Survival Analysis

References: Rupert G. Miller Jr., "Survival Analysis", Wiley 1981.

John Kalbfleisch and R.L. Prentice, "The Statistical Analysis of Failure Time Data", Wiley, 1980

Miller, Efron, et. al, "Biostatistics Casebook", Wiley, 1980.

Venables and Ripley, Chapter 11.

Cryer, J. D., "Time Series Analysis", Duxbury Press, 1986.[1]

Aim

For both of us (**you** and I, to learn something about the problems and methods of survival analysis. We will eventually take some data sets and attempt to do at least some rudimentary, and perhaps somewhat complicated, analyses on them.

Problem

We observer a **positive valued** random variable (the survivor time or failure time) and some co-variates. Typically the value of the random variable is the time to failure of some physical component or the time to death of a biological unit (or the time to relapse of some disease, etc.). There are all sorts of examples of waiting times which can be similarly analyzed—e.g., the length of time required to learn a new skill.

Early work in this area was some of the first (modern) statistics ever recorded. Graunt and Edmund Halley were responsible for the first useful **life tables** (estimate of the force of mortality) in about 166X. Medical and insurance applications carried the work steadily forward.

The modern era was ushered in by the need to test mechanical components for reliability. Rapid industrial expansion in the early 1940's was a major impetus.

The most novel feature of failure-time data is that notion of censoring—not all sample units (e.g., patients in a medical trial) will be observed to failure (e.g., death), yet even the "partially" observed data provide useful information on the lifetime distribution.

Example Carcinogenesis. Two groups of rats were given pre-treatments then exposed to a carcinogen. We would like to compare the pre-treatments. Presumably, we could transform the data and perform a two-sample t-test, except there is some censoring in the data. Some of the rats died of causes unrelated to the cancer, or survived the study.

Days to Vaginal Cancer Mortality in Rats

Group 1	143	164	188	188	190	192	206	209	213	216
	220	227	230	234	246	265	304	216*	244*	
Group 2	142	156	163	198	205	232	232	233	233	233
	233	239	240	261	280	280	296	296	323	204*
	344*									

The items with (*) are those that are censored. This is like a form of missing data except the "missing-ness" of the data contains information about the questions of interest.

[1]This is the most readable of these texts.

Example Accelerated Life Test. Many physical components are **designed** to be very reliable and have long lifetimes. (and then there are the other components). It is still of interest to know something about the expected lifetime of a component or a comparison between differing designs. When the lifetime is very long, it is often difficult to actually test the component. (Think about a computer chip. If you tested the lifetime in the standard way, the chip would be obsolete before you saw any failures).

In these circumstances and "accelerated life test" is often performed. This consists of an experiment where the components are subjected to *more* stress than would be expected under normal operating conditions with the aim of extrapolating or comparing.

For a certain type of motor designed to operate ate 130 degrees C, the following severely censored data were obtained.

150C All 10 motorettes without failure at 8064 hours

170C 1764, 2772, 3444, 3542, 3780, 4860, 5196, 3 motorettes without failure at 5448 hours

190C 408, 408, 1344, 1344, 1440, 5 motorettes without failure at 1680 hours

220C 408, 408, 504, 504, 504, 5 motorettes without failure at 528 hours

How would you analyze these data? Suppose you wanted to compare two types of motor?

African Children A Stanford psychiatrist wanted to know the age at which a group of African children learned to perform a particular task—so he wrote a grant proposal and went to Africa. When he arrived in the village some of the children could already perform the task, some children learnt while he was there and some hadn't learnt when he left. In this case there is both left and right censoring. For some individuals we know an age by which they could perform the task (left censoring); for some we know an age at which they could not perform the task, and for the rest we know the age at which they learnt the task.

Survival Functions and Hazard Rates

Let the random variable $T > 0$ denote survival time. Let T have density $f(t)$ and distribution

$$F(t) = \text{Prob}(T < t)$$

(Note that some books use $F(t) = \text{Prob}(T > t)$ which makes sense but is not standard).

The survival function is

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

and the hazard rate (function) is

$$\lambda(t) = \frac{f(t)}{1 - F(t)}$$

which can be interpreted as

$$\lambda(t) = \frac{\text{Prob}(t < T < d + \Delta t | T > t)}{\Delta t}$$

which can be read as

$$\text{Prob}\{\text{dying in}(t, t + \Delta t)|\text{lived to time } t\}$$

or the instantaneous force of mortality.

Now, notice that if we integrate $\lambda(t)$ we get

$$\begin{aligned}\int_0^t \lambda(u)du &= \int_0^t \frac{f(u)}{1-F(u)}du \\ &= -\log\left[1-F(u)\right]_0^t \\ &= -\log\left(S(t)\right)\end{aligned}$$

Thus

$$S(t) = \exp\{-\int_0^t \lambda(u)du\}$$

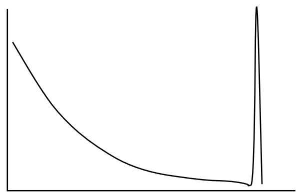
These various forms, definitions and **ideas** are very important. If you don't understand them now, stop me.

Types of Censoring

Let $T_1, \ldots, T_n$ be iid F.

- Type I Censoring. Let t_c be some preassigned fixed number, the fixed censoring time. Instead of seeing $T_1, \ldots, T_n$ we get to see

  $$Y_i = \min\left(T_i, t_c\right)$$

  so the density of Y might look like 

- Type II Censoring. Let $r < n$ be fixed. All we observe are $T(1), \ldots, T(r)$.

Both types of censoring are common—either we have only a certain amount of time in which to experiment (I) or we don't know how long we need to go but know how what percentage of failures we need to wait for (II).

- III Random Censoring. Let $C_1, \ldots, C_n$ be iid G were C_i is the censoring time associated with T_i. We observe

  $$Y_i = \min(T_i, C_i)$$

  and

  $$\delta_i = \mathrm{I}(Ti \leq C_i) = \begin{cases} 1 & \text{if } T_i \text{ not censored} \\ 0 & \text{if } T_i \text{ is censored} \end{cases}$$

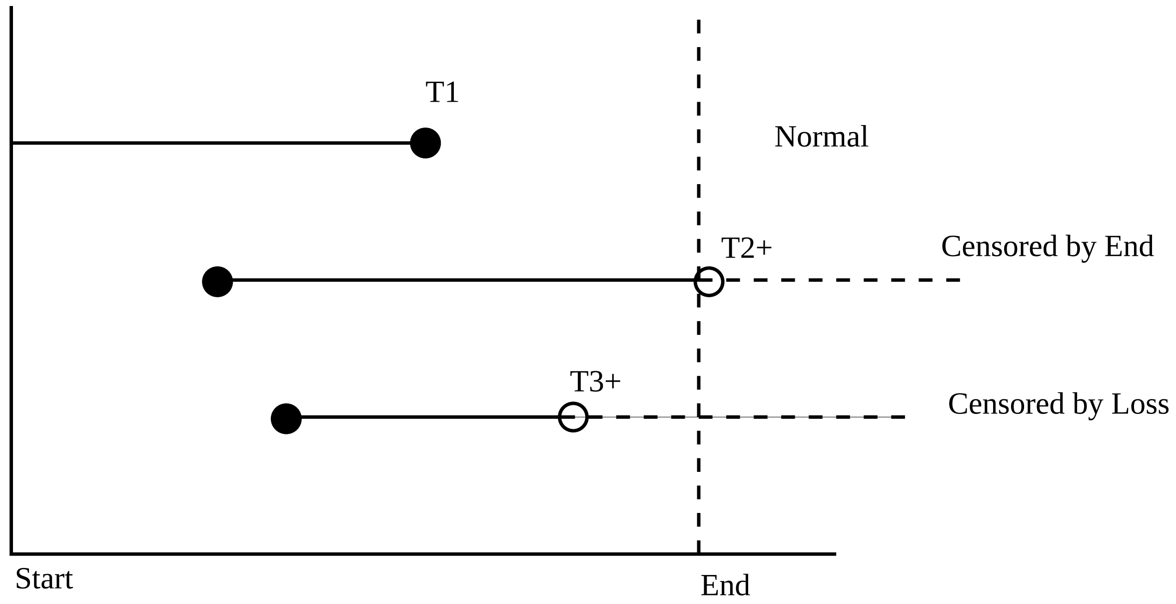
Many medical examples have people enter the study at different times by and the exit the study because of

1. loss of follow up (they move, etc.)

2. drop out (side effects, contrariness)

3. end of study.

The following picture shows what might happen in a study.



One might also imaging left censoring, where the start date is only know to be after a certain date.

Some distributions for Survival Analysis

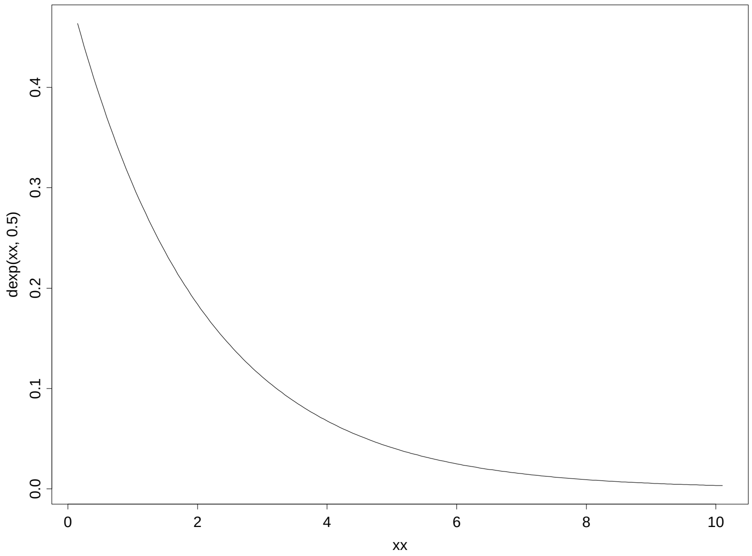
Exponential This is the basic model with constant risk, $\lambda(t) = \lambda > 0$. Thus

$$\int_0^t \lambda(u)du = \lambda \times t$$

and so

$$\begin{aligned} S(t) &= \exp \int_0^t \lambda(u)du = e^{-\lambda t} \\ F(t) &= 1 - e^{-\lambda t} \\ f(t) &= \lambda e^{-\lambda t} \end{aligned}$$

and thus $E(T) = 1/\lambda$ and $\text{Var}(T) = 1/\lambda^2$. Here is a plot of the density.



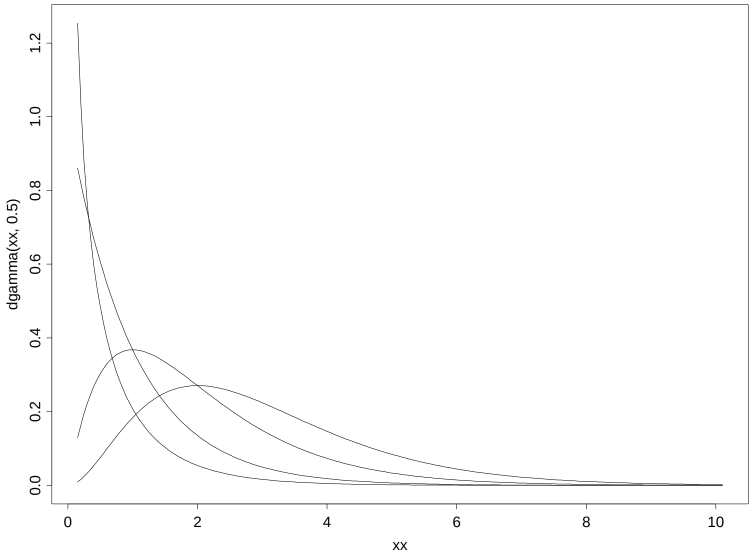
A reasonable thing to do might be to plot $\log(\hat{S}(t))$ against t. It should be a straight line.

Gamma One possible generalization of the exponential.

$$f(t) = \frac{\lambda^\alpha}{\Gamma(\alpha)} t^{\alpha-1} e^{\lambda t} \quad \alpha > 0 \quad \lambda > 0$$

$$\begin{aligned} F(t) &= \text{Incomplete Gamma Integral, no closed form} \\ \lambda(t) &= 1 - e^{-\lambda t} \end{aligned}$$

$E(T) = \alpha/\lambda$ and $\mathrm{Var}(T) = \alpha/\lambda^2$. Here is a plot of the density, for $\alpha = 0.5, 1, 2, 3$



You really already know a lot about the Gamma distribution because

$$\chi^2_k = \text{Gamma}(k/2, 1/2)$$

that is, $\alpha = k/2$ and $\lambda = 0.5$.

Weibull Another generalization of the exponential.

$$S(t) = e^{-(\lambda t)^\alpha} \quad \alpha > 0 \quad \lambda > 0$$

This can be forced to look a LOT like a gamma distribution. The Weibull is very important, but we will not dwell on it.

With complete (i.e., uncensored) data estimation of any of these distributions is relatively easy. Just write out the likelihood equations and differentiate. For some of the models there are even closed-form estimates. We will defer looking at the parametric ML estimates until we have looked at a non-parametric estimate of the survivor function.

Estimation of the Survivor Function

(Without parametric models).

Remember what the survivor function is, just the probability of surviving to be older than time t. A convenient and familiar object is the sample cumulative distribution function (cdf), written as

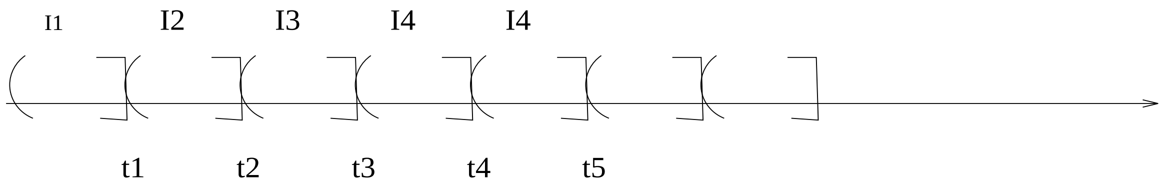
$$\bar{F}_n(x) = \frac{\#\text{ of obs} \leq x}{n}$$

Estimates of the sample survival function are obviously of the form, $1 - \bar{F}_n(x)$. In the analysis of survival data it is often useful to summarize the survival experience in terms of $\bar{F}_n(x)$ or $\bar{S}_n(x)$. (Why? Well it might direct us to an appropriate parametric model, or it might be interesting in its own right). When the data is censored we need to think of something other than the simple cdf.

Life Tables The classical method of estimating $S(t)$ is the so-called actuarial method which depends on the life table.

Let time be partitioned into a fixed sequence of intervals, $I_1, \ldots, I_k$ mostly of equal length. The intervals should be open at the left and closed at the right.

Consider the line



with

$$\begin{aligned}
n_i &= \#\text{ alive at beginning of } I_i \\
d_i &= \#\text{ died in } I_i \\
l_i &= \#\text{ lost to follow-up in } I_i \\
w_i &= \#\text{ withdrawn in } I_i \\
p_i &= P(\text{ surviving through } I_i |\text{ alive at start of } I_i
\end{aligned}$$

As an example, look at the data below, where we would like to estimate $S(5)$ say.

Years after diag	Alive at start	Died	Lost to Followup	Withdrawn
0–1	126	47	4	15
1–2	60	5	6	11
2–3	38	2	—	15
3–4	21	2	2	7
4–5	10	—	—	6

Reduced Sample Method To estimate $S(\tau_k)$ we only use those subjects who are at risk for the entire period $(0, \tau_k]$. That is then

$$\begin{array}{rcl} n &=& n_1 - \sum_{i=1}^k l_i - \sum_{i=1}^k w_i \\ d &=& \sum_{i=1}^k d_i \\ \text{and } S(\tau_k) &=& 1 - d/n \end{array} \qquad \begin{array}{rcl} n &=& 126 - 12 - 54 = 60 \\ d &=& 56 \\ &=& 1 - \frac{56}{60} = 0.078 \end{array}$$

What (if anything) is wrong with this? It throws away information, all the iformation in l_i and w_i, and it is thus an inefficient (and biased) estimate.

Actuarial Method The survival probability $S(\tau_k)$ can be broken into a product of probabilities,

$$\begin{aligned} S(\tau_k) &= P(T > \tau_k) \\ &= P(T > \tau_1) \cdot P(T > \tau_1 | T > \tau_2) \cdots P(T > \tau_k | T > \tau_{k-1}) \\ &= p_1 \cdot p_2 \cdot p_3 \cdots p_k \end{aligned}$$

where

$$p_i = P(T > \tau_i | T > \tau_{i-1})$$

The actuarial method estimates each p_i separately and multiplies to get $S(\tau_k)$. To estimate p_i one could use

$$1 - \frac{d_i}{n_i}$$

, but this ignores losses and withdrawals in interval I_i. A sensible(?) approach might be to assume that on average those that were lost or withdrew were at risk for 1/2 of the interval. Thus one might define the effective sample size as

$$n_i' = n_i - 1/2 \cdot (l_i - w_i)$$

with

$$\begin{aligned} \hat{q}' &= \frac{d_i}{n_i'} \\ \hat{p}_i &= 1 - q_i \\ \text{and } \hat{S}(\tau_k) &= \prod_{i=1}^k (\hat{p}_i) \end{aligned}$$

For the above data, this yields,

$$\hat{S}(5 \text{ years}) = 0.44$$

(You work this out!)

Variance of $\hat{S}(\tau_k)$ By assuming that

$$n_i'\hat{p}_i \approx \text{Binomial}(n_i', p_i)$$

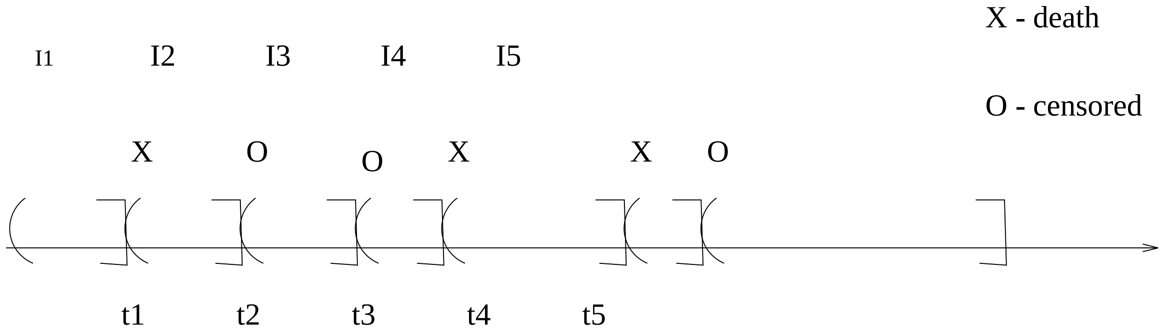
and some manipulations an estimate of the variance of the estimate can be found as

$$\widehat{\text{Var}}(\hat{S}(\tau_k)) = \hat{S}^2(\tau_k)\sum_{i=1}^{k}\frac{d_i}{n_i'(n_i'-d_i)}$$

This is known as Greenwood's formula (named after Major Greenwood).

The Kaplan-Meier Product Limit Estimate

This is very similar to the actuarial method we just saw, except that the intervals I_i are not fixed in advance. In fact the intervals are determined by when an event occurs. So a revised picture might be,



Recall that we observe $\begin{array}{ll} Y_1,\ldots,Y_n & \text{survival times, and} \\ \delta_1,\ldots,\delta_n & \begin{pmatrix} 1 & \text{censored} \\ 0 & \text{not censored} \end{pmatrix} \end{array}$

Order the Y_i into $\begin{array}{c} Y_{(1)},\ldots,Y_{(n)} \\ \delta_{(1)},\ldots,\delta_{(n)} \end{array}$ and carry only the δ_i

Let $\mathcal{R}(t)$ be the risk set at t and

$$\begin{aligned} n_i &= \#\text{ in } \mathcal{R}(Y_{(i)}) = \#\text{ alive just before } Y_{(i)} \\ d_i &= \#\text{died at } Y_{(i)} \\ p_i \text{ and } q_i &= \text{same as before—conditional probability of surviving} \end{aligned}$$

Then

$$\hat{p}_i = \begin{cases} 1 - 1/n_i & \text{if} \quad \delta_{(i)} = 1 \\ 1 & \text{if} \quad \delta_{(i)} = 0 \end{cases}$$

Thus, without ties, the PL estimate is

$$\begin{aligned} \hat{S}(t) &= \prod_{y(i)\le t}\hat{p}_i = \prod (1 - 1/n_i)^{\delta(i)} \\ &= \prod (1 - \frac{1}{(n-i+1)})^{\delta(i)} \end{aligned}$$

Notes

1. Tied Uncensored observations—do the obvious thing which is $p_i = 1 - \frac{d}{m}$ where d deaths occur.

2. Censored and uncensored tied data—consider the uncensored data to be just before the censored data. (Why in this order?)

Two Views of the PL estimate as an MLE

The PL estimate would not be nearly so popular if it could not be interpreted in other ways.

Let $t_1, \ldots, t_k$ be the observed failure times in a sample of size n_0 with survivor function, $S(t)$. Suppose d_j items fail at time t_j and m_j items are censored in $[t_j, t_{j+1})$ at known times t_{j_l}.

Let $n_j = (m_j + d_j) + \cdots + (m_k + d_k)$ be the number of items at risk at time just before t_j. The probability of failure at t_j is

$$S(t_j) - S(t_j + 0)$$

We assume that the contribution to the likelihood of a survival time censored at t_{j_l} is $P(T > t_{j_l}) = S(t_{j_l} + 0)$.

Thus,

$$\mathcal{L} = \prod_{j=0}^{k} \left\{ [S(t_j) - S(t_j + 0)]^{d_j} \cdot \prod_{l=1}^{m_j} S(t_{j_l} + 0) \right\}$$

Given the data, this can be viewed as a likelihood function on the space of all survivor functions, $S(t)$. There are problems with this definition of the MLE, but the intuition is good.

Clearly to maximize $\mathcal{L}$ one needs and S which is *discontinuous* at the observed failure times—otherwise $\mathcal{L}$ is zero.

Further, $\mathcal{L}$ is maximized when $S(t_{j_l} + 0) = S(t_j + 0)$

An estimate $\hat{S}(t)$ is then a discrete survivor function with hazard components $\hat{\lambda}_i$ at t_i.

Thus,

$$\hat{S}(t_j) = \prod_{l=1}^{j-1} (1 - \hat{\lambda}_l)$$

and

$$\hat{S}(t_j + 0) = \prod_{l=1}^{j} (1 - \hat{\lambda}_l)$$

where the λ are chosen to maximize the function

$$\begin{aligned} & \prod_{j=1}^{k} \left\{ \left[\prod_{l=1}^{j-1} (1 - \lambda_j)^{d_j} \right] \lambda_j^{d_j} \left[\prod_{l=1}^{j} (1 - \lambda_l)^{m_j} \right] \right\} \\ = & \prod_{j=1}^{k} \lambda_j^{d_j} (1 - \lambda_j)^{n_j - d_j} \end{aligned}$$

because

$$S(t_j) - S(t_j + 0) = \prod_{l=1}^{j-1} \lambda_j (1 - \lambda_j)$$

Obviously the maximum is achieved when $\hat{\lambda}_j = \frac{d_j}{n_j}$ and the product limit estimate is

$$\hat{F}(t) = \prod_{\substack{j \\ t_j < t}} \left(\frac{n_j - d_j}{n_j} \right)$$

Interpretation We are in effect making the conditional probability of failure at each t_j agree exactly with the observed conditional relative frequency of failure at t_j.

Problem: If we assume that $S(t)$ is continuous then $\hat{S}(t)$ is always bumpy, which might give you cause for concern. (This can all be fixed by appropriate technical mumbling).

Example

No Censoring Assume that $n = 5$ and $x =$ death and $o =$ censor with

$$\begin{array}{cccccc} & x & x & x & x & x \\ p_i & (1-\frac{1}{5}) & (1-\frac{1}{4}) & (1-\frac{1}{3}) & (1-\frac{1}{2}) & (1-\frac{1}{1}) \end{array}$$

where p_i is the p(survive to age i|alive before time$i$)

$$\begin{array}{rcccccc} S_i & = & \frac{4}{5} & \frac{4}{5}\cdot\frac{3}{4} & \frac{4}{5}\cdot\frac{3}{4}\cdot\frac{2}{3} & \ldots & \\ & = & \frac{4}{5} & \frac{3}{5} & \frac{2}{5} & \frac{1}{5} & 0 \end{array}$$

Thus, without censoring the PL estimate is the obvious one

Some censored data Let $n = 7$ and take the data

$$\begin{array}{ccccccc} & x & x & o & x & o & x \\ p_i & \frac{6}{7} & \frac{5}{6} & & \frac{3}{4} & & \frac{1}{2} \end{array}$$

The PL estimate is

$$\begin{array}{rcccccccc} S_i & = & \frac{6}{7} & \frac{6}{7}\cdot\frac{5}{6} & \frac{6}{7}\cdot\frac{5}{6} & \frac{6}{7}\cdot\frac{5}{6}\cdot\frac{3}{4} & \ldots & & \\ & = & \frac{6}{7} & \frac{5}{7} & \frac{5}{7} & \frac{15}{28} & \frac{15}{28} & \frac{15}{56} & \frac{15}{56} \end{array}$$

How does one compare two survival curves?

That is a very good question. One answer is to say, just use the Splus function `survdiff`, which has a description of

> Test if there is a difference between two or more survival curves using the G-rho family of tests. Or, test for a single curve against a known alternative.

But, to really understand what this is all about, we need a digression.

Digression on some Nonparametric Techniques

Wilcoxen or Mann-Whitney Test

Suppose we have samples from two distributions F and G and we wish to compare the distributions. Let $X_1, \ldots X_m$ and $Y_1, \ldots Y_m$ denote the random samples from the two distributions.

One test of the difference between the distributions proceeds as

1. Rank all the data—i.e., mix the two samples.
2. Let $R(X_i)$ denote the ranks of the F sample in the combined data.
3. Form $S = \sum_{i=1}^{m} R(X_i)$, the rank sum.

Now note that S will be large if the F-distribution has larger values than the G-distribution and vice-versa. It is easy to show that

$$\mathrm{E}(S) = \frac{m(n+m+1)}{2}$$

and

$$\mathrm{Var}(S) = \frac{m(n+m+1)n}{2}$$

(This comes from assuming that $F = G$ in which case the ranks are just sampled form a discrete uniform distribution.

For small values of n and m there are tables of exact tail probabilities for S (or some linear transformation of it).

For large samples, just take a normal approximation

$$z = \frac{S - \mathrm{E}(S)}{\sqrt{\mathrm{Var}(S)}}$$

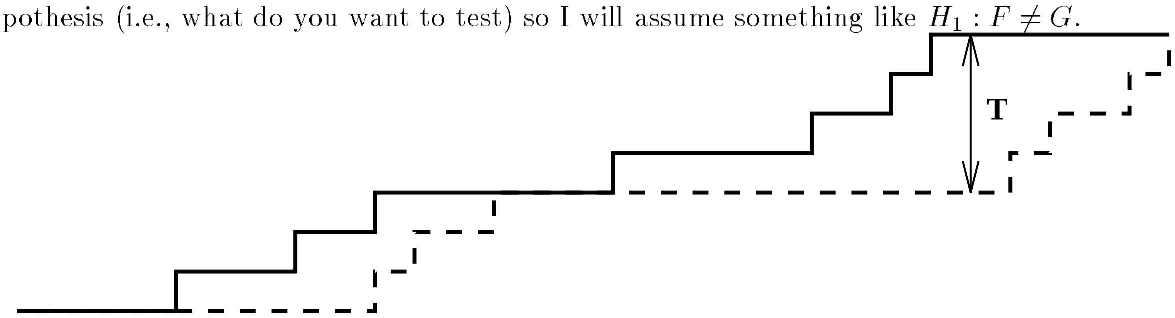
The test makes minimal assumptions and works quite well in certain situations? However, what is the test sensitive to? When normality can be assumed, the 2-sample t-test is a competitor and like most specific purpose tools does better than the Wilcoxon test.

You can find this test as `wilcox.test` in Splus.

Kolmogorov-Smirnov test

The previous test is not very sensitive to differences in the distribution shape, as opposed to solely location differences. A test which is sometimes useful here is the (Kolmogorov-)Smirnov test.

Let $\hat{F}$ and $\hat{G}$ be estimates of two distribution functions. (For censored data one might think about using the Kaplan-Meier estimates). The exact form of the test depends on the alternate hypothesis (i.e., what do you want to test) so I will assume something like $H_1 : F \neq G$.



Now form $T = \max|F - G|$, and look up the results in a table, OR just use the Splus function `ks.gof`.

The test is sensitive to all sorts of departures, particularly changes in tails and different variances.

End Digression

So how does one compare two Kaplan-Meier estimates? One of the things that `survdiff` does is a (version) of Gehan's test, which itself is just a modification of the Wilcoxon test.

Instead of using S, the rank-sum, we need to develop another equivalent statistic.

Let

$$U(X_i, Y_j) = \begin{cases} +1 & \text{if} \quad X_i > Y_j \\ 0 & \text{if} \quad X_i = Y_j \\ -1 & \text{if} \quad X_i < Y_j \end{cases}$$

Let $U = \sum_{i=1}^{m} \sum_{j=1}^{n} U_{i,j}$, which can be read as the average number of times that $X > Y$. (Would $\sum_{i=1}^{m} \sum_{j=1}^{n} |U_{i,j}|$ make sense?).

It isn't hard to see that

$$S = \frac{m(m+n+1)}{2} = \frac{1}{2}U$$

thus $\mathrm{E}(U) = 0$ and $\mathrm{Var}(U) = \frac{nm(n+m+1)}{3}$.

For censored data, the Gehan test suggests defining $U_{i,j}$ in a different form.

Our basic data is (X_i, δ_i) and (Y_i, ϵ_i) where δ and ϵ are the indicators of censoring.

The null hypothesis is $F_1 = F_2$ (distribution) and $G_1 = G_2$ (censoring), with alternate interest in $F_1 \neq F_2$. (I am not sure what happens when the censoring is different).

Now define

$$U_{i,j} = \begin{cases} +1 & \text{if} & \begin{array}{l} \text{we know that } t_i > u_j \\ \text{i.e., } (x_i > y_j, \epsilon_j = 1) \text{ or } (x_i = y_j, \delta_i = 0, \epsilon_i = 1) \end{array} \\ 0 & \text{otherwise} & \\ -1 & \text{if} & \begin{array}{l} \text{we know that } t_i < u_j \\ \text{i.e., } (x_i < y_j, \delta_j = 1) \text{ or } (x_i = y_j, \delta_i = 1, \epsilon_i = 0) \end{array} \end{cases}$$

and $U = \sum_{i=1}^{m} \sum_{j=1}^{n} U_{i,j}$

It is possible to show (via some permutation theory) that $\mathrm{E}(U) = 0$ and $\mathrm{Var}(U) =$ a big mess, but there is a computational formula, and packages.

Digression on the Mantel-Haenszel Test

Suppose there are two populations and each individual in each population can have one of two characteristics. For example, two populations of cancer patients with different treatments who can either survive beyond a year or die within the year. We could write out the table of data as,

	Dead	Alive	
Pop1	a	b	n_1
Pop2	c	d	n_2
	m_1	m_2	n

Let

$$\begin{aligned} p_1 &= P(\text{Dead}|\text{Pop1}) \\ p_2 &= P(\text{Dead}|\text{Pop2}) \end{aligned}$$

A test of $p_1 = p_2$ can be based on (thing that we already know about homogeneity of proportions) or the statistic

$$\chi^2 = \left[\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}} \right]^2 = \frac{n(ad - bc)^2}{n_1 n_2 m_1 m_2}$$

where $\hat{p}_1 = \frac{a}{n_1}$, $\hat{p}_2 = \frac{c}{n_2}$, and $\hat{p} = \frac{m_1}{n}$. The χ^2 is approximately distributed as a χ_1^2 random variable.

The intuition for this test comes (again) from the 2-sample t-test, but the test itself is derived from a hypergeometric distribution for the contingency table.

$$P(A = a) = \frac{\binom{n_1}{a}\binom{n_2}{m_1 - a}}{\binom{n}{m_1}}$$

This test assume that the n and m margins are BOTH fixed in advance.

Under the hypergeometric distribution (which is our null)

$$\mathrm{E}(A) = \frac{n_1 m_1}{n}$$

and

$$\mathrm{Var}(A) = \frac{n_1 n_2 m_1 m_2}{n^2(n-1)}$$

whence

$$ad - bc = n(a - \mathrm{E}(A)$$

and

$$n_1 n_2 m_1 m_2 = n^2(n-1)\mathrm{Var}(A)$$

thus

$$\chi^2 = \frac{n}{n-1} \left[\frac{a - \mathrm{E}(A)}{\sqrt{\mathrm{Var}(A)}} \right]^2$$

For the example above, the hypergeometric assumption is suspect. If I collected the data by finding m_1 deaths and m_2 survivors and then recorded their populations, then it is probably OK. On the other hand, if I just start out with two populations and watch for deaths a binomial (or product binomial) seem more reasonable. But the particular form of the χ^2 test is (I think) just the same as the binomial one would have been, modulo a fudge about n or $n - 1$.

The real reason that the hypergeometric is used here is based on the statistical notion of ancilarity

Continue Digression, a sequence of 2×2 tables

The same setup as before, but now there are K hospitals. Because there may be differences between hospitals we don't wish to combine all the 2×2 tables. Rather we would like to test

$$p_1^1 = p_2^1, p_1^2 = p_2^2, \ldots, p_1^k = p_2^k,$$

i.e., equality in each population, but not overall equality.

The data are now

	Dead	Alive	
Pop1	a_1	b_1	$n_1 1$
Pop2	c_1	d_1	$n_1 2$
	$m_1 1$	$m_1 2$	n_1

$\cdots$

	Dead	Alive	
Pop1	a_k	b_k	$n_k 1$
Pop2	c_k	d_k	$n_k 2$
	$m_k 1$	$m_k 2$	n_k

The Mantel-Haenszel statistic for this problem is

$$\mathrm{MH} = \frac{\sum_{i=1}^{k}(a_i - \mathrm{E}(A_i))}{\sqrt{\sum_{i=1}^{k}(\mathrm{Var}(A_i))}}$$

If the tables are independent then MH tends towards a $N(0,1)$ when either k is fixed and $n_i \to \infty$ or $k \to \infty$ and the tables are iid.

Of course we could use our contingency table techniques to answer these questions—but you end up with slightly different tests.

When used in survival analysis, neither of these conditions for asymptotic normality holds—but the result still does.

End Digression

Kaplan-Meier in Survival Analysis

In survival analysis this test is used in the following manner. Let $\begin{matrix} X_1, \ldots, X_n \\ \delta_1, \ldots, \delta_n \end{matrix}$ be one population and $\begin{matrix} Y_1, \ldots, Y_n \\ \epsilon_1, \ldots, \epsilon_n \end{matrix}$ be the other. Consider $Z_{(1)}, \ldots, Z_{(n+m)}$ to be the combined ordered populations. For each $Z_{(i)}$ which corresponds to a death (in either population) construct a 2×2 table and compute the MH test for this sequence of tables. (E.G., at time (i) we have

	Dead	Alive	
Pop1	a	b	n_1
Pop2	c	d	n_2

and at time (i+1) we have

	Dead	Alive	
Pop1	$a+1$	$b-1$	n_1
Pop2	c	d	n_2

etc.,)

In other words, test $F_1 = F_2$ (which is the same as $S_1 = S_2$) by using the MH test. (Independence of tables is violated here, but asymptotic normality of the test can still be pushed through).

What next

Before proceeding to other (maybe more complicated) matters, we need to ask if any of these ideas are useful or used in practice. The best I can do is present a section of the book "Biostatistics Casebook" by Miller, Efron, et. al. which shows a very simple yet convincing application.

Some handwritten material was inserted

Some examples, from Venables and Ripley

Look at the leukaemia data, which is the survival times and white blood counts for a group of leukaemia patients. The data has 33 leukaemia patients, and variables.

```
wbc:   white blood count
ag:    a test result, present or absent
time:      survival time in weeks
```

The Sources are D.R. Cox and D. Oakes (1984) *Analysis of Survival Data*, Chapman & Hall, p. 9, who themselves took the data from P. Feigl & M. Zelen (1965) *Estimation of exponential survival probabilities with concomitant information.* Biometrics, 21, 826-838.

The data are,

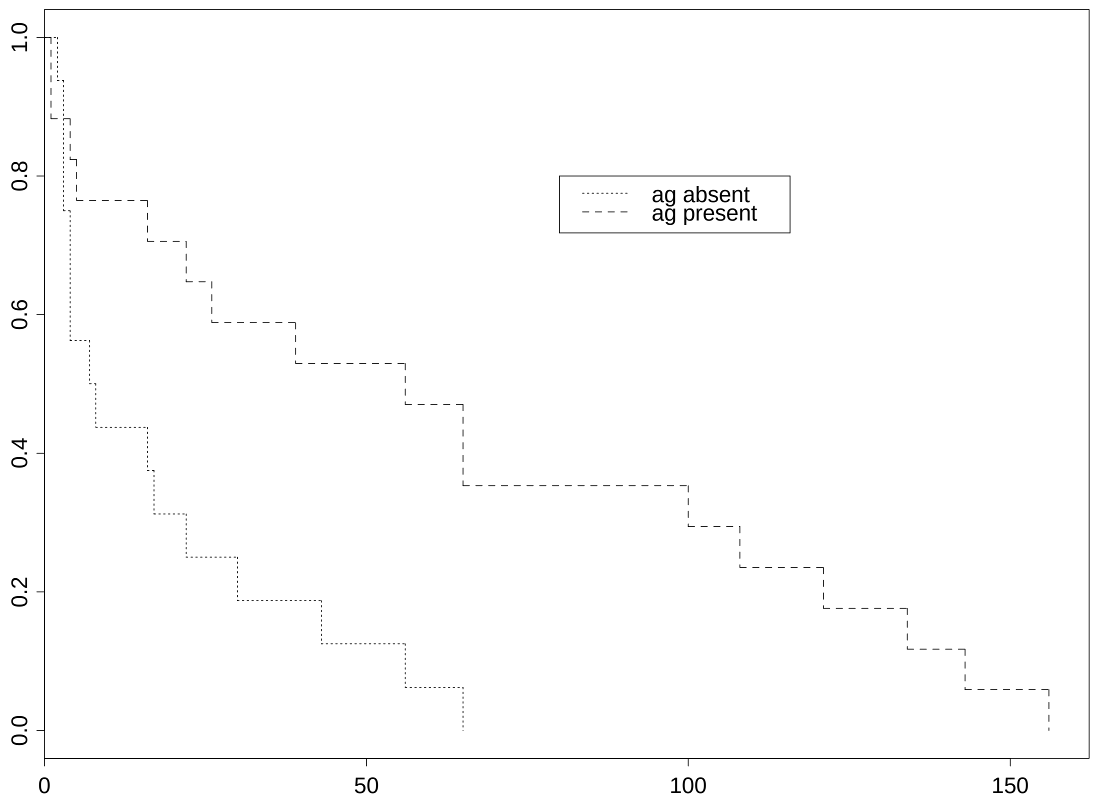
```
> leuk
      wbc      ag time
 1   2300 present   65
 2    750 present  156
 3   4300 present  100
 4   2600 present  134
 5   6000 present   16
 6  10500 present  108
 7  10000 present  121
 8  17000 present    4
 9   5400 present   39
10   7000 present  143
11   9400 present   56
12  32000 present   26
13  35000 present   22
14 100000 present    1
15 100000 present    1
16  52000 present    5
17 100000 present   65
18   4400  absent   56
19   3000  absent   65
20   4000  absent   17
21   1500  absent    7
22   9000  absent   16
23   5300  absent   22
24  10000  absent    3
25  19000  absent    4
26  27000  absent    2
27  28000  absent    3
28  31000  absent    8
29  26000  absent    4
30  21000  absent    3
31  79000  absent   30
```

```
32 100000   absent    4
33 100000   absent   43
```

The first thing one might wish to do with these data is plot the two survival curves, via

```
> attach(leuk)
> plot(survfit(Surv(time) ~ ag), lty=c(2,3))
```

where the `survfit` function, omputes an estimate of a survival curve for censored data using either the Kaplan-Meier or the Fleming-Herrington method or computes the predicted survivor function for a Cox proportional hazards model. In our case the model is just the two levels of `ag`. The picture looks like,



There are clearly differences in the curves, but if they are significant or not is not obvious. See V & R (page 269) for some suggestions on how to test this.

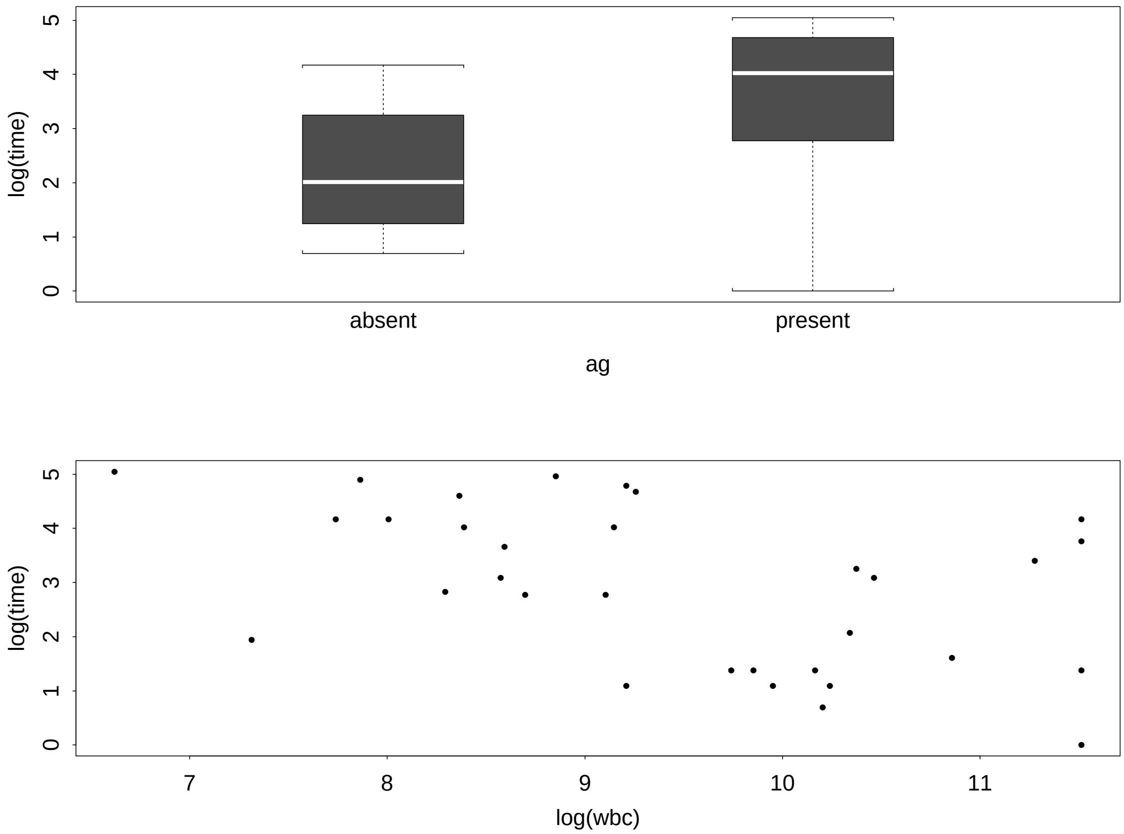
There are now many ways we could proceed with these data, including

- Parametric models based on Gamma errors and the `glm` function,
- Parametric models, based on the exponential and Weibull distributions using `survreg`, and
- The family of Cox models.

We will **not** do a lot of the theory, but will attempt to push the example through with just enough to make sense of what we are doing.

The next example

```
lwbc <- log(wbc)
par(mfrow=c(2,1))
plot(log(time) ~ ag + log(wbc))
options(contrasts=c("contr.treatment","contr.poly"))
leuk.glm <- glm(time ~ ag*lwbc, Gamma(log))
 print("## summary(leuk.glm, dispersion=1)")
print(summary(leuk.glm, dispersion=1)) # force exponential data summary
print("## anova(leuk.glm))")
print(anova(leuk.glm))
# drop interaction term
leuk.glm <- update(leuk.glm,  ~ .-ag:lwbc)
print("## summary(leuk.glm, dispersion=1))")
print(summary(leuk.glm, dispersion=1))
# compare with natural link
leuk.glmi <- glm(time ~ ag*wbc, Gamma(inverse))
print("## summary(leuk.glmi, dispersion=1))")
print(summary(leuk.glmi, dispersion=1))
print("## anova(leuk.glmi))")
print(anova(leuk.glmi))
```



```
[1] "## summary(leuk.glm, dispersion=1)"

Call: glm(formula = time ~ ag * lwbc, family = Gamma(log))
Deviance Residuals:
       Min       1Q     Median        3Q      Max
 -1.992227 -1.21155 -0.3270025 0.2160396 1.564344

Coefficients:
                 Value Std. Error   t value
(Intercept)  4.3291270  1.9661806  2.201795
         ag  4.1481711  2.5702903  1.613892
       lwbc -0.1525468  0.2026831 -0.752637
    ag:lwbc -0.3291857  0.2668885 -1.233420

(Dispersion Parameter for Gamma family taken to be 1 )

    Null Deviance: 58.13839 on 32 degrees of freedom

Residual Deviance: 38.55468 on 29 degrees of freedom

Number of Fisher Scoring Iterations: 6

Correlation of Coefficients:
        (Intercept)         ag       lwbc
     ag -0.7649644
   lwbc -0.9918835   0.7587556
ag:lwbc  0.7532658  -0.9907268 -0.7594297
[1] "## anova(leuk.glm))"
Analysis of Deviance Table

Gamma model

Response: time

Terms added sequentially (first to last)
        Df Deviance Resid. Df Resid. Dev
   NULL                    32   58.13839
     ag  1 11.94009        31   46.19829
   lwbc  1  5.87920        30   40.31909
ag:lwbc  1  1.76441        29   38.55468
[1] "## summary(leuk.glm, dispersion=1))"

Call: glm(formula = time ~ ag + lwbc, family = Gamma(log))
Deviance Residuals:
       Min        1Q     Median        3Q      Max
 -2.174855 -1.266119 -0.4246385 0.4961872 1.904819
```

```
Coefficients:
                 Value Std. Error   t value
(Intercept)  5.8134506  1.2931888  4.495438
         ag  1.0182978  0.3492227  2.915898
       lwbc -0.3042296  0.1318634 -2.307157

(Dispersion Parameter for Gamma family taken to be 1 )

    Null Deviance: 58.13839 on 32 degrees of freedom

Residual Deviance: 40.31909 on 30 degrees of freedom

Number of Fisher Scoring Iterations: 6

Correlation of Coefficients:
     (Intercept)         ag
  ag -0.2090766
lwbc -0.9811356   0.0720421
[1] "## summary(leuk.glmi, dispersion=1))"

Call: glm(formula = time ~ ag * wbc, family = Gamma(inverse))
Deviance Residuals:
       Min        1Q     Median        3Q      Max
 -1.890629 -1.299155 -0.4143898 0.3559901 1.909283

Coefficients:
                    Value   Std. Error     t value
(Intercept)  5.635111e-02 1.887885e-02  2.98488018
         ag -4.780969e-02 1.932473e-02 -2.47401586
        wbc -2.030302e-08 4.239835e-07 -0.04788634
     ag:wbc  5.926747e-07 5.232047e-07  1.13277789

(Dispersion Parameter for Gamma family taken to be 1 )

    Null Deviance: 58.13839 on 32 degrees of freedom

Residual Deviance: 39.73216 on 29 degrees of freedom

Number of Fisher Scoring Iterations: 6

Correlation of Coefficients:
       (Intercept)         ag        wbc
    ag -0.9769269
   wbc -0.6746535   0.6590872
ag:wbc  0.5467114  -0.6022141 -0.8103588
[1] "## anova(leuk.glmi))"
Analysis of Deviance Table
```

```
Gamma model

Response: time

Terms added sequentially (first to last)
       Df Deviance Resid. Df Resid. Dev
  NULL                    32   58.13839
    ag  1 11.94009        31   46.19829
   wbc  1  5.31608        30   40.88221
ag:wbc  1  1.15005        29   39.73216
```