------
no. of subjects              48
no. of records               48           1           1           1           1

(first) entry time                        0           0           0           0
(final) exit time                      15.5           1        12.5          39

subjects with gap             0
time on gap if gap            0
time at risk                744        15.5           1        12.5          39

failures                     31   .6458333           0           1           1
------------------------------------------------------------------------------
```

Survival Probability of data set

```
sts graph
```



Basic Survival Analysis Theory

- We are interested in the Survivorship function $S(t)$
- The Survivorship function is a function of the probability of surviving plotted against time.
- We use the *cancer.dta* provided with STATA
- We graph the survivorship function

Computation of $S(t)$

- Suppose the study time is divided into periods, the number of which is designated by the letter, t.
- The survivorship probability is computed by multiplying a proportion of people surviving for each period of the study.
- If we subtract the conditional probability of the failure event for each period from one, we obtain that quantity.
- The product of these quantities constitutes the survivorship function.

Survival Function

The survival probability is equal to the product of 1 minus the conditional probability of the event of interest.

$$S(t) = \prod_{t=1}^{T}(1 - h_i(t))$$

where
$S(t)$: estimated survivorship function at time t
$h(t)$: conditional probability of event at time t

The nature of the data

- The data are non-normal in distribution.
- They are right skewed.
- There may be varying degrees of censoring in the data.
- We have to use a nonparametric test to determine whether the survival curves are statistically different from one another.

Examining the Survival Probability

```
. sts list

        failure _d:  died
  analysis time _t:  studytime
```

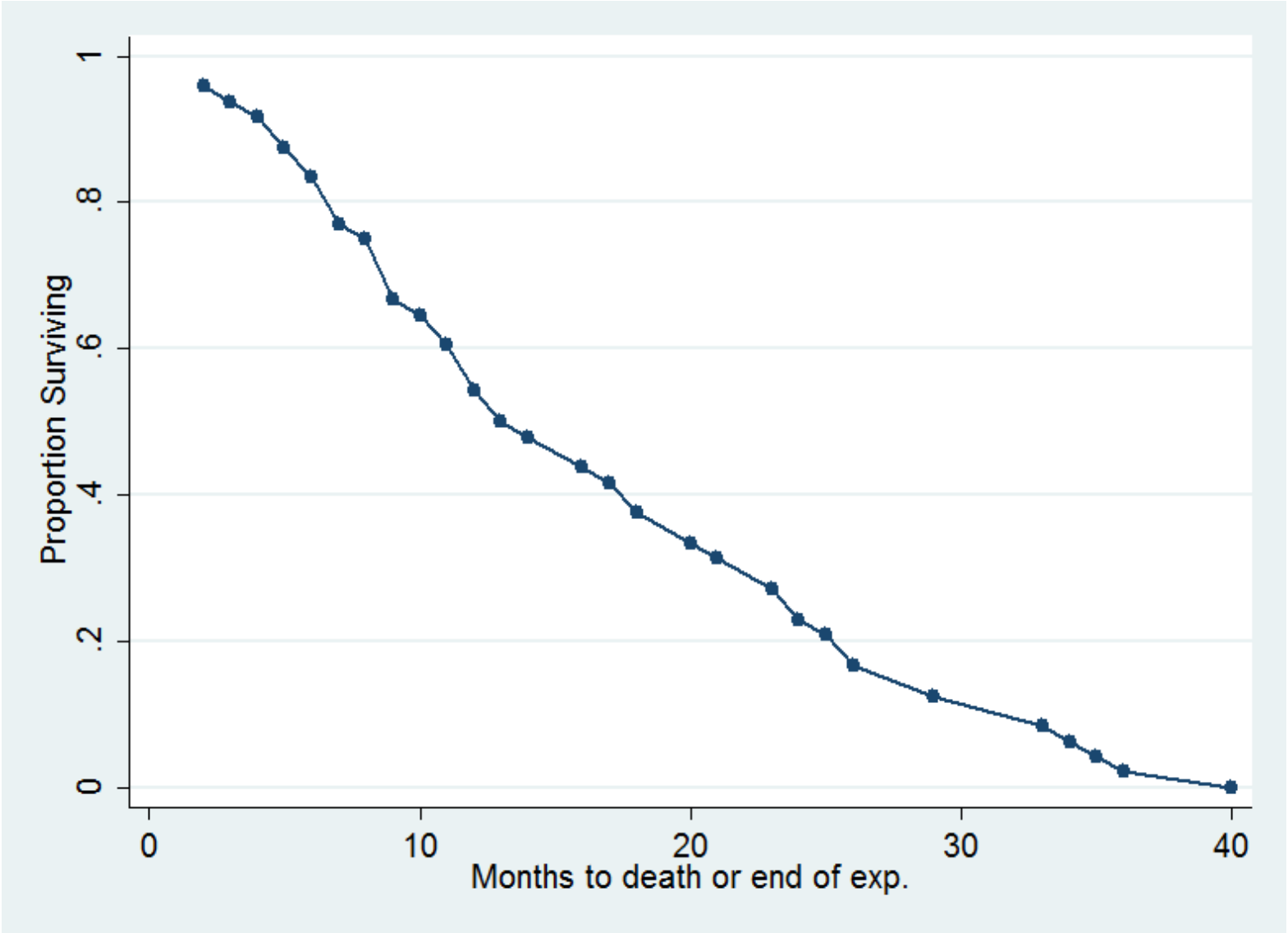
Time	Beg. Total	Fail	Net Lost	Survivor Function	Std. Error	[95% Conf.	Int.]
1	48	2	0	0.9583	0.0288	0.8435	0.9894
2	46	1	0	0.9375	0.0349	0.8186	0.9794
3	45	1	0	0.9167	0.0399	0.7930	0.9679
4	44	2	0	0.8750	0.0477	0.7427	0.9418
5	42	2	0	0.8333	0.0538	0.6943	0.9129
6	40	2	1	0.7917	0.0586	0.6474	0.8820
7	37	1	0	0.7703	0.0608	0.6236	0.8656
8	36	3	1	0.7061	0.0661	0.5546	0.8143
9	32	0	1	0.7061	0.0661	0.5546	0.8143
10	31	1	1	0.6833	0.0678	0.5302	0.7957
11	29	2	1	0.6362	0.0708	0.4807	0.7564
.							
.							
39	1	0	1	0.1918	0.0791	0.0676	0.3634

The Life Tables Analysis

```
. ltable studytime
```

Interval		Beg. Total	Deaths	Lost	Survival	Std. Error	[95% Conf.	Int.]
1	2	48	2	0	0.9583	0.0288	0.8435	0.9894
2	3	46	1	0	0.9375	0.0349	0.8186	0.9794
3	4	45	1	0	0.9167	0.0399	0.7930	0.9679
4	5	44	2	0	0.8750	0.0477	0.7427	0.9418
5	6	42	2	0	0.8333	0.0538	0.6943	0.9129
6	7	40	3	0	0.7708	0.0607	0.6245	0.8660
7	8	37	1	0	0.7500	0.0625	0.6020	0.8495
8	9	36	4	0	0.6667	0.0680	0.5148	0.7807
9	10	32	1	0	0.6458	0.0690	0.4936	0.7628
10	11	31	2	0	0.6042	0.0706	0.4521	0.7262
11	12	29	3	0	0.5417	0.0719	0.3917	0.6696
12	13	26	2	0	0.5000	0.0722	0.3526	0.6307
13	14	24	1	0	0.4792	0.0721	0.3334	0.6110
15	16	23	2	0	0.4375	0.0716	0.2956	0.5707
.								
.								
39	40	1	1	0	0.0000	.	.	.

Graphing the survival probability

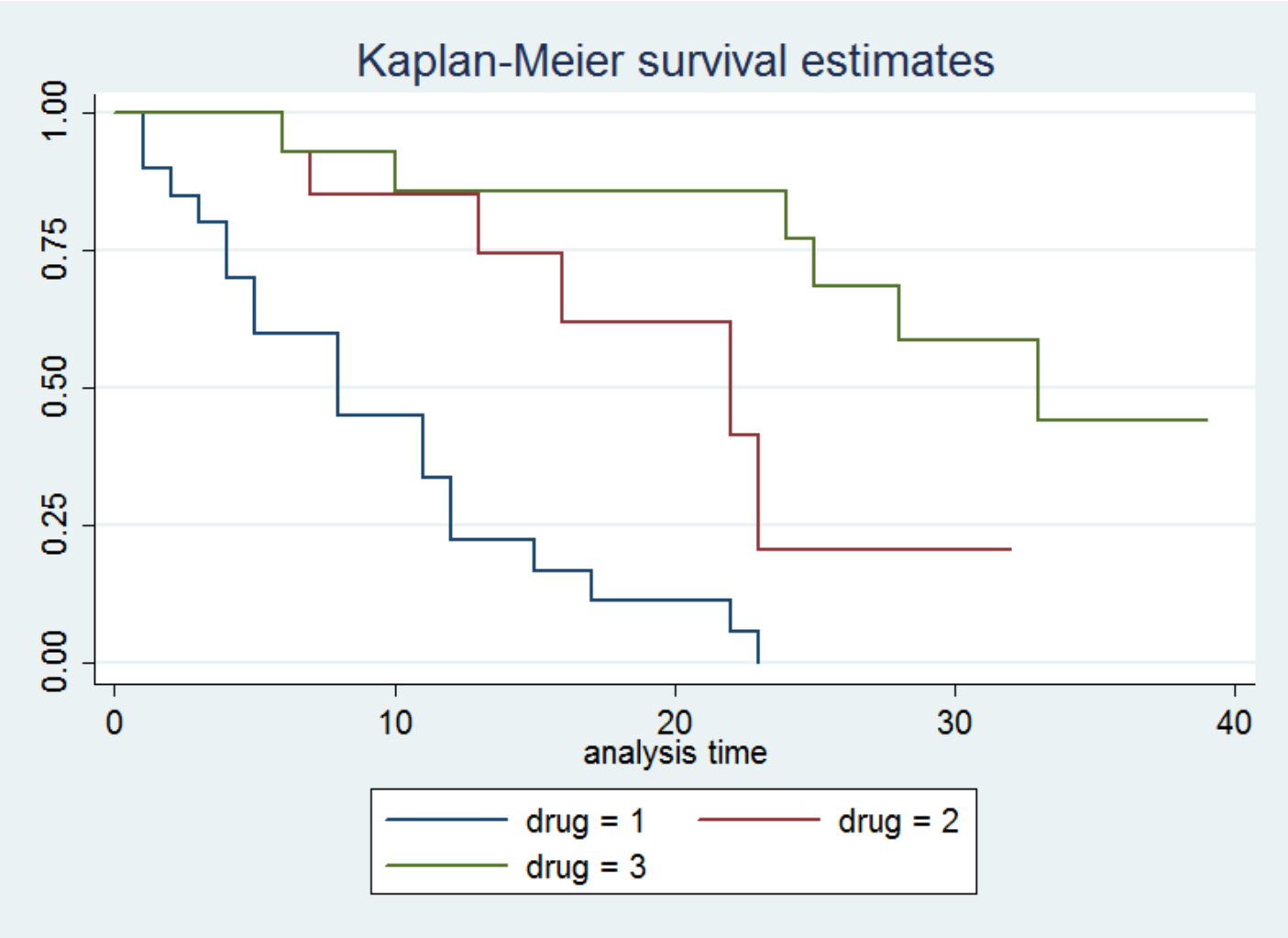


Hypothesis testing

- We need to develop tests that determine whether the survival rates are now statistically significantly different from one another
- If we were conducting a cancer clinical trial and were trying to slow down the impending death of terminally ill patients, we might test three different drugs.
- The drugs in the three treatment arms of this clinical trial, we designate as drugs 1, 2, and 3. We plot the survival functions of the three groups

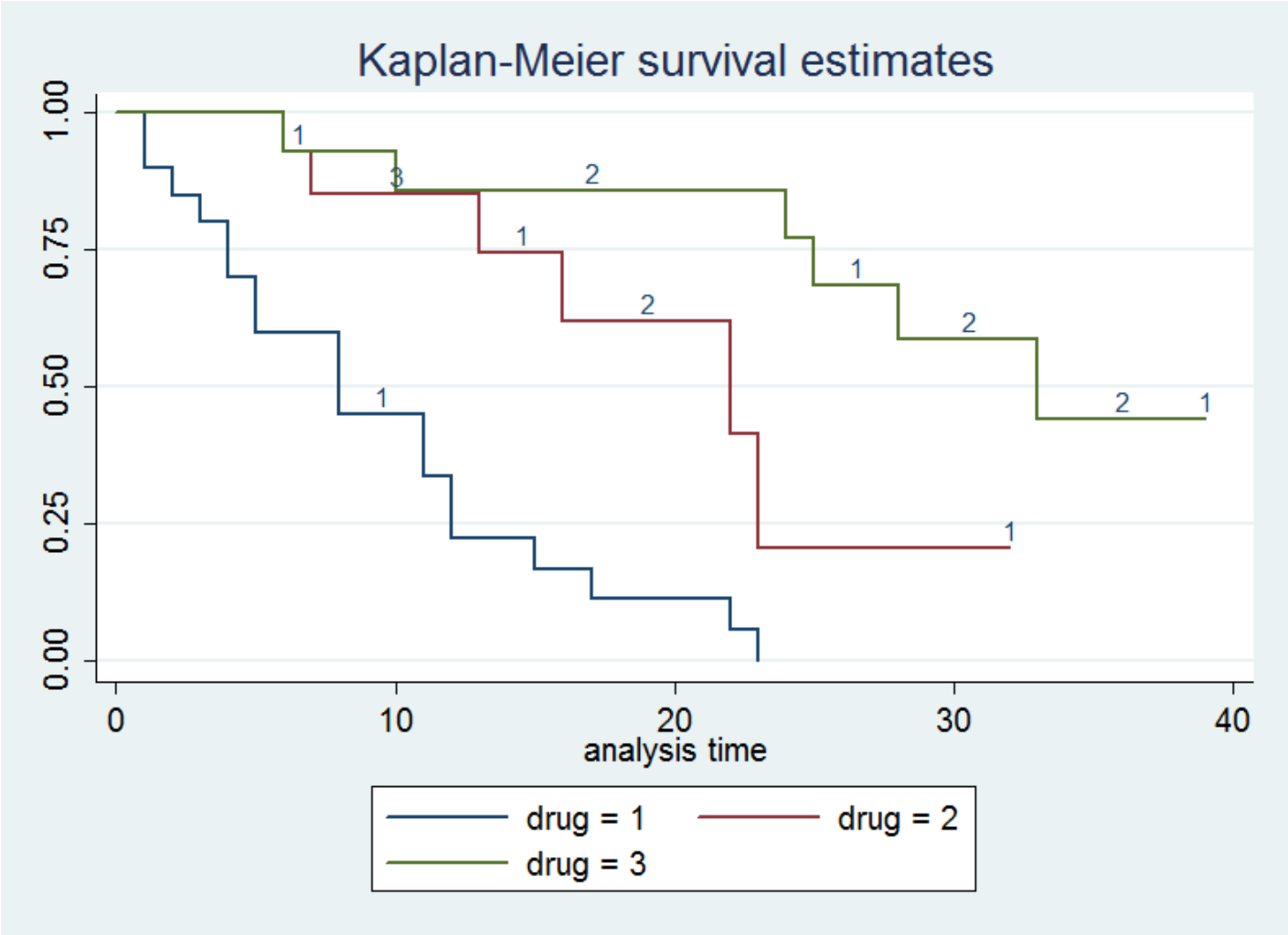
Analyzing stratified survival rates

```
sts graph, by(drug)
```



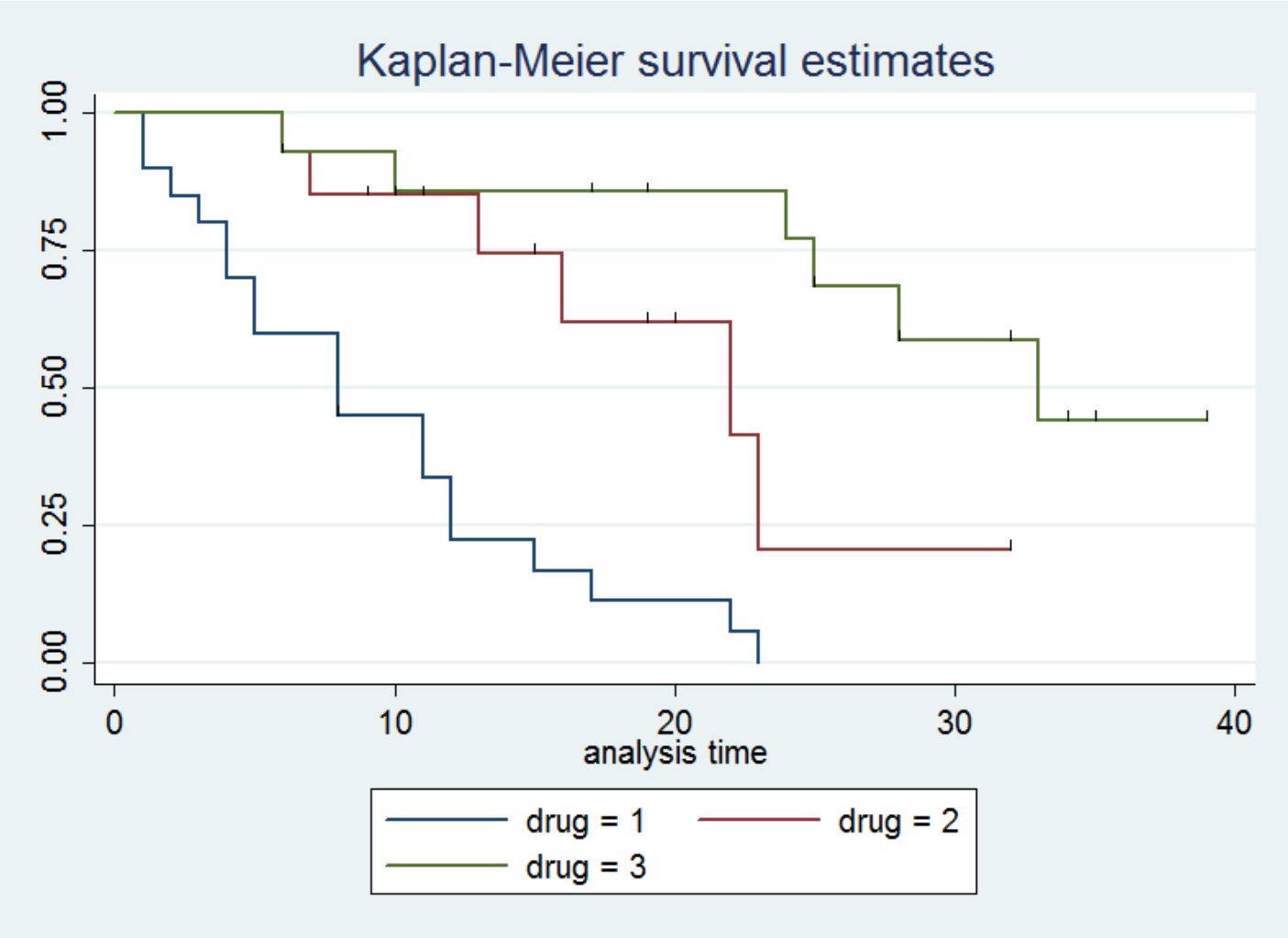
One can also identify the times of failure events

```
sts graph, by (drug) lost
```



Identifying the censored times

```
sts graph, by(drug)  censored(single)
```



Logrank Test

```
. sts test studytime, logrank strata(drug)

Stratified log-rank test for equality of survivor functions

           |   Events         Events
studytime |  observed    expected(*)
----------+--------------------------
1          |         2            0.20
2          |         1            0.16
3          |         1            0.21
4          |         2            0.68
.
.
.
39         |         0            0.75
----------+--------------------------
Total      |        31           31.00

(*) sum over calculations within drug

                chi2(27) =      85.14
                 Pr>chi2 =     0.0000
```

The hazard rate

- The hazard rate is the conditional probability of the death, failure, or event under study, provided the patient has survived up to an including that time period.
- Sometimes the hazard rate is called the intensity function, the failure rate
- When it is applied to continuous data, it is sometimes referred to as the instantaneous failure rate

Cox Regression

The Cox model presumes that the ratio of the hazard rate to a baseline hazard rate is an exponential function of the parameter vector.

$$\frac{h(t)}{h_0(t)} = exp(X'\beta) = e^{\beta_1 x_1 + \beta_1 x_1 + ... + \beta_p x_p}$$

We have to assume that the proportional hazard remains constant

Proportional Hazards model

```
. stcox age drug, nohr

Cox regression -- Breslow method for ties

No. of subjects =           48                  Number of obs    =        48
No. of failures =           31
Time at risk    =          744
                                                LR chi2(2)       =     36.29
Log likelihood  =   -81.765061                  Prob > chi2      =    0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .1100654   .0361337     3.05   0.002      .0392447    .1808862
        drug |  -1.535422   .3143056    -4.89   0.000     -2.15145   -.9193943
------------------------------------------------------------------------------
```

Hazards Ratio

```
. stcox age drug

Cox regression -- Breslow method for ties

No. of subjects =           48                  Number of obs    =        48
No. of failures =           31
Time at risk    =          744
                                                LR chi2(2)       =     36.29
Log likelihood  =   -81.765061                  Prob > chi2      =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.116351   .0403379     3.05   0.002     1.040025    1.198279
        drug |   .2153648   .0676904    -4.89   0.000     .1163154    .3987605
------------------------------------------------------------------------------
```