Statistics for Life Sciences - Survival Analysis

Nathan Harmston

2022-03-29

Survival Analysis

from an **exposure** to an **event**

Time-to-Event Analysis

In survival analysis, we are interested not only in outcomes, but also in the time it takes for them to occur. This time variable can be referred to as failure time, survival time or event time.

Instances where you would use survival analysis in biological research

- Time from exposure to onset of symptoms
- Time from cancer treatment until death
- Time until seizure freedom after taking an anti-epileptic drug
- lifespan of flies on distinct sugar diets

Survival Analysis

from an **exposure** to an **event**

Time-to-Event Analysis

In survival analysis, we are interested not only in outcomes, but also in the time it takes for them to occur. This time variable can be referred to as failure time, survival time or event time.

Instances where you would use survival analysis in biological research

- Time from exposure to onset of symptoms
- Time from cancer treatment until death
- Time until seizure freedom after taking an anti-epileptic drug
- lifespan of flies on distinct sugar diets

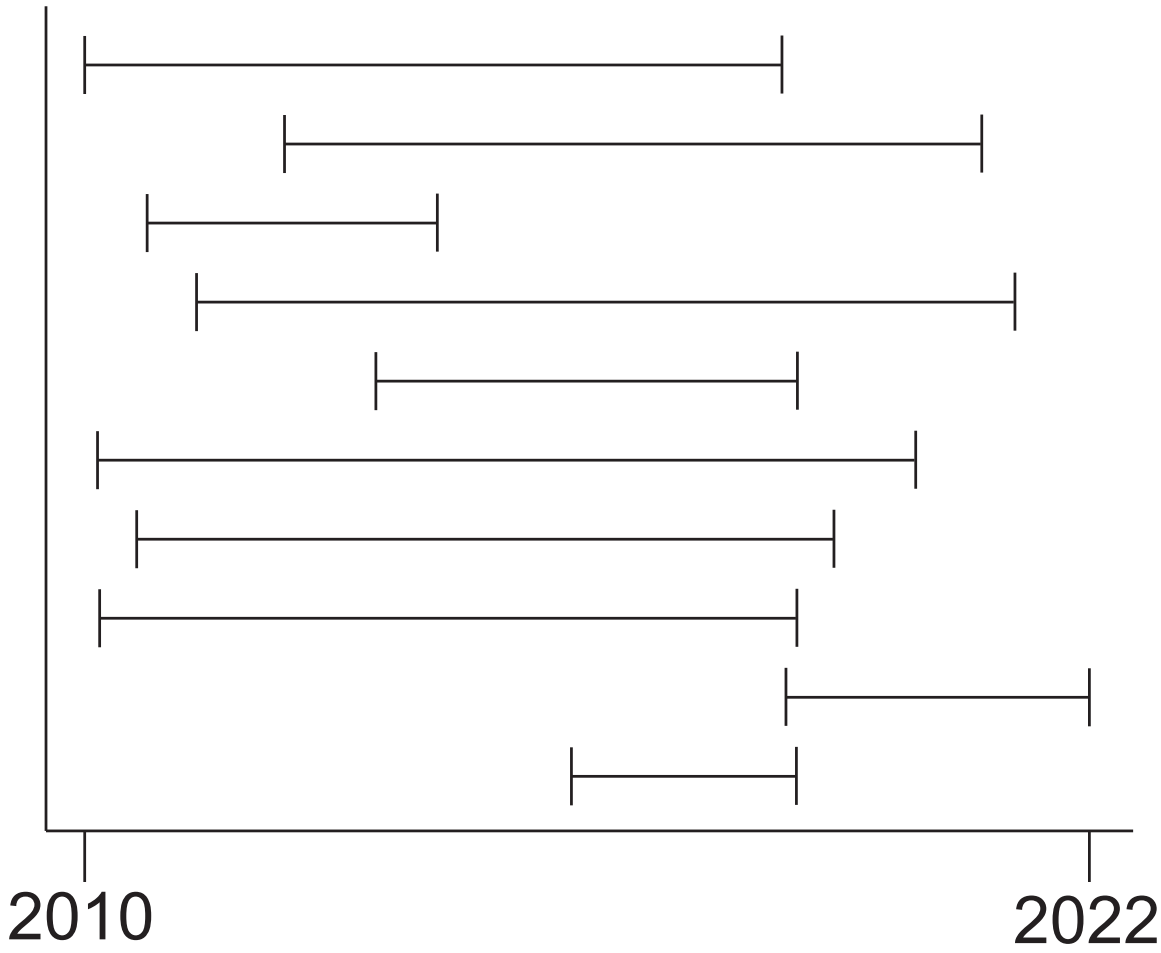
marriage .... divorce

Survival Time or Time-to-Event response

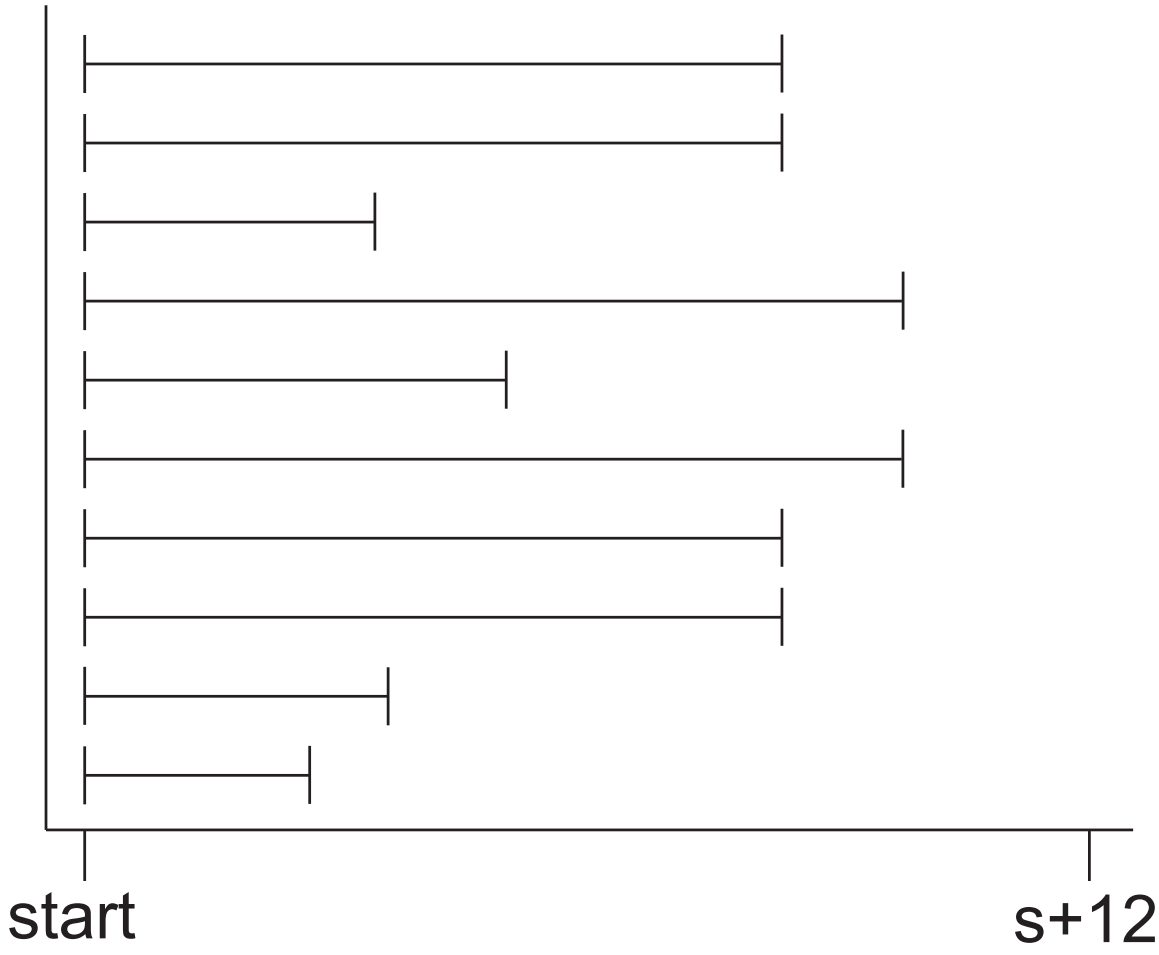
Features of the time-to-event response

- Usually continuous
- must be a one time event
- Is always ≥ 0 (no time traveling allowed)
- Partially or incompletely observed responses are **censored**.

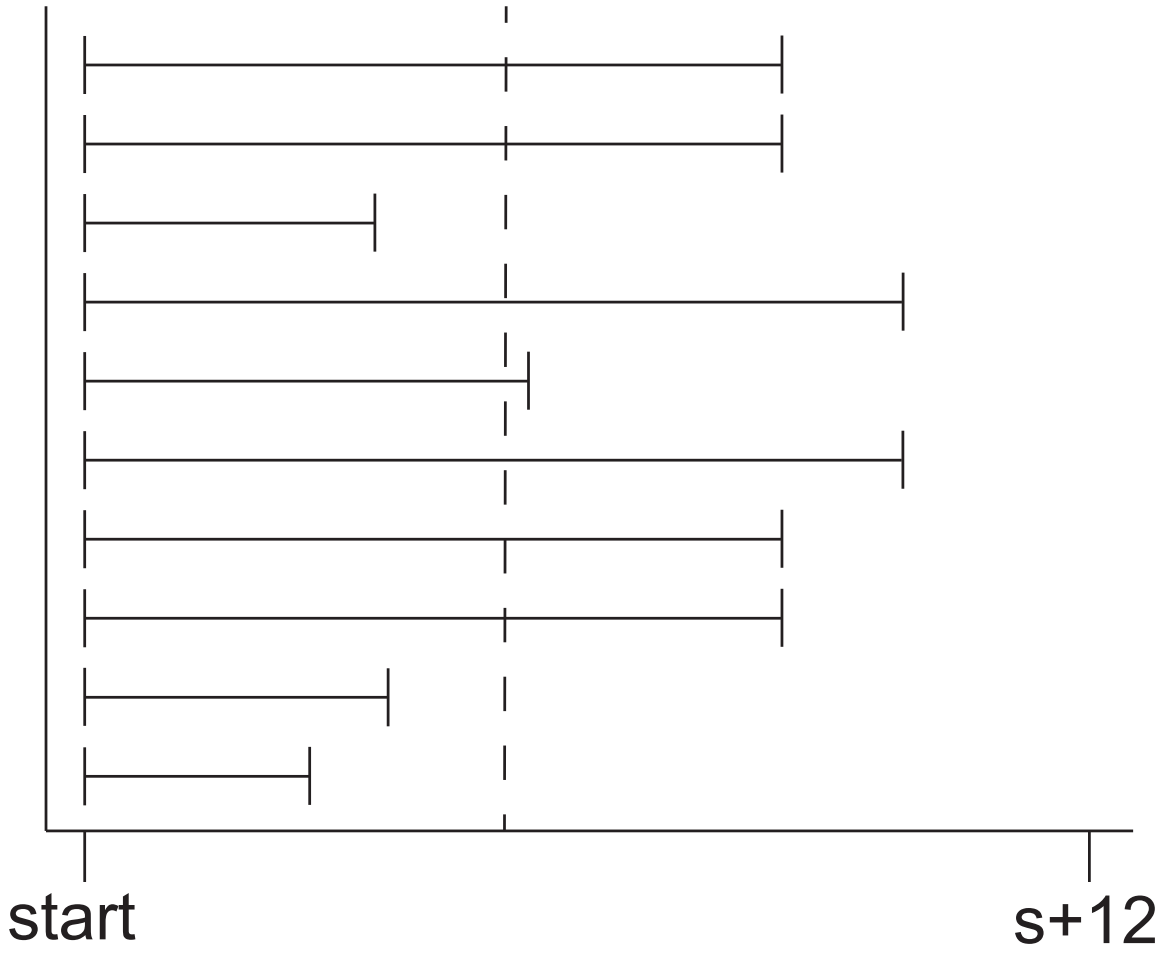
Survival analysis



Survival analysis

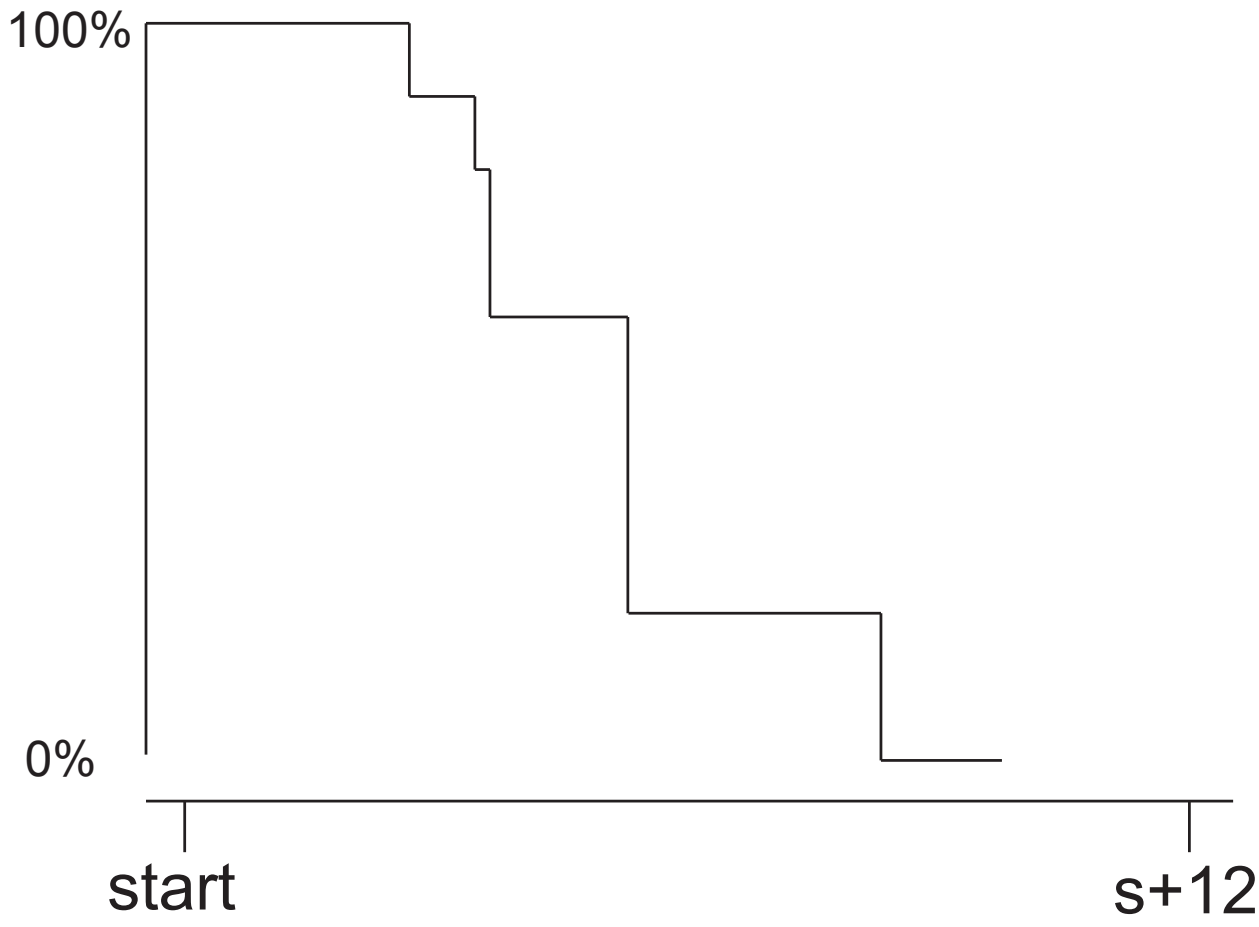


Survival analysis



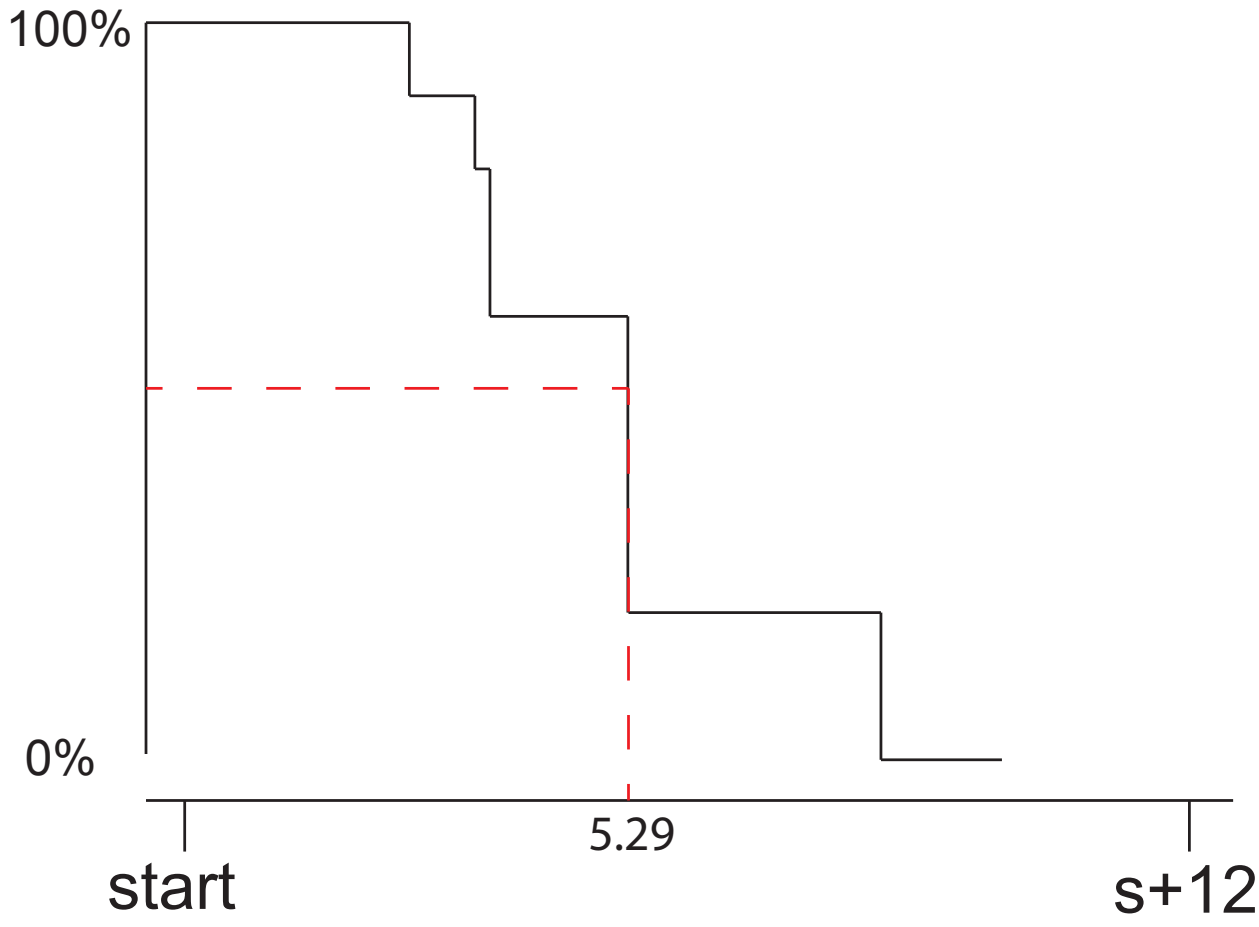
Survival analysis

Kaplan-Meier Curve



Survival analysis

Median survival time

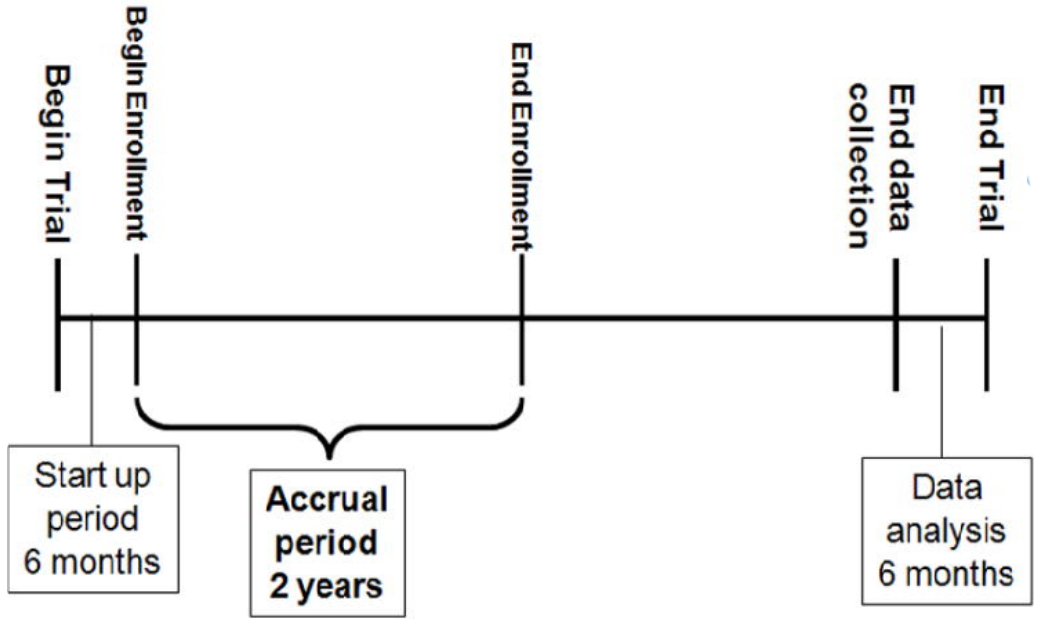


Censoring

Censoring occurs when we don't know the exact time-to-event, but we have some information about the event time. To be able to include a subject in the analysis, the censoring must be independent of the survival mechanism.

For in example, a subject can be censored from a study because they've moved and can no longer participate. A subject leaving the study because of an adverse reaction, on the other hand, would not be censored but removed.

Censoring



A PRACTICAL GUIDE TO UNDERSTANDING KAPLAN-MEIER CURVES. 2010

Censoring

Types of censoring

- left: $t_i < x$ - we may not know the precise start time
- interval: $x_i <$t$_i < x_2$ - where we only know an event occurred in an interval of time
- right: $x < t_i$

Types of censoring

There are three main reasons why a subject would be censored

- the subject does not have the event of interest during the study
- the subject is lost to follow-up during the study
- the subject withdraws from the study

These are examples of right censoring, where the we may not know the precise time of event.

Types of Right Censoring

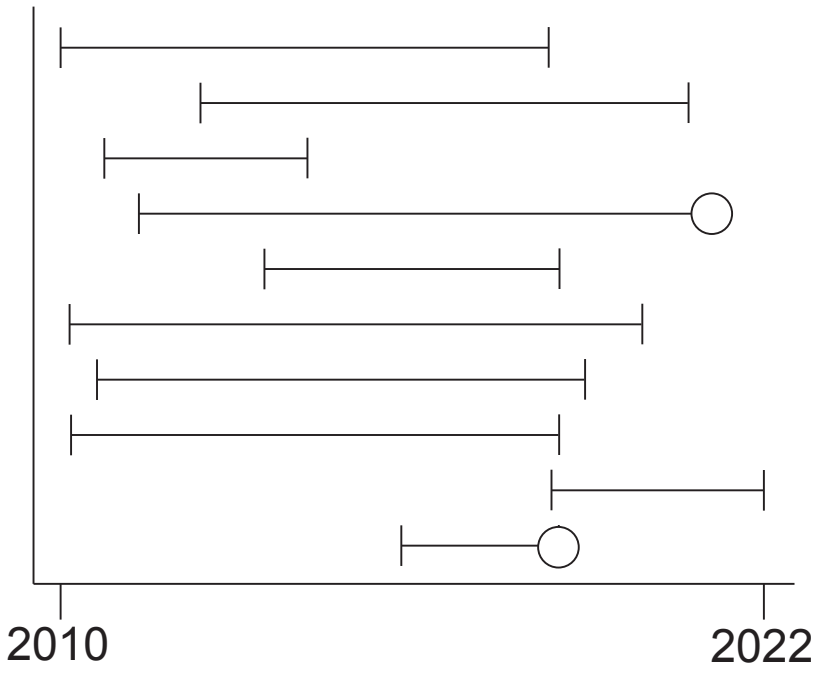
- **Fixed type I:** happens when the study ends after a pre-specified amount of time, C. Subjects who have not experienced an event during the course of the study will be censored at the end, or time C.
- **Random type I:** happens when the study ends after a pre-specified amount of time, C, but subjects are censored before the study ends due to other factors like withdrawing from the study
- **Type II:** happens when the endpoint of the study is not determined by a pre-specified time, but ends when a pre-specified number of events has occurred.

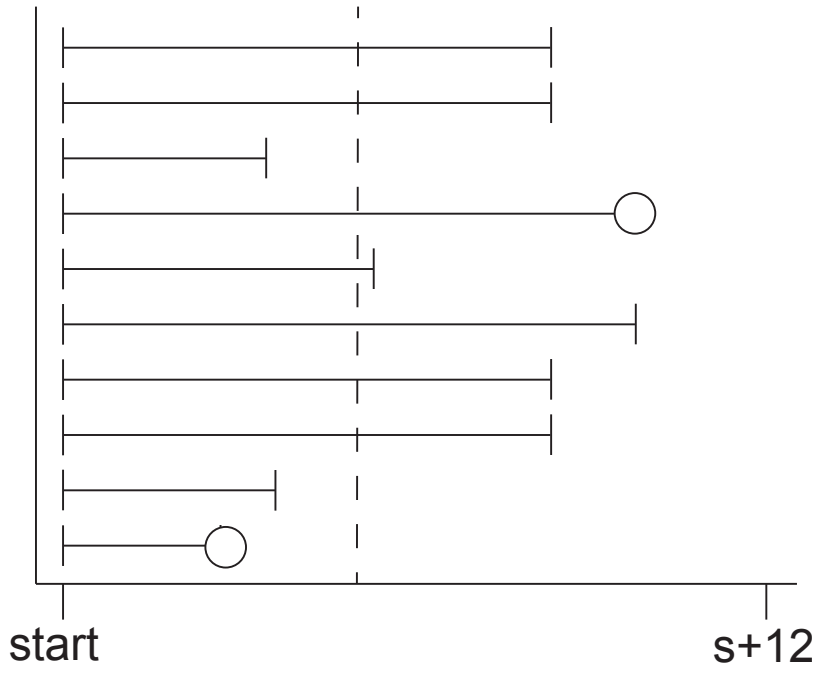
Reasons it's important

If, in the course of a study, no censoring occurs, standard regression could be used, but it would likely be inadequate because

- the time-to-event response is always positive with a skewed distribution
- we lose information on the probability of surviving past a certain time point

Censoring





What does it look like?

- T_i is the response for the ith subject
- C_i is the censoring time for the ith subject
- Y_i is the observed response, $Y_i = min(T_i, C_i)$
- δ_i is the event indicator where

$$\delta = \left\{ \begin{array}{l} 1, \text{ if the event was observed } (T_i \leq C_i) \\ 0, \text{ if the event was censored } (T_i > C_i) \end{array} \right\}$$

Truncation of data

left truncation

If individuals in the population are not observed if the event occurs before time t, then the data are said to be left truncated at t.

individuals with very short survival evade sampling - leads to a bias in our analysis

right truncation

observations with long time to event are excluded from the analysis
biased towards individuals that have a short time to event

Survival Function

- T is the time-to-event response variable and $T \geq 0$
- The survival function is $S(t)$

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

where the $S(t)$ gives the probability that a subject will survive past time t

Survival Function and t

As time t ranges from 0 to ∞, the survival function follows these properties

- It is non-increasing
- At the start, $t = 0$ and $S(t) = 1$, meaning the probability of surviving at time 0 is 1 (everyone's still alive at the start)
- As t approaches ∞, $S(t) = 0$, meaning the probability of surviving for an infinite amount of time is 0 (everyone dies at some point)

On a long enough time line, the survival rate for everyone drops to zero.- Chuck Palahniuk

How do we estimate the survival function?

Depending on the dataset, we can use either a parametric or non-parametric estimation of $S(t)$.

- If we cannot or do not want to make an assumption about the underlying distribution of the data, we can use an estimator like the **Kaplan Meier** estimator - non parametric
- If we can assume a distribution, we can use maximum likelihood estimation to estimate $S(t)$ - parametric
  - exponential
  - Weibull
  - Gamma
  - log-normal

Ovarian Cancer

This dataset from a cohort of ovarian cancer patients. It contains clinical information, including age, treatment group, presence of residual disease, performance, blood pressure*, cholesterol levels*, **if the subjects were censored or not and the time subjects were tracked until they either died or were lost to follow-up.** The patients were followed up for $\sim$ 3.5 years after treatment.

Age - age at treatment
Resid Disease - was there residual disease after treatment
Rx - which drug were they put on, A or B
ECOG - quality of life
BP - blood pressure at start of treatment
Chol - cholesterol levels at start of treatment
Death - 1 if the subject died or 0 if they were censored
Time - time until death or censoring

Survival Data - What does it look like?

Ovarian Cancer

	Age	Resid Dis	Rx	ECOG	BP	Chol	**Death**	**Time**
1	72.33	yes	A	good	117.83	13.58	**1**	**59**
2	74.49	yes	A	good	114.00	7.78	**1**	**115**
3	66.47	yes	A	bad	117.55	10.95	**1**	**156**
4	74.50	yes	A	bad	113.50	22.50	**1**	**268**
5	43.14	yes	A	good	139.19	22.11	**1**	**329**
6	63.22	no	B	bad	124.80	8.46	**1**	**353**
7	64.42	yes	B	good	118.09	23.19	**1**	**365**
8	58.31	no	B	good	130.09	26.51	**0**	**377**
.	.	.	.	.	.	.	**.**	**.**
.	.	.	.	.	.	.	**.**	**.**
25	44.21	yes	B	good	138.34	26.25	**0**	**1206**
26	59.59	no	B	bad	129.18	22.93	**0**	**1227**

Survival Analysis - Ovarian Cancer

In this cohort of ovarian cancer patients,

- What is the survival rate over time for all subjects in the cohort?
- is there a difference in survival rates between drug treatment A and drug treatment B?
- is there a difference in survival rates between patients with residual disease and those that do not have residual disease ?

Survival Function

How do we estimate the survival function?

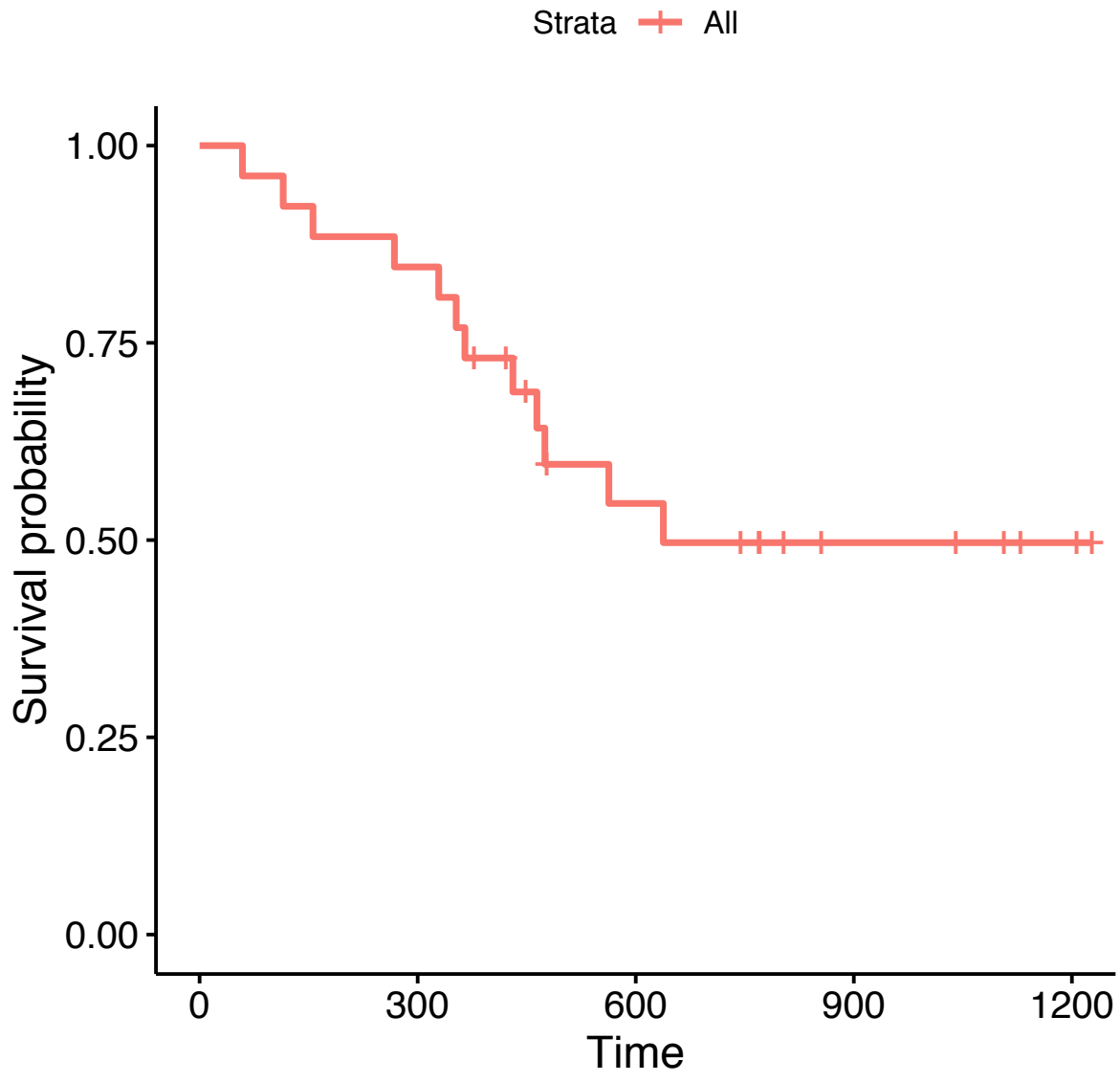
$$\prod_{t_i < t} \frac{n_i - d_i}{n_i} \tag{1}$$

n = number at risk d = number of deaths/events

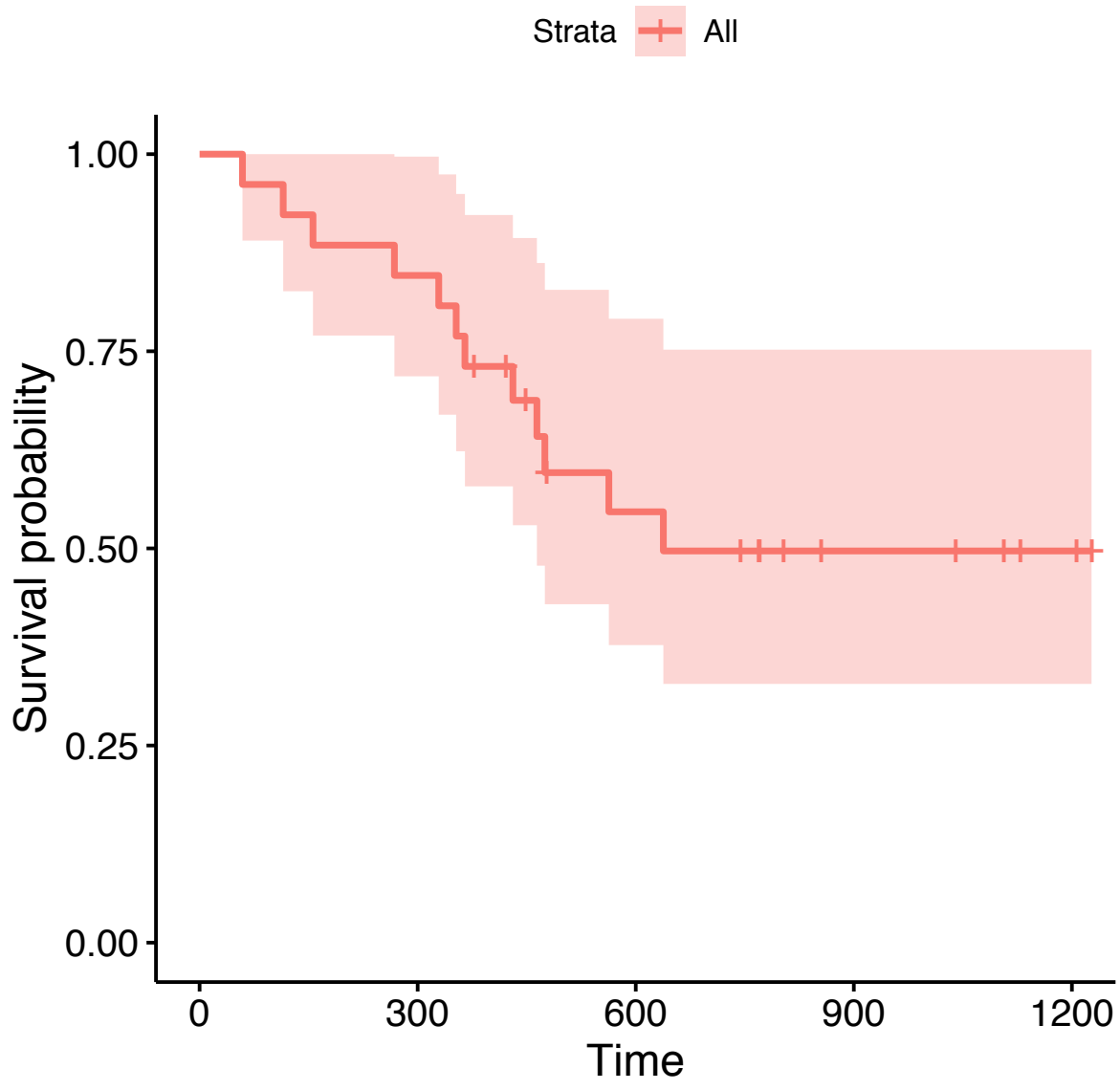
Non-parametric estimation with Kaplan Meier

time	n.risk	deaths	censored	cumulative survival
59	26	1	0	$25/26 = 0.962$
115	25	1	0	$24/25 \times 0.962 = 0.923$
156	24	1	0	$23/24 \times 0.923 = 0.885$
268	23	1	0	$22/23 \times 0.885 = 0.846$
329	22	1	0	$21/23 \times 0.846 = 0.808$
353	21	1	0	$20/21 \times 0.808 = 0.769$
365	20	1	0	$19/20 \times 0.769 = 0.730$
377	19	0	1	$19/19 \times 0.730 = 0.730$
421	18	0	1	$18/18 \times 0.730 = 0.730$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

Survival Function - Ovarian Cancer



Survival Function - Ovarian Cancer



Survival Data - Cancer Treatment

Ovarian Cancer

	Age	Resid Dis	**Rx**	ECOG	BP	Chol	Death	Time
1	72.33	yes	**A**	good	117.83	13.58	1	59
2	74.49	yes	**A**	good	114.00	7.78	1	115
3	66.47	yes	**A**	bad	117.55	10.95	1	156
4	74.50	yes	**A**	bad	113.50	22.50	1	268
5	43.14	yes	**A**	good	139.19	22.11	1	329
6	63.22	no	**B**	bad	124.80	8.46	1	353
7	64.42	yes	**B**	good	118.09	23.19	1	365
8	58.31	no	**B**	good	130.09	26.51	0	377
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
25	44.21	yes	**B**	good	138.34	26.25	0	1206
26	59.59	no	**B**	bad	129.18	22.93	0	1227

Survival Function - log rank test

survival curves differ - but is this sufficient to conclude that in individuals in population A have worse survival than population?

Log rank test

compare the survival curves of two or more groups

H_0: there is no difference between the populations in the probability of an event (death) at any time point.
H_1: there is a difference between the populations in the probability of an event (death) at any time point.

The logrank test - Bland et al 2004

Survival Function - log rank test

look at each time period calculate the number of expected deaths for each group

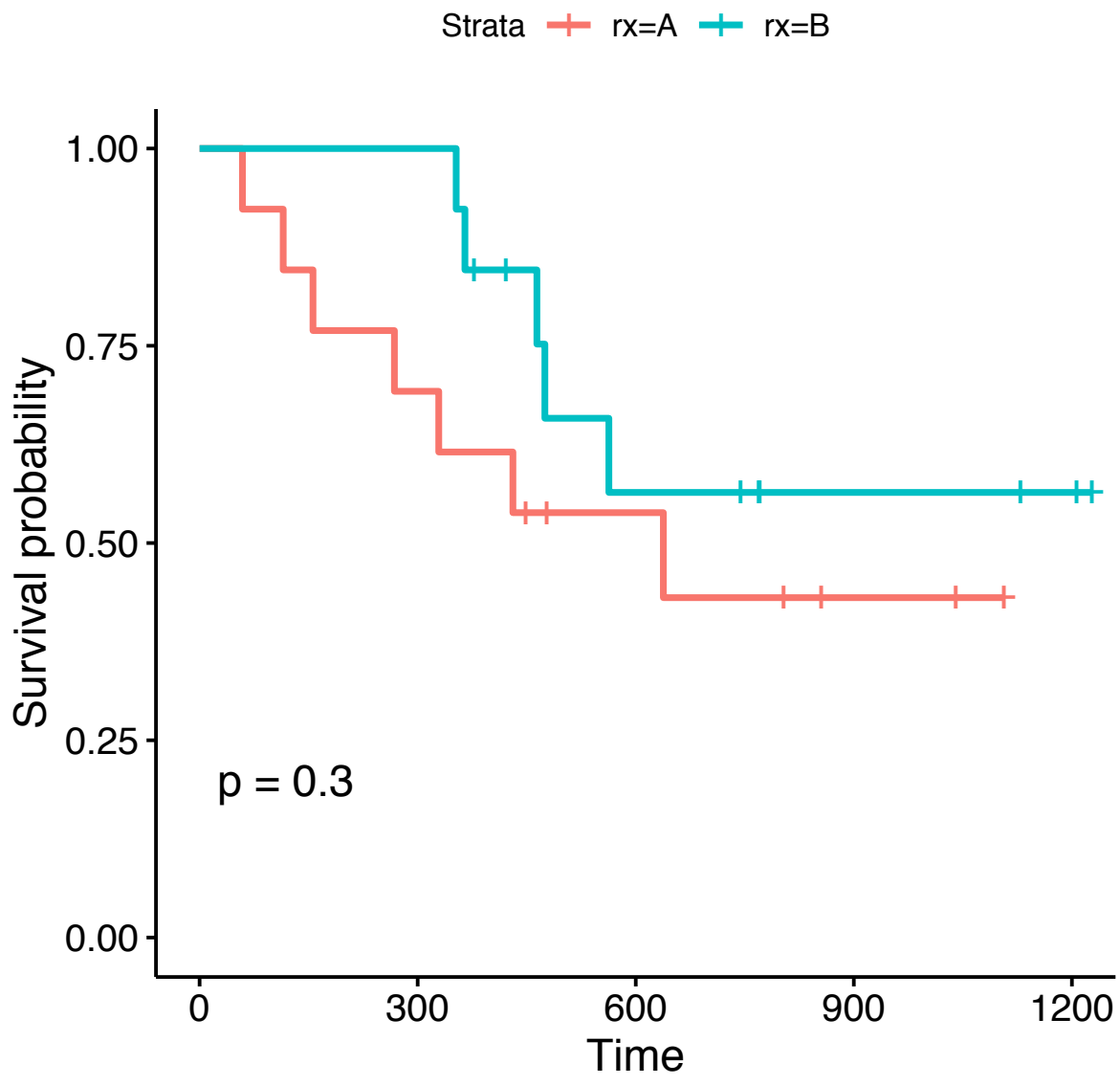
$$\frac{\text{number at risk}}{\text{total}} * \text{total deaths} \tag{2}$$

look at each time period calculate the number of expected deaths for each group

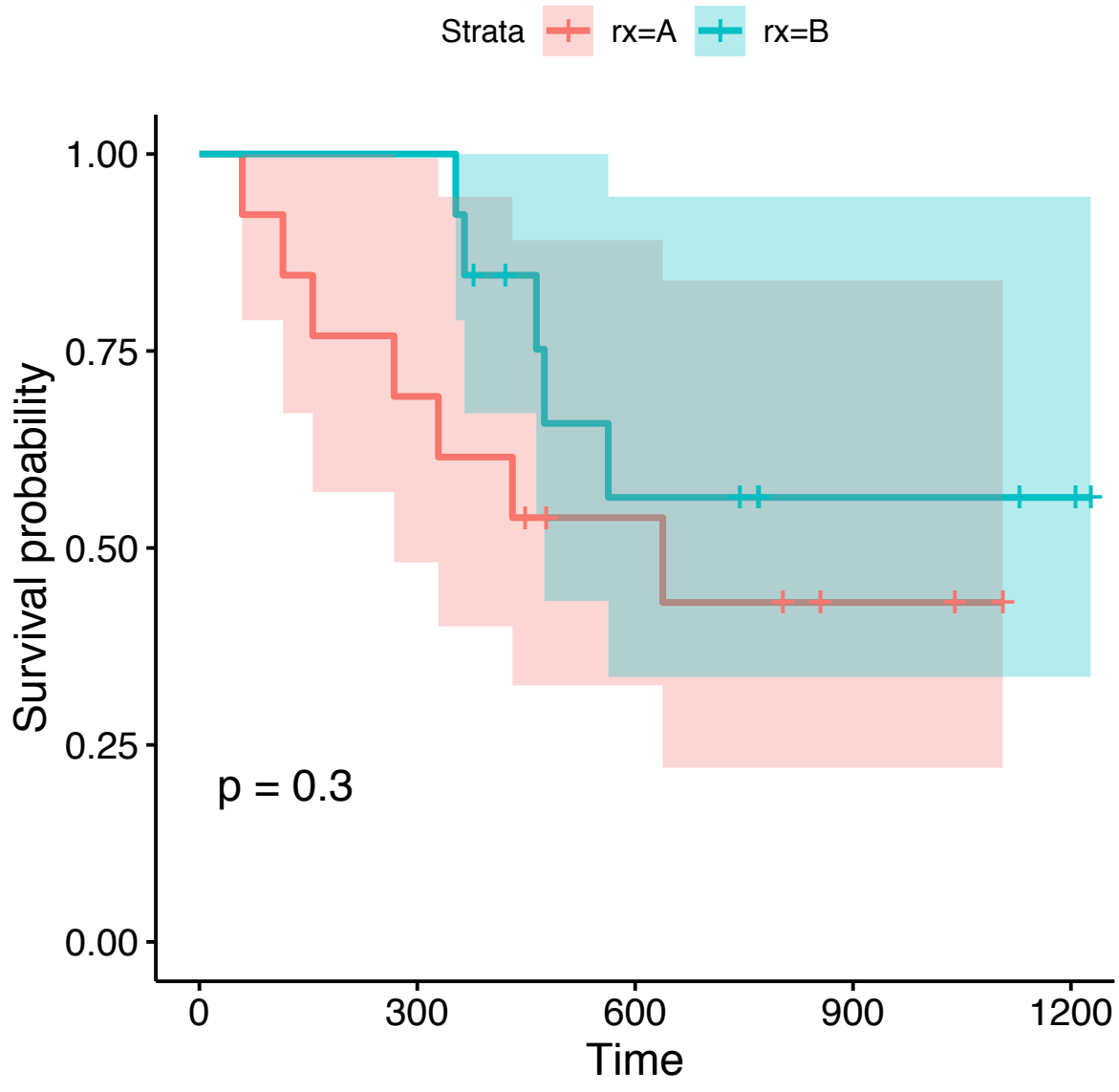
$$\frac{\text{number at risk}}{\text{total}} * \text{total deaths} \tag{2}$$

construct a chi^2 statistic

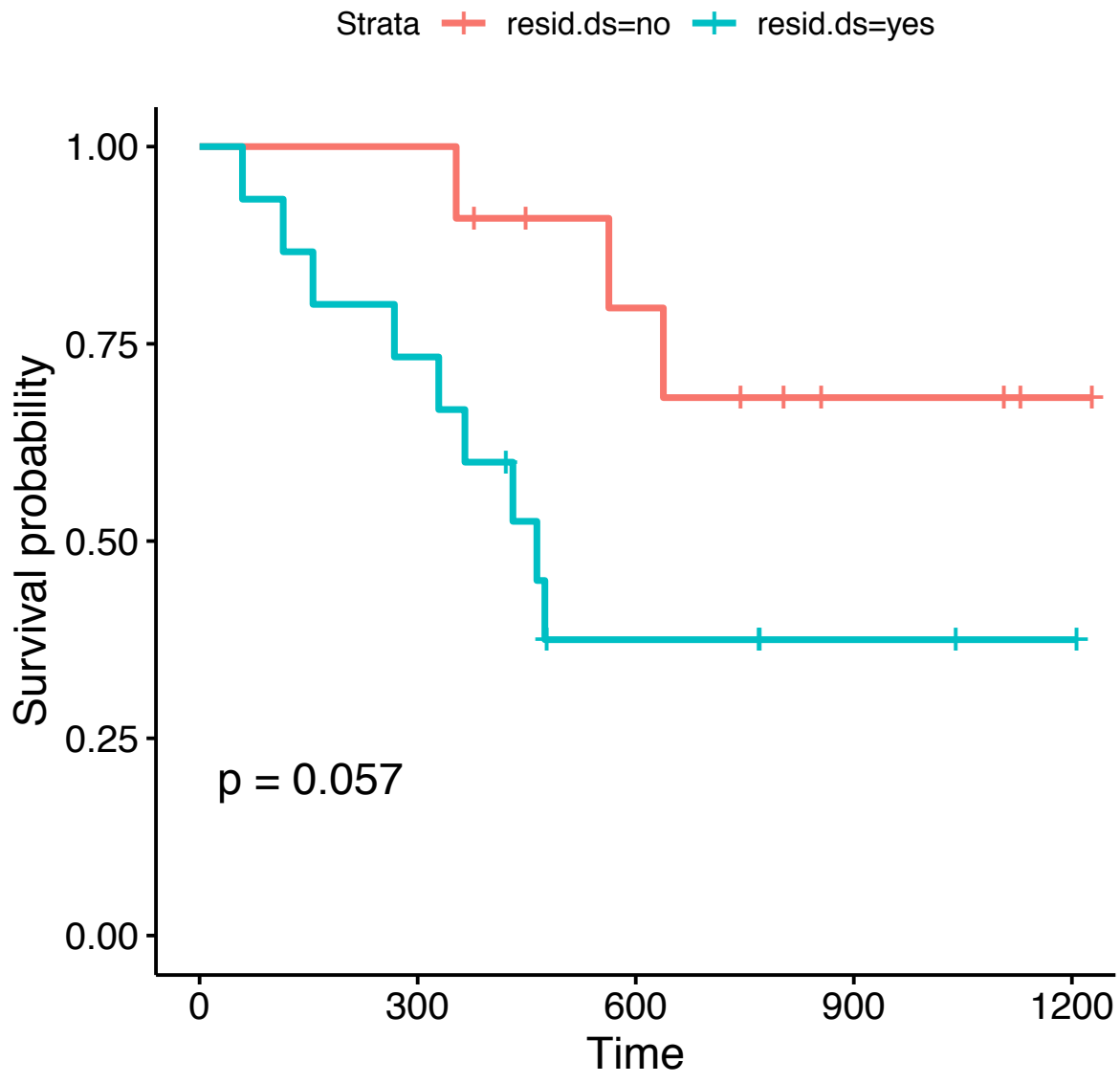
Survival Function - Ovarian Cancer Treatment



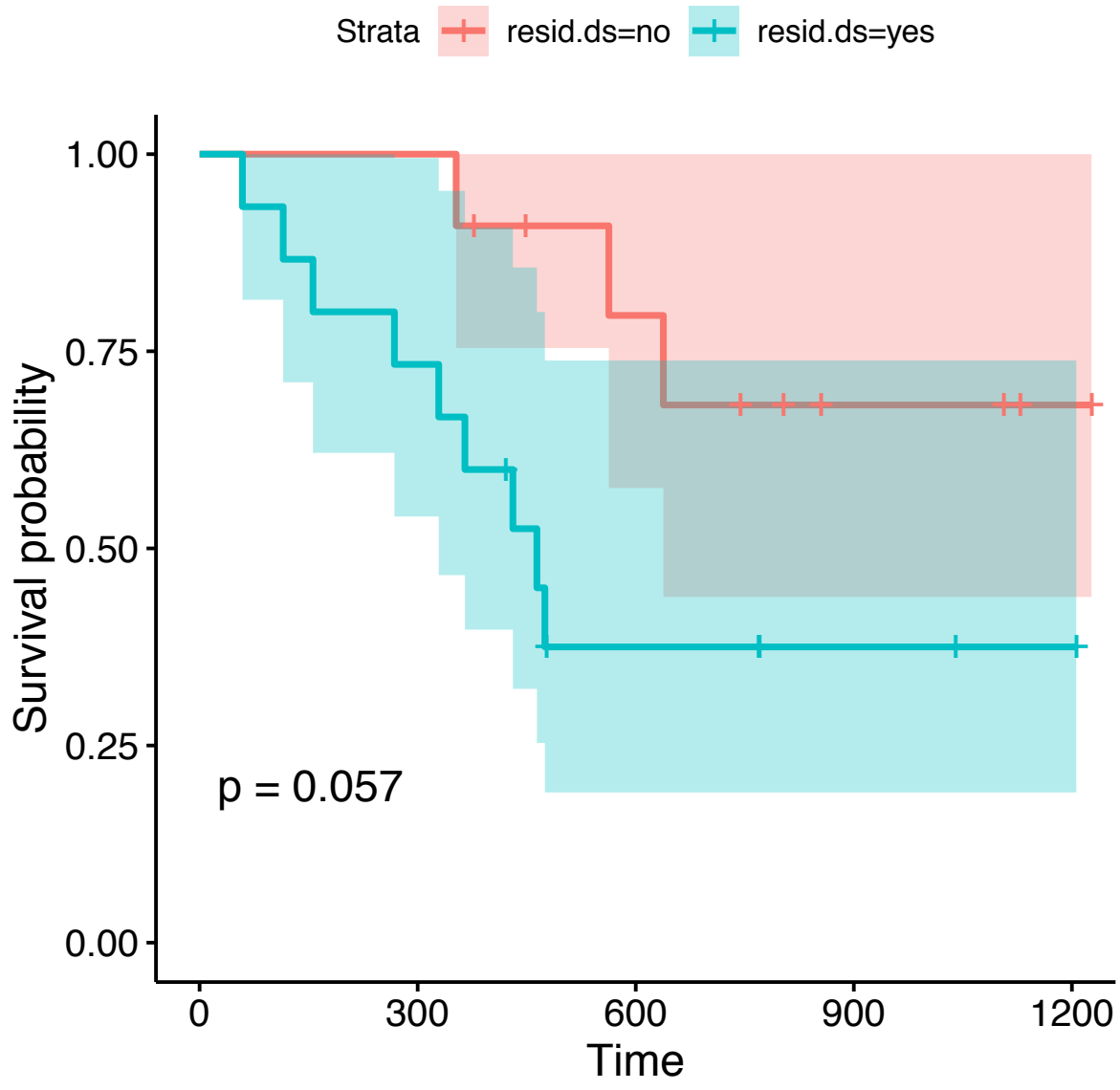
Survival Function - Ovarian Cancer Treatment



Survival Function - Ovarian Cancer Residual Disease



Survival Function - Ovarian Cancer Residual Disease



Five year survival rate

a lot of cancer survival studies talk about 5-year survival

draw a line at x=5 years and find where the line intersects the five-year percentage survival

```
summary(survfit(Surv(time, status) ~ 1, data = ovarian),
 times = 365.25*5)
```

```
summary(survfit(Surv(time, status) ~ 1, data = ovarian),
 times = 365.25)
```

Summary - Survival Analysis

- Models the relationship between time-to-event variable, like time to death, and other independent variables, like drug treatment
- Model the overall survival probability of a cohort or population over time
- Calculate significant differences in the survival probabilities of a population with respect to factor, like drug treatment
- Uses a Kaplan Meier or distribution estimator to fit the model

- Friday
- hazard functions
- Cox regression - Cox proportional hazards regression