Wikipedia
The Free Encyclopedia

Survival analysis

Survival analysis is a branch of statistics for analyzing the expected duration of time until one event occurs, such as death in biological organisms and failure in mechanical systems. This topic is called **reliability theory**, **reliability analysis** or reliability engineering in engineering, **duration analysis** or **duration modelling** in economics, and **event history analysis** in sociology. Survival analysis attempts to answer certain questions, such as what is the proportion of a population which will survive past a certain time? Of those that survive, at what rate will they die or fail? Can multiple causes of death or failure be taken into account? How do particular circumstances or characteristics increase or decrease the probability of survival?

To answer such questions, it is necessary to define "lifetime". In the case of biological survival, death is unambiguous, but for mechanical reliability, failure may not be well-defined, for there may well be mechanical systems in which failure is partial, a matter of degree, or not otherwise localized in time. Even in biological problems, some events (for example, heart attack or other organ failure) may have the same ambiguity. The theory outlined below assumes well-defined events at specific times; other cases may be better treated by models which explicitly account for ambiguous events.

More generally, survival analysis involves the modelling of time to event data; in this context, death or failure is considered an "event" in the survival analysis literature – traditionally only a single event occurs for each subject, after which the organism or mechanism is dead or broken. *Recurring event* or *repeated event* models relax that assumption. The study of recurring events is relevant in systems reliability, and in many areas of social sciences and medical research.

Introduction to survival analysis

Survival analysis is used in several ways:

- To describe the survival times of members of a group
  - Life tables
  - Kaplan–Meier curves
  - Survival function
  - Hazard function
- To compare the survival times of two or more groups
  - Log-rank test
- To describe the effect of categorical or quantitative variables on survival
  - Cox proportional hazards regression
  - Parametric survival models
  - Survival trees
  - Survival random forests

Definitions of common terms in survival analysis

The following terms are commonly used in survival analyses:

- Event: Death, disease occurrence, disease recurrence, recovery, or other experience of interest
- Time: The time from the beginning of an observation period (such as surgery or beginning treatment) to (i) an event, or (ii) end of the study, or (iii) loss of contact or withdrawal from the study.
- Censoring / Censored observation: Censoring occurs when we have some information about individual survival time, but we do not know the survival time exactly. The subject is censored in the sense that nothing is observed or known about that subject after the time of censoring. A censored subject may or may not have an event after the end of observation time.
- Survival function S(t): The probability that a subject survives longer than time t.

Example: Acute myelogenous leukemia survival data

This example uses the Acute Myelogenous Leukemia survival data set "aml" from the "survival" package in R. The data set is from Miller (1997)[1] and the question is whether the standard course of chemotherapy should be extended ('maintained') for additional cycles.

The aml data set sorted by survival time is shown in the box.

Aml data set sorted by survival time

observation	**time (weeks)**	**status**	**x**
12	5	1	Nonmaintained
13	5	1	Nonmaintained
14	8	1	Nonmaintained
15	8	1	Nonmaintained
1	9	1	Maintained
16	12	1	Nonmaintained
2	13	1	Maintained
3	13	0	Maintained
17	16	0	Nonmaintained
4	18	1	Maintained
5	23	1	Maintained
18	23	1	Nonmaintained
19	27	1	Nonmaintained
6	28	0	Maintained
20	30	1	Nonmaintained
7	31	1	Maintained
21	33	1	Nonmaintained
8	34	1	Maintained
22	43	1	Nonmaintained
9	45	0	Maintained
23	45	1	Nonmaintained
10	48	1	Maintained
11	161	0	Maintained

- Time is indicated by the variable "time", which is the survival or censoring time
- Event (recurrence of aml cancer) is indicated by the variable "status". 0 = no event (censored), 1 = event (recurrence)
- Treatment group: the variable "x" indicates if maintenance chemotherapy was given

The last observation (11), at 161 weeks, is censored. Censoring indicates that the patient did not have an event (no recurrence of aml cancer). Another subject, observation 3, was censored at 13 weeks (indicated by status=0). This subject was in the study for only 13 weeks, and the aml cancer did not recur during those 13 weeks. It is possible that this patient was enrolled near the end of the study, so that they could be observed for only 13 weeks. It is also possible that the patient was enrolled early in the study, but was lost to follow up or withdrew from the study. The table shows that other subjects were censored at 16, 28, and 45 weeks (observations 17, 6, and 9 with status=0). The remaining subjects all experienced events (recurrence of aml cancer) while in the study. The question of interest is whether recurrence occurs later in maintained patients than in non-maintained patients.

Kaplan–Meier plot for the aml data

The survival function $S(t)$, is the probability that a subject survives longer than time t. $S(t)$ is theoretically a smooth curve, but it is usually estimated using the Kaplan–Meier (KM) curve. The graph shows the KM plot for the aml data and can be interpreted as follows:

- The *x* axis is time, from zero (when observation began) to the last observed time point.
- The *y* axis is the proportion of subjects surviving. At time zero, 100% of the subjects are alive without an event.
- The solid line (similar to a staircase) shows the progression of event occurrences.
- A vertical drop indicates an event. In the aml table shown above, two subjects had events at five weeks, two had events at eight weeks, one had an event at nine weeks, and so on. These events at five weeks, eight weeks and so on are indicated by the vertical drops in the KM plot at those time points.
- At the far right end of the KM plot there is a tick mark at 161 weeks. The vertical tick mark indicates that a patient was censored at this time. In the aml data table five subjects were censored, at 13, 16, 28, 45 and 161 weeks. There are five tick marks in the KM plot, corresponding to these censored observations.

Life table for the aml data

A life table summarizes survival data in terms of the number of events and the proportion surviving at each event time point. The life table for the aml data, created using the R software, is shown.

Life Table for the aml Data

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
5	23	2	0.913	0.0588	0.8049	1
8	21	2	0.8261	0.079	0.6848	0.996
9	19	1	0.7826	0.086	0.631	0.971
12	18	1	0.7391	0.0916	0.5798	0.942
13	17	1	0.6957	0.0959	0.5309	0.912
18	14	1	0.646	0.1011	0.4753	0.878
23	13	2	0.5466	0.1073	0.3721	0.803
27	11	1	0.4969	0.1084	0.324	0.762
30	9	1	0.4417	0.1095	0.2717	0.718
31	8	1	0.3865	0.1089	0.2225	0.671
33	7	1	0.3313	0.1064	0.1765	0.622
34	6	1	0.2761	0.102	0.1338	0.569
43	5	1	0.2208	0.0954	0.0947	0.515
45	4	1	0.1656	0.086	0.0598	0.458
48	2	1	0.0828	0.0727	0.0148	0.462

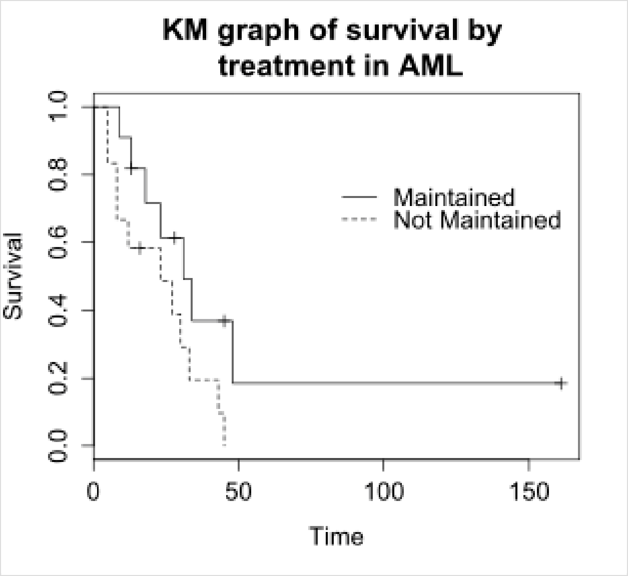
The life table summarizes the events and the proportion surviving at each event time point. The columns in the life table have the following interpretation:

- time gives the time points at which events occur.
- n.risk is the number of subjects at risk immediately before the time point, t. Being "at risk" means that the subject has not had an event before time t, and is not censored before or at time t.
- n.event is the number of subjects who have events at time t.
- survival is the proportion surviving, as determined using the Kaplan–Meier product-limit estimate.
- std.err is the standard error of the estimated survival. The standard error of the Kaplan–Meier product-limit estimate it is calculated using Greenwood's formula, and depends on the number at risk (n.risk in the table), the number of deaths (n.event in the table) and the proportion surviving (survival in the table).
- lower 95% CI and upper 95% CI are the lower and upper 95% confidence bounds for the proportion surviving.

Log-rank test: Testing for differences in survival in the aml data

The log-rank test compares the survival times of two or more groups. This example uses a log-rank test for a difference in survival in the maintained versus non-maintained treatment groups in the aml data. The graph shows KM plots for the aml data broken out by treatment group, which is indicated by the variable "x" in the data.

The null hypothesis for a log-rank test is that the groups have the same survival. The expected number of subjects surviving at each time point in each is adjusted for the number of subjects at risk in the groups at each event time. The log-rank test determines if the observed number of events in each group is significantly different from the expected number. The formal test is based on a chi-squared statistic. When the log-rank statistic is large, it is evidence for a difference in the survival times between the groups. The log-rank statistic approximately has a Chi-squared distribution with one degree of freedom, and the p-value is calculated using the Chi-squared test.



Kaplan–Meier graph by treatment group in aml

For the example data, the log-rank test for difference in survival gives a p-value of p=0.0653, indicating that the treatment groups do not differ significantly in survival, assuming an alpha level of 0.05. The sample size of 23 subjects is modest, so there is little power to detect differences between the treatment groups. The chi-squared test is based on asymptotic approximation, so the p-value should be regarded with caution for small sample sizes.

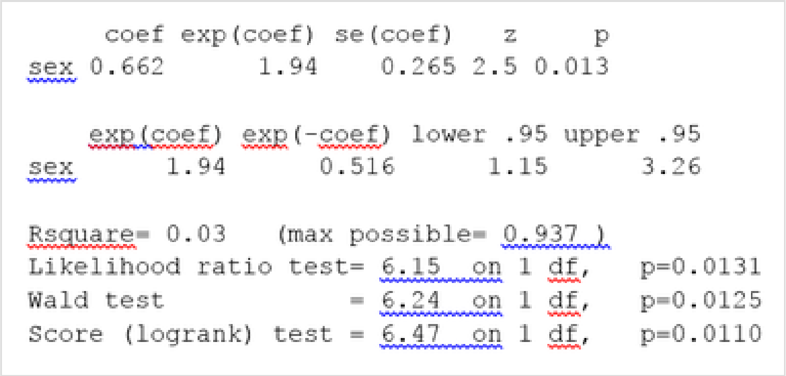
Cox proportional hazards (PH) regression analysis

Kaplan–Meier curves and log-rank tests are most useful when the predictor variable is categorical (e.g., drug vs. placebo), or takes a small number of values (e.g., drug doses 0, 20, 50, and 100 mg/day) that can be treated as categorical. The log-rank test and KM curves don't work easily with quantitative predictors such as gene expression, white blood count, or age. For quantitative predictor variables, an alternative method is Cox proportional hazards regression analysis. Cox PH models work also with categorical predictor variables, which are encoded as {0,1} indicator or dummy variables. The log-rank test is a special case of a Cox PH analysis, and can be performed using Cox PH software.

Example: Cox proportional hazards regression analysis for melanoma

This example uses the melanoma data set from Dalgaard Chapter 14. [2]

Data are in the R package ISwR. The Cox proportional hazards regression using R gives the results shown in the box.



```
      coef exp(coef) se(coef)   z     p
sex 0.662      1.94    0.265 2.5 0.013

     exp(coef) exp(-coef) lower .95 upper .95
sex       1.94      0.516      1.15      3.26

Rsquare= 0.03   (max possible= 0.937 )
Likelihood ratio test= 6.15  on 1 df,   p=0.0131
Wald test            = 6.24  on 1 df,   p=0.0125
Score (logrank) test = 6.47  on 1 df,   p=0.0110
```

Cox proportional hazards regression output for melanoma data. Predictor variable is sex 1: female, 2: male.

The Cox regression results are interpreted as follows.

- Sex is encoded as a numeric vector (1: female, 2: male). The R summary for the Cox model gives the hazard ratio (HR) for the second group relative to the first group, that is, male versus female.
- coef = 0.662 is the estimated logarithm of the hazard ratio for males versus females.
- exp(coef) = 1.94 = exp(0.662) - The log of the hazard ratio (coef= 0.662) is transformed to the hazard ratio using exp(coef). The summary for the Cox model gives the hazard ratio for the second group relative to the first group, that is, male versus female. The estimated hazard ratio of 1.94 indicates that males have higher risk of death (lower survival rates) than females, in these data.
- se(coef) = 0.265 is the standard error of the log hazard ratio.
- z = 2.5 = coef/se(coef) = 0.662/0.265. Dividing the coef by its standard error gives the z score.
- p=0.013. The p-value corresponding to z=2.5 for sex is p=0.013, indicating that there is a significant difference in survival as a function of sex.

The summary output also gives upper and lower 95% confidence intervals for the hazard ratio: lower 95% bound = 1.15; upper 95% bound = 3.26.

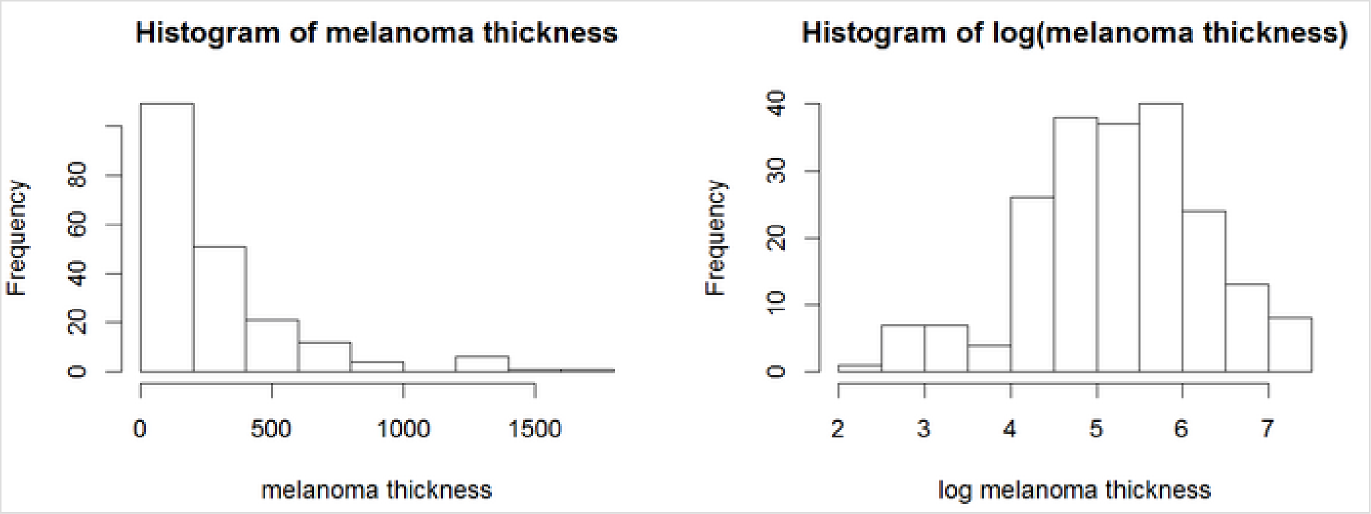
Finally, the output gives p-values for three alternative tests for overall significance of the model:

- Likelihood ratio test = 6.15 on 1 df, p=0.0131
- Wald test = 6.24 on 1 df, p=0.0125
- Score (log-rank) test = 6.47 on 1 df, p=0.0110

These three tests are asymptotically equivalent. For large enough N, they will give similar results. For small N, they may differ somewhat. The last row, "Score (logrank) test" is the result for the log-rank test, with p=0.011, the same result as the log-rank test, because the log-rank test is a special case of a Cox PH regression. The Likelihood ratio test has better behavior for small sample sizes, so it is generally preferred.

Cox model using a covariate in the melanoma data

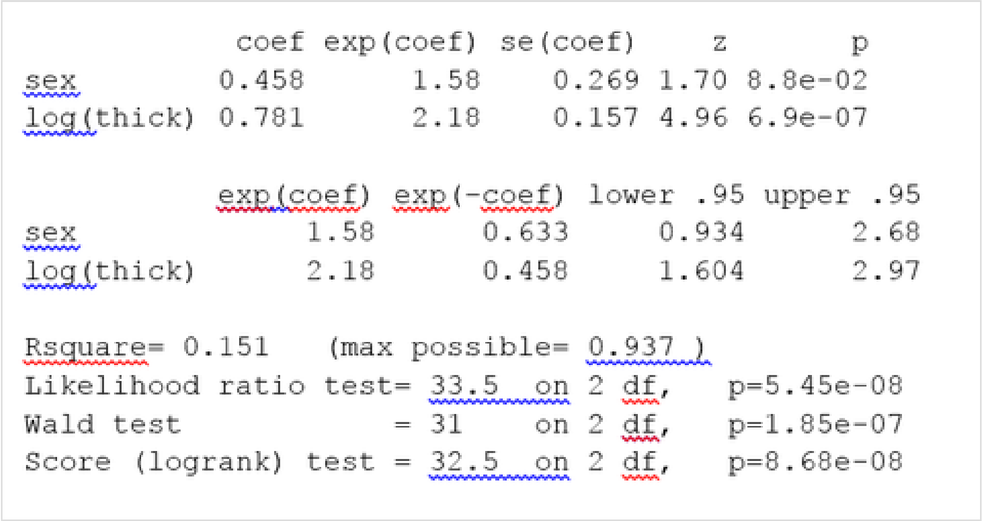
The Cox model extends the log-rank test by allowing the inclusion of additional covariates.[3] This example use the melanoma data set where the predictor variables include a continuous covariate, the thickness of the tumor (variable name = "thick").



Histograms of melanoma tumor thickness

In the histograms, the thickness values are positively skewed and do not have a Gaussian-like, Symmetric probability distribution. Regression models, including the Cox model, generally give more reliable results with normally-distributed variables. For this example we may use a logarithmic transform. The log of the thickness of the tumor looks to be more normally distributed, so the Cox models will use log thickness. The Cox PH analysis gives the results in the box.

The p-value for all three overall tests (likelihood, Wald, and score) are significant, indicating that the model is significant. The p-value for log(thick) is 6.9e-07, with a hazard ratio HR = exp(coef) = 2.18, indicating a strong relationship between the thickness of the tumor and increased risk of death.



```
              coef exp(coef) se(coef)    z       p
sex          0.458      1.58    0.269 1.70 8.8e-02
log(thick)   0.781      2.18    0.157 4.96 6.9e-07

             exp(coef) exp(-coef) lower .95 upper .95
sex               1.58      0.633     0.934      2.68
log(thick)        2.18      0.458     1.604      2.97

Rsquare= 0.151   (max possible= 0.937 )
Likelihood ratio test= 33.5  on 2 df,   p=5.45e-08
Wald test            = 31    on 2 df,   p=1.85e-07
Score (logrank) test = 32.5  on 2 df,   p=8.68e-08
```

Cox PH output for melanoma data set with covariate log tumor thickness

By contrast, the p-value for sex is now p=0.088. The hazard ratio HR = exp(coef) = 1.58, with a 95% confidence interval of 0.934 to 2.68. Because the confidence interval for HR includes 1, these results indicate that sex makes a smaller contribution to the difference in the HR after controlling for the thickness of the tumor, and only trend toward significance. Examination of graphs of log(thickness) by sex and a t-test of log(thickness) by sex both indicate that there is a significant difference between men and women in the thickness of the tumor when they first see the clinician.

The Cox model assumes that the hazards are proportional. The proportional hazard assumption may be tested using the R function cox.zph(). A p-value which is less than 0.05 indicates that the hazards are not proportional. For the melanoma data we obtain p=0.222. Hence, we cannot reject the null hypothesis of the hazards being proportional. Additional tests and graphs for examining a Cox model are described in the textbooks cited.

Extensions to Cox models

Cox models can be extended to deal with variations on the simple analysis.

- Stratification. The subjects can be divided into strata, where subjects within a stratum are expected to be relatively more similar to each other than to randomly chosen subjects from other strata. The regression parameters are assumed to be the same across the strata, but a different baseline hazard may exist for each stratum. Stratification is useful for analyses using matched subjects, for dealing with patient subsets, such as different clinics, and for dealing with violations of the proportional hazard assumption.
- Time-varying covariates. Some variables, such as gender and treatment group, generally stay the same in a clinical trial. Other clinical variables, such as serum protein levels or dose of concomitant medications may change over the course of a study. Cox models may be extended for such time-varying covariates.

Tree-structured survival models

The Cox PH regression model is a linear model. It is similar to linear regression and logistic regression. Specifically, these methods assume that a single line, curve, plane, or surface is sufficient to separate groups (alive, dead) or to estimate a quantitative response (survival time).

In some cases alternative partitions give more accurate classification or quantitative estimates. One set of alternative methods are tree-structured survival models,[4][5][6] including survival random forests.[7] Tree-structured survival models may give more accurate predictions than Cox models. Examining both types of models for a given data set is a reasonable strategy.

Example survival tree analysis

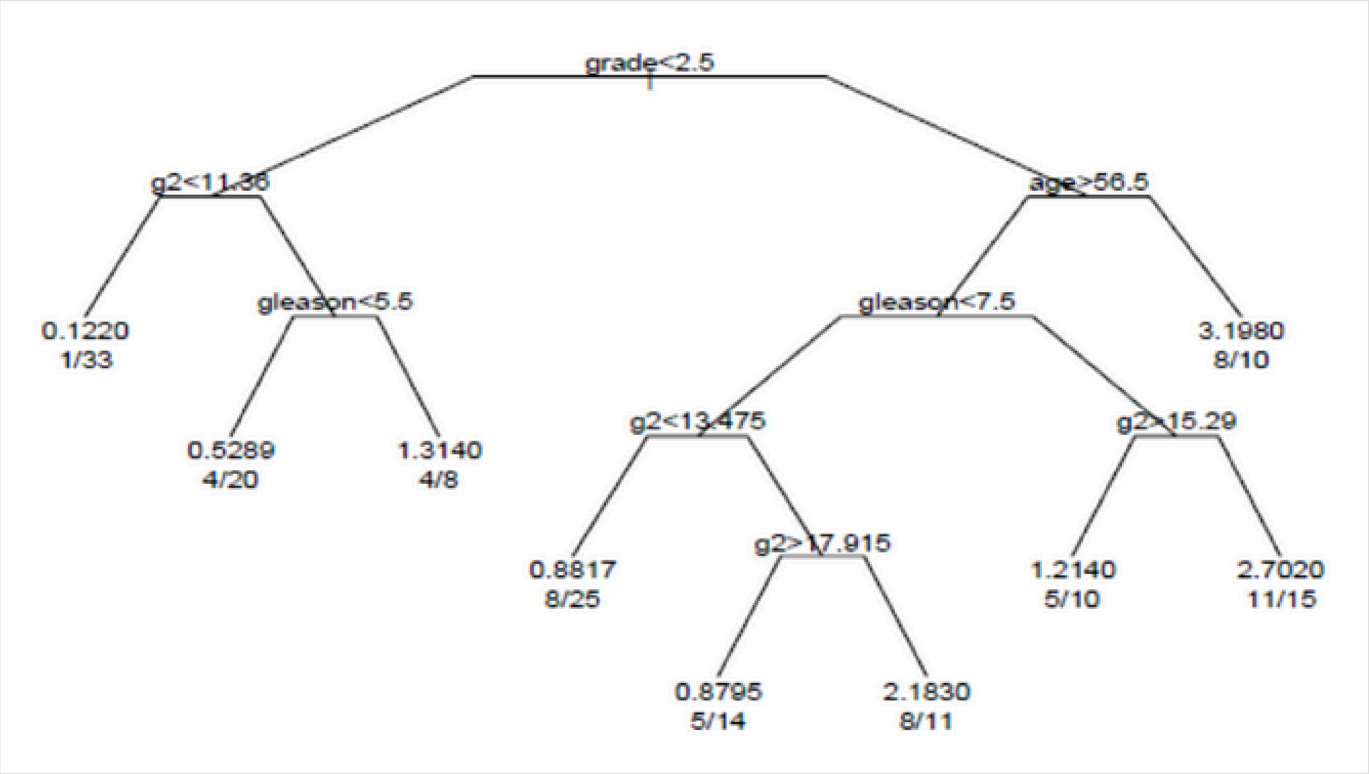
This example of a survival tree analysis uses the R package "rpart".[8] The example is based on 146 stage C prostate cancer patients in the data set stagec in rpart. Rpart and the stagec example are described in Atkinson and Therneau (1997),[9] which is also distributed as a vignette of the rpart package.[8]

The variables in stages are:

- **pgtime**: time to progression, or last follow-up free of progression
- **pgstat**: status at last follow-up (1=progressed, 0=censored)
- **age**: age at diagnosis
- **eet**: early endocrine therapy (1=no, 0=yes)

- **ploidy**: diploid/tetraploid/aneuploid DNA pattern
- **g2**: % of cells in G2 phase
- **grade**: tumor grade (1-4)
- **gleason**: Gleason grade (3-10)

The survival tree produced by the analysis is shown in the figure.



Survival tree for prostate cancer data set

Each branch in the tree indicates a split on the value of a variable. For example, the root of the tree splits subjects with grade < 2.5 versus subjects with grade 2.5 or greater. The terminal nodes indicate the number of subjects in the node, the number of subjects who have events, and the relative event rate compared to the root. In the node on the far left, the values 1/33 indicate that one of the 33 subjects in the node had an event, and that the relative event rate is 0.122. In the node on the far right bottom, the values 11/15 indicate that 11 of 15 subjects in the node had an event, and the relative event rate is 2.7.

Survival random forests

An alternative to building a single survival tree is to build many survival trees, where each tree is constructed using a sample of the data, and average the trees to predict survival.[7] This is the method underlying the survival random forest models. Survival random forest analysis is available in the R package "randomForestSRC".[10]

The randomForestSRC package includes an example survival random forest analysis using the data set pbc. This data is from the Mayo Clinic Primary Biliary Cirrhosis (PBC) trial of the liver conducted between 1974 and 1984. In the example, the random forest survival model gives more accurate predictions of survival than the Cox PH model. The prediction errors are estimated by bootstrap re-sampling.

Deep Learning survival models

Recent advancements in deep representation learning have been extended to survival estimation. The DeepSurv[11] model proposes to replace the log-linear parameterization of the CoxPH model with a multi-layer perceptron. Further extensions like Deep Survival Machines[12] and Deep Cox Mixtures[13] involve the use of latent variable mixture models to model the time-to-event distribution as a mixture of parametric or semi-parametric distributions while jointly learning representations of the input covariates. Deep learning approaches have shown superior performance especially on complex input data modalities such as images and clinical time-series.

General formulation

Survival function

The object of primary interest is the **survival function**, conventionally denoted *S*, which is defined as

$$S(t) = \Pr(T > t)$$

where *t* is some time, *T* is a random variable denoting the time of death, and "Pr" stands for probability. That is, the survival function is the probability that the time of death is later than some specified time *t*. The survival function is also called the *survivor function* or *survivorship function* in problems of biological survival, and the *reliability function* in mechanical survival problems. In the latter case, the reliability function is denoted *R*(*t*).

Usually one assumes *S*(0) = 1, although it could be less than 1 if there is the possibility of immediate death or failure.

The survival function must be non-increasing: *S*(*u*) ≤ *S*(*t*) if *u* ≥ *t*. This property follows directly because *T*>*u* implies *T*>*t*. This reflects the notion that survival to a later age is possible only if all younger ages are attained. Given this property, the lifetime distribution function and event density (*F* and *f* below) are well-defined.

The survival function is usually assumed to approach zero as age increases without bound (i.e., *S*(*t*) → 0 as *t* → ∞), although the limit could be greater than zero if eternal life is possible. For instance, we could apply survival analysis to a mixture of stable and unstable carbon isotopes; unstable isotopes would decay sooner or later, but the stable isotopes would last indefinitely.

Lifetime distribution function and event density

Related quantities are defined in terms of the survival function.

The **lifetime distribution function**, conventionally denoted *F*, is defined as the complement of the survival function,

$$F(t) = \Pr(T \le t) = 1 - S(t).$$

If *F* is differentiable then the derivative, which is the density function of the lifetime distribution, is conventionally denoted *f*,

$$f(t) = F'(t) = \frac{d}{dt}F(t).$$

The function *f* is sometimes called the **event density**; it is the rate of death or failure events per unit time.

The survival function can be expressed in terms of probability distribution and probability density functions

$$S(t) = \Pr(T > t) = \int_t^\infty f(u)\,du = 1 - F(t).$$

Similarly, a survival event density function can be defined as

$$s(t) = S'(t) = \frac{d}{dt}S(t) = \frac{d}{dt}\int_t^\infty f(u)\,du = \frac{d}{dt}[1 - F(t)] = -f(t).$$

In other fields, such as statistical physics, the survival event density function is known as the first passage time density.

Hazard function and cumulative hazard function

The **hazard function**, conventionally denoted λ or h, is defined as the event rate at time t conditional on survival until time t or later (that is, $T \geq t$). Suppose that an item has survived for a time t and we desire the probability that it will not survive for an additional time dt:

$$h(t) = \lim_{dt \to 0} \frac{\Pr(t \leq T < t + dt)}{dt \cdot S(t)} = \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)}.$$

Force of mortality is a synonym of *hazard function* which is used particularly in demography and actuarial science, where it is denoted by μ. The term *hazard rate* is another synonym.

The force of mortality of the survival function is defined as $\mu(x) = -\frac{d}{dx}\ln(S(x)) = \frac{f(x)}{S(x)}$

The force of mortality is also called the force of failure. It is the probability density function of the distribution of mortality.

In actuarial science, the hazard rate is the rate of death for lives aged x. For a life aged x, the force of mortality t years later is the force of mortality for a $(x + t)$–year old. The hazard rate is also called the failure rate. Hazard rate and failure rate are names used in reliability theory.

Any function h is a hazard function if and only if it satisfies the following properties:

1. $\forall x \geq 0\,(h(x) \geq 0)$,
2. $\int_0^\infty h(x)dx = \infty$.

In fact, the hazard rate is usually more informative about the underlying mechanism of failure than the other representations of a lifetime distribution.

The hazard function must be non-negative, $\lambda(t) \geq 0$, and its integral over $[0, \infty]$ must be infinite, but is not otherwise constrained; it may be increasing or decreasing, non-monotonic, or discontinuous. An example is the bathtub curve hazard function, which is large for small values of t, decreasing to some minimum, and thereafter increasing again; this can model the property of some mechanical systems to either fail soon after operation, or much later, as the system ages.

The hazard function can alternatively be represented in terms of the **cumulative hazard function**, conventionally denoted Λ or H:

$$\Lambda(t) = -\log S(t)$$

so transposing signs and exponentiating

$$S(t) = \exp(-\Lambda(t))$$

or differentiating (with the chain rule)

$$\frac{d}{dt}\Lambda(t) = -\frac{S'(t)}{S(t)} = \lambda(t).$$

The name "cumulative hazard function" is derived from the fact that

$$\Lambda(t) = \int_0^t \lambda(u)\,du$$

which is the "accumulation" of the hazard over time.

From the definition of $\Lambda(t)$, we see that it increases without bound as *t* tends to infinity (assuming that $S(t)$ tends to zero). This implies that $\lambda(t)$ must not decrease too quickly, since, by definition, the cumulative hazard has to diverge. For example, $\exp(-t)$ is not the hazard function of any survival distribution, because its integral converges to 1.

The survival function $S(t)$, the cumulative hazard function $\Lambda(t)$, the density $f(t)$, the hazard function $\lambda(t)$, and the lifetime distribution function $F(t)$ are related through

$$S(t) = \exp[-\Lambda(t)] = \frac{f(t)}{\lambda(t)} = 1 - F(t), \quad t > 0.$$

Quantities derived from the survival distribution

Future lifetime at a given time t_0 is the time remaining until death, given survival to age t_0. Thus, it is $T - t_0$ in the present notation. The **expected future lifetime** is the expected value of future lifetime. The probability of death at or before age $t_0 + t$, given survival until age t_0, is just

$$P(T \leq t_0 + t \mid T > t_0) = \frac{P(t_0 < T \leq t_0 + t)}{P(T > t_0)} = \frac{F(t_0 + t) - F(t_0)}{S(t_0)}.$$

Therefore, the probability density of future lifetime is

$$\frac{d}{dt}\frac{F(t_0+t)-F(t_0)}{S(t_0)}=\frac{f(t_0+t)}{S(t_0)}$$

and the expected future lifetime is

$$\frac{1}{S(t_0)}\int_0^\infty t\,f(t_0+t)\,dt=\frac{1}{S(t_0)}\int_{t_0}^\infty S(t)\,dt,$$

where the second expression is obtained using integration by parts.

For $t_0 = 0$, that is, at birth, this reduces to the expected lifetime.

In reliability problems, the expected lifetime is called the *mean time to failure*, and the expected future lifetime is called the *mean residual lifetime*.

As the probability of an individual surviving until age *t* or later is *S*(*t*), by definition, the expected number of survivors at age *t* out of an initial population of *n* newborns is *n* × *S*(*t*), assuming the same survival function for all individuals. Thus the expected proportion of survivors is *S*(*t*). If the survival of different individuals is independent, the number of survivors at age *t* has a binomial distribution with parameters *n* and *S*(*t*), and the variance of the proportion of survivors is *S*(*t*) × (1-*S*(*t*))/*n*.

The age at which a specified proportion of survivors remain can be found by solving the equation *S*(*t*) = *q* for *t*, where *q* is the quantile in question. Typically one is interested in the **median lifetime**, for which *q* = 1/2, or other quantiles such as *q* = 0.90 or *q* = 0.99.

Censoring

Censoring is a form of missing data problem in which time to event is not observed for reasons such as termination of study before all recruited subjects have shown the event of interest or the subject has left the study prior to experiencing an event. Censoring is common in survival analysis.

If only the lower limit *l* for the true event time *T* is known such that *T* > *l*, this is called *right censoring*. Right censoring will occur, for example, for those subjects whose birth date is known but who are still alive when they are lost to follow-up or when the study ends. We generally encounter right-censored data.

If the event of interest has already happened before the subject is included in the study but it is not known when it occurred, the data is said to be *left-censored*.[14] When it can only be said that the event happened between two observations or examinations, this is *interval censoring*.

Left censoring occurs for example when a permanent tooth has already emerged prior to the start of a dental study that aims to estimate its emergence distribution. In the same study, an emergence time is interval-censored when the permanent tooth is present in the mouth at the current examination but not yet at the previous examination. Interval censoring often occurs in HIV/AIDS studies. Indeed, time to HIV seroconversion can be determined only by a laboratory assessment which is usually initiated after a visit to the physician. Then one can only conclude that HIV seroconversion has happened between two examinations. The same is true for the diagnosis of AIDS, which is based on clinical symptoms and needs to be confirmed by a medical examination.

It may also happen that subjects with a lifetime less than some threshold may not be observed at all: this is called *truncation*. Note that truncation is different from left censoring, since for a left censored datum, we know the subject exists, but for a truncated datum, we may be completely unaware of the subject. Truncation is also common. In a so-called *delayed entry* study, subjects are not observed at all until they have reached a certain age. For example, people may not be observed until they have reached the age to enter school. Any deceased subjects in the pre-school age group would be unknown. Left-truncated data are common in actuarial work for life insurance and pensions.[15]

Left-censored data can occur when a person's survival time becomes incomplete on the left side of the follow-up period for the person. For example, in an epidemiological example, we may monitor a patient for an infectious disorder starting from the time when he or she is tested positive for the infection. Although we may know the right-hand side of the duration of interest, we may never know the exact time of exposure to the infectious agent.[16]

Fitting parameters to data

Survival models can be usefully viewed as ordinary regression models in which the response variable is time. However, computing the likelihood function (needed for fitting parameters or making other kinds of inferences) is complicated by the censoring. The likelihood function for a survival model, in the presence of censored data, is formulated as follows. By definition the likelihood function is the conditional probability of the data given the parameters of the model. It is customary to assume that the data are independent given the parameters. Then the likelihood function is the product of the likelihood of each datum. It is convenient to partition the data into four categories: uncensored, left censored, right censored, and interval censored. These are denoted "unc.", "l.c.", "r.c.", and "i.c." in the equation below.

$$L(\theta)=\prod_{T_i\in unc.}\Pr(T=T_i\mid\theta)\prod_{i\in l.c.}\Pr(T<T_i\mid\theta)\prod_{i\in r.c.}\Pr(T>T_i\mid\theta)\prod_{i\in i.c.}\Pr(T_{i,l}<T<T_{i,r}\mid\theta).$$

For uncensored data, with T_i equal to the age at death, we have

$$\Pr(T=T_i\mid\theta)=f(T_i\mid\theta).$$

For left-censored data, such that the age at death is known to be less than T_i, we have

$$\Pr(T<T_i\mid\theta)=F(T_i\mid\theta)=1-S(T_i\mid\theta).$$

For right-censored data, such that the age at death is known to be greater than T_i, we have

$$\Pr(T>T_i\mid\theta)=1-F(T_i\mid\theta)=S(T_i\mid\theta).$$

For an interval censored datum, such that the age at death is known to be less than $T_{i,r}$ and greater than $T_{i,l}$, we have

$$\Pr(T_{i,l} < T < T_{i,r} \mid \theta) = S(T_{i,l} \mid \theta) - S(T_{i,r} \mid \theta).$$

An important application where interval-censored data arises is current status data, where an event T_i is known not to have occurred before an observation time and to have occurred before the next observation time.

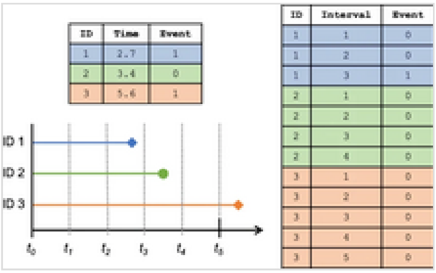
Non-parametric estimation

The Kaplan–Meier estimator can be used to estimate the survival function. The Nelson–Aalen estimator can be used to provide a non-parametric estimate of the cumulative hazard rate function. These estimators require lifetime data. Periodic case (cohort) and death (and recovery) counts are statistically sufficient to make nonparametric maximum likelihood and least squares estimates of survival functions, without lifetime data.

Discrete-time survival models

While many parametric models assume a continuous-time, discrete-time survival models can be mapped to a binary classification problem. In a discrete-time survival model the survival period is artificially resampled in intervals where for each interval a binary target indicator is recorded if the event takes place in a certain time horizon.[17] If a binary classifier (potentially enhanced with a different likelihood to take more structure of the problem into account) is calibrated, then the classifier score is the hazard function (i.e. the conditional probability of failure).[17]

Discrete-time survival models are connected to empirical likelihood.[18][19]



ID	Time	Event
1	2.7	1
2	3.4	0
3	5.6	1

ID	Interval	Event
1	1	0
1	2	0
1	3	1
2	1	0
2	2	0
2	3	0
2	4	0
3	1	0
3	2	0
3	3	0
3	4	0
3	5	0

Description of the transformation of continuous-time survival data to discrete-time survival data. Individual 4 is censored and for individual 5 the event happens outside the observation window 5.

Goodness of fit

The goodness of fit of survival models can be assessed using scoring rules.[20]

Computer software for survival analysis

The textbook by Kleinbaum has examples of survival analyses using SAS, R, and other packages.[21] The textbooks by Brostrom,[22] Dalgaard[2] and Tableman and Kim[23] give examples of survival analyses using R (or using S, and which run in R).

Distributions used in survival analysis

- Exponential distribution
- Weibull distribution
- Log-logistic distribution

- Gamma distribution
- Exponential-logarithmic distribution
- Generalized gamma distribution
- Hypertabastic distribution

Applications

- Credit risk[24][25]
- False conviction rate of inmates sentenced to death[26]
- Lead times for metallic components in the aerospace industry[27]
- Predictors of criminal recidivism[28]
- Survival distribution of radio-tagged animals[29]
- Time-to-violent death of Roman emperors[30]
- Intertrade waiting times of electronically traded shares on a stock exchange[31]

See also

- Accelerated failure time model
- Bayesian survival analysis
- Cell survival curve
- Censoring (statistics)
- Chance-constrained portfolio selection
- Failure rate
- Frequency of exceedance
- Kaplan–Meier estimator
- Logrank test
- Maximum likelihood
- Mortality rate
- MTBF
- Proportional hazards models
- Reliability theory
- Residence time (statistics)
- Sequence analysis in social sciences
- Survival function
- Survival rate
- Discrete-time proportional hazards

References

1. Miller, Rupert G. (1997), *Survival analysis*, John Wiley & Sons, ISBN 0-471-25218-2
2. Dalgaard, Peter (2008), *Introductory Statistics with R* (Second ed.), Springer, ISBN 978-0387790534

3. Saegusa, Takumi; Di, Chongzhi; Chen, Ying Qing (September 2014). "Hypothesis testing for an extended cox model with time-varying coefficients" (https://academic.oup.com/biometrics/article/70/3/619-628/7419905). *Biometrics*. **70** (3): 619–628. doi:10.1111/biom.12185 (https://doi.org/10.1111%2Fbiom.12185). ISSN 0006-341X (https://search.worldcat.org/issn/0006-341X). PMC 4247822 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4247822).
4. Segal, Mark Robert (1988). "Regression Trees for Censored Data" (https://www.jstor.org/stable/2531894). *Biometrics*. **44** (1): 35–47. doi:10.2307/2531894 (https://doi.org/10.2307%2F2531894). JSTOR 2531894 (https://www.jstor.org/stable/2531894). S2CID 60974957 (https://api.semanticscholar.org/CorpusID:60974957).
5. Leblanc, Michael; Crowley, John (1993). "Survival Trees by Goodness of Split" (http://www.tandfonline.com/doi/abs/10.1080/01621459.1993.10476296). *Journal of the American Statistical Association*. **88** (422): 457–467. doi:10.1080/01621459.1993.10476296 (https://doi.org/10.1080%2F01621459.1993.10476296). ISSN 0162-1459 (https://search.worldcat.org/issn/0162-1459).
6. Ritschard, Gilbert; Gabadinho, Alexis; Muller, Nicolas S.; Studer, Matthias (2008). "Mining event histories: a social science perspective" (http://www.inderscience.com/link.php?id=22538). *International Journal of Data Mining, Modelling and Management*. **1** (1): 68. doi:10.1504/IJDMMM.2008.022538 (https://doi.org/10.1504%2FIJDMMM.2008.022538). ISSN 1759-1163 (https://search.worldcat.org/issn/1759-1163).
7. Ishwaran, Hemant; Kogalur, Udaya B.; Blackstone, Eugene H.; Lauer, Michael S. (2008-09-01). "Random survival forests" (https://doi.org/10.1214%2F08-AOAS169). *The Annals of Applied Statistics*. **2** (3). arXiv:0811.1645 (https://arxiv.org/abs/0811.1645). doi:10.1214/08-AOAS169 (https://doi.org/10.1214%2F08-AOAS169). ISSN 1932-6157 (https://search.worldcat.org/issn/1932-6157). S2CID 2003897 (https://api.semanticscholar.org/CorpusID:2003897).
8. Therneau, Terry J.; Atkinson, Elizabeth J. "rpart: Recursive Partitioning and Regression Trees" (https://CRAN.R-project.org/package=rpart). *CRAN*. Retrieved November 12, 2021.
9. Atkinson, Elizabeth J.; Therneau, Terry J. (1997). *An introduction to recursive partitioning using the RPART routines* (https://www.researchgate.net/publication/235665541). Mayo Foundation.
10. Ishwaran, Hemant; Kogalur, Udaya B. "randomForestSRC: Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC)" (https://CRAN.R-project.org/package=randomForestSRC). *CRAN*. Retrieved November 12, 2021.
11. Singh, Jared; Katzman, L. (2018). "DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network". *BMC Medical Research Methodology*.
12. Nagpal, Chirag (2021). "Deep survival machines: Fully parametric survival regression and representation learning for censored data with competing risks". *IEEE Journal of Biomedical and Health Informatics*. **25** (8): 3163–3175. arXiv:2003.01176 (https://arxiv.org/abs/2003.01176). doi:10.1109/JBHI.2021.3052441 (https://doi.org/10.1109%2FJBHI.2021.3052441). PMID 33460387 (https://pubmed.ncbi.nlm.nih.gov/33460387). S2CID 211817982 (https://api.semanticscholar.org/CorpusID:211817982).
13. Nagpal, Chirag (2021). "Deep Cox mixtures for survival regression". *Machine Learning for Healthcare Conference*. arXiv:2101.06536 (https://arxiv.org/abs/2101.06536).
14. Darity, William A. Jr., ed. (2008). "Censoring, Left and Right" (http://ic.galegroup.com/ic/uhic/ReferenceDetailsPage/ReferenceDetailsWindow?disableHighlighting=false&displayGroupName=Reference&currPage=&scanId=&query=&prodId=UHIC&search_within_results=&p=UHIC%3AWHIC&mode=view&catId=&limiter=&display-query=&displayGroups=&contentModules=&action=e&sortBy=&documentId=GALE%7CCX3045300295&windowstate=normal&activityType=&failOverType=&commentary=&source=Bookmark&u=mlin_w_amhercol&jsid=0938fef854cc86b83b5fe8a2c4bcb54b). *International Encyclopedia of the Social Sciences*. Vol. 1 (2nd ed.). Macmillan. pp. 473–474. Retrieved 6 November 2016.
15. Richards, S. J. (2012). "A handbook of parametric survival models for actuarial use". *Scandinavian Actuarial Journal*. **2012** (4): 233–257. doi:10.1080/03461238.2010.506688 (https://doi.org/10.1080%2F03461238.2010.506688). S2CID 119577304 (https://api.semanticscholar.org/CorpusID:119577304).

16. Singh, R.; Mukhopadhyay, K. (2011). "Survival analysis in clinical trials: Basics and must know areas" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3227332). *Perspect Clin Res*. **2** (4): 145–148. doi:10.4103/2229-3485.86872 (https://doi.org/10.4103%2F2229-3485.86872). PMC 3227332 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3227332). PMID 22145125 (https://pubmed.ncbi.nlm.nih.gov/22145125).
17. Suresh, K., Severn, C. & Ghosh, D. Survival prediction models: an introduction to discrete-time modeling. BMC Med Res Methodol 22, 207 (2022). https://doi.org/10.1186/s12874-022-01679-6 , https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-022-01679-6
18. Empirical Likelihood in Survival Analysis, Gang Li (U.S.A.), Runze Li (U.S.A.), and Mai Zhou (U.S.A.), Contemporary Multivariate Analysis and Design of Experiments. March 2005, 337-349, https://www.ms.uky.edu/~mai/research/llz.pdf
19. The Empirical Distribution Function with Arbitrarily Grouped, Censored and Truncated Data, Bruce W. Turnbull, Journal of the Royal Statistical Society. Series B (Methodological) Vol. 38, No. 3 (1976), pp. 290-295 (6 pages), https://apps.dtic.mil/sti/tr/pdf/ADA030940.pdf
20. Proper Scoring Rules for Survival Analysis, Hiroki Yanagisawa, https://arxiv.org/abs/2305.00621v3
21. Kleinbaum, David G.; Klein, Mitchel (2012), *Survival analysis: A Self-learning text* (Third ed.), Springer, ISBN 978-1441966452
22. Brostrom, Göran (2012), *Event History Analysis with R* (First ed.), Chapman & Hall/CRC, ISBN 978-1439831649
23. Tableman, Mara; Kim, Jong Sung (2003), *Survival Analysis Using S* (First ed.), Chapman and Hall/CRC, ISBN 978-1584884088
24. Stepanova, Maria; Thomas, Lyn (2002-04-01). "Survival Analysis Methods for Personal Loan Data". *Operations Research*. **50** (2): 277–289. doi:10.1287/opre.50.2.277.426 (https://doi.org/10.1287%2Fopre.50.2.277.426). ISSN 0030-364X (https://search.worldcat.org/issn/0030-364X).
25. Glennon, Dennis; Nigro, Peter (2005). "Measuring the Default Risk of Small Business Loans: A Survival Analysis Approach". *Journal of Money, Credit and Banking*. **37** (5): 923–947. doi:10.1353/mcb.2005.0051 (https://doi.org/10.1353%2Fmcb.2005.0051). ISSN 0022-2879 (https://search.worldcat.org/issn/0022-2879). JSTOR 3839153 (https://www.jstor.org/stable/3839153). S2CID 154615623 (https://api.semanticscholar.org/CorpusID:154615623).
26. Kennedy, Edward H.; Hu, Chen; O’Brien, Barbara; Gross, Samuel R. (2014-05-20). "Rate of false conviction of criminal defendants who are sentenced to death" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4034186). *Proceedings of the National Academy of Sciences*. **111** (20): 7230–7235. Bibcode:2014PNAS..111.7230G (https://ui.adsabs.harvard.edu/abs/2014PNAS..111.7230G). doi:10.1073/pnas.1306417111 (https://doi.org/10.1073%2Fpnas.1306417111). ISSN 0027-8424 (https://search.worldcat.org/issn/0027-8424). PMC 4034186 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4034186). PMID 24778209 (https://pubmed.ncbi.nlm.nih.gov/24778209).
27. de Cos Juez, F. J.; García Nieto, P. J.; Martínez Torres, J.; Taboada Castro, J. (2010-10-01). "Analysis of lead times of metallic components in the aerospace industry through a supported vector machine model" (https://doi.org/10.1016%2Fj.mcm.2010.03.017). *Mathematical and Computer Modelling*. Mathematical Models in Medicine, Business & Engineering 2009. **52** (7): 1177–1184. doi:10.1016/j.mcm.2010.03.017 (https://doi.org/10.1016%2Fj.mcm.2010.03.017). ISSN 0895-7177 (https://search.worldcat.org/issn/0895-7177).
28. Spivak, Andrew L.; Damphousse, Kelly R. (2006). "Who Returns to Prison? A Survival Analysis of Recidivism among Adult Offenders Released in Oklahoma, 1985 – 2004". *Justice Research and Policy*. **8** (2): 57–88. doi:10.3818/jrp.8.2.2006.57 (https://doi.org/10.3818%2Fjrp.8.2.2006.57). ISSN 1525-1071 (https://search.worldcat.org/issn/1525-1071). S2CID 144566819 (https://api.semanticscholar.org/CorpusID:144566819).
29. Pollock, Kenneth H.; Winterstein, Scott R.; Bunck, Christine M.; Curtis, Paul D. (1989). "Survival Analysis in Telemetry Studies: The Staggered Entry Design" (http://www.lib.ncsu.edu/resolver/1840.4/8416). *The Journal of Wildlife Management*. **53** (1): 7–15. doi:10.2307/3801296 (https://doi.org/10.2307%2F3801296). ISSN 0022-541X (https://search.worldcat.org/issn/0022-541X). JSTOR 3801296 (https://www.jstor.org/stable/3801296).

30. Saleh, Joseph Homer (2019-12-23). "Statistical reliability analysis for a most dangerous occupation: Roman emperor" (https://doi.org/10.1057%2Fs41599-019-0366-y). *Palgrave Communications*. **5** (1): 1–7. doi:10.1057/s41599-019-0366-y (https://doi.org/10.1057%2Fs41599-019-0366-y). ISSN 2055-1045 (https://search.worldcat.org/issn/2055-1045).
31. Kreer, Markus; Kizilersu, Ayse; Thomas, Anthony W. (2022). "Censored expectation maximization algorithm for mixtures: Application to intertrade waiting times" (https://www.sciencedirect.com/science/article/pii/S0378437121007299). *Physica A: Statistical Mechanics and Its Applications*. **587** (1): 126456. Bibcode:2022PhyA..58726456K (https://ui.adsabs.harvard.edu/abs/2022PhyA..58726456K). doi:10.1016/j.physa.2021.126456 (https://doi.org/10.1016%2Fj.physa.2021.126456). ISSN 0378-4371 (https://search.worldcat.org/issn/0378-4371). S2CID 244198364 (https://api.semanticscholar.org/CorpusID:244198364).

Further reading

- Collett, David (2003). *Modelling Survival Data in Medical Research* (Second ed.). Boca Raton: Chapman & Hall/CRC. ISBN 1584883251.
- Elandt-Johnson, Regina; Johnson, Norman (1999). *Survival Models and Data Analysis*. New York: John Wiley & Sons. ISBN 0471349925.
- Kalbfleisch, J. D.; Prentice, Ross L. (2002). *The statistical analysis of failure time data*. New York: John Wiley & Sons. ISBN 047136357X.
- Lawless, Jerald F. (2003). *Statistical Models and Methods for Lifetime Data* (2nd ed.). Hoboken: John Wiley and Sons. ISBN 0471372153.
- Rausand, M.; Hoyland, A. (2004). *System Reliability Theory: Models, Statistical Methods, and Applications*. Hoboken: John Wiley & Sons. ISBN 047147133X.

External links

- Therneau, Terry. "A Package for Survival Analysis in S" (https://web.archive.org/web/20060907234826/http://www.mayo.edu/hsr/people/therneau/survival.ps). Archived from the original (http://www.mayo.edu/hsr/people/therneau/survival.ps) on 2006-09-07. via Dr. Therneau's page on the Mayo Clinic website (https://web.archive.org/web/20130209163950/http://mayoresearch.mayo.edu/mayo/research/biostat/therneau.cfm)
- "Engineering Statistics Handbook" (http://www.itl.nist.gov/div898/handbook/). NIST/SEMATEK.
- SOCR, Survival analysis applet (http://www.socr.ucla.edu/htmls/ana/Survival_Analysis.html) and interactive learning activity (http://wiki.stat.ucla.edu/socr/index.php/SOCR_EduMaterials_AnalysisActivities_Survival).
- Survival/Failure Time Analysis (http://www.statsoft.com/textbook/stsurvan.html) @ Statistics' Textbook Page (http://www.statsoft.com/textbook/)
- Survival Analysis in R (http://www.netstorm.be/home/survival)
- Lifelines, a Python package for survival analysis (http://lifelines.readthedocs.org/en/latest/)
- Survival Analysis in NAG Fortran Library (http://www.nag.co.uk/numeric/fl/nagdoc_fl24/html/G12/g12conts.html)

Retrieved from "https://en.wikipedia.org/w/index.php?title=Survival_analysis&oldid=1255969085"