Chapter 2

Studies with Censored Data

Herein, we present three studies dealing with censored data. Two of them are on the HIV/AIDS epidemic, an area where survival analysis is a habitual tool for data evaluation, whereas the third one is on the shelf lives of food products. This is an area, where the methodology of survival analysis has recently been introduced.

In Section 2.1, the analysis of a data set on the survival of Tuberculosis (TB) patients co-infected with HIV is summarized. The time of interest has been the elapsed time from TB diagnosis until death. Since at the end of the study many individuals have still been alive, right-censoring has been present in many cases. Hence, standard statistical procedures for right-censored data have been applied.

In the subsequent section, the data of the study on injecting drugs users (IDU) in Badalona are presented. This includes a descriptive analysis, as well as a summary of all the censoring patterns observed in the data. These censoring patterns in the variables of interest have motivated the development of a parametric survival model with an interval-censored covariate to determine the predictive factors of the survival time. This model shall be presented and discussed in detail in Chapter 3, its application to the data set is the content of Chapter 6.

Finally, Section 2.3 deals with a completely different area where interval-censored data arise: the shelf lives of food products. The shelf life of a food product is defined as the storage time at which the product is rejected by the consumers. Hence, the survival function at time t is the probability that a consumer accepts the product beyond that time. In order to estimate the distribution function of a food product's shelf life, sensorial studies are carried out: consumers are given samples of the product stored during different times and have to judge whether or not they would normally eat this sample. The observed data consist of intervals into which the unobserved exact shelf lives fall. The use of methods for interval-censored data in this field, introduced by Hough, Langohr, Gómez, and Curia (2003), is a novel approach to evaluate data on the acceptance of food products. The study of Hough et al., presented in Section 2.3, deals specifically with the shelf life of a commercial whole fat, strawberry-flavored yogurt.

2.1 Survival of HIV-infected Tuberculosis patients

This study is on the survival of HIV-infected Tuberculosis patients in Barcelona, carried out by the local Municipal Health Institute, the *Institut Municipal de la Salut de Barcelona*. At the beginning of the nineties, among these patients, a high degree of lethality during the first few weeks of the TB treatment has been observed. It has therefore been the study objective to examine the survival of these patients and study the possible predictive factors of an early death after the start of an anti-Tuberculosis treatment.

Tuberculosis has nearly been eradicated in Europe during the second half of the past century. However, in the course of the HIV/AIDS epidemic, this infectious disease has spread again among the HIV infected patients due to their debilitated immune systems (Hoffmann and Kamps 2003). For this reason and according to the definition of AIDS by the Centers for Disease Control (CDC) in the USA, extrapulmonary and pulmonary TB are considered AIDS-indicating diseases since 1986 (Centers for Disease Control 1987) and since 1993 (Centers for Disease Control 1992), respectively. Sepkowitz and Raffalli (1994) point out that HIV and TB have a mutual negative influence on each other: HIV infection is the main risk factor for the outbreak of TB, which itself accelerates the course of AIDS.

In the first of the following sections, the sources of the given data set are presented. This is followed by a descriptive analysis in Section 2.1.2. Furthermore, patients are divided into short-term and long-term survivors in order to find possible predictors of the relatively high lethality after the start of the anti-Tuberculosis treatment (Section 2.1.3). Finally, a multivariate Cox model is applied to find out, which factors are predictive of the survival of HIV-infected TB patients in general.

The description of the data analysis follows the work of Falqués, Langohr, Gómez, Olalla, Jansà, and Caylà (1999). The only difference remains in the use of the statistical software S-Plus (instead of SPSS) for the adjustment and validation of the Cox model in Section 2.1.4.

2.1.1 Data sources

The basis of the evaluated data set is the Tuberculosis register of the *Institut Municipal de la Salut de Barcelona* (IMSB), which includes the data corresponding to all TB cases in Barcelona. This register is a result of the Program for the prevention and control of Tuberculosis, the *Programa de prevenció i control de la tuberculosi*, which exists since 1986. Its aim is to obtain a survey, as complete as possible, on the spread of this disease in Barcelona. Doctors and hospitals are involved in the program passing all the relevant information on TB cases to the IMSB. These data are mainly recorded by interviewing the patients about sociodemographic data and medical history and by analyzing blood samples.

The mentioned program forms part of the System of active epidemiologic vigilance, the *Sistema de vigilància epidemiològica activa*, that has also created a register on all cases of HIV and

AIDS in Barcelona. For the present study, both registers have been looked through in order to record all persons who have been infected with HIV and have been under an anti-Tuberculosis treatment. In case of multiple entries in the TB register due to several episodes of Tuberculosis, only the first record has been considered for the study. The resulting data set comprises all cases of HIV-infected TB patients older than 15 years who have started anti-Tuberculosis treatment between January 1st 1988 and December 31st 1993 and have followed it until the prescribed end (if they did not die before). The end of the follow-up time is September 30th 1995. The total number of cases in the study is 1135.

2.1.2 Description of the data set

List of variables

The original data set has comprised more than hundred variables, including many medical markers. For the present study, this number has been reduced to about ten variables of interest as possible predictors of the survival time. Besides the age and gender of the patients and the information on former imprisonment, the variables used for the data analysis are the following:

- **Risk group for HIV transmission**
  Either injecting drug users, homosexual men, both, heterosexuals or hemophiliacs.

- **Percentage of T CD4+ lymphocytes**
  Up to 14% of T CD4+ lymphocytes in the blood or more.

- **Tuberculin test**
  Result of the tuberculin test for presence of TB.

- **AIDS diagnosis**
  Presence of AIDS-indicating disease at start of anti-TB treatment according to the CDC's AIDS definition of 1986.

- **X-ray pattern**
  Condition of the lungs: normal, cavitary abnormal, or non-cavitary abnormal.

- **Bacteriology**
  Result of microbiological examination for presence of the TB bacterium, the *Mycobacterium tuberculosis*: positive microbiology, only culture-positive, or negative.

- **Location of the Tuberculosis**
  Pulmonary, extrapulmonary, or mixed form of TB.

Survival times

The main interest of the study has been to detect possible predictive variables of the survival time of the patients. This time has been measured (in days) from the date of the anti-Tuberculosis treatment's start until the date of the patient's death. Survival times of the patients have been right-censored if they either have been alive on September 30th 1995, the final day of the study, or if they have moved to another town and follow-up has been lost thereby.

Descriptive characteristics

The mean age of the 1135 patients has been 32.4 years (standard deviation: 8.8) at the beginning of the anti-TB treatment being men (32.6; 8.3) older then women (30.3; 8.3). The distribution of each of the above mentioned categorial variables is shown in Table 2.3 on page 25. The cohort consist mainly of men (82.6%), injecting drugs is the most frequent risk behavior for HIV transmission, and more than two thirds of the patients have already developed AIDS by the beginning of the anti-Tuberculosis treatment. These patients have either developed an extrapulmonary Tuberculosis or another AIDS indicating disease according to the 1986 definition of the CDC. Somewhat striking is the high proportion of missing values for the variables 'Tuberculin test' (49.8%) and 'Percentage of CD4 lymphocytes' (53.6%). Possible reasons for that and its implications for the data analysis will be discussed in Section 2.1.4 below. An extensive treatise of these aspects can also be found in Gómez and Serrat (1999).

Some significant changes have been within the study cohort throughout the six years of study period. While the age mean in 1988 is 30.4 (3.0), it amounts to 34.7 (4.0) in 1993. In the same period, the proportion of injecting drug users decreases from 81.2% to 64.6%, whereas the proportions of homosexuals and heterosexuals increase from 11.0% to 18.7% and from 0% to 11.0%, respectively. For the former variable, we have applied the t-test, for the latter the χ^2-test for homogeneity using a 95% significance level.

2.1.3 Short-term survivors

The mentioned high lethality observed among HIV-infected TB patients has motivated the comparison of the group of short-term survivors with the remaining patients. The break point chosen has been nine months given the fact that this is the period standard TB therapies last. For that analysis, a total of 102 patients has been disregarded because of a right-censored survival time of less than nine months. These are patients, that have left Barcelona during the first nine months of the study period. In total, there are 247 (23.9% of 1033) short-term survivors and 786 (76.1%) long-term survivors.

In order to compare both groups with respect to the other variables, the t-test has been applied for the continuous variable 'Age', whereas the χ^2-test for homogeneity has been chosen for the categorical variables. A significant difference (at a 95% significance level) is observed regarding the age of the patients: among the short-term survivors, the mean age has been 36.0 years (standard deviation: 10.8), among the long-term survivors, this values amounts to only 31.5 (7.6). In Table 2.4 on page 26, the comparison of both groups is summarized. According to this univariate analysis, the following subgroups have shown better survival: former prisoners, injecting drug users, patients with higher percentage of T CD4+ lymphocytes, patients with a positive tuberculin test result, and AIDS-free patients.

The better survival of former prisoners and injecting drug users might be surprising at first sight, but can mainly be explained by the better medical care patients received in the prisons. This is particularly important for the anti-TB treatment which requires a disciplined follow-up for several months. Whereas more than 30% of the intravenous drug users have been imprisoned formerly, not even three percent of homosexual men and heterosexuals have been in jail.

2.1.4 Application of the Cox model

The previous analysis can give an idea on possible predictors of the survival time of HIV-infected TB patients, however, it is an univariate approach. For a multivariate analysis of the survival time, we have applied the proportional hazards model of Cox (Cox 1972):

$$\lambda(t;\boldsymbol{z}) = \lambda_0(t)\exp(\boldsymbol{\beta}'\boldsymbol{z}), \tag{2.1}$$

where λ_0 is the underlying baseline hazard function and $\boldsymbol{z}$ is the multivariate covariate vector. The procedure to select an appropriate model has been a mixture of forward and backward selection: variables are successively included in the model as long as the model fit improves significantly at a 95% level. With each new variable in the model, the variables already included in the model are checked to still be significant choosing an exclusion criterium of 10%. That is, if any of the p-values in the new model exceeds 0.1, the corresponding variables are excluded. This procedure is continued uill the model fit cannot be significantly improved anymore.

Selection of variables

Three of the possible covariates for model (2.1) —the variables of Table 2.3 plus ‘Age’— have been disregarded for different reasons. First, when proving the condition of proportional hazards for each variable univariately, the variable ‘X-ray pattern’ shows non-proportional hazards and is therefore not considered. Any transformation of this variable might achieve the required proportionality of risks, however, the previous analysis summarized in Table 2.4 has not shown any influence of this variable on short- and long-time survival. Secondly, the variable ‘Location of TB’ is disregarded for the same reason and also due to its correlation with the variable ‘AIDS’. Following the 1986 CDC’s AIDS definition, any case of HIV infection in combination with extrapulmonary TB is considered an AIDS case. Finally, comparing the survival of patients with known and unknown tuberculin test result, the log-rank test shows significant differences

($p < 0.001$) being the survival of patients with missing test result nearly the same as the one of patients with a negative test result ($p = 0.52$). That is, patients with given tuberculin test result are most probably not representative for the whole cohort. An explanation for this observation is that the tuberculin test, in many cases, has not been applied to patients in a bad health shape, who normally show a negative test result. Hence, the inclusion of this variable into the model would introduce a non-ignorable bias since only individuals without any missing value are considered for the model construction.

In contrast with the tuberculin test, for missing observations of the variable 'T CD4+ Lymphocytes' we can assume missing at random since the survival curves of patients with and without observations show a very similar behavior ($p = 0.96$). However, considering the high percentage of missing values (53.6%; see Table 2.3), two models are adjusted one disregarding this variable and a second one including it.

Model adjustment

Two proportional hazards models are adjusted, for the first of which the percentage of the T CD4+ lymphocytes in the blood is excluded in order to increase the sample size. Disregarding all patients with at least one missing observation in the remaining variables, in case of model 1, 791 patients are considered, whereas for model 2, the sample size amounts to 384. The values of the estimated parameters and their standard errors for both model 1 and model 2 are summarized in Table 2.1. This table includes also the relative risk, estimated by $\exp(\hat{\beta})$, corresponding to the significant variables and their 95% confidence interval.

Table 2.1: Parameter estimates of Cox models

			Model 1 ($n = 791$)			
Variables	$\hat{\beta}$	$s.e.(\hat{\beta})$	p**-Value**	**Rel. Risk**	$\mathbf{CI}_{95\%}$**(RR)**	
AIDS	1.192	0.121	< 0.001	3.29	$[2.60, 4.17]$	
Age	0.03	0.006	< 0.001	1.03	$[1.02, 1.04]$	
			Model 2 ($n = 384$)			
Variables	$\hat{\beta}$	$s.e.(\hat{\beta})$	p**-Value**	**Rel. Risk**	$\mathbf{CI}_{95\%}$**(RR)**	**Reference**[a]
AIDS	1.881	0.284	< 0.001	6.56	$[3.76, 11.45]$	CD4 high
				1.36	$[0.92, 2.01]$	CD4 low
CD4	2.024	0.322	< 0.001	7.57	$[4.03, 14.21]$	No AIDS
				1.56	$[1.16, 2.11]$	AIDS
AIDS$*$CD4	-1.576	0.348	< 0.001			

[a] Reference category for relative risk

Two variables are included in model 1: AIDS and the age of the patients. For the former variable, the estimated relative risk amounts to 3.3, that is, the risk of dying of a patient with AIDS at the beginning of the anti-TB treatment is 3.3 times higher than the risk of an AIDS-free patient of the same age. Regarding age, the relative risk of 1.03 indicates that the risk of dying augments with increasing the age. Given the same diagnosis of AIDS, a patient of age a (years) has 1.03 times the risk of dying of a patient with age $a - 1$ and $\exp(10 \cdot 0.03) \approx 1.35$ times the risk of a patient of age $a - 10$.

When including the variable 'T CD4+ lymphocytes', the effect of the age is superseded by this variable. Besides, the interaction of CD4+ cells and AIDS diagnosis results significant. This interaction implies that the relative risk for each of the two variables depends on the levels of the other. For example, comparing two individuals with a high CD4 cell count, the one with AIDS has 6.56 times more risk of dying than the AIDS-free individual. On the other hand, if both individuals have a low CD4 cell count, the relative risk for AIDS diagnosis amounts to only 1.36.

The validity of both models has been examined by means of the S-Plus function `cox.zph`, which applies a Kolmogorov-based test on the score residuals proposed by Schoenfeld (Mathsoft 1999). The null hypothesis is that proportional hazards hold; then, these residuals are randomly distributed (Collett 1994). According to the test results for both models, summarized in Table 2.2 below, there is no clear evidence against the assumption of proportional hazards since none of the p-values is smaller than 0.2.

Table 2.2: Verification of proportional hazards

Model 1		**Model 2**	
Variables	**p-Value**[a]	**Variables**	**p-Value**[a]
AIDS	0.238	AIDS	0.123
Age	0.242	CD4	0.828
		AIDS∗CD4	0.23
Global	0.257	Overall	0.217

[a] test based on score residuals proposed by Schoenfeld

2.1.5 Conclusions

According to the analysis of the cohort of HIV-infected Tuberculosis patients in Barcelona, AIDS diagnosis and the level of T CD4+ lymphocytes in the blood are the main predictors of the survival time. Following the results of model 2, a patient with AIDS diagnosis and a low CD4 cell count has $\exp(1.881 + 2.024 - 1.576) \approx 10.3$ times the risk of dying of an AIDS-free patient with a high CD4 cell level. If the CD4 cell count is not considered for the Cox model, the age of the patients is one of the significant predictors implying that the risk of dying augments with increasing age. For both models, we can assume that the assumption of proportional hazards is justified.

In neither of the two models the variable 'Former imprisonment' has shown significance. That is, the observed better survival of former prisoners within the first nine months (Section 2.1.3), is most probably due to the fact that these are significantly younger (mean age: 29.9, standard deviation: 5.6) than patients who have not been in jail (33.0, 9.0). The p-value of the corresponding t-test amounts to less than 0.001.

Nowadays, the survival of the HIV-infected patients has improved very much both in length and quality because of the highly active anti-retroviral therapies. Therefore, the obtained results might now be of less importance in the highly industrialized countries. However, in many parts of the world, these new treatments are not available for the big majority of the affected population, and it is therefore very important to know which HIV-infected TB patients are at high risk of dying.

Table 2.3: Characteristics of TB patients co-infected with HIV

Variable	**Categories**	**Number**	**Percentage**
Gender	Male	938	82.6
	Female	197	17.4
Former imprisonment	Yes	273	24.1
	No	862	75.9
Risk group	Injecting drug users	833	73.4
	Homosexuals	156	13.7
	Hemophiliacs	10	0.9
	Heterosexuals	41	3.6
	IDU & Homosexuals	29	2.6
	Unknown	66	5.8
Percentage of T CD4+	$\leq$ 14%	288	25.4
Lymphocytes	$>$ 14%	239	21.0
	Unknown	608	53.6
Tuberculin test	Positive	245	21.6
	Negative	325	28.6
	Unknown	565	49.8
AIDS[a]	Yes	780	68.7
	No	355	31.3
X-ray pattern	Normal	181	15.9
	Cavitary abnormal	160	14.1
	Non-cavitary abnormal	740	65.2
	Unknown	54	4.8
Bacteriology	Positive microscopy	427	37.6
	Only culture-positive	361	31.8
	Negative	209	18.4
	Not determined	100	8.8
	Others[b]	38	3.4
Location of TB	Pulmonary	529	46.6
	Extrapulmonary	393	34.6
	Mixed	201	17.7
	Unknown	12	1.1
Total		1135	100.0

[a] according to the 1986 AIDS definition of the CDC

[b] mainly diagnosed by clinical-radiological criteria or the ADA test

Table 2.4: Comparison of short- and long-term survivors

		Survivors				
		Short-term		**Long-term**		
Variable	**Categories**	n	%[a]	n	%[a]	p**-Value**[b]
Gender	Male	204	82.6	650	82.7	0.963
	Female	43	17.4	136	17.3	
Former imprisonment	Yes	27	10.9	213	27.1	< 0.001
	No	220	89.1	573	72.9	
Risk group	Injecting drug users	154	73.3	615	84.2	0.001
	Homosexuals	43	20.5	91	12.5	
	Heterosexuals	13	6.2	24	3.3	
	Unknown	37		56		
Percentage of T CD4+ Lymphocytes	$\leq 14\%$	82	69.5	182	49.5	< 0.001
	$> 14\%$	36	30.5	186	50.5	
	Unknown	129		418		
Tuberculin test	Positive	21	22.6	198	46.5	< 0.001
	Negative	72	77.4	228	53.5	
	Unknown	154		360		
AIDS[c]	Yes	215	87.0	498	63.4	< 0.001
	No	32	13.0	288	36.6	
X-ray pattern	Normal	37	15.5	128	17.2	0.190
	Cavitary abnormal	27	11.4	116	15.5	
	Non-cavitary abnormal	174	73.1	503	67.3	
	Unknown	9		39		
Bacteriology	Positive microscopy	44	20.8	125	17.9	0.645
	Only culture-positive	127	59.9	432	61.9	
	Negative	41	19.3	141	20.2	
	Unknown	35		88		
Location of TB	Pulmonary	122	50.4	367	46.9	0.602
	Extrapulmonary	81	33.5	273	34.9	
	Mixed	39	16.1	142	18.2	
	Unknown	5		4		
Total		247	100.0	786	100.0	

[a] of non-missing data

[b] the χ^2-tests for homogeneity have not included the category ‘Unknown’

[c] according to the 1986 AIDS definition of the CDC

2.2 The data set on injecting drug users in Badalona

The data which have mainly motivated the present PhD thesis come from the detoxication unit of the *Hospital Trias i Pujol*, also known as *Hospital Can Ruti*, in Badalona (Spain). The study population consists of intravenous drug users from Badalona and surroundings, many of whom became infected with HIV mainly because of sharing their syringes with others.

In this section, we first illustrate the study objective (Section 2.2.1) and then present the data set with more detail (Section 2.2.2). One important aspect of the data is the censoring in the variables of interest which will be looked at in the closing Section 2.2.3.

2.2.1 Objective of the study

The motivating epidemiological question regarding the data set has been:

> Does the length of the elapsed time from first potential HIV exposure by injecting drugs until HIV infection have any influence on the subsequent AIDS incubation period?

In the remainder of this thesis, both times will be denoted by Z and Y, respectively, as illustrated in Figure 2.1, where 0 corresponds to the moment of first intravenous drug use. In case of individuals, who started injecting drug use before 1978, Z is measured from January 1st 1978, because it is assumed that HIV was not spread before that year in Spain.

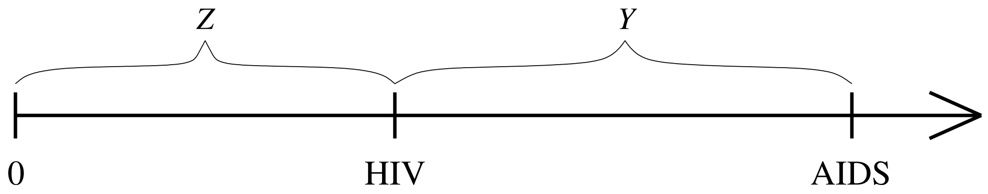


Figure 2.1: Pattern of disease stages

Actually, routine HIV tests, available since 1985, cannot detect the HIV infection during the first days and weeks after its occurrence, because HIV antibodies are not produced immediately after the infection with HIV. The moment the antibodies are produced and can be detected the first time in the blood is called seroconversion. It is estimated that the median time from HIV infection until seroconversion lasts about two months (Brookmeyer and Gail 1994; Brookmeyer and Quinn 1995). Since this period is relatively short compared with the subsequent AIDS incubation period, we assume that the results and conclusions of the present study are hardly altered by the use of seroconversion instead of HIV infection.

2.2.2 Descriptive analysis of the data set

The original data set contains the data of a total of 370 injecting drug users from Badalona and nearby municipalities like Santa Coloma or Sant Adrià. No information concerning the moment of HIV infection has been given in case of nine individuals, for which reason they have been removed from the data set. Hence, the study population consists of 361 injecting drug users.

Variables of interest

Besides the individuals' age and gender, we have had the following information at our disposal:

- Date of first injecting drug use,
- Dates of last seronegative and first seropositive HIV test result, respectively,
- Indicator of AIDS onset with date and, in case of AIDS, the indicating disease,
- Indicator of death with corresponding date, and in case of death, the death cause.

Whereas the date of first injecting drug use has been available for all individuals, data regarding HIV test results, AIDS onset, and death have been partly missing (see Section 2.2.3). Given the mentioned dates, we have calculated the following times: time from first injecting drug use until the last seronegative and first seropositive test result, respectively, as well as the subsequent times until date of AIDS onset. The chosen time unit are months a value of i being equal to the ith month since the corresponding starting point.

Sociodemographic data

Among the 361 intravenous drug users from Badalona and surroundings, 307 (85.0%) are men. As shown in Table 2.5, 67% of the individuals have been 20 years or younger, with median equal to 19 years, when consuming intravenous drugs the first time. The youngest individual has not even been ten years old when starting injecting drugs, whereas only 14 of them (3.9%) have been older than 30 years. No correlation ($p > 0.9$; χ^2-test) is observed between gender and age at first intravenous drug use.

Table 2.5: Age at first iv drug use

Age group	***n***	**%**
≤ 15	57	15.8
$16 - 20$	185	51.2
$21 - 25$	81	22.4
$26 - 30$	24	6.7
≥ 31	14	3.9
Total	361	100.0

We observe a broad range of years during which individuals have started intravenous drug consumption: it spans from 1974 to 1997 with peaks in the mid-eighties; see Figure 2.2 on the following page. As mentioned before, we consider the January 1st 1978 as the earliest time of

HIV exposure by injecting drugs, and measure Z from that date in case of the 15 individuals who have started injecting drugs in the years before 1978.

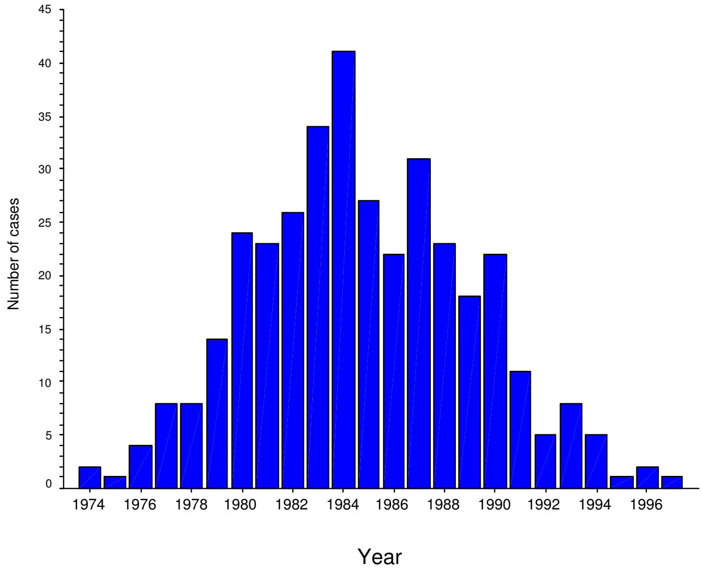


Figure 2.2: Year of first injecting drug use

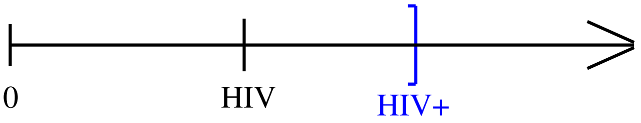
2.2.3 Presence of censored data

The motivating question concerning our data might have been easily answered if the corresponding times could have been observed exactly. However, that has not been the case: both the time from first injecting drug use until HIV infection and the AIDS incubation period are mainly censored.

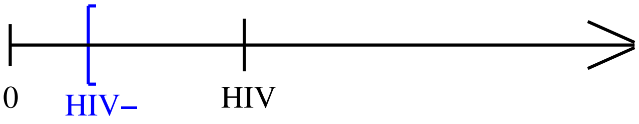
Interval-censored times until HIV infection

Injecting drug users coming to the detoxication unit of the Hospital Can Ruti have been tested for HIV routinely since 1985. Since neither the moment of HIV infection nor the seroconversion can be observed exactly, we deal with censored times until HIV infection. We distinguish the following three cases of interval-censored observations:

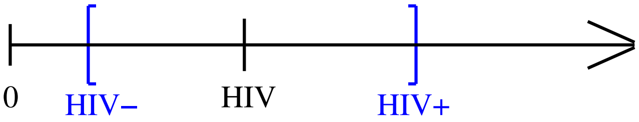
Seropositive cases A seropositive time HIV+ is observed but no seronegative observation is available. Hence, the time until HIV infection is known to lie in $(0, \text{HIV+}]$. In our data set, we have 225 (62.3%) seropositive individuals.



Seronegative cases A seronegative observation HIV– time is given but no seropositive test result; thus, $Z \in [\text{HIV}-, \infty)$. 98 individuals (27.2%) show that observational pattern.



Seroconverters Both a last seronegative and a first seropositive test result are available, that is $Z \in [\text{HIV}-, \text{HIV}+]$. There are 38 (10.5%) seroconverters in the data set.



Consequently, times until HIV infection are mainly current status data, whereas interval-censoring case 2 is present in about 10% of all cases.

It is important to note, that in the remainder we do not consider the case of long-time survivors with respect to time until HIV infection. That means, we assume that all injecting drug users will become infected with HIV. This might not be true for all the 98 seronegative cases, nonetheless we believe that the possible bias caused by that assumption can be neglected.

Doubly-censored AIDS incubation period

The AIDS incubation periods in the data set are doubly-censored due to the fact that, on one hand, the time origin, that is HIV infection, is interval-censored and, on the other hand, the moment of AIDS onset is partly left- and right-censored. The reason for this and the corresponding number of cases can be seen in the Tables 2.6 and 2.7.

In Table 2.6, we show the number and proportion of AIDS cases among the study cohort. In a total of 82 (22.7%) cases, AIDS has been diagnosed, that is, AIDS incubation periods are uncensored. In contrast with that, 182 cases (50.4%) are right-censored as these individuals have been AIDS-free at their last visit in the hospital. We do not have any information on possible deaths after these diagnoses. Moreover, there are 97 injecting drug users for whom information about development from HIV infection till AIDS

Table 2.6: AIDS diagnosis

Diagnosis	n	%
AIDS	82	22.7
No AIDS	182	50.4
Missing	97	26.9
Total	361	100.0

Table 2.7: Death causes of individuals with missing AIDS onset

Cause of Death	n	%
AIDS	14	31.8
Overdose	20	45.5
Accident/Violence	3	6.8
Others	1	2.3
Unknown	6	13.6
Total	44	100.0

is missing. However, 44 (45.4%) of these individuals have died by the end of the study period with causes of death given in Table 2.7. We see that 14 of the 44 individuals died because of AIDS, which implies that AIDS must have been developed before the corresponding date of death. Hence, these are left-censored observations for time until AIDS onset. In case of the remaining 30 individuals it is unclear, whether or not they have developed AIDS before their death.

In Table 2.8 on page 32, all observed censoring patterns of the times until HIV infection and the subsequent AIDS incubation period are summarized. Note that the 20 individuals with a seropositive observation but no information about AIDS onset are considered right-censored observations concerning the AIDS incubation time. This is because HIV infection has occurred before the date of the seropositive observation, but AIDS has not yet been diagnosed.

Following the illustration in De Gruttola and Lagakos (1989), in Figure 2.3, we show the possible combinations of HIV infection times and AIDS incubation period according to the censoring in the latter variable. Assume HIV infection falls into the interval $[2, 4]$ and the observation for AIDS onset is equal to 7. Then, for any value $z \in [2, 4]$ of HIV infection, if AIDS onset is observed exactly, the AIDS incubation period is equal to $7 - z$. If AIDS onset is right-censored, the incubation periods are larger or equal to $7 - z$, and if left-censoring is given, the possible incubation times fall into the interval $[4 - z, 7 - z]$. Hence, the possible combinations of times until HIV infection and AIDS incubation periods either lie on a straight line or fall into an infinite or finite parallelogram.

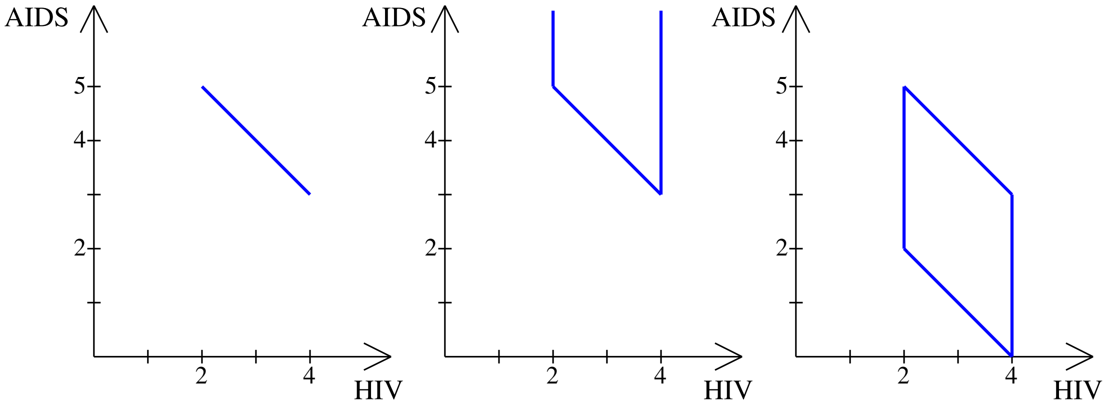
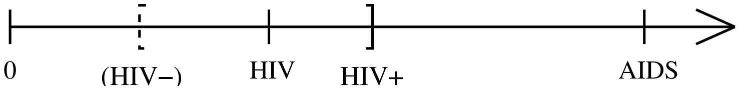
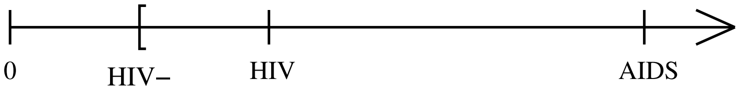
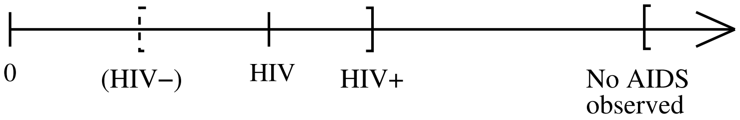
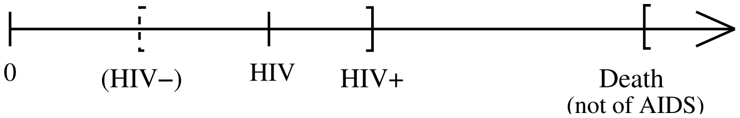
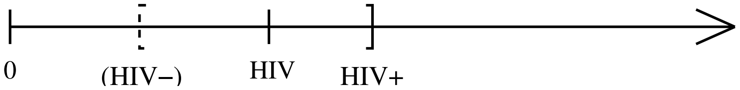
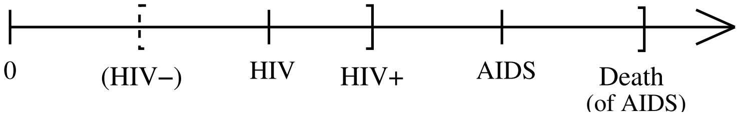
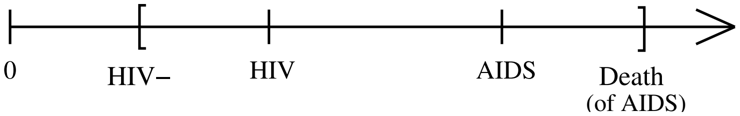
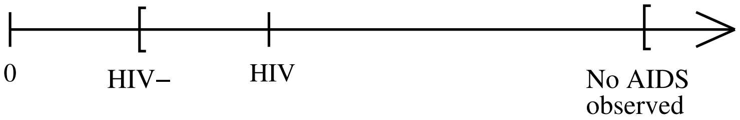
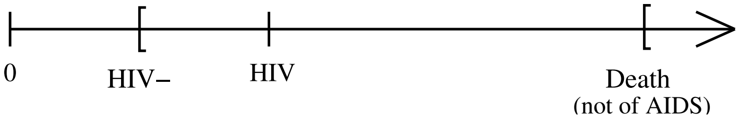
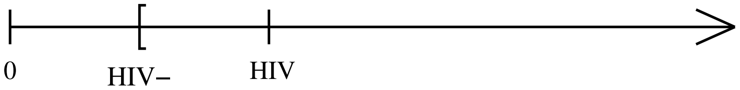


Figure 2.3: Possible combinations of time until HIV infection and AIDS incubation period given an exact, a right-censored, and a left-censored date of AIDS diagnosis

A way to evaluate the data of this study by means of a parametric survival model shall be presented in detail in the Chapters 3 and 6. Its particularity lies in the fact that the covariate is interval-censored. Whereas the following chapter deals with the theoretical background, in Chapter 6, the application of this method to the given data set is presented, including the estimation results under different model assumptions.

Table 2.8: Frequencies of observed censoring patterns

Observational pattern[a]	**Frequency**	**%**
	80	22.2
	2	0.5
Subtotal: exact observations of AIDS onset	82	22.7
	136	37.7
	14	3.9
	20[b]	5.5
Subtotal: right-censored observations of AIDS onset	170	47.1
	13	3.6
	1	0.3
Subtotal: left-censored observations of AIDS onset	14	3.9
	46	12.7
	10	2.8
	39	10.8
Subtotal: missing observations of AIDS onset	95	26.3
Total	361	100.0

[a] 'HIV' and 'AIDS' denote the exact moments of HIV infection and AIDS onset, 'HIV-' and 'HIV+' denote the seronegative and seropositive observations, '(HIV-)' stands for a possible seronegative observation.

[b] the seropositive observation coincides with right-censoring of AIDS onset at HIV+

2.3 Study on shelf life of yogurt

As mentioned at the beginning of the present chapter, the use of survival analysis is a new approach for the evaluation of data on the shelf lives of food products. In this area, the time of interest is the storage time of a food product under given circumstances until it is rejected by the consumers. Hence, the survival function, $S(t)$, is defined as the probability of the consumers accepting the product beyond storage time t, and $F(t) = 1 - S(t)$ is the probability that the consumers reject it before t.

As Hough, Langohr, Gómez, and Curia (2003), who have introduced the application of survival analysis to the evaluation of shelf lives of foods, point out,

> "...food products do not have shelf lives on their own, rather they will depend on the interaction of the food with the consumer."

That is, a product can be microbiologically safe to eat, but might be rejected due to its sensorial properties such as wrinkled apples or soft bananas. Thus, the hazard is not focused on the deterioration of the product but on the rejection by the consumers.

In order to estimate the distribution function of the shelf lives of food products, sensorial studies are carried out. In these studies, consumers have to try several samples of the product, each stored under the same conditions but for different time periods unknown to the panel of consumers. They have to answer with either yes or no to whether they would normally consume this product. Since the exact shelf life of the food product cannot be observed exactly, interval-censored data arise.

In the following, these kind of data and the corresponding methods for their evaluation are presented with more detail following the work of Hough et al., whose study deals with the shelf lives of a commercial whole fat, strawberry-flavored yogurt. Besides, the advantages of the applied methods over the logistic regression approach are discussed.

2.3.1 Type of data

For an illustration of the obtained data, see the following Table 2.9. Therein, $t_1 < t_2 < \cdots < t_6$ denote six different storage times of a study and $+/-$ stand for the consumers' judgements (yes/no). The last column contains the resulting intervals for each of the five subjects. According to Hough et al., these intervals are semi-open, but depending on the study and the chosen time units, the observed intervals can be interpreted as closed intervals. However, the methodology presented in Section 2.3.2 applies equally to both types of intervals.

Subject 1 shows the expected observational pattern: the samples are accepted up to a certain storage time, after which all samples are rejected. In the given example, these times are equal to t_3 and t_4, that is, the exact moment of rejection lies in the interval $(t_3, t_4]$. Subject 2 presents a right-censored observation since all samples have been accepted; hence, rejection lies beyond t_6.

Table 2.9: Illustration of shelf life data

	Storage times						
Consumer	t_1	t_2	t_3	t_4	t_5	t_6	**Interval**
1	+	+	+	−	−	−	$(t_3, t_4]$
2	+	+	+	+	+	+	(t_6, ∞)
3	−	−	−	−	−	−	$(0, t_1]$
4	+	+	+	−	+	−	$(t_3, t_6]$
5	−	+	−	−	−	−	$(0, t_3]$

In case of Subject 3 it is the other way round: all samples are rejected. Consequently, the exact moment of rejection occurs before t_1. If t_1 was equal to 0, a subject with rejection at t_1 might be removed from the sample, because he or she might not like fresh products and would therefore not be suited for the study.

Subjects 4 and 5 show a somewhat inconsistent pattern: the first rejection is followed by an acceptance and the second time of rejection. Hough et al. propose to deal with these inconsistencies by choosing broader intervals, $(t_3, t_6]$ and $(0, t_3]$, in order to account for the uncertainty of what actually happens between these times. Another possibility would be to ignore the observations after the first rejection and to consider the intervals $(t_3, t_4]$ (Subject 4) and $(0, t_1]$ (Subject 5). This choice would reduce the variance of $\hat{S}(t)$, however it would probably overestimate $F(t)$.

2.3.2 Application of survival analysis to shelf life data

In order to make inference on the distribution function of the shelf lives, $F(t)$, known methods for interval-censored data can be applied. Inference is based on the likelihood function (1.1) on page 3, which can be written in the following way:

$$L(F) = \prod_{i \in \mathcal{R}} \big(1 - F(t_i)\big) \prod_{i \in \mathcal{L}} F(t_i) \prod_{i \in \mathcal{I}} \big(F(t_{r_i}) - F(t_{l_i})\big), \tag{2.2}$$

where $\mathcal{R}$ denotes the set of the right-censored observations, $\mathcal{L}$ the one of the left-censored observations, and $\mathcal{I}$ the set of interval-censored times.

To obtain a nonparametric estimate of $F(t)$, the Turnbull estimator of Section 1.1.1 is the adequate tool. This estimation is easily accomplished, for example, by the use of the function `kaplanMeier` of the statistical software package S-Plus.

An alternative to the Turnbull estimator is the use of a parametric model of the form:

$$\ln(T) = \mu + \sigma W, \tag{2.3}$$

where μ is a constant, σ the scale parameter and W the error term distribution. For example, if T follows a Weibull distribution, W is the extreme value or Gumbel distribution; if T follows a log normal distribution, W is the standard normal distribution. The expression for $F(t)$ determined by the chosen distribution is plugged into (2.2) and numerical methods are applied to obtain the unknown distribution parameters. For details on that model, see, for example, Klein and Moeschberger (1997, Chap. 12). Model (2.3) does not include any possible covariates, but these can easily be incorporated; in Section 3.1 ahead, some general aspects on such a log linear regression model will be summarized.

The fit of this model to the interval-censored shelf life data permits estimating, for example, the mean or median storage time until rejection under different parametric assumptions. Other values of interest are the quantiles of the distribution, that is in this case, the moments at which certain percentages of consumers reject the product. Estimates based on model (2.3) are more precise than the ones based on the Turnbull estimator as long as the parametric choice is adequate.

For the fit of model (2.3) to the observed data, S-Plus offers the functions `censorReg` and `probplot6.censorReg`. The former fits the model to the data given a specified distribution for T, whose choice can be based on the probability plots for different parametric choices drawn by the latter function. An illustration of these methods and their implementation in S-Plus can be found in Garitta, Gómez, Hough, Langohr, and Serrat (2003).

2.3.3 Results of study on shelf life of strawberry-flavored yogurt

Herein, the evaluation of the data of 46 consumers on commercial whole fat, stirred, strawberry-flavored yogurt is presented. The main interest has been to estimate the median shelf life of that type of yogurt as well as several quantiles both based on the use of model (2.3). The whole data set of this sensorial study is shown in Table A.1 on page 121.

Both the Weibull and the log normal distribution have shown a reasonably good fit to the nonparametric estimate and are therefore chosen to estimate the mentioned quantiles. These are shown in Table 2.10 below, in which the (interpolated) estimates of the Turnbull estimator are added in the last column.

The table shows that the quantiles for both distributions coincide quite well, also with the Turnbull estimates, until the 75%-quantile, whereas the differences are much bigger at the 90%-quantile. This is due to the facts that both distributions can fit the data well up to the maximum storage time, which has been 48 hours in the present study, and that right-censored data beyond that time are present. Here, rejection occurs in about 17% of all cases beyond 48 hours and therefore the Turnbull estimator cannot estimate the 90%-quantile. For this quantile, the log normal distribution furnishes a higher value than the Weibull distribution since the former is a heavy tailed distribution. Generally, the higher the percentage of rejection beyond the maximum storage time, the bigger the possible differences between parametric fits beyond that value.

Table 2.10: Quantiles of shelf life of strawberry-flavored yogurt

	Distributions				
Quantiles	**Log normal**[a]		**Weibull**[a]		**Turnbull**
0.1	6.0	$[3.9, 9.4]$	5.0	$[2.5, 10.0]$	5.8
0.25	10.6	$[7.5, 14.9]$	11.1	$[7.0, 17.5]$	9.7
0.5	19.8	$[14.8, 26.6]$	22.2	$[16.5, 29.8]$	20.1
0.75	37.1	$[26.5, 52.0]$	38.4	$[29.6, 49.9]$	39.6
0.9	65.3	$[42.2, 101.0]$	57.3	$[42.0, 78.4]$	—

[a] Estimated quantiles and 95% confidence intervals

2.3.4 Choice of storage times

Since the Turnbull estimator may put positive probability mass only on the interval endpoints, it is obvious that a finer grid of storage times can improve the nonparametric and parametric fit and hence the inference based on these. However, if the values of the rejection curve between the chosen storage times were not of interest, the data could be treated as grouped data and the estimator for the survival curve for life tables could be applied; see, for example, Collett (1994, Sec. 2.1.1).

The choice of the maximum storage time depends on the quantiles of interest. These should be covered by the nonparametric fit. For example, if one wants to estimate only the 10%- or 20%-quantiles, he or she will choose the storage times such that these will definitely be estimated nonparametrically. Parametric fits can most probably be find to coincide well with the Turnbull estimator within this range and possible differences beyond it will not be of interest.

2.3.5 Survival analysis vs. logistic regression

Vaisey-Genser, Malcomson, Przybylski, Eskin, and Armstrong in 1994 present a study on consumers' acceptance of canola oils. Their approach to estimate the shelf lives is the use of logistic regression. For this purpose, they fit a logistic curve to the probabilities of rejection at each of the storage times. To illustrate this, in case of the example in Table 2.9, these probabilities would amount to: $\mathrm{P}(T \leq t_1) = 0.4, \mathrm{P}(T \leq t_2) = 0.2, \ldots, \mathrm{P}(T \leq t_6) = 0.8$.

This procedure is asymptotically equivalent to the maximum likelihood estimation of the parameters of the following logistic regression model as long as the probabilities are neither equal to 0 nor equal to 1 (Hosmer and Lemeshow 1989):

$$p = \mathrm{P}(Y = 1) = \frac{\exp(\alpha + \beta t)}{1 + \exp(\alpha + \beta t)}, \tag{2.4}$$

where T is the storage time until rejection of the product and the response variable Y is defined as follows:

$$Y = \begin{cases} 1 & \text{Rejection} \\ 0 & \text{Acceptance} \end{cases}.$$

That is, in model (2.4), the rejection of the food product is the response variable and the storage time is a covariant, whereas in the survival analysis approach by means of model (2.3) the shelf life is the response variable.

For the following reasons, we believe that the use of the survival analysis methodology is a more powerful tool to evaluate the shelf lives of food products and, hence, the adequate approach to estimate the shelf lives of food products:

1. If a consumer is given several samples, model (2.4) would have to account for the dependencies of the observations, which would increase the variance of the parameter estimates. In contrast with that, using survival analysis, all the observations of a consumer reduce to a single interval.

   This problem of dependent observations could be overcome with current status shelf life data, that is, when each consumer is given only a single sample instead of one sample of each storage time. However, the number of consumers would have to be multiplied, in order to obtain the same number of observations for each storage time.

2. The fit of model (2.3) allows for several distributions, such as Weibull, log normal or log logistic, whereas the rejection curve under model (2.4) is always a logistic curve. In Section B.1 on page 123, we show that the logistic regression fit is equivalent to the use of a logistic survival model for current status data.

3. The logistic regression model implies a probability for the rejection at time zero equal to $\mathrm{P}(Y = 1|T = 0) = \frac{\alpha}{1+\alpha} \geq 0$, whereas for logarithmic distributions such as the Weibull, log normal or log logistic distribution, we have $p(0) = 0$. The latter corresponds with the assumption that the fresh sample is not to be rejected.

This last property concerning the rejection curves is illustrated in Figure 2.4 on the following page. Given the data on the shelf life of strawberry-flavored yogurt, the figure shows the estimation of the rejection curves for three different parametric assumptions: the Weibull, the log normal and the logistic distribution. In case of the latter distribution, the proportion of rejection of the fresh sample is about 10%.

So far, the survival analysis approach for shelf life data has not considered any other covariates. However, model (2.3) can easily accommodate such variables:

$$\ln(T) = \mu + \boldsymbol{\beta}'\boldsymbol{X} + \sigma W,$$

where $\boldsymbol{X}$ represents the vector of the model covariates and $\boldsymbol{\beta}$ quantifies their effect on the shelf life T. Typical covariates applied to trials on the shelf lives on yogurt are, for example, flavor or the fat degree.

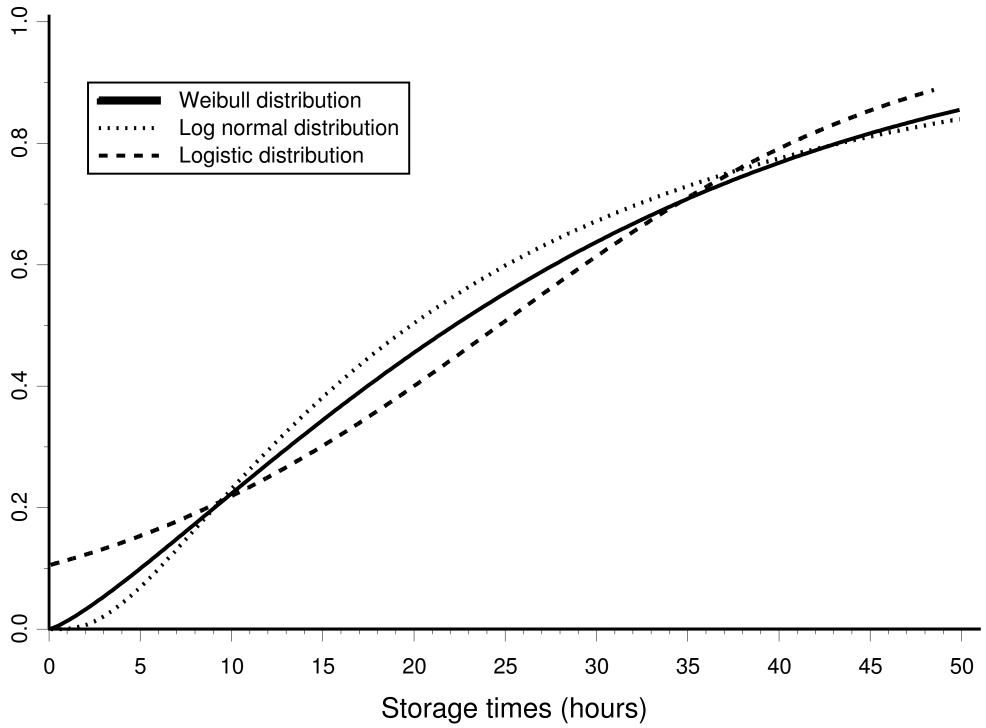


Figure 2.4: Probability of rejection of strawberry-flavored yogurt