Survival Analysis

Polychronis Economou

BIOSTATISTICS – DATA ANALYTICS

Biomedical Engineering

Outline

Outline

Time to event data

Survival Analysis (or Reliability Analysis depending on scientific field applied) typically focuses on time to event data, i.e. the time until the occurrence of the event of interest.

The **response** T of interest is often referred to as a failure time, survival time or event time and it is a **positive valued** random variable.

Examples

- time to death
- time to onset (or relapse) of a disease
- length of stay in a hospital
- time until tumor recurrence
- time until a machine part fails

↪ money paid by health insurance

↪ kilometers until a car breaks down

- duration of a strike

What is the main target of survival analysis?

The main target of survival analysis is the estimation of the so called survival function (or reliability function)

$$\begin{aligned} S(t) &= P(T > t) \\ &= 1 - F(t) \end{aligned}$$

where $F(t)$ is the cumulative distribution function of T.

Survival function describes the probability that an individual will survive beyond a specified time.

Alternative notations: $R(t), \bar{F}(t)$.

Why we need the survival analysis?

Data characteristics

In survival analysis subjects are usually followed over a specified time period and the focus is on the time at which the event of interest occurs.

As a result the time to event data often present a characteristic feature, known as **censoring**, which broadly speaking is when the time to event (lifetime) is incompletely determined for some subjects, i.e. for some subjects we may know that their survival time have occurred within certain intervals, whereas, for other subjects, we will know their exact time of event.

Another feature of time to event data that may be present in some survival studies is that of **truncation**. Truncation occurs when only individuals who experience some event can be observed by the investigator.

Censoring

There are various categories of censoring, such as

- right censoring
- left censoring
- interval censoring

Censoring

A **right censored observation** is one that is known only to be larger than some value (for example lifetime>70 years).

A **left censored observation** is one that is known only to be less than some value (for example I start smoking at some age earlier than 15 but I can not remember exactly)

An **interval censored observation** is reported as being within a specified interval (for examples a tumor recurrence in a patient may be known to fall only the interval between visits).

Right Censoring

Let

- C_i be the (right) censoring time for the i^{th} individual
- F_i is its Failure time

then we observe

$$T_i = min(F_i, C_i)$$

There are three types of right censoring

- **type I**
- **type II**
- **random (generalized) type I**

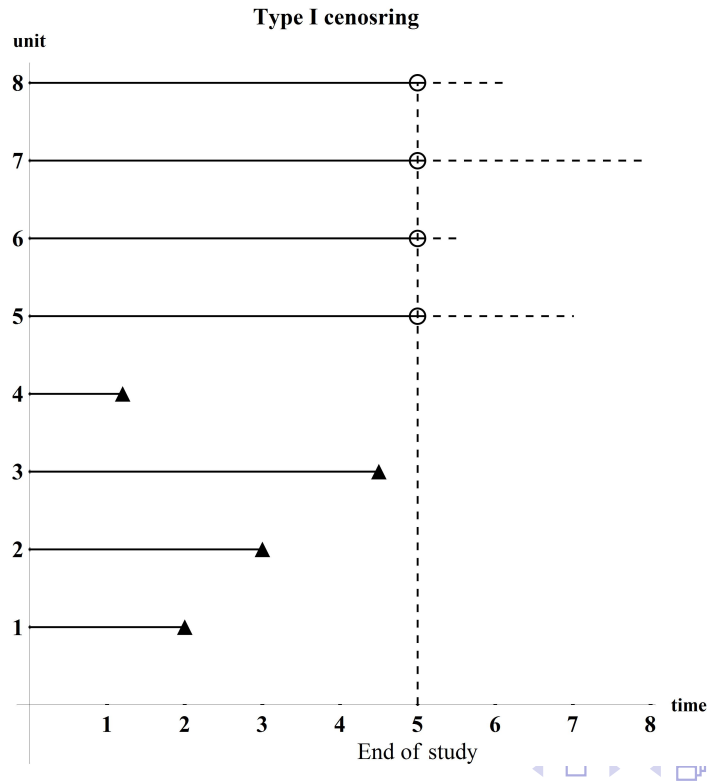
Right Censoring

Type I Right Censoring

Type I occurs if an experiment has a set number of subjects or items and stops the experiment at a predetermined time. Any subjects remaining are right censored at this point.

In this case C_i is predetermined and common to all individuals ($C_i = C$)

- number of failures is random

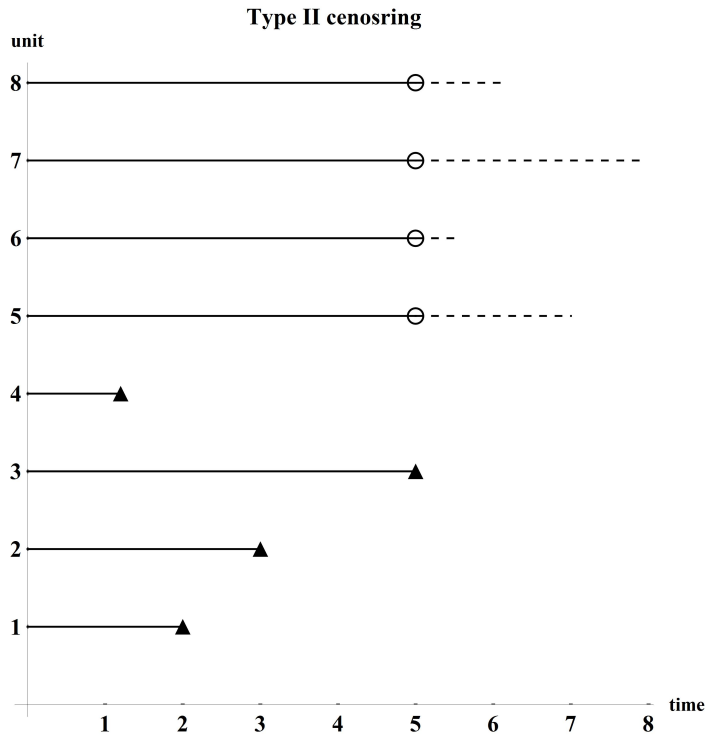


Right Censoring

Type II Right Censoring

Type II occurs if an experiment has a set number of subjects or items and stops the experiment when a predetermined number K are observed to have failed; the remaining subjects are then right-censored.

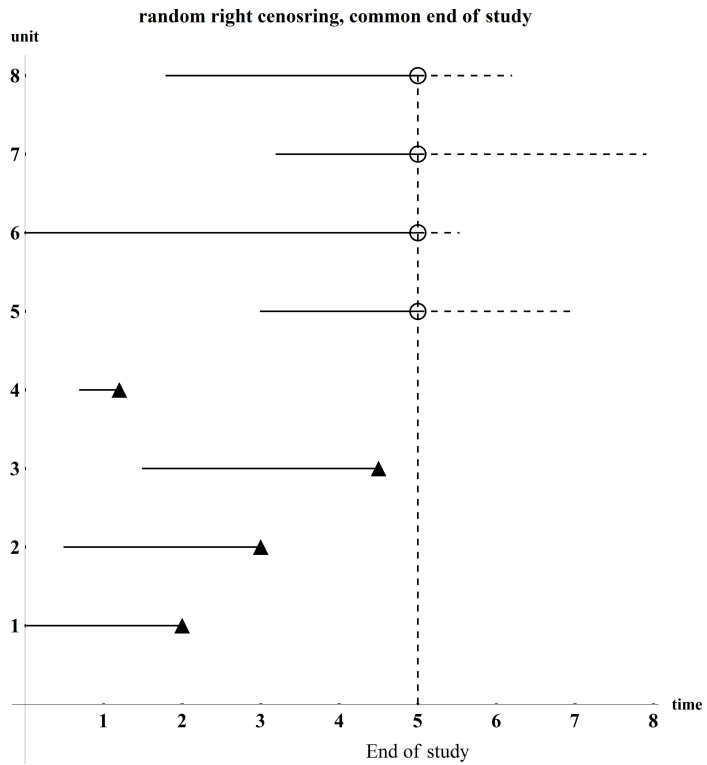
- Time to the k^{th} failures is random

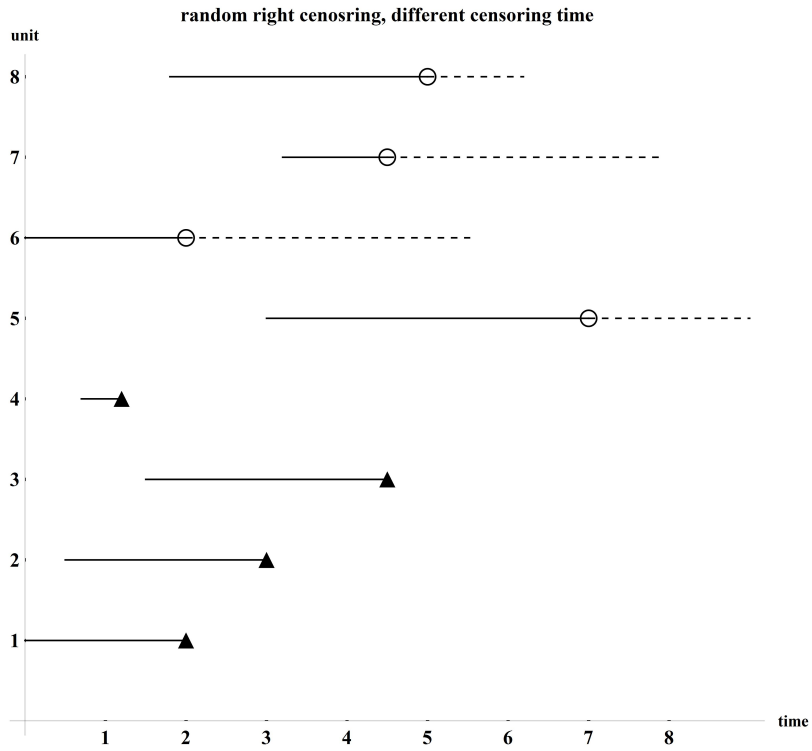


Right Censoring

Random (generalized) type I Right Censoring

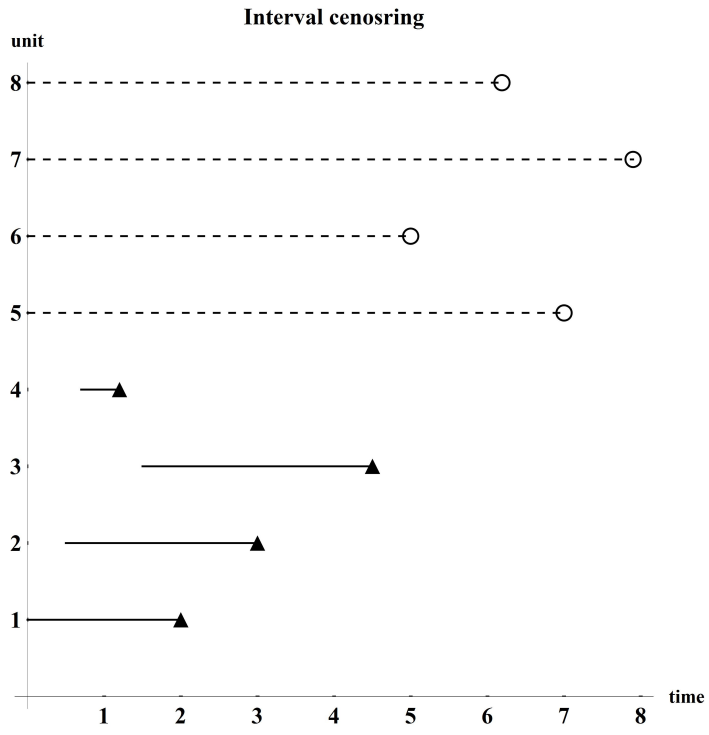
Random (generalized) type I is when each subject has a different censoring time





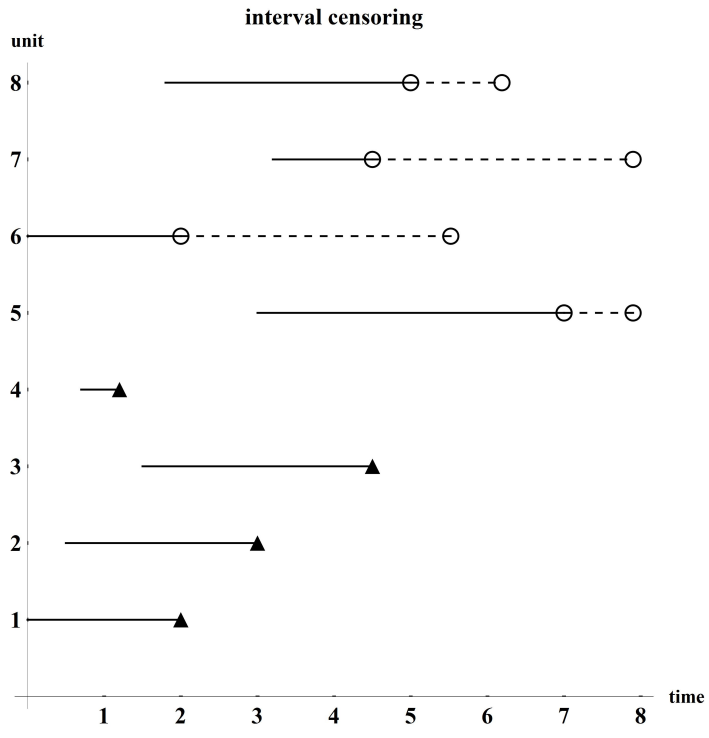
Left Censoring

A **left censored observation** is one that is known only to be less than some value (for example I start smoking at some age earlier than 15 but I can not remember exactly)



Interval Censoring

An **interval censored observation** is reported as being within a specified interval (for examples a tumor recurrence in a patient may be known to fall only the interval between visits).



Non-informative Censoring

In any case censoring is assumed to **statistically independent** of the failure time.

Truncation

Truncation is similar to but distinct from the concept of censoring.

Truncation occurs when the subjects have been at risk before entering the study.

This means that for a portion of the population **the event of interest may have occurred but could not be observed** and as result is unknown if or not has occurred.

As a consequence, the investigator is not aware of the existence of these individuals.

There are **two types** of truncation

- left and
- right

Data sets

In most of the cases we deal with **right censoring** & sometimes **left truncation.**

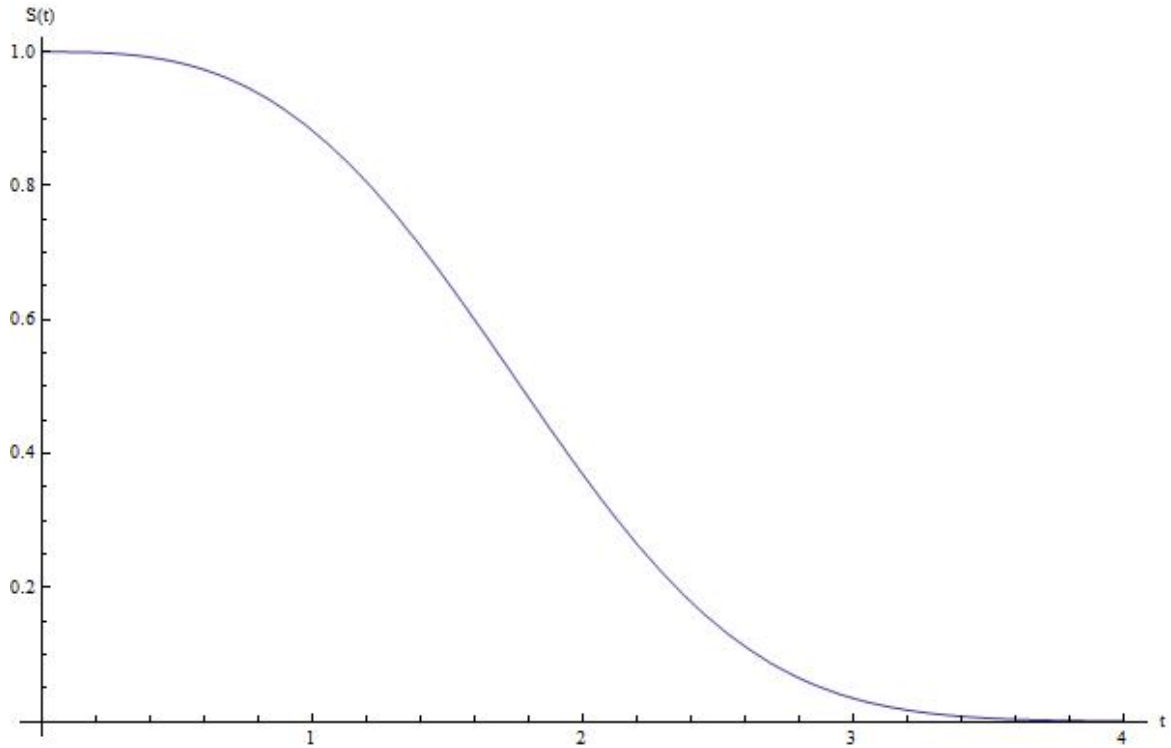
Outline

Survival function

T denotes the time to event (lifetime) variable, $T \geq 0$

The **survival function** is defined as

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



Survival function

Interpretation and Properties

Interpretation

The survival function gives the probability that a subject will survive past time t.

Properties

The survival function has the following properties

- It is non-increasing
- At time $t = 0$, S(t) $= 1$.
- At time $t = \infty$, $S(t) = S(\infty) = 0$. As time goes to infinity, the survival curve goes to 0.

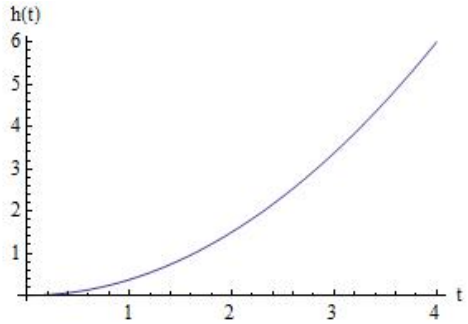
Hazard function

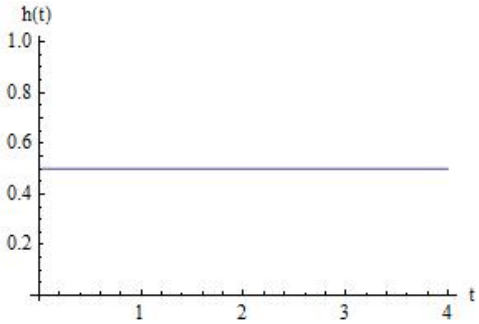
The hazard function $h(t)$ (also known as the failure rate, hazard rate, or force of mortality) is **the instantaneous rate of failure** (experiencing the event) **at time t given that an individual is alive at time t**.

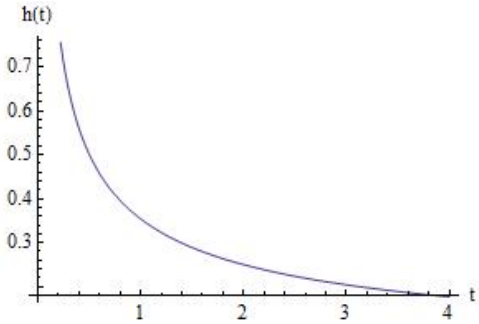
$$
\begin{aligned}
h(t) &= \lim_{dt \to 0} \frac{\Pr(t \leq T < t + dt | T \geq t)}{dt} \\
&= \lim_{dt \to 0} \frac{\Pr(t \leq T < t + dt)}{dt \cdot S(t)} \\
&= \lim_{dt \to 0} \frac{F(t + dt) - F(t)}{dt \cdot S(t)} \\
&= \frac{F'(t)}{S(t)} \\
&= \frac{f(t)}{S(t)} =
\end{aligned}
$$

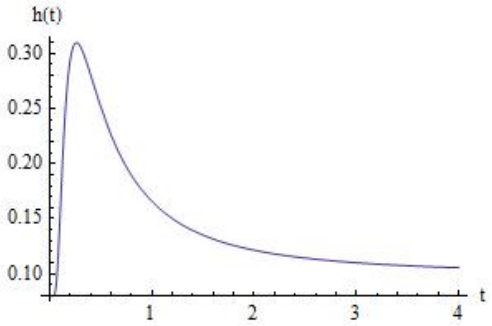
Hazard function

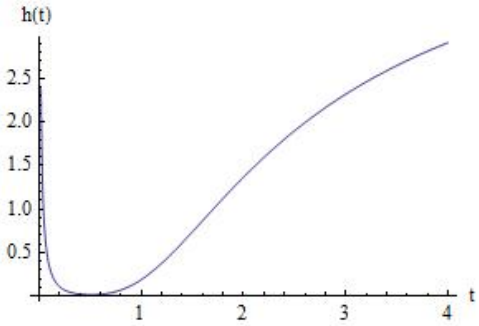
hazard patterns











Cumulative hazard function

The cumulative hazard $H(t)$ describes the accumulated risk up to time t.

$$H(t) = \int_0^t h(u)du$$

Mean Residual lifetime

The mean residual life function is defined as the expected value of the remaining lifetimes after a fixed time point t

$$m(t) = E(T - t | T > t) = \frac{\int_t^\infty S(u)du}{S(t)}$$

which exists for all t if and only if $m(0) = E(T)$ is finite.

Examples

Weibull

- $S(t) = e^{-\left(\frac{t}{b}\right)^a}$
- $h(t) = \frac{a}{b} \cdot \left(\frac{t}{b}\right)^{a-1}$
- $H(t) = \left(\frac{t}{b}\right)^a$
- $m(t) = \frac{\int_t^\infty e^{-\left(\frac{u}{b}\right)^a} du}{S(t)}$

Exercise: Compute the $m(t)$ for the exponential distribution (Recall that Weibull$(1, b) \equiv$ Exponential(b)). How can you interpret the result?

Relations between the functions

If we know any one of the functions $f(t)$, $F(t)$, $S(t)$, $H(t)$, $h(t)$ or $m(t)$, we can derive the other functions.
For example,

- $h(t) = -\frac{d}{dt} log S(t)$
- $S(t) = e^{-H(t)}$
- $m(t) = \int_0^\infty e^{H(t)-H(t+x)} dx$
- ...

Outline

Estimating $S(t)$ – Non-Parametric models
No censoring

The empirical estimate of the survival function $S(t)$ is simply the proportion of individuals with event times greater than t

$$\hat{S}(t) = \frac{\#\text{individuals with } T > t}{\text{total sample size}}$$

Estimating $S(t)$ – Non-Parametric models
With censoring

Kaplan and Meier (1958)[1] proposed a non-parametrically estimation of $S(t)$ in the presence of censoring. The method is based on the ideas of conditional probability.
Let

- $t_{(1)} < t_{(2)} < \ldots < t_{(m)}$ denote the distinct times in which an event was observed,
- d_i the number of events that occurred at time $t_{(i)}$ and
- r_i the size of the risk set at time $t_{(i)}$

The Kaplan-Meier estimate for a survival function, also called product-limit estimate, is given by

$$\hat{S}(t) = \begin{cases} 1 & \text{if } t < t_{(1)} \\ \prod_{j:t_{(j)} \leq t} \left(1 - \frac{d_j}{n_j}\right) & \text{if } t \geq t_{(1)} \end{cases}$$

[1] E. L. Kaplan and P. Meier (1972). Nonparametric Estimation from Incomplete Observations. Journal of the American Statistical Association, Vol. 53, No. 282 (Jun., 1958), pp. 457-. 481.
1984 citations per year – the most cited paper in Statistics and number 11 among all papers of all time.

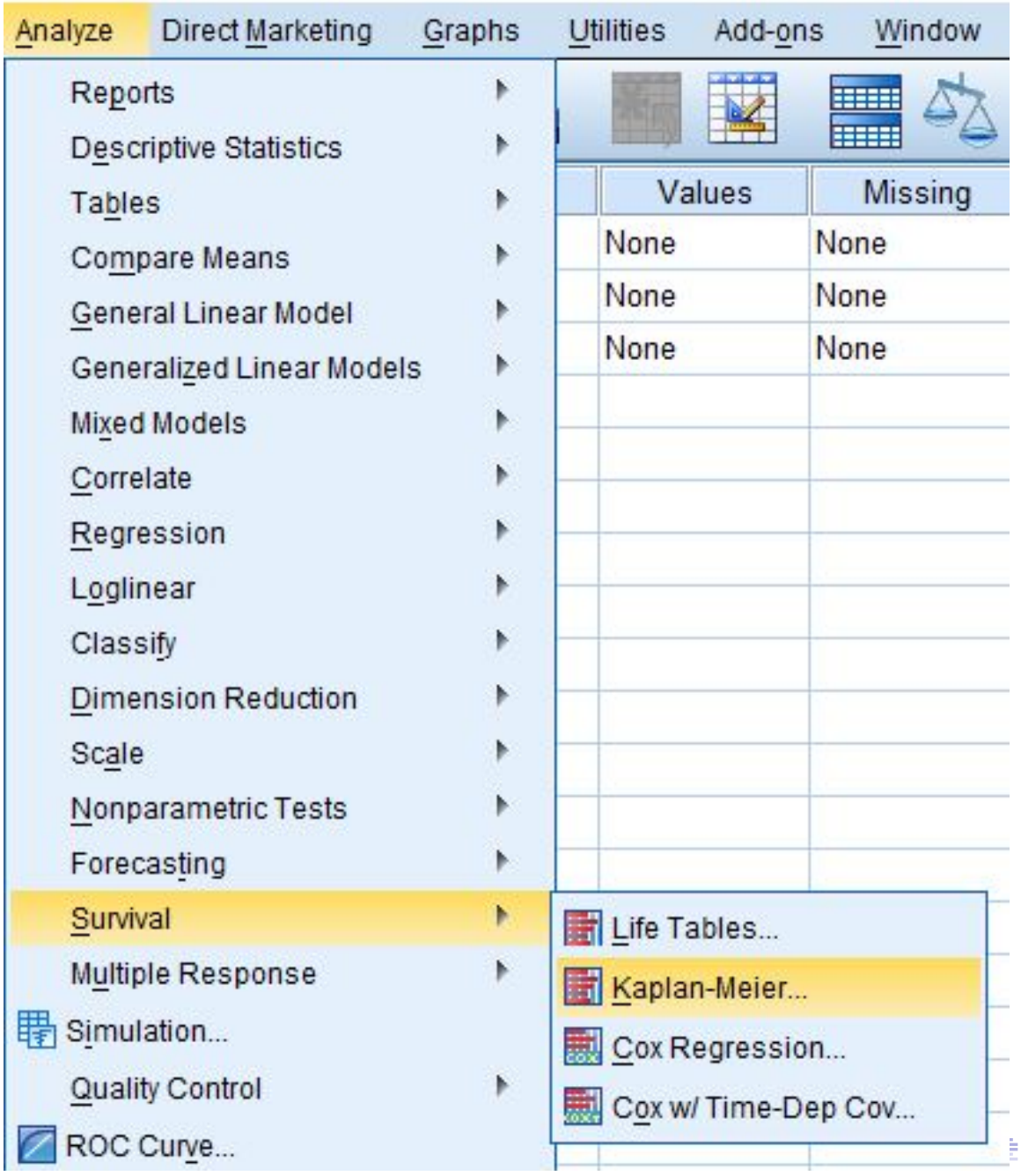
Example – Gastric cancer

The table shows the survival times of two groups of 45 patients suffering from gastric cancer. Group 1 received chemotherapy and radiation. Group 2 just received chemotherapy. An asterisk indicates censoring.

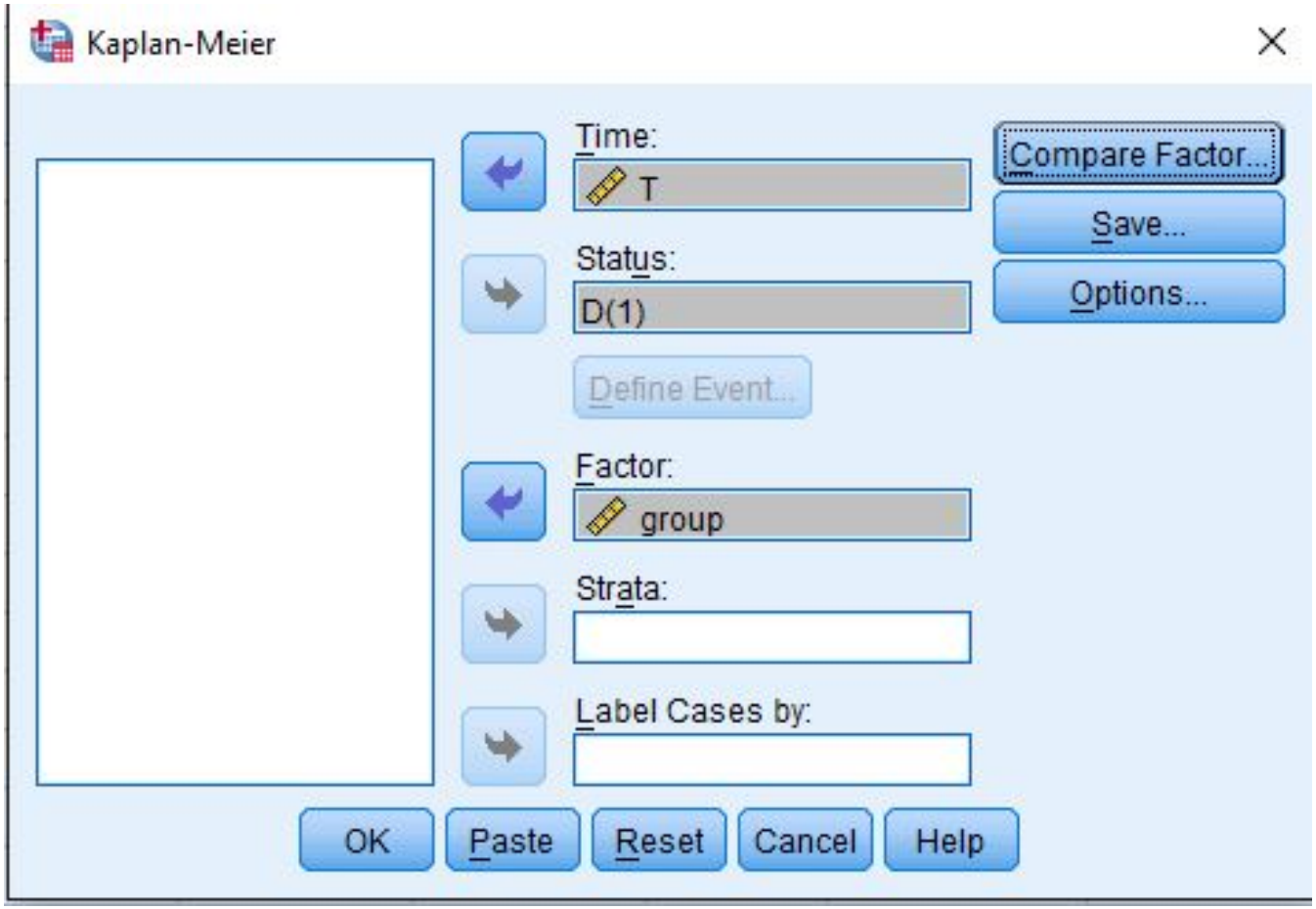
Group 1								
1	63	105	129	182	216	250	262	301
301	342	354	356	358	380	383	383	388
394	408	460	489	499	523	524	535	562
569	675	676	748	778	786	797	955	968
1000	1245	1271	1420	1551	1694	2363	2754*	2950*
Group 2								
17	42	44	48	60	72	74	95	103
108	122	144	167	170	183	185	193	195
197	208	234	235	254	307	315	401	445
464	484	528	542	567	577	580	795	855
1366	1577	2060	2412*	2486*	2796*	2802*	2934*	2988*

Gamerman, D. (1991) Dynamic Bayesian models for survival data. Applied Statistics, 40, 63-79.

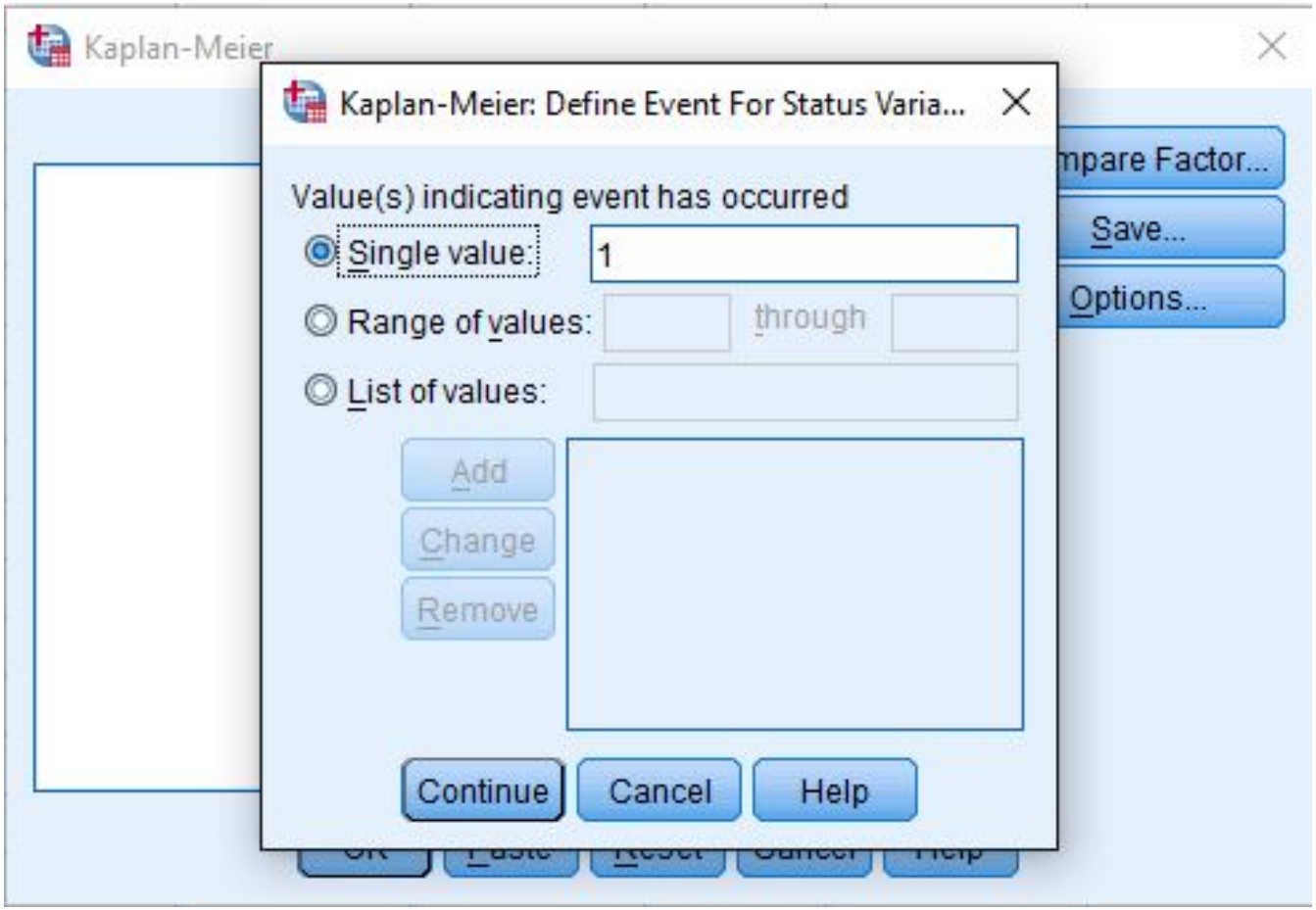
Example - SPSS



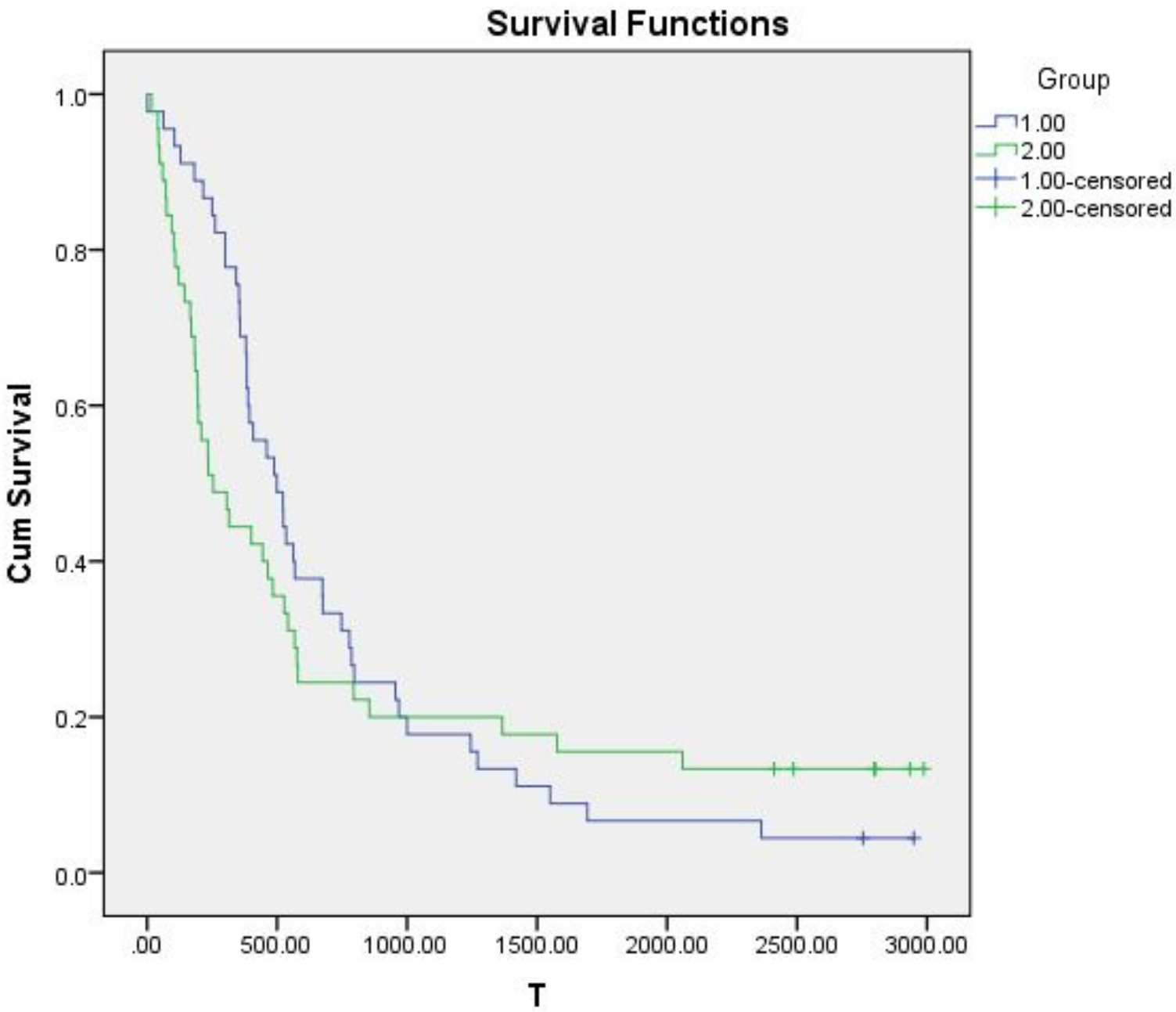
Example - SPSS



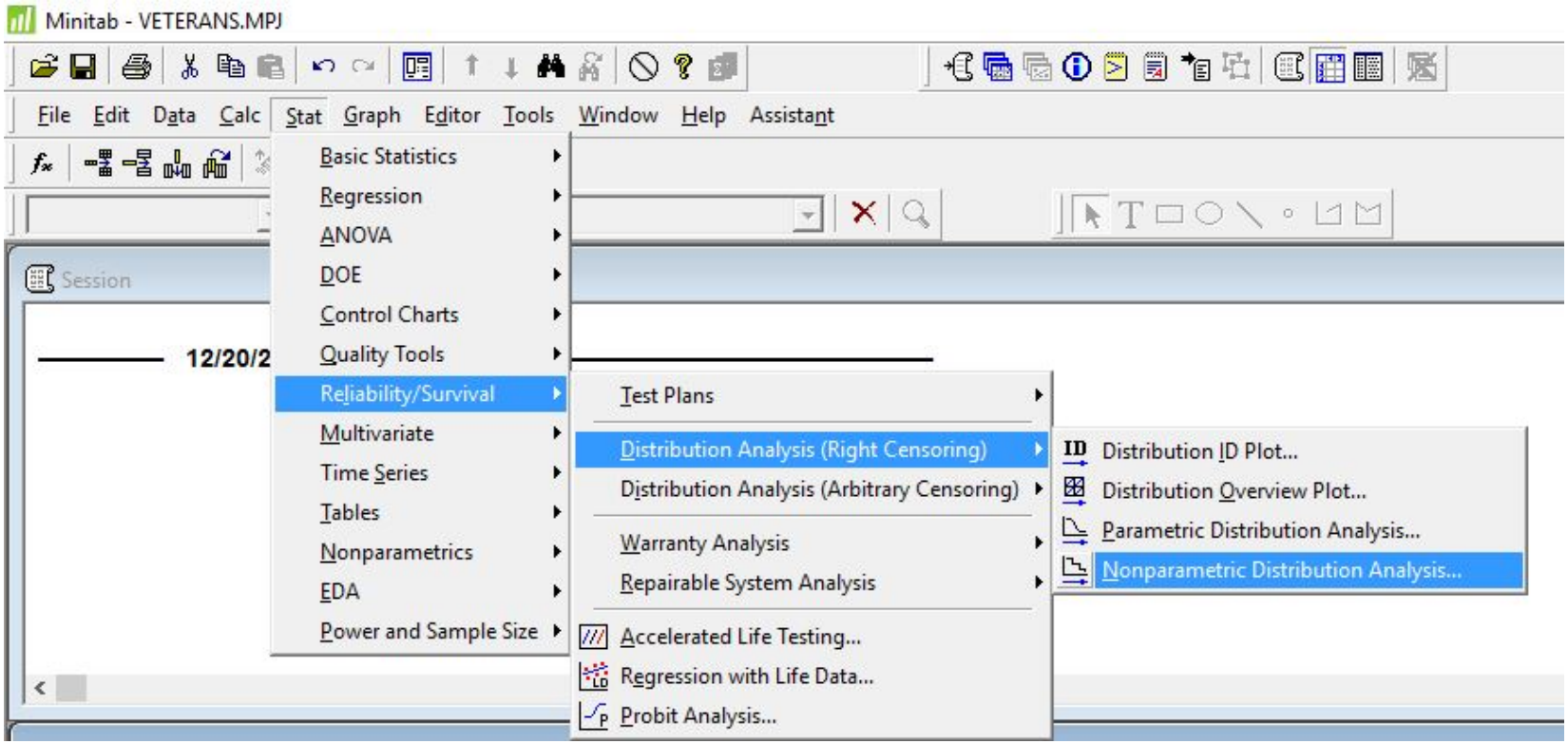
Example - SPSS



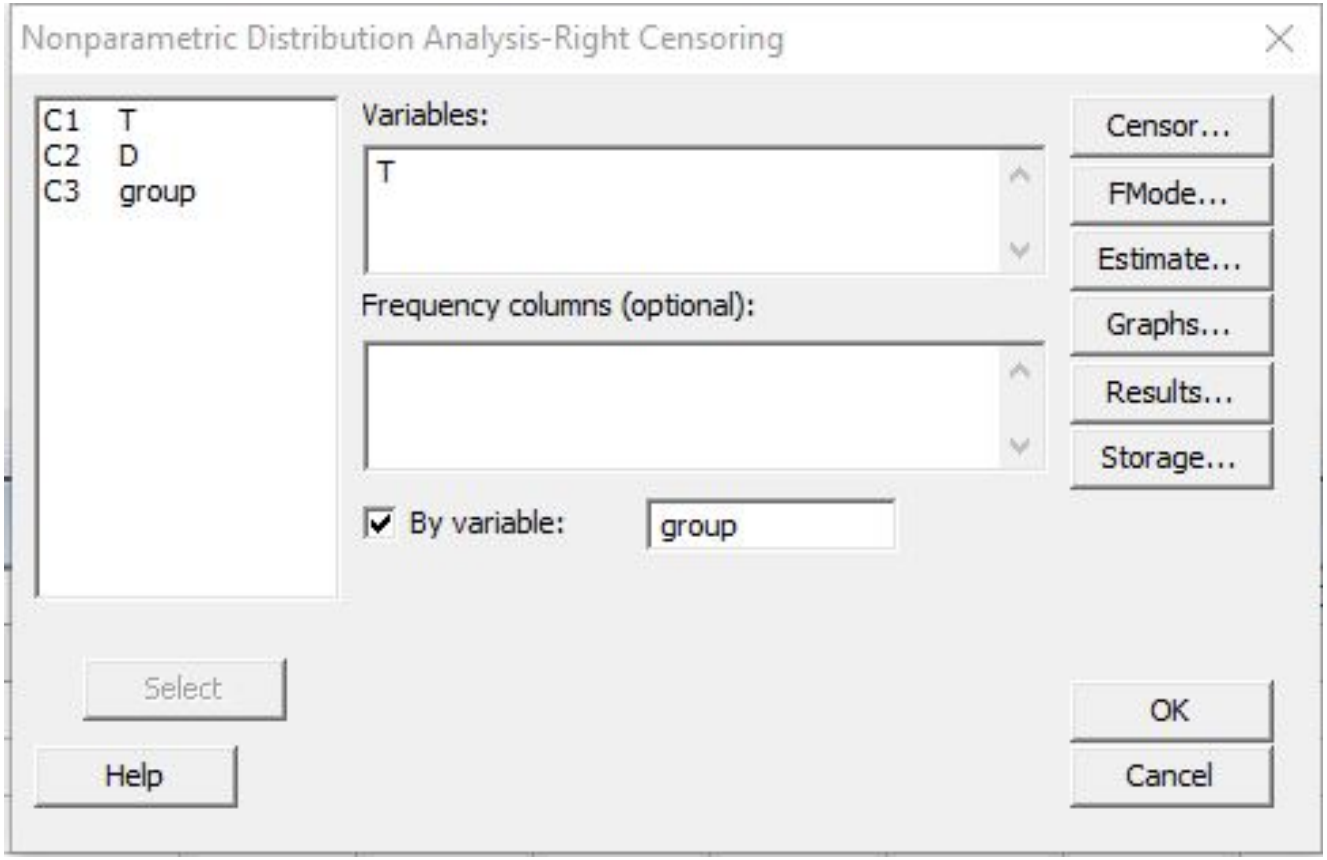
Example - SPSS



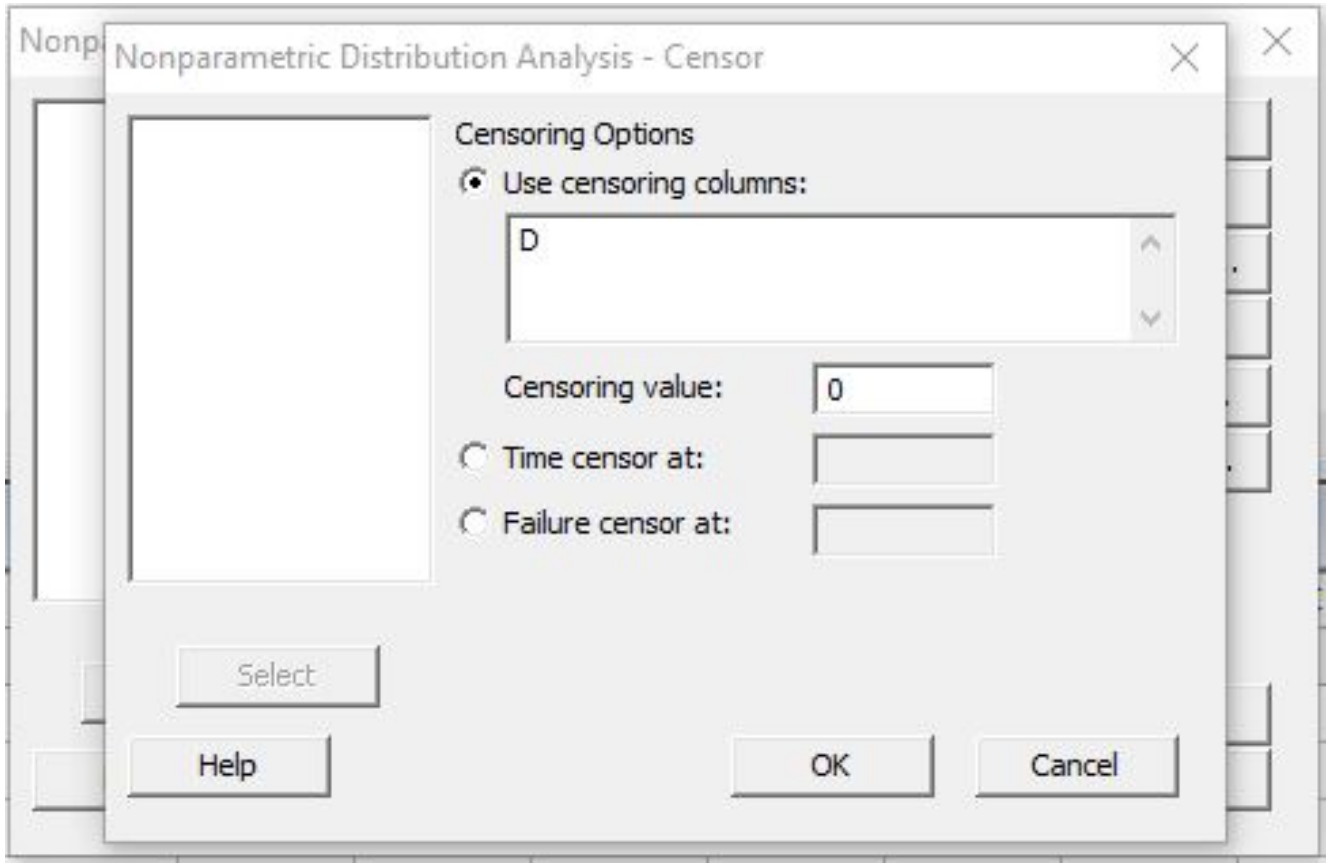
Example - MINITAB



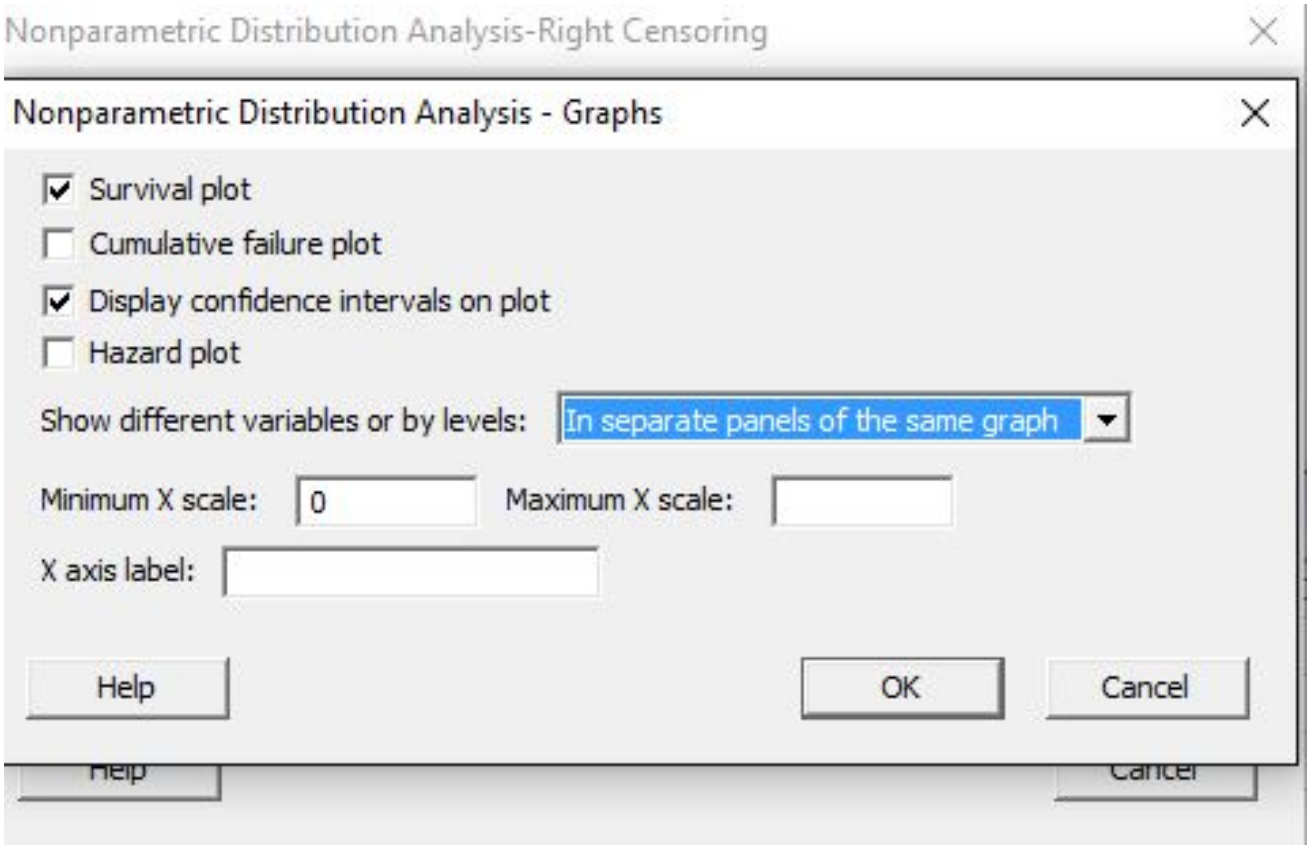
Example - MINITAB



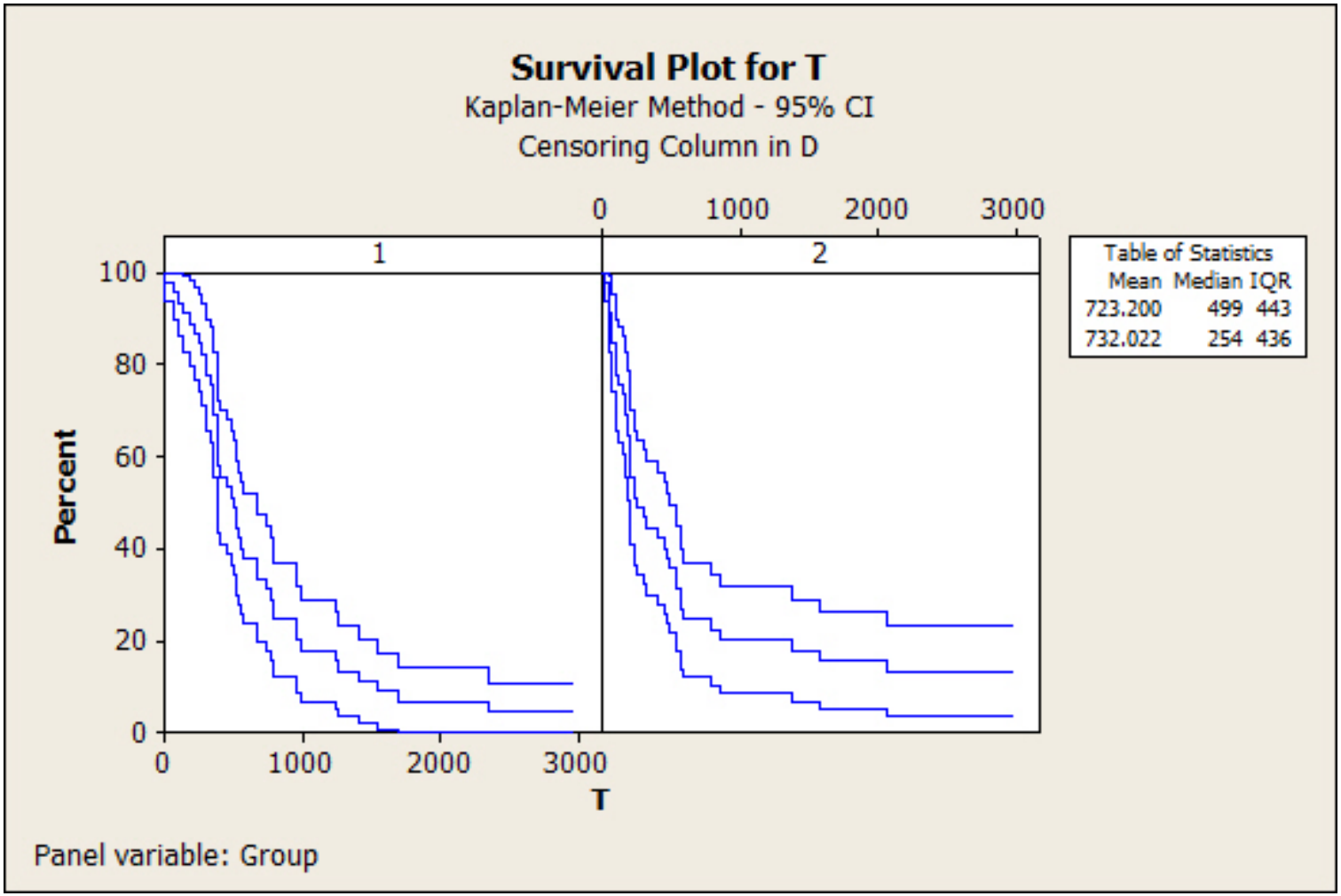
Example - MINITAB



Example - MINITAB



Example - MINITAB



The confidence intervals are calculated using a normal approximation and an estimation of the standard error of $\hat{S}(t)$ (Greenwood's formula).

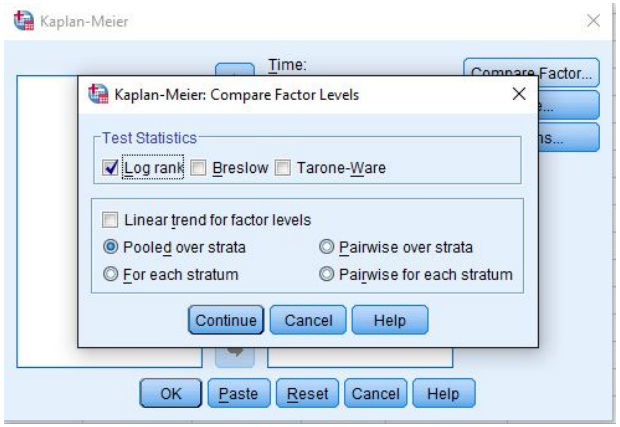
Log-Rank test - Example

In the previous example an interesting task could be to compare the survival times of the two groups. This can be done with the log-rank test.

The **log-rank test** is a hypothesis test to compare the survival distributions of two samples.

$H_0 : S_1(t) = S_2(t)$ (No differences between the two groups)
$H_1 :$ not H_0

SPSS



Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	.225	1	.635

Test of equality of survival distributions for the different levels of Group.

Graphical tests

The survival function for the Weibull distribution is given by

$$S(t) = e^{-\left(\frac{t}{\eta}\right)^{\beta}}$$

from which we have that

$$\log S(t) = -\left(\frac{t}{\eta}\right)^{\beta}$$

and

$$\log(-\log S(t)) = \beta \log t - \beta \log \eta$$

This implies that if $T \sim Weibull(\eta, \beta)$ then $\log(-\log S(t))$ is linearly related to $\log t$. Consequently, a plot of

$$\log(-\log \hat{S}(t)) \quad vs \quad \log t$$

should be a straight line, where $\hat{S}(t)$ is a non-parametric estimator of the survivor function, such as the Kaplan–Meier.
Similar plots can be constructed and for other models (see next slide).

Graphical tests

distribution	plot		
Exponential	$-\log \hat{S}(t)$	vs	t
Weibull	$\log(-\log \hat{S}(t))$	vs	$\log t$
Gumbel	$\log(-\log \hat{S}(t))$	vs	t
lognormal	$\Phi^{(-1)}(1-\hat{S}(t))$	vs	$\log t$
Gamma	$\Phi^{(-1)}(1-\hat{S}(t))$	vs	$\sqrt{t}$
Pareto	$-\log \hat{S}(t)$	vs	$\log t$
loglogistic	$-\log\left(\frac{1-\hat{S}(t)}{\hat{S}(t)}\right)$	vs	$\log t$

Probability plot

The probability plot is another graphical technique for assessing whether or not a given data set follows a specific distribution.

The data are plotted against a theoretical distribution in such a way that the points should form approximately a straight line.

Departures from this straight line indicate departures from the specified distribution.

Example – Veterans' Administration Lung Cancer study – MINITAB

Description

Randomized trial of two treatment regimens for lung cancer.

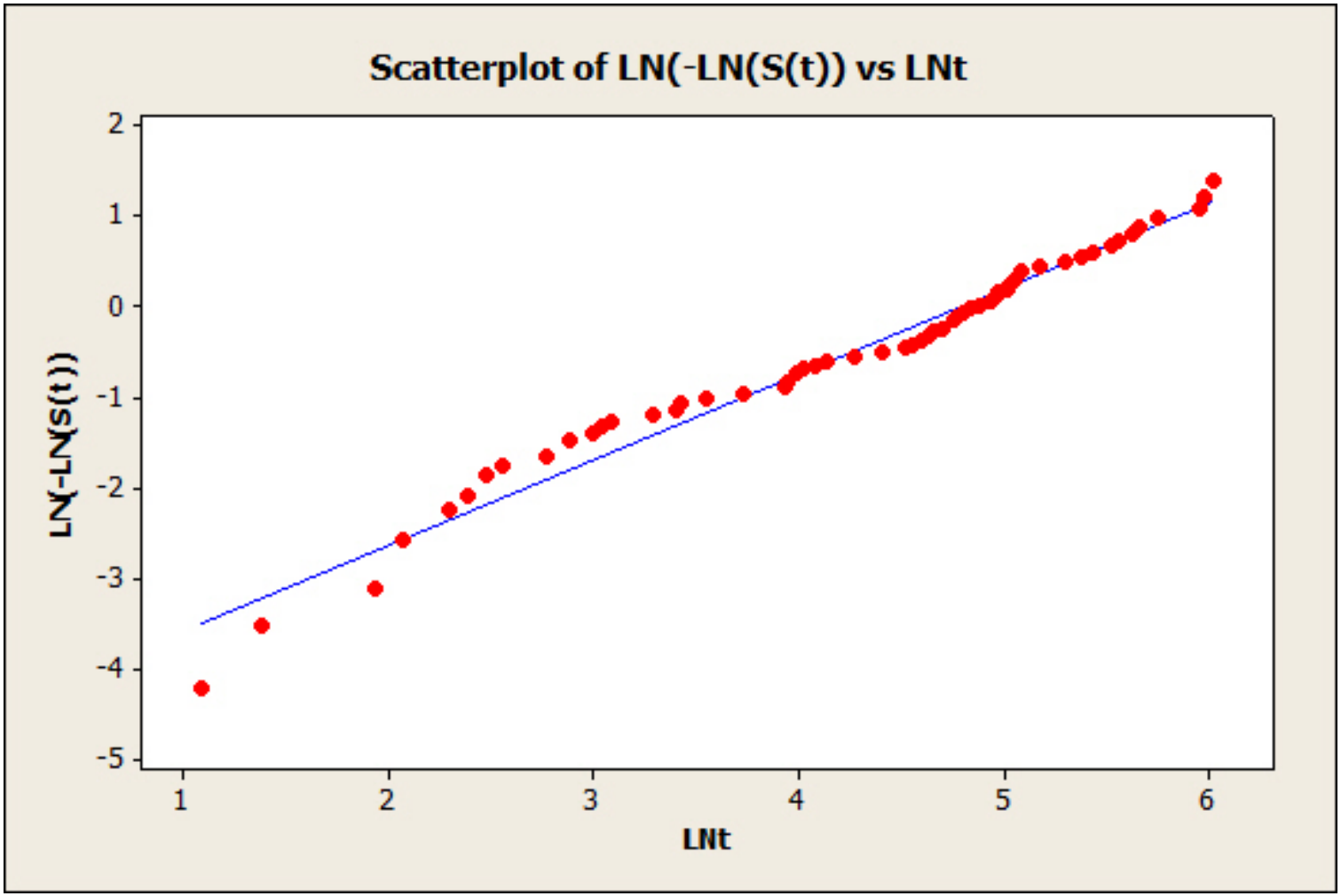
- trt: 1=standard 2=test
- celltype: 1=squamous, 2=smallcell, 3=adeno, 4=large
- time: survival time
- status: censoring status
- karno: Karnofsky performance score (100=good)
- diagtime: months from diagnosis to randomisation
- age: in years
- prior: prior therapy 0=no, 1=yes

D. Kalbfleisch and R.L. Prentice (1980), The Statistical Analysis of Failure Time Data. Wiley, New York.

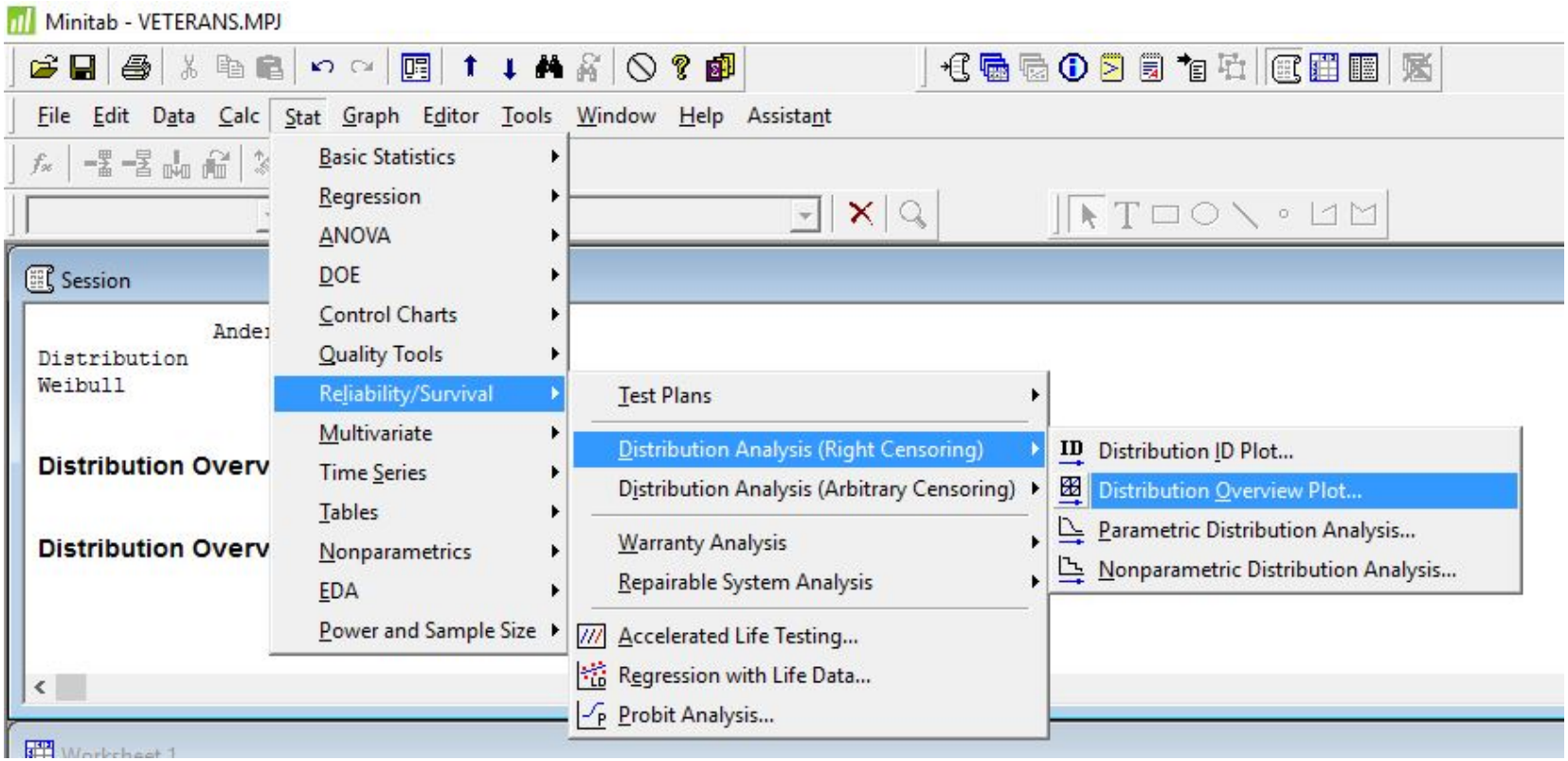
Example – Veterans' Administration Lung Cancer study – MINITAB

Treatment 1=standard

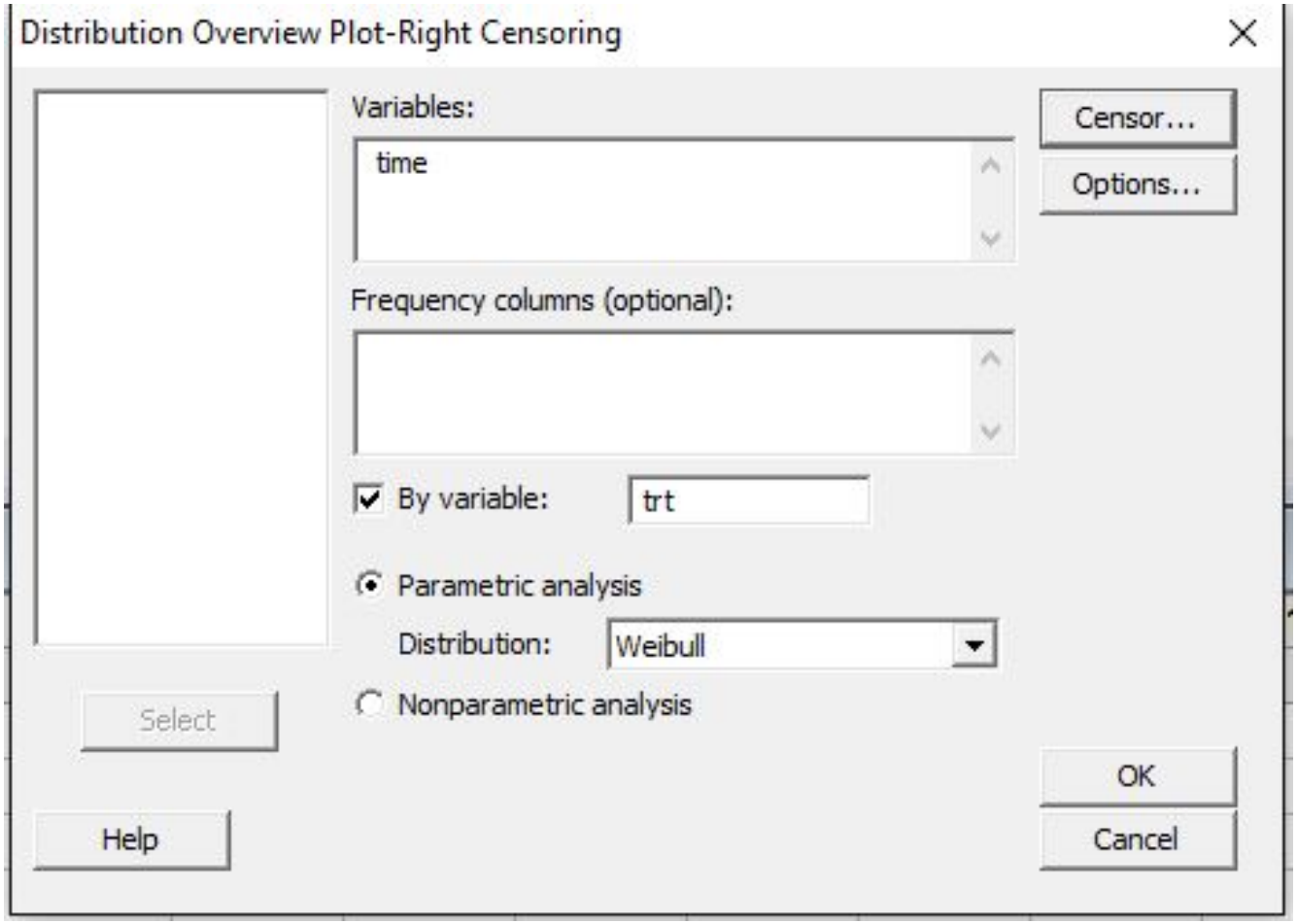
The plot $\log(-\log \hat{S}(t))$ *vs* $\log t$ (Weibull distribution)



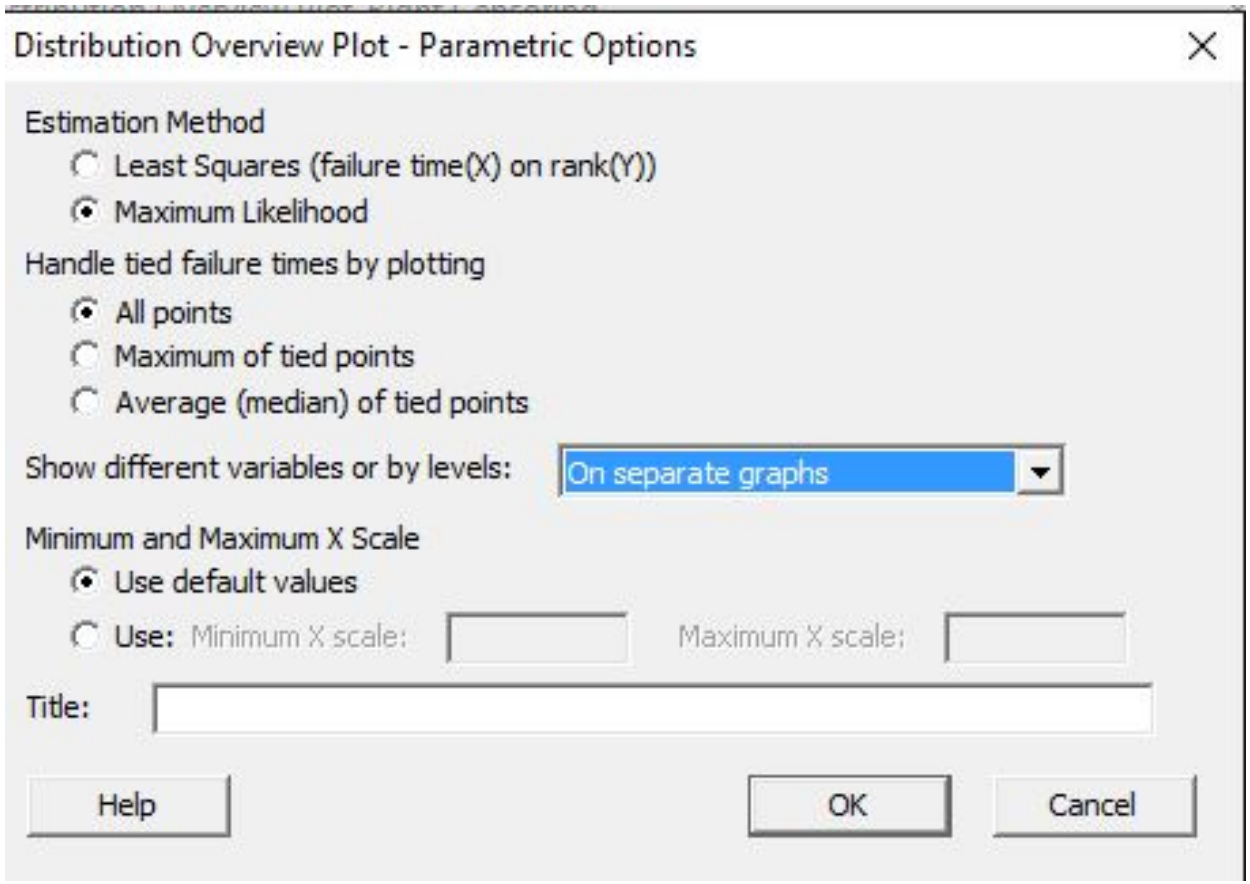
Example – Veterans' Administration Lung Cancer study – MINITAB



Example – Veterans' Administration Lung Cancer study – MINITAB



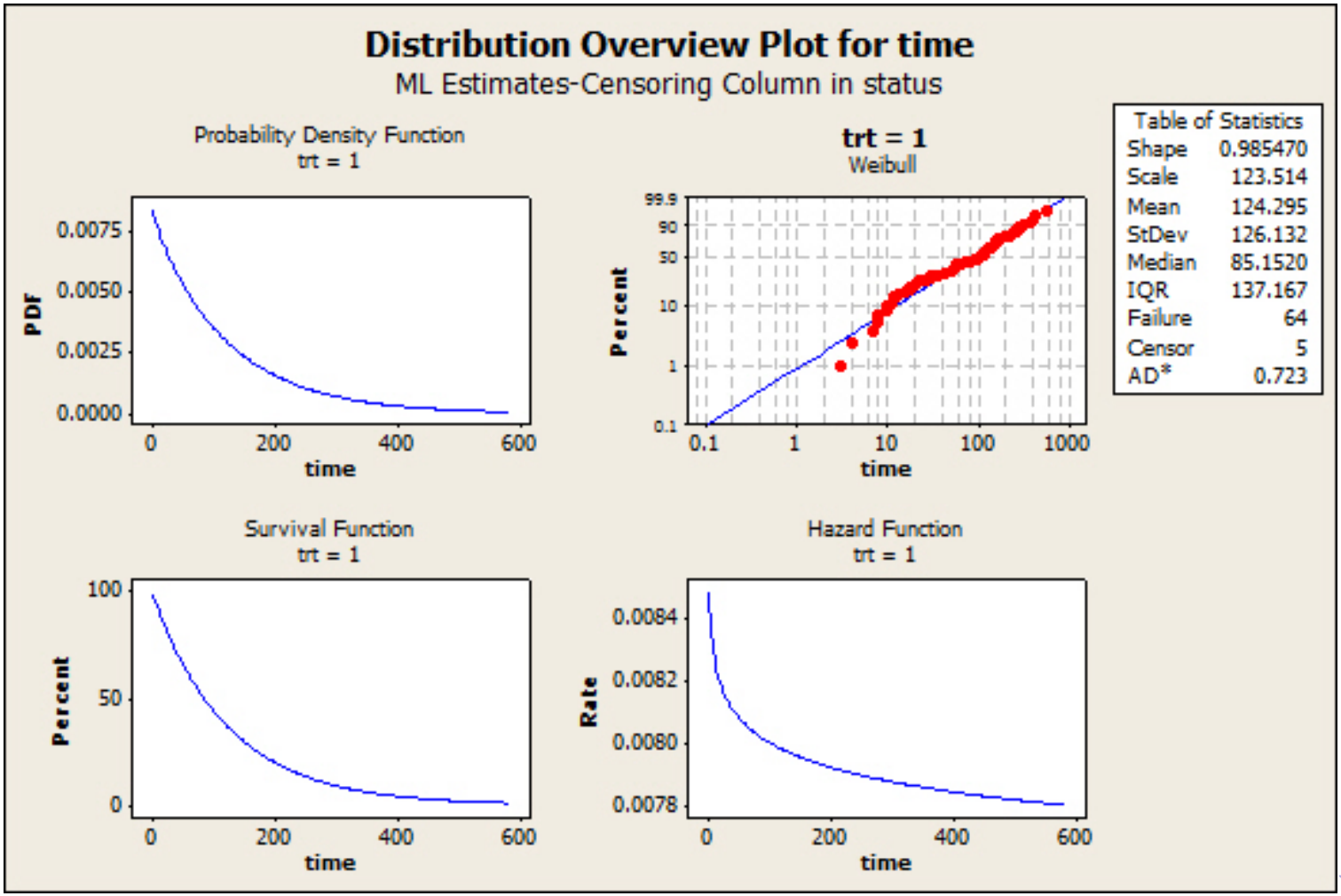
Example – Veterans' Administration Lung Cancer study – MINITAB



Example – Veterans' Administration Lung Cancer study – MINITAB

Treatment 1=standard

Probability plot (upper right) and other related graphs for the Weibull distribution for the Veterans' Administration Lung Cancer data (Treatment 1=standard).



Estimation

In the previous graph the estimators of the parameters of the Weibull distribution are given among others statistics.
The methods used in order to estimate the parameters are

- the method of moments
- the maximum-likelihood estimation
- least squares estimation

Here we will only present (briefly) the maximum-likelihood estimation.

Flashback – Maximum likelihood estimators No censoring

The idea behind the Maximum likelihood estimators is to find the value of the parameters θ which maximizes the likelihood of getting the data that we, in fact, observed.

Suppose we have a random sample $X = (X_1, X_2, \ldots, X_n)$ from a random variable with pdf $f(x;\theta)$. Then, the joint probability mass (or density) function of $X_1, X_2, \ldots, X_n$ denoted by $L(\theta)$ is given by

$$\begin{aligned} L(\theta) &= f(x_1;\theta) \cdot f(x_2;\theta) \cdots f(x_n;\theta) \\ &= \prod_{i=1}^{n} f(x_i;\theta) \end{aligned}$$

The MLEs of θ are obtained by maximizing $L(\theta)$ or equivalently the log-likelihood function

$$\ell(\theta) = \sum_{i=1}^{n} \log f(x_i;\theta)$$

with respect to θ.

Maximum-likelihood estimation
With right censored observations

Suppose then that we have n individuals with lifetimes distributed accordingly to a random variable T with Survival function $S(t)$ (pdf $f(t)$ and hazard $h(t)$).
$\hookrightarrow$ Suppose that individual i is observed for a time t_i. If the individual died at t_i, its contribution to the likelihood function is the density at that duration

$$L_i = f(t_i)$$

.
$\hookrightarrow$ On the other hand if the individual is still alive at t_i (censored observation) all we know is that its lifetime exceeds t_i. The probability of this event is $S(t_i)$ and so its contribution to the likelihood is

$$L_i = S(t_i).$$

So, the **likelihood function** is given by

$$L(\theta) = \prod_{i=1}^{n} (f(x_i;\theta))^{\delta_i} (S(x_i;\theta))^{1-\delta_i}$$

where δ_i is the censored indicator, taking the value zero if the i^{th} observation is censored and the value one if not.

Example – Exponential

Let $T \sim Exp(\lambda)$ then $f(t;\lambda) = \lambda e^{-\lambda t}$ and $S(t;\lambda) = e^{-\lambda t}$.

likelihood function $L(\lambda) = \prod_{i=1}^{n} \left(\lambda e^{-\lambda t_i}\right)^{\delta_i} \left(e^{-\lambda - t_i}\right)^{1-\delta_i}$

log-likelihood function

$$\ell(\lambda) = \sum_{i=1}^{n} \Big(\delta_i(\log\lambda - \lambda t_i) - (1-\delta_i)\lambda t_i\Big) = \sum_{i=1}^{n} (\delta_i \log\lambda - \lambda t_i)$$

$$= \left(\sum_{i=1}^{n} \delta_i\right) \log\lambda - \lambda \sum_{i=1}^{n} t_i$$

$$= k\log\lambda - \lambda \sum_{i=1}^{n} t_i$$

where k is the number of non censored observations.

MLE

$$\frac{\partial \ell(\lambda)}{\partial \lambda} = \frac{k}{\lambda} - \sum_{i=1}^{n} t_i = 0 \Rightarrow \hat{\lambda} = k / \sum_{i=1}^{n} t_i$$

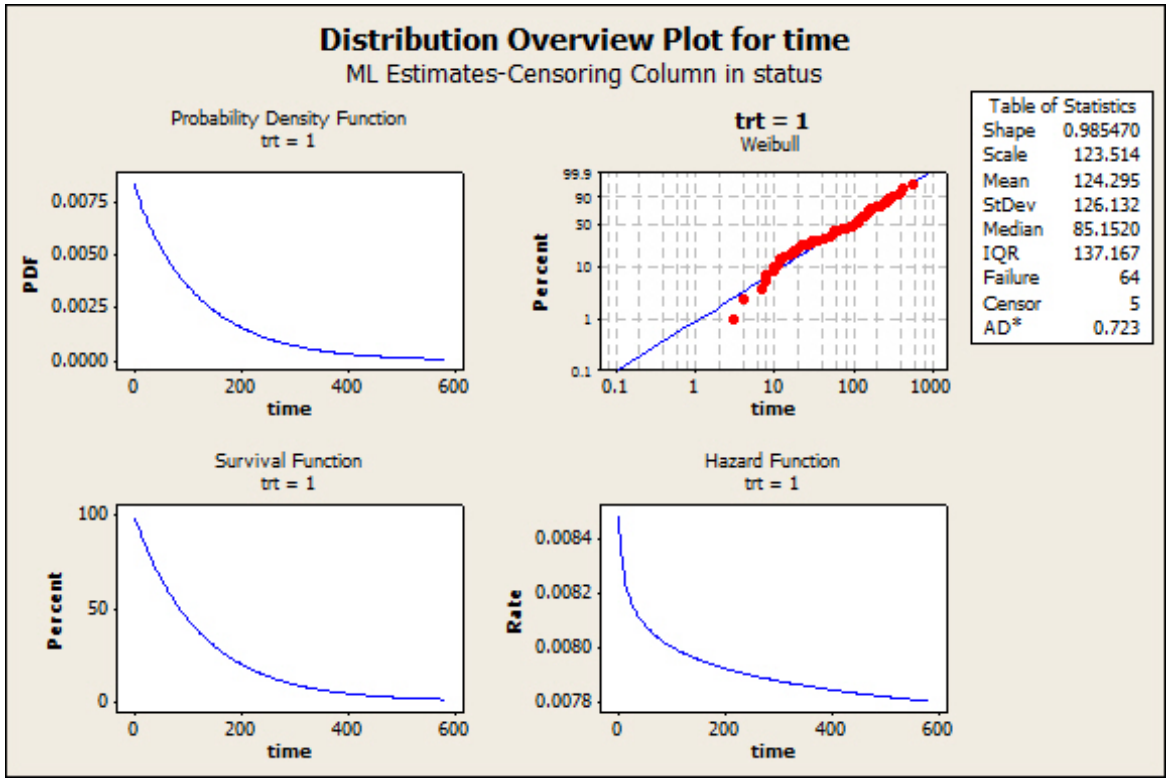
Maximum-likelihood estimation
With right censored observations

In most of the cases the MLEs has no closed-form solution and we must use numerical methods in order to maximize the log-likelihood.

Example – Veterans' Administration Lung Cancer study – MINITAB

Treatment 1=standard

The MLEs of η and β are $\hat{\eta} = 123.514$ and $\hat{\beta} = 0.985470$ respectively.



Recall that the survival function of Weibull is given by $S(t) = e^{-\left(\frac{t}{\eta}\right)^{\beta}}$.

Goodness of fit tests – Anderson-Darling statistic

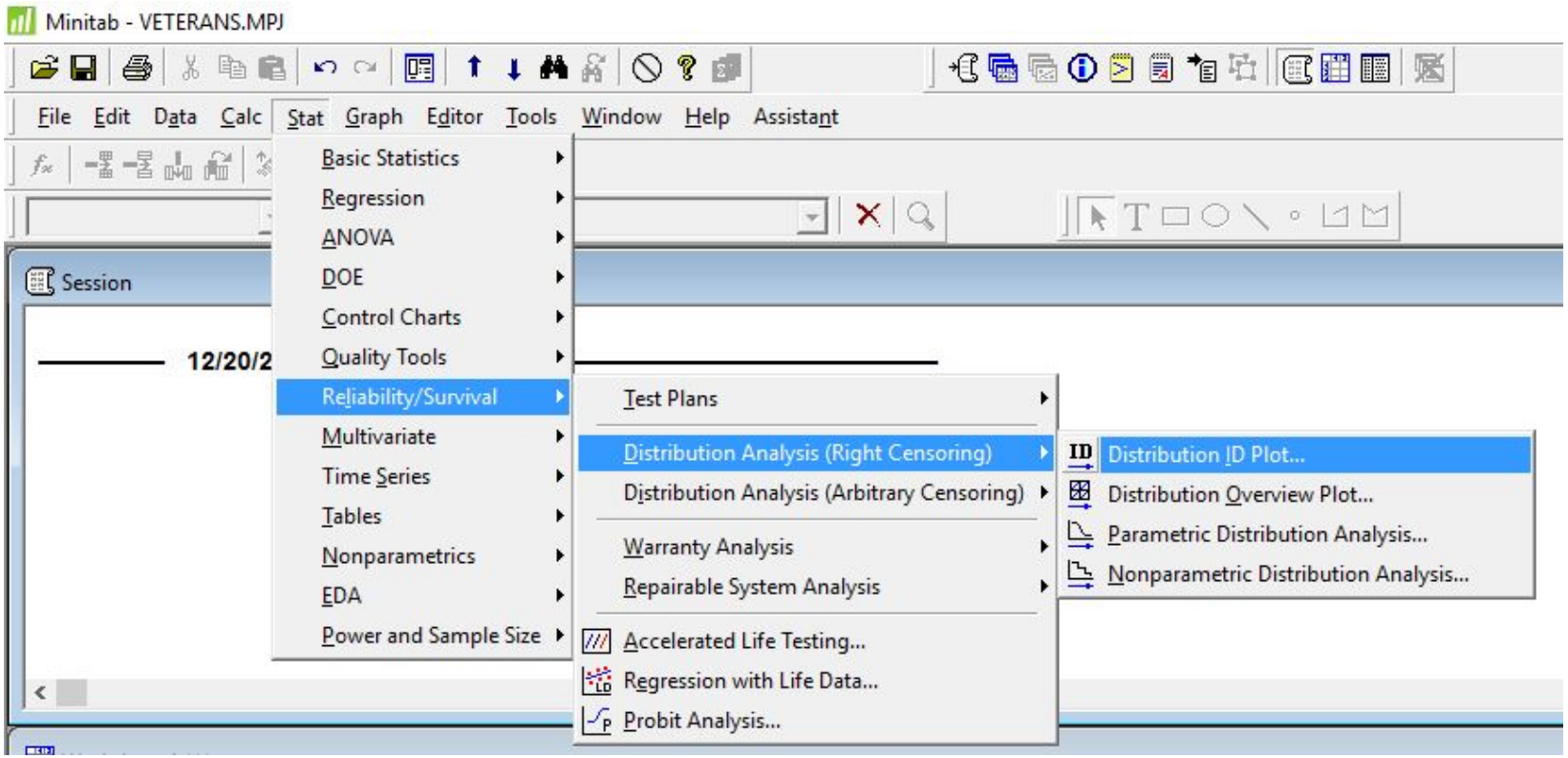
The Anderson-Darling statistic measures how well the data follow a particular distribution. For a specified data set and distribution, the better the distribution fits the data, the smaller this statistic will be.

The hypotheses for the Anderson-Darling test are:
H_0 : The data follow a specified distribution
H_1 : The data do not follow a specified distribution

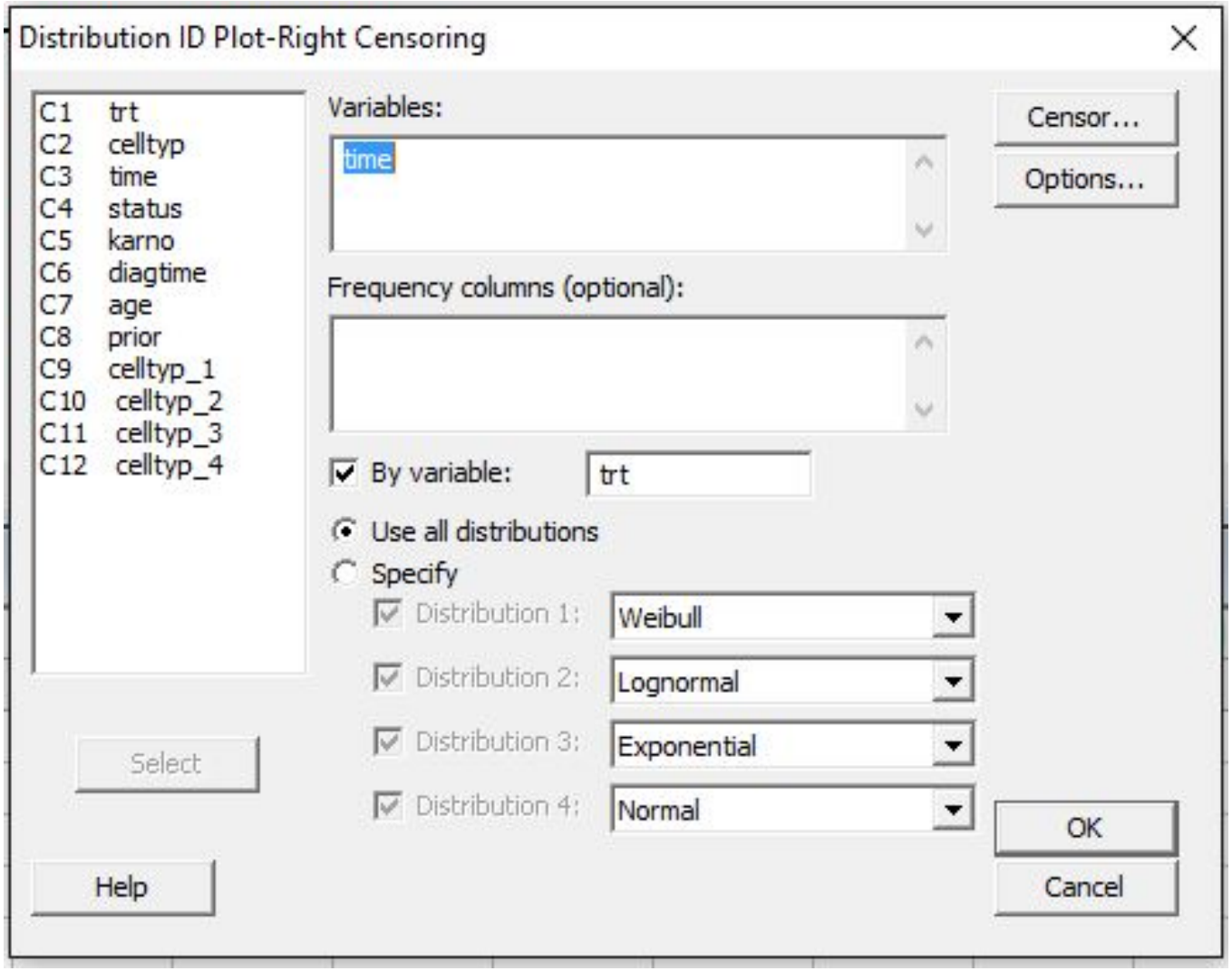
Unfortunately, the p-value for the Anderson-Darling statistic can not be calculated for every distribution when there are censored observations.

Since lower values of Anderson-Darling identify a better fitting distribution, we can evaluate and compare the Anderson-Darling statistic for many distributions and choose the one with the smaller value.

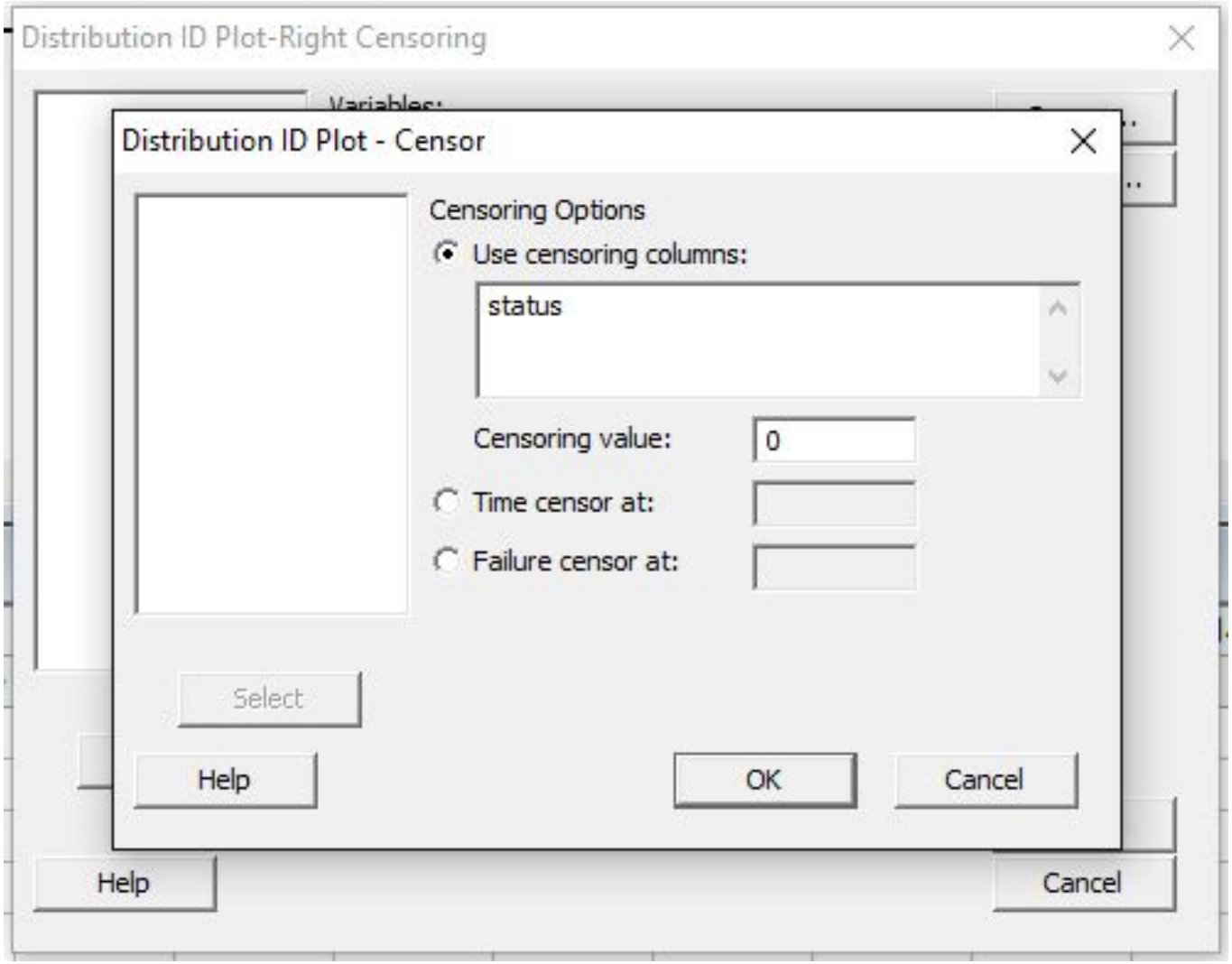
Example – Veterans' Administration Lung Cancer study – MINITAB



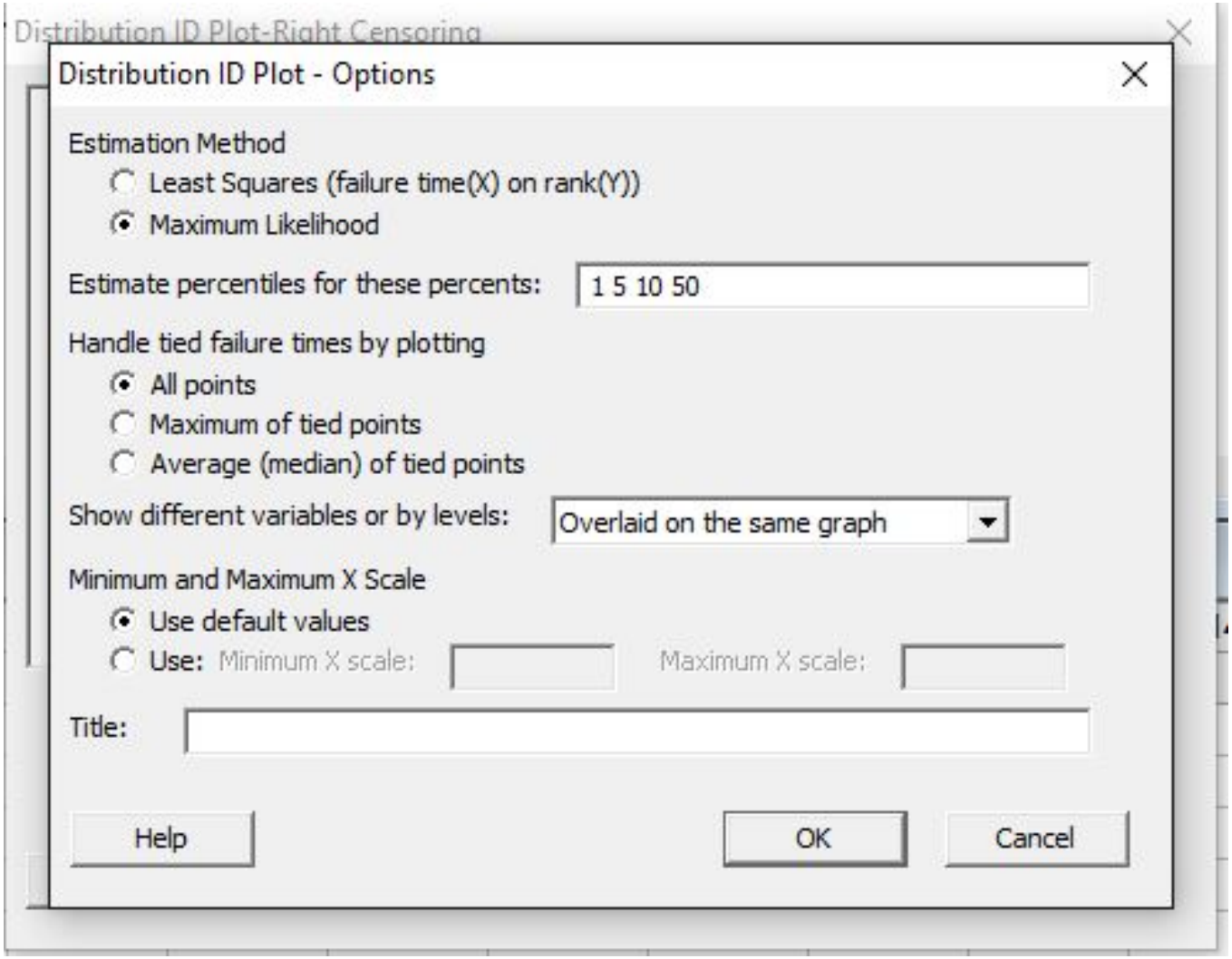
Example – Veterans' Administration Lung Cancer study – MINITAB



Example – Veterans' Administration Lung Cancer study – MINITAB



Example – Veterans' Administration Lung Cancer study – MINITAB



Example – Veterans' Administration Lung Cancer study – MINITAB

Treatment 1=standard

Goodness-of-Fit

Distribution	Anderson-Darling (adj)
Weibull	0.723
Lognormal	1.334
Exponential	0.751
Loglogistic	1.209

Outline

Regression models

Up to this point we have only consider homogeneous populations (with an exception, when we compared the survival function of two groups with the long-rank test), where all the individuals are governed by the same survival function $S(t)$.

In this last section of the lecture we will allow describe heterogeneous populations with the help of covariates (or explanatory variables) that may affect the survival time.

There are several regression models in order to incorporate the covariates in a model, but the most frequently used are the

- accelerate life regression models
- proportional hazards regression model[2]

An important subclass of the latter model is the so called Cox model.

[2] D.R. Cox (1972). Regression Models and Life-Tables. Journal of the Royal Statistical Society, Series B 34 (2), pp 187–220.
1342 citations per year – the second most cited paper in Statistics and number 24 among all papers of all time.

Accelerate life regression models

Let $T_{\mathbf{x}}$ be a random variable representing the survival time of an individual with covariates $\mathbf{x}$.

Since $T_{\mathbf{x}}$ must be non-negative, we might consider modeling its logarithm using a conventional linear model

$$\log T_{\mathbf{x}} = \mathbf{x}'\beta + \epsilon$$

where ϵ is a suitable error term, with a distribution to be specified.

The previous model can be written as

$$T_{\mathbf{x}} = e^{\mathbf{x}'\beta} T_0$$

where T_0 stands for the exponentiated error term ϵ.

Accelerate life regression models

The survival function $S(t;\mathbf{x})$ of $T_{\mathbf{x}}$ can be expressed as

$$\begin{aligned} S(t;\mathbf{x}) &= P(T_{\mathbf{x}} > t) = P(e^{\mathbf{x}'\beta} T_0 > t) \\ &= P(T_0 > te^{-\mathbf{x}'\beta}) \\ &= S_0(te^{-\mathbf{x}'\beta}) \end{aligned}$$

where $S_0(t)$ is the survival function of T_0, which will serve as a reference group.

Interpretation – Example
Consider a model with a constant and a dummy variable x representing a factor with two levels taking the value 1, if an individual belongs to group one, and the value zero, if he belongs to group zero.

Let $\beta = \log(2)$ so that the corresponding multiplicative effect $e^{\beta x}$ be 2.

Then the survival function $S_1(t)$ of group one is given by

$$S_1(t) = S_0(t/2)$$

which implies that the probability that a member of group one will be alive at age t is exactly the same as the probability that a member of group zero will be alive at age $t/2$

Accelerate life regression models - Example

Let $T_0 \sim Weibull(\eta, \beta)$ then

$$S(t; \mathbf{x}) = S_0(te^{-\mathbf{x}'\boldsymbol{\beta}}) = e^{-\left(\frac{te^{-\mathbf{x}'\boldsymbol{\beta}}}{\eta}\right)^{\beta}} = e^{-\left(\frac{t}{\eta e^{\mathbf{x}'\boldsymbol{\beta}}}\right)^{\beta}}$$

from which we have that $T_{\mathbf{x}}$ follows again a Weibull distribution with the same shape parameter β but with different scale parameter

$$T_{\mathbf{x}} \sim Weibull(\eta e^{\mathbf{x}'\boldsymbol{\beta}}, \beta)$$

Accelerate life regression models - Model fitting

The parameters of the model, i.e.

- the regression coefficients β and
- the parameters of the distribution of T_0

are estimated using the Maximum Likelihood Estimation procedure. (see following slides for comments and the use of statistical programs).

Proportional hazards regression model

Cox (1972) introduced a family of regression models based on the hazard function.

The simplest member of the family is the proportional hazards model, where the hazard at time t for an individual with covariates $\mathbf{x}$ (not including a constant) is assumed to be

$$h(t;\mathbf{x}) = e^{\mathbf{x}'\beta} h_0(t)$$

where $h_0(t)$ is a baseline hazard function that describes the risk for individuals with $\mathbf{x} = 0$, who serve as a reference group, and $e^{\mathbf{x}'\beta}$ is the relative risk, a proportionate increase or decrease of risk, associated with the set of covariates $\mathbf{x}$.
Note that the proportionate increase or decrease of risk is the same at all t.

Proportional hazards regression model

From the definition of the model

$$h(t;\mathbf{x}) = e^{\mathbf{x}'\beta} h_0(t)$$

we obtain

$$S(t;\mathbf{x}) = (S_0(t))^{e^{\mathbf{x}'\beta}}$$

Interpretation – Example

In the two level factor example with a relative risk of $e^{\mathbf{x}'\beta} = 2$, the probability that a member of group one will be alive at any given t is the square of the probability that a member of group zero would be alive at the same time.

Proportional hazards regression model – Example

Show that the Weibull distribution is closed under the Proportional hazards regression model.

In reality the Accelerate lifetime and the Proportional hazards regression models are equivalent for the Weibull distribution (Cox and Oakes (1984)).

Proportional hazards regression model – Model fitting

There are two approaches to fitting the Proportional hazards regression model:

1. the parametric approach
   - a specific functional form for the baseline hazard $h_0(t)$ is assumed and
     - the regression coefficients β and
     - the parameters of $h_0(t)$

     are estimated using the Maximum Likelihood Estimation procedure.
2. the semi-parametric approach
   - We focus only on the estimation of the regression coefficients leaving the baseline hazard $h_0(t)$ completely unspecified.

   This approach relies on a partial likelihood function proposed by Cox (1972) in his original paper and it is usually refereed as the Cox model.

Example – Veterans' Administration Lung Cancer study

Description

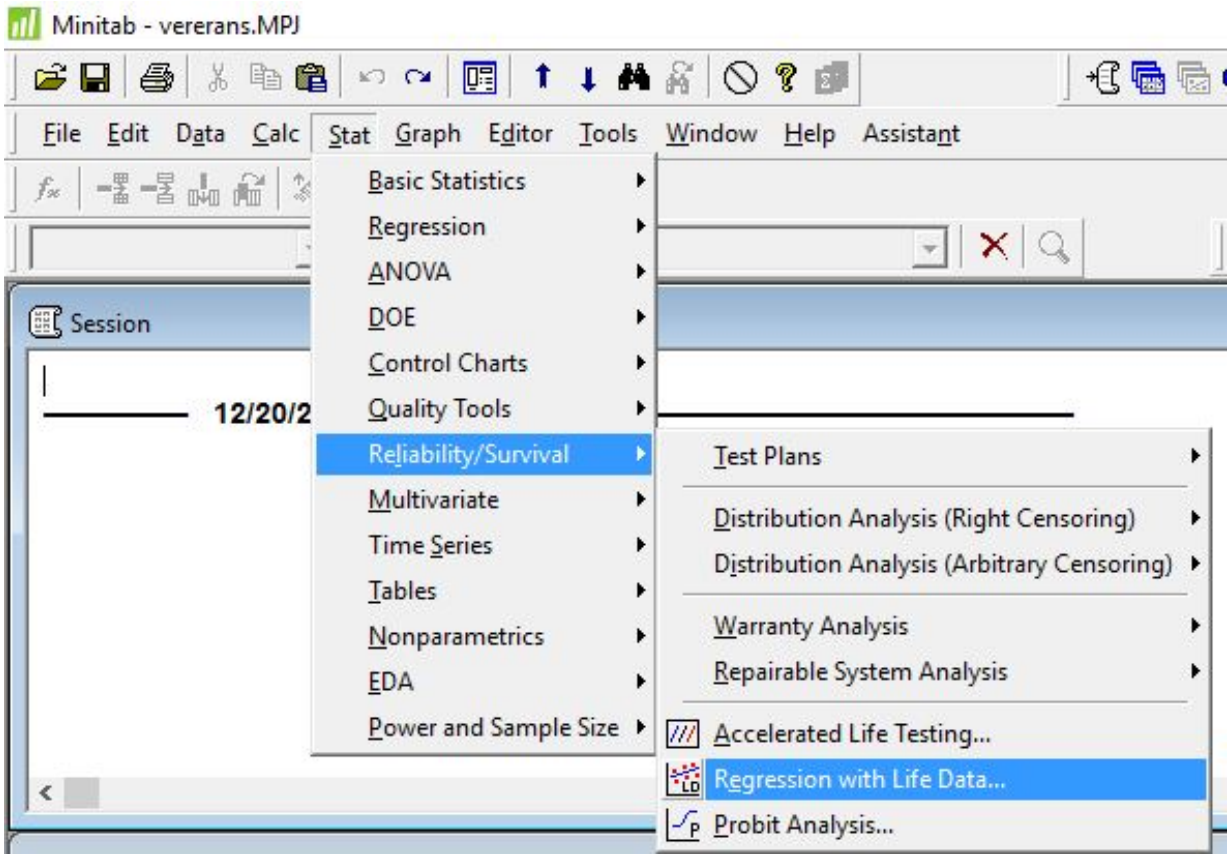
Randomized trial of two treatment regimens for lung cancer.

- trt: 1=standard 2=test
- celltype: 1=squamous, 2=smallcell, 3=adeno, 4=large
- time: survival time
- status: censoring status
- karno: Karnofsky performance score (100=good)
- diagtime: months from diagnosis to randomisation
- age: in years
- prior: prior therapy 0=no, 1=yes

D. Kalbfleisch and R.L. Prentice (1980), The Statistical Analysis of Failure Time Data. Wiley, New York.

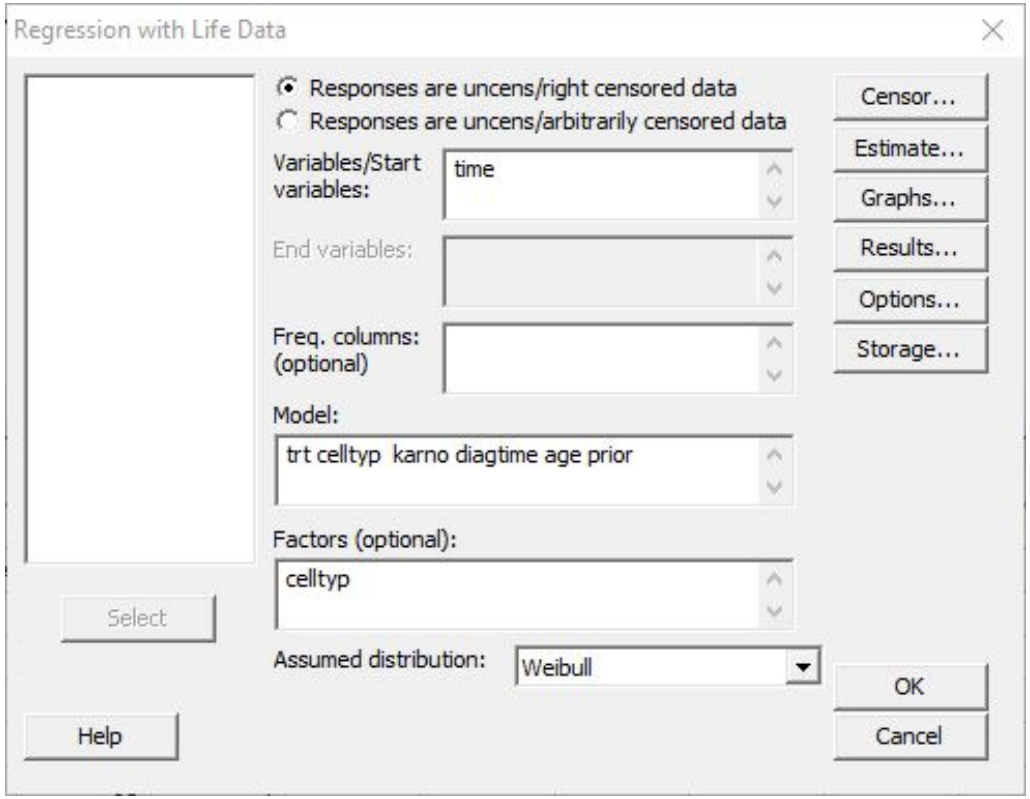
Example – Veterans' Administration Lung Cancer study – MINITAB

Proportional hazards regression model



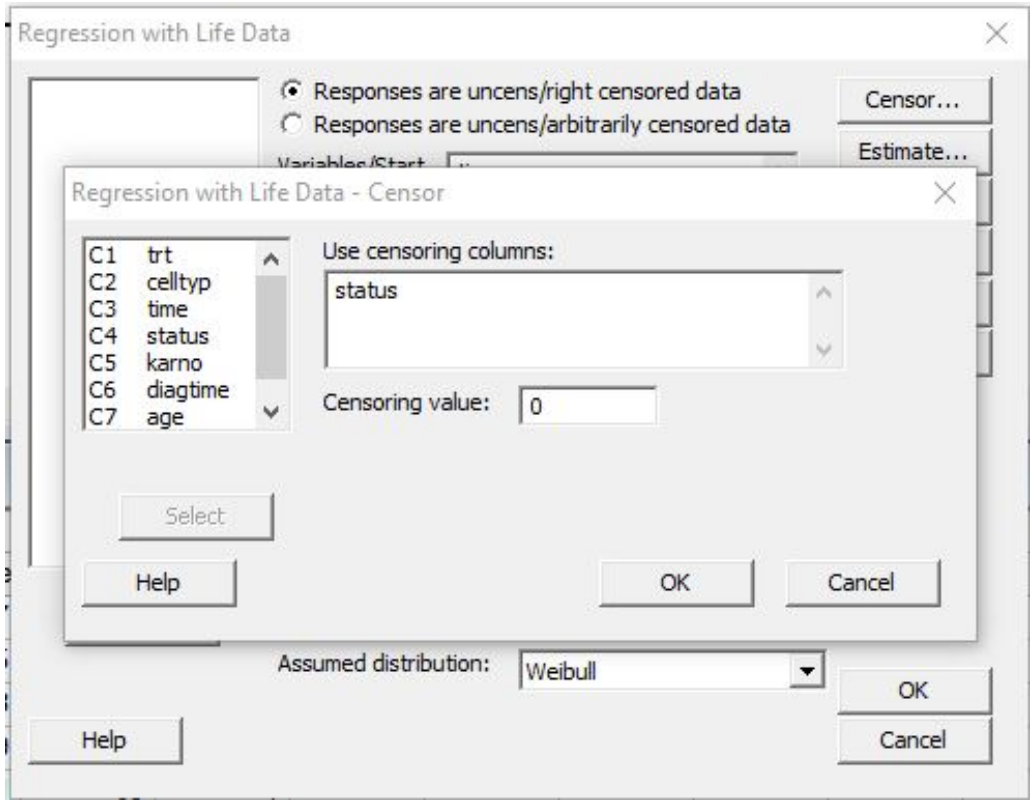
Example – Veterans' Administration Lung Cancer study – MINITAB

Proportional hazards regression model



Example – Veterans' Administration Lung Cancer study – MINITAB

Proportional hazards regression model



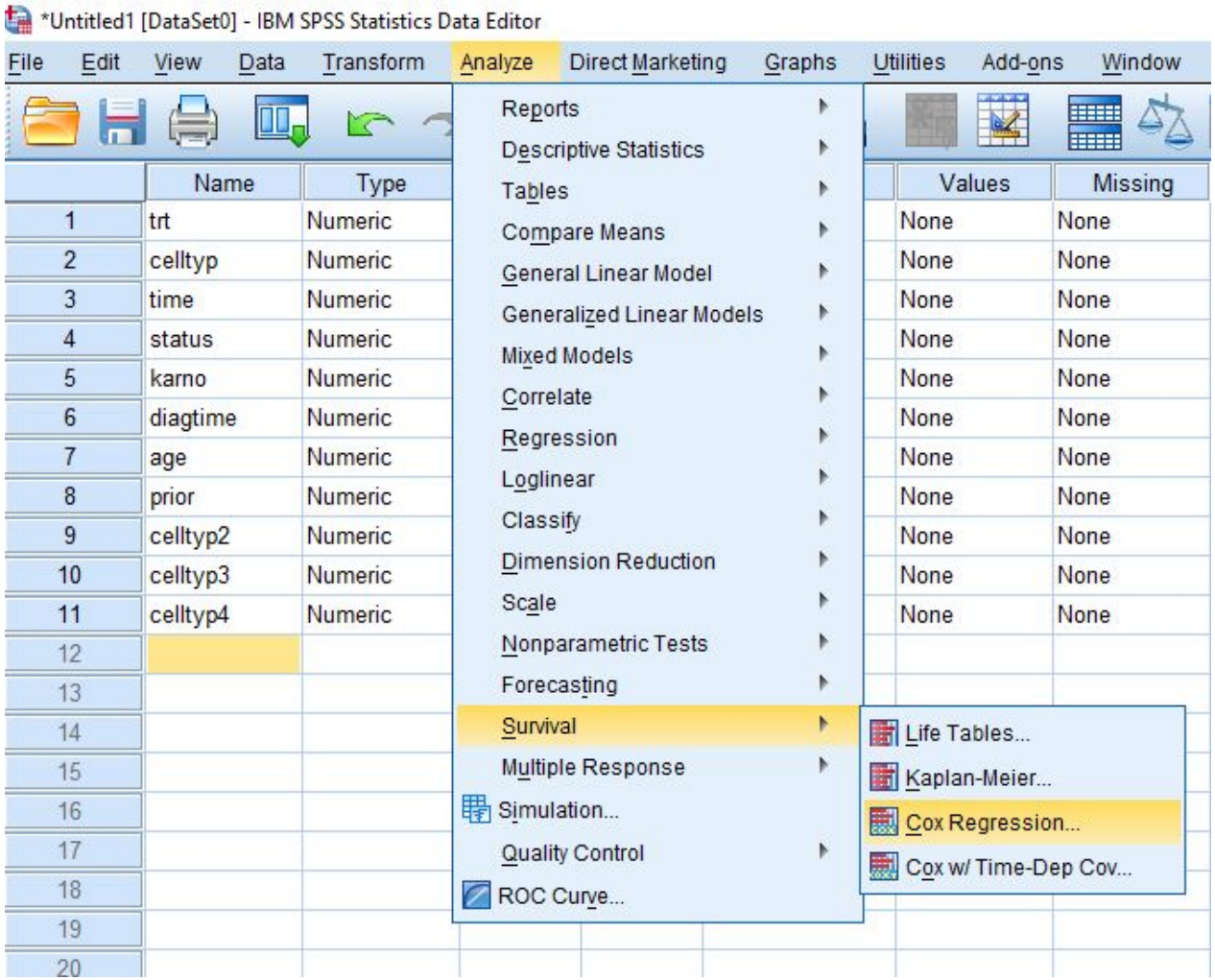
Example – Veterans' Administration Lung Cancer study – MINITAB

Proportional hazards regression model

```
Regression with Life Data: time versus trt, celltyp, ...

Response Variable: time

Censoring Information  Count
Uncensored value         128
Right censored value       9

Censoring value: status = 0

Estimation Method: Maximum Likelihood

Distribution:   Weibull

Regression Table
                        Standard                    95.0% Normal CI
Predictor         Coef      Error       Z      P       Lower      Upper
Intercept      3.49054   0.691171    5.05  0.000     2.13587    4.84521
trt         -0.228523   0.186844   -1.22  0.221   -0.594730   0.137684
celltyp
 2          -0.826185   0.246312   -3.35  0.001    -1.30895  -0.343422
 3           -1.13273   0.257598   -4.40  0.000    -1.63761  -0.627842
 4          -0.397681   0.254749   -1.56  0.119   -0.896981   0.101619
karno       0.0300683  0.0048279    6.23  0.000   0.0206059  0.0395308
diagtime   -0.0004688  0.0083614   -0.06  0.955  -0.0168569  0.0159193
age         0.0060992  0.0085534    0.71  0.476  -0.0106651  0.0228635
prior      -0.0438976   0.212279   -0.21  0.836   -0.459956   0.372161
Shape         1.07745  0.0714475                  0.946136    1.22699

Log-Likelihood = -715.551

Anderson-Darling (adjusted) Goodness-of-Fit
```

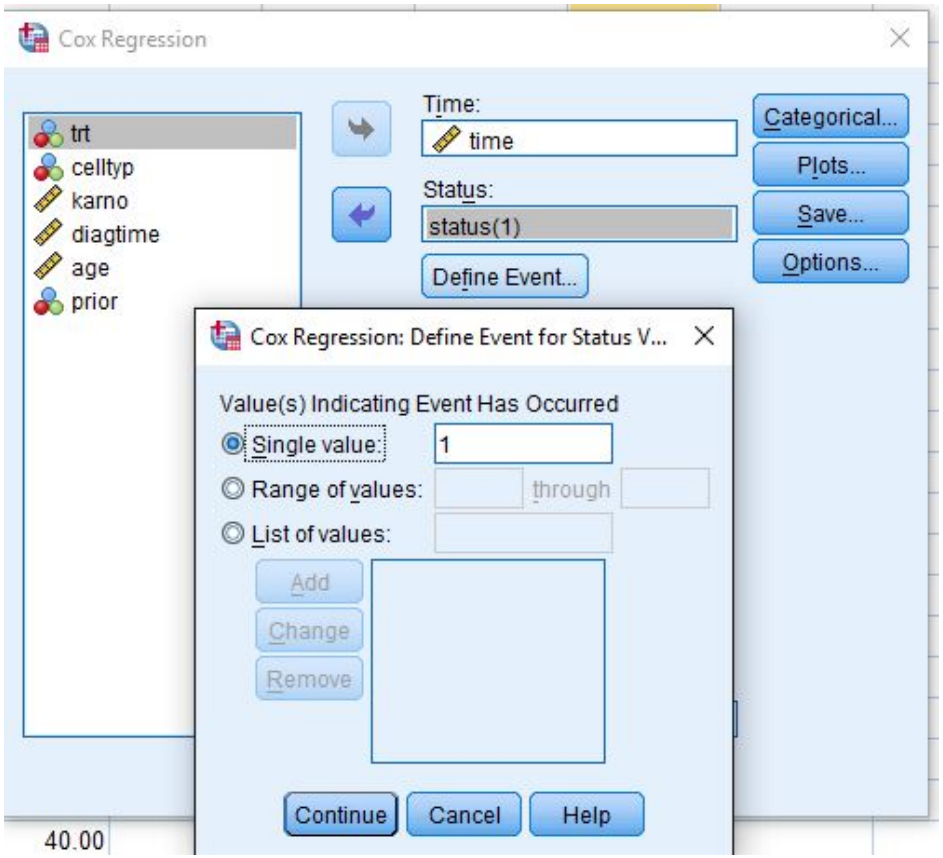
Example – Veterans' Administration Lung Cancer study – SPSS

Cox model



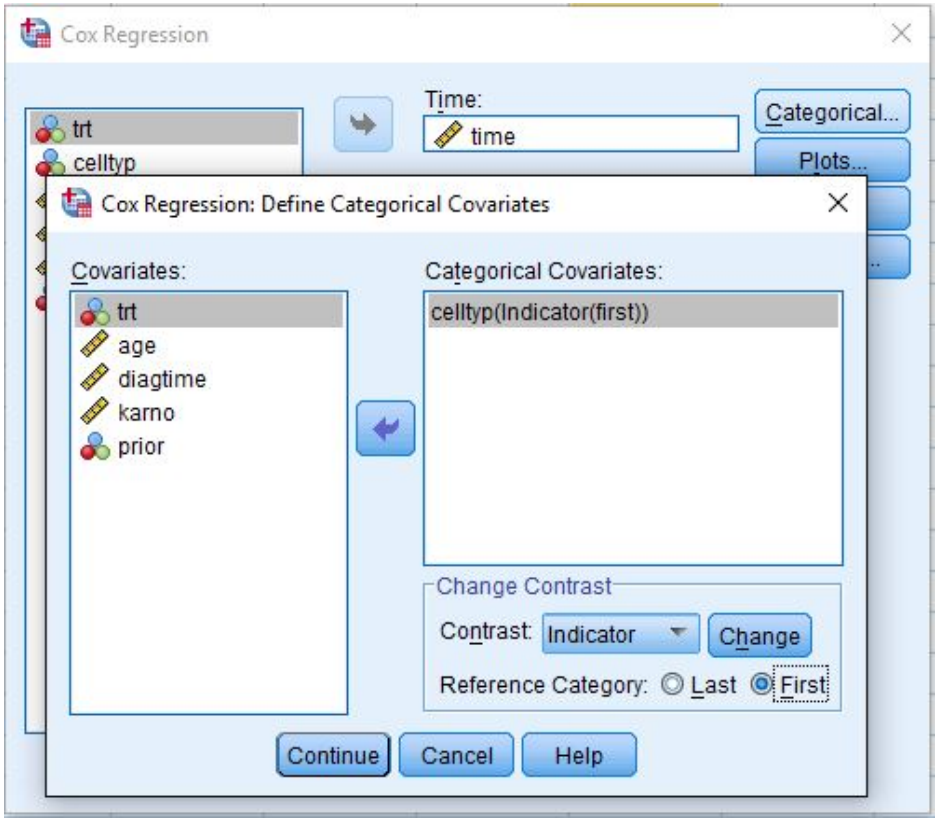
Example – Veterans' Administration Lung Cancer study – SPSS

Cox model



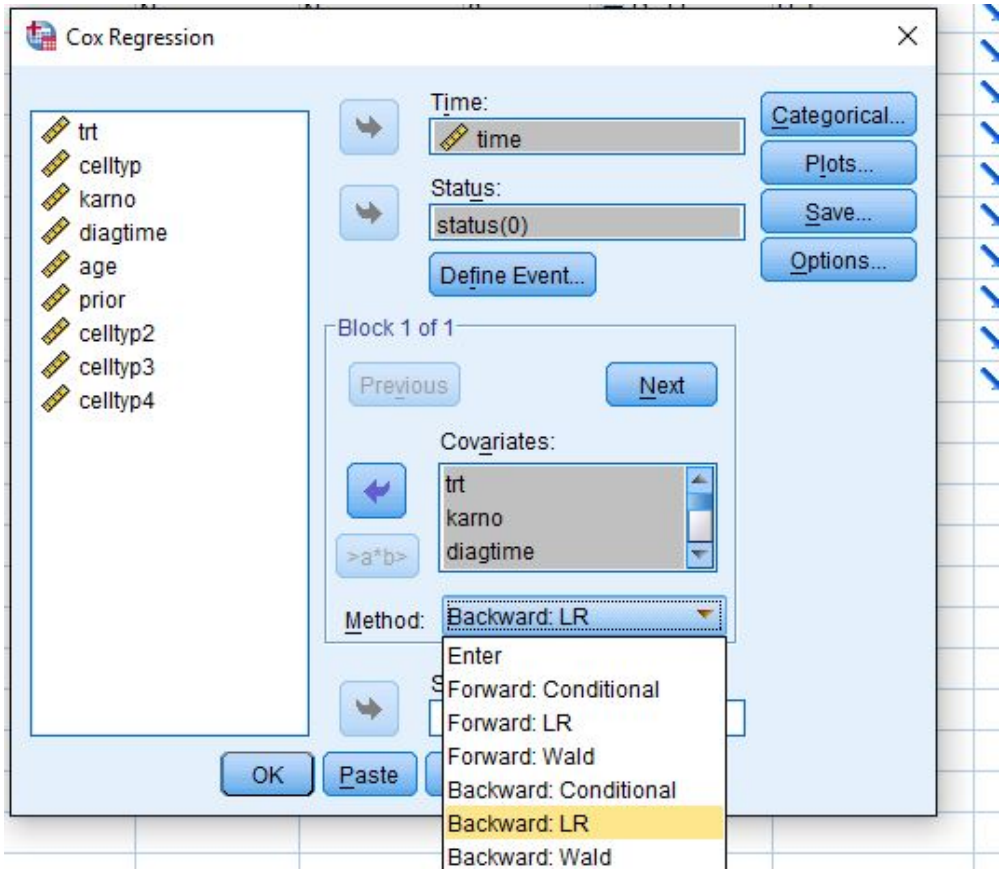
Example – Veterans' Administration Lung Cancer study – SPSS

Cox model



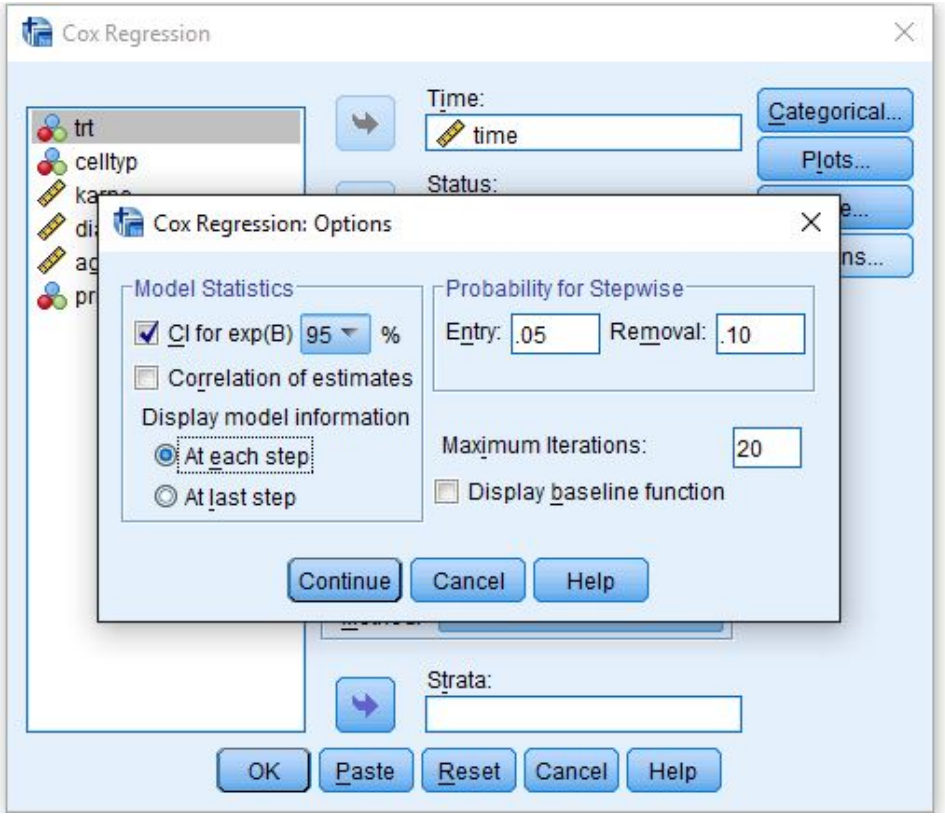
Example – Veterans' Administration Lung Cancer study – SPSS

Cox model



Example – Veterans' Administration Lung Cancer study – SPSS

Cox model



Example – Veterans' Administration Lung Cancer study – SPSS

Cox model

Block 1: Method = Backward Stepwise (Likelihood Ratio)

Omnibus Tests of Model Coefficients[f]

Step	-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
		Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
1[a]	950.359	65.917	8	.000	61.409	8	.000	61.409	8	.000
2[b]	950.359	65.821	7	.000	.000	1	.992	61.409	7	.000
3[c]	950.476	65.810	6	.000	.117	1	.732	61.292	6	.000
4[d]	951.352	65.747	5	.000	.876	1	.349	60.416	5	.000
5[e]	952.997	63.219	4	.000	1.645	1	.200	58.771	4	.000

a. Variable(s) Entered at Step Number 1: trt karno diagtime age prior celltyp

b. Variable Removed at Step Number 2: diagtime

c. Variable Removed at Step Number 3: prior

d. Variable Removed at Step Number 4: age

e. Variable Removed at Step Number 5: trt

f. Beginning Block Number 1. Method = Backward Stepwise (Likelihood Ratio)

Example – Veterans' Administration Lung Cancer study – SPSS

Cox model

Variables in the Equation

		B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
								Lower	Upper
Step 1	trt	.290	.207	1.958	1	.162	1.336	.890	2.006
	karno	-.033	.006	35.112	1	.000	.968	.958	.978
	diagtime	.000	.009	.000	1	.992	1.000	.982	1.018
	age	-.009	.009	.844	1	.358	.991	.974	1.010
	prior	.072	.232	.097	1	.755	1.075	.682	1.694
	celltyp			17.916	3	.000			
	celltyp(1)	.856	.275	9.687	1	.002	2.355	1.373	4.038
	celltyp(2)	1.188	.301	15.610	1	.000	3.281	1.820	5.917
	celltyp(3)	.400	.283	1.999	1	.157	1.491	.857	2.595
Step 2	trt	.290	.206	1.974	1	.160	1.336	.892	2.001
	karno	-.033	.005	36.036	1	.000	.968	.958	.978
	age	-.009	.009	.853	1	.356	.991	.974	1.010
	...	...	...	...	...	...	...	...	...
Step 5	karno	-.031	.005	35.612	1	.000	.970	.960	.979
	celltyp			17.080	3	.001			
	celltyp(1)	.712	.253	7.939	1	.005	2.038	1.242	3.345
	celltyp(2)	1.151	.293	15.441	1	.000	3.161	1.780	5.611
	celltyp(3)	.325	.277	1.381	1	.240	1.384	.805	2.381

Outline

Reading

- Biostatistics, A Foundation for Analysis in the Health Sciences, W.W. Daniel and C.L. Cross

  `http://informatika.uvlf.sk/subory/prezentacie%20zas/book%201.pdf`

  Chapter 14

- Rodriguez, G. (2007). Lecture Notes on Generalized Linear Models.

  `data.princeton.edu/wws509/notes/c7.pdf`
  Chapter 7

- Research Method II: Multivariate Analysis, Journal of Tropical Pediatrics

  `www.oxfordjournals.org/tropej/online/ma_chap12.pdf`

  Chapter 12

Assignment

1. Show that the Weibull distribution is closed under the Proportional Hazards regression model.
2. Select a survival (or reliability) data set from literature or from an online data sets archive (see websites at the next slide)[3]. The selection of your data set is up to you, but it has to be related with your MSc (Biomedical Engineering)[4],
   The data set should include at least 2 covariates. At least one of the covariates should be a binary categorical variable (e.g. gender, prior treatment, etc)
   - Give a short description of the data set (include some descriptive statistics and a reference)
   - Compare the survival times of the two groups (defined by the binary categorical variable) with the long-rank test.
   - Fit a Proportional Hazards regression model (Cox model) using all the available covariates. Interpret the results (Which covariates seem to be significantly important? How do they influence the hazard function? etc)
   - Find the 'best' model using the Forward LR selection or/and the Backward LR elimination techniques. Interpret the results.

[3] If you are having problems finding a data set, please do not hesitate to contact with me.

[4] There will be not much strictness about what makes a data set related to your MSc. But if your interests are pushing the envelope, let's chat before you get too far.

Data sets

Websites [5]

- `http://www.statsci.org/datasets.html`
- `https://datahub.io`
- `https://vincentarelbundock.github.io/Rdatasets/datasets.html`

Books [6]

- Handbook of Small Data Sets by Hand, Daly, Lunn, McConway and Ostrowski.

  `http://www.statsci.org/data/books/handetal.zip`
- Klein, J. P. and Moeschberger, M. L. (2003). Survival Analysis: Techniques for Censored and Truncated Data, 2nd edition. New York : Springer Verlag.
- Cox, D. R.; Oakes, D. (1984). Analysis of Survival Data. Chapman and Hall, London – New York.
- Crowder, M. J., Kimber, A. C., Smith, R. L., and Sweeting, T. J. (1991). Statistical Analysis of Reliability Data. Chapman Hall, London

[5] Please note that you will need to do some research in order to find a suitable data set for the Assignment.

[6] You can find the first two books in the Library