Lecturer: James Degnan
Office: SMLC 342
Office hours: MW 12:00–1:00 or by appointment
E-mail: jamdeg@unm.edu
Please include STAT474 or STAT574 in the subject line of the email to make sure I don't overlook your email.

Textbook: *Survivial Analysis Techniques for Censored and Truncated Data* 2nd edition, by Klein and Moeschberger

Assessment: Grading will be based homework (roughly 5 assignments in the semester) (40%), three in-class tests (20% each).

We will mostly use R for computing. Solutions to homework can be done in any software package, but it will be easier for me to grade and give partial credit for homework done in R and SAS.

Homework

For turning in computer-based homework, **turn in all computer code used as an appendix only. Do not include computer code as part of your solutions**. Figures and tables can be generated from computer output, but solutions must be discussed separately from the output, and the results in the Figures and Tables should be cited in the homework solutions. This will be discussed further in class.

Late homework will be penalized 10% per day. All homework must be printed (not emailed) and turned in either in class or to my office. Sliding homework under my door is fine.

Topics in the course

The main topic for the course is survival analysis, which concerns estimating the probability of events occurring within a certain period of time, for example the probability that a cancer recurs within t years after treatment.

Other topics toward the end of the course might be covered as well such as genome-wide association tests, analysis of gene expression data, and controlling false discovery rates.

Survival analysis: applications

Survival data sets typically include the amount of time until some observed event, such as death, recurrence of a cancer or illness, cardiac failure and so forth. You could also have less depressing examples, such as the time until weaning for breast-feeding.

The techniques can be used for non-medical data, such as

- time until a battery fails
- time until wings in airplanes develop cracks
- time until a car needs the transmission replaced
- In behavioral applications, you might have a certain intervention (such as rehabilitation for drug-users, or time in a prison system), and time until the behavior (drug use, recidivism for prisons)
- in learning experiments, the time until a certain skill is learned

Survival data

Two goals of survival analysis include (1) being able to estimate how long until something occurs (how long a patient with a diagnosis is likely to live), and (2) being able to compare treatments. By looking at survival rates for two different treatments (e.g., chemotherapy versus chemotherapy plus radiation therapy), we might try to determine which treatment is better. This latter problem can involve hypothesis testing as well as estimation.

In industrial applications, you might try to determine whether a higher quality component has significantly better failure time rates compared to a cheaper component.

Survivial data: censoring

A key feature of survival data is that we often can't observe when something recurs. There are several reasons for this:

- an individual might drop out of a study (maybe they move to a different hospital system)
- the condition might not recur before the study ends, although the condition might recur sometime after the study ends
- something might prevent the condition from recurring, such as an individual dying from another cause

When the end-point of an observation can't be observed, then the data is called *censored*$^+$

Survival data: example with censoring

For this example, there were 101 patients with acute myelogenous leukemia reported to the International Bone Marrow Registry. Fifty-one patients received autologous bone marrow transplants, in which their own bone marrow was reinfused after chemotherapy. An additional 50 patients were given bone marrow transplants from a sibling. The question is to determine which method leads to better survival times. Here are the survival times in months:

The leukemia-free survival times for the 50 allo transplant patients were 0.030, 0.493, 0.855, 1.184, 1.283, 1.480, 1.776, 2.138, 2.500, 2.763, 2.993, 3.224, 3.421, 4.178, 4.441^{+}, 5.691, 5.855^{+}, 6.941^{+}, 6.941, 7.993^{+}, 8.882, 8.882, 9.145^{+}, 11.480, 11.513, 12.105^{+}, 12.796, 12.993^{+}, 13.849^{+}, 16.612^{+}, 17.138^{+}, 20.066, 20.329^{+}, 22.368^{+}, 26.776^{+}, 28.717^{+}, 28.717^{+}, 32.928^{+}, 33.783^{+}, 34.211^{+}, 34.770^{+}, 39.539^{+}, 41.118^{+}, 45.033^{+}, 46.053^{+}, 46.941^{+}, 48.289^{+}, 57.401^{+}, 58.322^{+}, 60.625^{+};

and, for the 51 auto patients, 0.658, 0.822, 1.414, 2.500, 3.322, 3.816, 4.737, 4.836^{+}, 4.934, 5.033, 5.757, 5.855, 5.987, 6.151, 6.217, 6.447^{+}, 8.651, 8.717, 9.441^{+}, 10.329, 11.480, 12.007, 12.007^{+}, 12.237^{+}, 12.401^{+}, 13.059^{+}, 14.474^{+}, 15.000^{+}, 15.461, 15.757, 16.480, 16.711, 17.204^{+}, 17.237, 17.303^{+}, 17.664^{+}, 18.092^{+}, 18.092, 18.750^{+}, 20.625^{+}, 23.158, 27.730^{+}, 31.184^{+}, 32.434^{+}, 35.921^{+}, 42.237^{+}, 44.638^{+}, 46.480^{+}, 47.467^{+}, 48.322^{+},56.086

Here the $^{+}$ symbol indicates censoring, meaning that at the end of the study, the patient hadn't died, and the amount of time given is the amount of time since the bone-marrow transplant.

The previous example is typical of survival data. Note that for this type of data, it might have been decided that they would keep track of all patients within the registry who had bone marrow transplants within a certain period of time, and that the study would end at another period of time. It's possible that not all patients were in the study for the same period of time (because some patients will have had bone marrow transplants closer to the end of the study than others).

A value of 12.007 means that the patient died 12.007 months after the transplant. Whereas 12.007^+ means that either the study ended 12.007 months after the patient had the transplant and the patient still hadn't died, or the patient dropped out of the study at that point but hadn't died. In other words, 12.007 means that the patient survived only 12.007 months whereas 12.007^+ means that the patient survived at least 12.007 months, and we don't know how much longer the patient survived.
You might notice that censored observations seem to be more common for larger values. Is this reasonable?

Censored data in the computer

The notation of putting a plus sign or an asterisk (e.g., 12.007^* instead of 12.007^+) is very common for indicating censored observation and is fairly compact, but is more difficult to work with in the computer. An easier approach in the computer is to to have two variables: one for the survival time, one to indicate whether or not the observation is censored. Yet another variable would be used to indicate the treatment group. Of course, other variables could be used, such as the sex of the patient and other covariates such as age, ethnicity, height, weight, time of first diagnosis, etc.

Another way of organizing the data might look like this:

Censored data in the computer

```
group time censored
allo 0.030 0
allo 0.493 0
...
allo 58.322 1
allo 60.625 1
auto 0.658 0
auto 0.822 0
...
auto 12.007 0
auto 12.007 1
...
auto 48.322 1
auto 56.086 0
```

The survival function

It's also possible to have data sets with no censoring. This might be more common in industrial applications where a product is tested until it fails, and all products fail eventually. Still it is conceptually useful to think about what would happen if we had all observations without any censoring.

What we would like to know is the probability of surviving longer than any given amount of time. If X is the random time of death, or failure or other event, we can define

$$S(x) = P(X > x)$$

as the survival function. Here X is a random variable and x is a particular value for the random variable.

The survival function and cumulative distribution function (CDF)

If we substract the survival function from 1, we get

$$1 - S(x) = 1 - P(X > x) = P(X \leq x) = F(x)$$

where $F(x)$ is the cumulative distribution function (CDF) of the random variable X. In survival analysis, it is traditional to work with $S(x)$ instead of $F(x)$.

The CDF

Since the survival function and CDF are so closely related, it would be good to review cumulative distribution functions.

What properties do you remember of CDFs?

Properties of CDFs

For any CDF $F(x)$

- Because $F(x)$ is a probability, $0 \leq F(x) \leq 1$ for all $x \in (-\infty, \infty)$
- $F(x)$ is nonidecreasing. That is, if $x > y$, then $F(x) \geq F(y)$.
- $\lim_{x \to -\infty} F(x) = 0$
- $\lim_{x \to \infty} F(x) = 1$
- $F(x)$ is right-continuous, i.e. $\lim_{x \to x_0^+} = F(x_0)$
- $P(X = x_0) = F(x_0) - \lim_{x \to x_0^-} F(x)$
- $P(a \leq X < b) = F(b) - F(a)$

The last property is useful for distributions that are either discrete or have a mixture of discrete and continuous components. It says that the probability of a particular value, x_0 is the height of the jump of $F(x)$ at $x = x_0$.

Properties of CDFs

If X is a continuous random variable (for example, normal or exponential) with density function $f(x)$, then the following are also true:

$$F(x) = \int_{-\infty}^{x} f(u)\, du \quad f(x) = \frac{d}{dx} F(x)$$

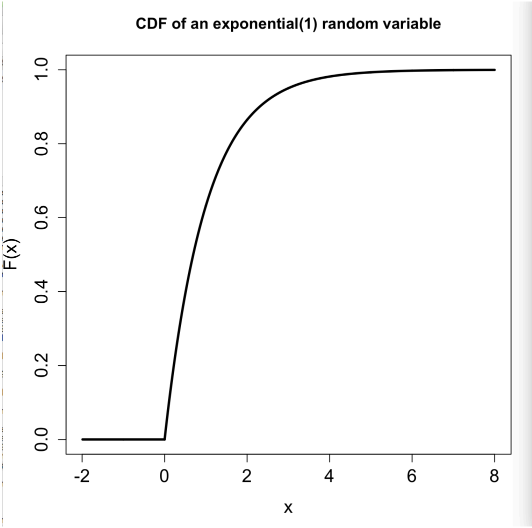
If X is a discrete random variable (for example, binomial or Poisson), the following are true:

$$F(x) = \sum_{-\infty}^{\infty} P(X = x) \quad P(X = x_0) = F(x_0) - \lim_{x \to x_0^-} F(x)$$

Examples of CDFs

If X is a continuous random variable, then $F(x)$ typically looks somewhat S shaped or like a square root function or the right-hand side of a parabola that is chopped off at 1. More complicated curves are also possible. Here is an example of a continuous CDF.

Examples of CDFs:exponential



Plotting the exponential CDF

For distributions that are not too complicated mathematically, you can graph them by hand, but otherwise you typically need software. It is useful to be able to use software to do this, so we'll go over how to do this in R (which is much easier than SAS...)

To get R on your computer, go to `http://www.r-project.org/` to download and install R. It will ask you to pick a "mirror" location where to download from.

You can also use R online (without downloading) at http://www.tutorialspoint.com/execute_r_online.php

Brief R tutorial

We'll go online to start a brief tutorial. Here are some concepts:

- Entering data as a vector, e.g., `x <- c(3,6,2)`
- Generating sequences: `x <- 1:10`, `x <- seq(1,10,2)`
- plotting: `plot(x,sqrt(x),type="l")`
- Using built-in functions
  - `x <- pexp(4)` Probability that exponential(1) R.V. is less than or equal to 4
  - `x <- dexp(4)` density (y-axis) value for exponential(1) at x=4
  - `x <- rexp(1,4)` generate one random exponential with rate 4
  - `x <- rexp(4,1)` generate four random exponentials with rate 1
  - `x <- sum(c(3,4,5))`

R tutorials

There are many R tutorials online and on Youtube. A fairly comprehensive place to go is `http://cran.r-project.org/doc/manuals/r-release/R-intro.html`, but you might try searching on specific topics for shorter introductions.

Plotting the exponential CDF

Once you have an R session started, you can enter the following at the prompt to generate the plot.

```
> x <- seq(-2,8,.1)
> x
  [1] -2.0 -1.9 -1.8 -1.7 -1.6 -1.5 -1.4 -1.3 -1.2 -1.1 -1.0 -0.9 -0.8 -0.7 -0.6
 [16] -0.5 -0.4 -0.3 -0.2 -0.1  0.0  0.1  0.2  0.3  0.4  0.5  0.6  0.7  0.8  0.9
 [31]  1.0  1.1  1.2  1.3  1.4  1.5  1.6  1.7  1.8  1.9  2.0  2.1  2.2  2.3  2.4
 [46]  2.5  2.6  2.7  2.8  2.9  3.0  3.1  3.2  3.3  3.4  3.5  3.6  3.7  3.8  3.9
 [61]  4.0  4.1  4.2  4.3  4.4  4.5  4.6  4.7  4.8  4.9  5.0  5.1  5.2  5.3  5.4
 [76]  5.5  5.6  5.7  5.8  5.9  6.0  6.1  6.2  6.3  6.4  6.5  6.6  6.7  6.8  6.9
 [91]  7.0  7.1  7.2  7.3  7.4  7.5  7.6  7.7  7.8  7.9  8.0
> plot(x,sapply(x,pexp),type="l",lwd=3,ylab="F(x)",xlab="x"
,cex.lab=1.4,cex.axis=1.4,main=title)
```

Plotting the exponential CDF

There's a fair bit of R code in the previous example, which I'll try to explain piece by piece.

The first line, `x <- seq(-2,8,.1)` generates a sequence of values starting with -2 and ending with 8 in increments of 0.1. You don't have to determine the number of points, but it is 101 values. The purpose of this is to give values on the *x*-axis. It was a somewhat arbitrary decision on my part to plot from -2 to 8 as opposed to 0 to 10 or 0 to 5 or -1 to 13.7 etc., but I picked values to make the plot look nice.

Plotting the exponential CDF

To evaluate the CDF, you want to plug in the different x values into $F(x)$, and the resulting values are the y-axis values. If you know the formula for $F(x)$, then you can use this formula to plug in the different values. For well-known distributions such as the exponential, R has built in functions for evaluating the density, the cumulative distribution function, and quantiles (the inverse of the cumulative distribution function). These built in functions are very handy.

If you want to know the density function when $x = 1.5$, you can type `dexp(1.5)`, and this gives you the value. If want the CDF when $x = 1.5$, you can type `pexp(1.5)`. To generate a single random exponential variable, you can type `rexp(1)`, and to use the inverse CDF, you can use `qexp()`, where the q stands for quantile. For example, `qexp(.5,1)` is the median of an exponential with rate 1.

Plotting the exponential CDF

Here you can combine several CDFs onto one plot to compare different rate parameters.

```
> x <- seq(-2,8,.1)
> plot(x,sapply(x,pexp),type="l",lwd=3,ylab="F(x)",xlab="x"
,cex.lab=1.4,cex.axis=1.4,main=title)
> points(x,sapply(x,pexp,rate=2),type="l",lwd=3,col="orange")
> points(x,sapply(x,pexp,rate=3),type="l",lwd=3,col="pink")
> points(x,sapply(x,pexp,rate=0.5),type="l",lwd=3,col="brown")
```

Plotting the exponential survival function

The survival function $S(x) = 1 - F(x)$. To plot the survival function, you can plot 1 minus the y-coordinate for the CDF. Try this at home (or in a computer lab).

```
> x <- seq(-2,8,.1)
> plot(x,1-pexp(x),type="l",lwd=3,ylab="F(x)",xlab="x"
,cex.lab=1.4,cex.axis=1.4)
> points(x,1-pexp(x,rate=2),type="l",lwd=3,col="orange")
> points(x,1-pexp(x,rate=3),type="l",lwd=3,col="pink")
> points(x,1-pexp(x,rate=0.5),type="l",lwd=3,col="brown")
> legend(4,1.0,legend=c(".5","1","2","3"),fill=c("brown","black",
"orange","pink"),cex=1.4,title="rate")
```

Several other distributions follow the same notation in R, so that `dnorm()`, `pnorm()`, `qnorm()`, and `rnorm()` generate density values, CDF values, quantiles, and random variables from a normal distribution. You can also apply this to Poisson, binomial, gamma, beta, uniform, and other distributions.

In addition to the default parameters such as rate 1 for an exponential, you can also specify other parameters for these distributions. The CDF at $x = 0.2$ for an exponential distribution with rate 10 (mean $= 1/10$) is given in R by `pexp(.2,10)`. This returns the value 0.8646647, which means that there is roughly an 84.5% chance that a random exponential with rate 10 (or mean of 0.1) is less than 0.2.

The exponential distribution in R

The exponential distribution is sometimes parameterized using the mean, and sometimes the rate, depending on the book or the software, so it can be important to check what the software is doing. The mean is 1 divided by the rate, so that if the rate is high, the mean is low. You can check what R is doing by either typing

```
help(pexp)
```

and reading the description or trying some values. For example if you type `rexp(100,10)`, you are generating 100 random variables with mean 0.1, so you should see small values.

The gamma distribution is another one that can be parameterized different ways depending on the book or the software.

CDFs for discrete random variables

If you have a discrete random variable, then the CDF is not continuous but looks like a step function. An example is the function

$$f(x) = P(X = x) = \begin{cases} 1/2 & x = 0 \\ 1/5 & x = 1 \\ 1/4 & x = 3 \\ 1/20 & x = 4 \\ 0 & \text{otherwise} \end{cases}$$

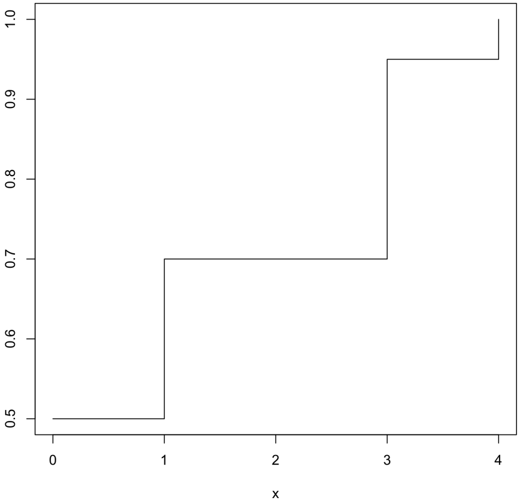
To get the CDF, we need the cumulative probabilities. For example
$P(X \le 0) = 1/2$
$P(X \le 1) = 1/2 + 1/5 = 7/10$
$P(X \le 2) = 7/10$
$P(X \le 3) = 7/10 + 1/4 = 19/20,$
$P(X \le 4) = 1.$
Note that $P(X < 4) = P(X \le 3) = 19/20$ and $P(X < 2) = P(X \le 1) = 7/10.$

Plotting the CDF in R

To plot the CDF in R, you can type

```
> y <- c(.5,.7,.7,.95,1)
> x <- c(0,1,2,3,4)
> plot(x,y,type="s")
```

Plotting the CDF in R



Back to theory: cdf

For distributions that aren't too complicated, you can also use theory and definitions to work with CDFs, PDFs (probability density) functions, and survival functions.

For an exponential distribution, the density is

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

The CDF is

$$F(x) = \int_0^x \lambda e^{-\lambda\, u}\, du = 1 - e^{-\lambda\, x}$$

The survival function is

$$S(x) = 1 - F(x) = e^{-\lambda\, x}$$

Survival function and CDF

For this case, if the rate parameter $\lambda = 1$, then the survival function happens to be the same as the density function. But this is highly unusual. For other rates for the exponential, $S(x)$ is proportional to the density.

In general, the survival function is related to the density by the following

$$\frac{d}{dx}S(x) = \frac{d}{dx}1 - F(x) = -f(x)$$

In other words, the derivative of the survival function is minus the density.

Properties of $S(x)$

Similar to the CDF, the survival function has these properties:

- nonincreasing: for $x > y$, $S(x) \leq S(y)$
- Is it right-continous or left-continuous?
- has limits of 1 and 0

Hazard functions

A concept from survival analysis is the Hazard function.

$$h(x) = \frac{f(x)}{S(x)}$$

You can also think of this as

$$h(x) = -\frac{d}{dx} \ln S(x)$$

There is also a cumulative hazard function

$$H(x) = \int_0^x h(u)\ du = -\ln S(x)$$

Thus,

$$S(x) = e^{-H(x)}$$

Hazard function

You can interpret the hazard function as the instaneous rate of failure at a time x given that failure hasn't occurred yet at time x. Theoretically, the derivative is involved because it is the limit of a difference quotient involving a conditional probability:

$$\begin{aligned} h(x) &= \lim_{\Delta x \to 0} \frac{P(x \leq X \leq x + \Delta x | X \geq x)}{\Delta x} \\ &= \frac{1}{S(x)} \lim_{\Delta x \to 0} \frac{P(x \leq X \leq x + \Delta x)}{\Delta x} \\ &= \frac{1}{S(x)} \cdot f(x) \end{aligned}$$

You can think of $h(x)\Delta x$ as the approximate probability of experiencing the event in the next Δx units of time.

Properties of hazard functions

Hazard functions can have many different properties. They can be increasing over time, meaning that as something ages, it is more likely to fail. Or they can be constant. A decreasing hazard can be used for something that is likely to fail initially, but once it is working is likely to continue working, such as for a transplant.

A “bathtub” hazard is useful if something is likely to fail early or late but not in the middle. For example, humans are more likely to die in infancy or in old age than for intermediate ages.

The hazard for the exponential distirbution

For the exponential distribution, the density for $x > 0$ is

$$f(x) = \lambda e^{-\lambda x}$$

and the survival function is

$$S(x) = e^{-\lambda x}$$

Therefore the hazard function is

$$h(x) = \frac{\lambda e^{-\lambda x}}{e^{-\lambda x}} = \lambda$$

So the hazard function is constant. This means that failure is no more or less likely after waiting for a time. If lightbulbs have exponential hazards, then it has the same the same probability of going out in the next moment whether it has been used for 100 hours or 1000 hours.

The Weibul distribution

A flexible family of distributions that can have either decreasing, increasing, or constant hazard functions is the Weibul distribution:

$$f(x) = \alpha\lambda x^{\alpha-1} e^{-\lambda\, x^{\alpha}}$$

where $\alpha, \lambda > 0$, $x \geq 0$ (and is otherwise 0). If $\alpha = 1$, then the Weibul reduces to the exponential. For this distribution

$$S(x) = e^{-\lambda\, x^{\alpha}}$$

$$h(x) = \alpha\lambda x^{\alpha-1}$$

The hazard function is decreasing for $\alpha < 1$, increasing for $\alpha > 1$ and constant for $\alpha = 1$.

Hazard functions for discrete distributions

For a discrete distribution, the hazard function is

$$h(x_j) = P(X = x_j | X \geq x_j) = \frac{P(X = x_j)}{P(X \geq x_j)} = \frac{P(X = x_j)}{P(X > x_{j-1})} = \frac{P(X = x_j)}{S(x_{j-1})}$$

Letting $S(x_0) = 1$. Note that

$$P(X = x_j) = P(X > x_{j-1}) - P(X > x_j) = S(x_{j-1}) - S(x_j)$$

so the hazard can be written entirely interms of the survival function

$$h(x) = \frac{S(x_{j-1}) - S(x_j)}{S(x_{j-1})} = 1 - \frac{S(x_j)}{S(x_{j-1})}$$

Hazard functions for discrete distributions

That survival function for a discrete distribution can also be written as

$$S(x) = \prod_{x_j \le x} \frac{S(x_j)}{S(x_{j-1})}$$

why is this? If you multiply the product, you have a telescoping product, and most terms cancel out. This relates the survival function to the hazard as follows

$$S(x) = \prod_{x_j \le x} [1 - h(x_j)]$$

You can also think of this as the probability of surviving more than x units of time as being the probability of survivng the first unit of time, times the probability of surviving the second unit of time given that you've survived the first, times the probability of surviving the third given that you've survived the first two, etc.

Homework (due 2/1/16): first of two problems

Only include R code as an appendix. For questions asking for plots, you can just turn in a plot without any commentary.

1. Let X have a log normal distribution with parameters μ and σ.

(a) Plot the survival function $S(x)$ for all combinations of $\mu = 1, 2, 5$ and $\sigma = 2, 4$. Put all curves on the same plot and use color to distinguish levels of μ and solid versus dashes curves to distinguish σ values. Make sure that the plot has a legend and appropriate axis labels.

(b) Also plot the hazard function $h(x)$ for the same parameter combinations and use the same color scheme and legend to distinguish parameters. Also include a legend and have appropriate axis labels.

Homework: second of two problems

2. Suppose the density function for X is

$$f(x) = \begin{cases} \frac{k}{x^4} & x > 1 \\ 0 & \text{otherwise} \end{cases}$$

Then answer the following questions:

(a) Find k so that density integrates to 1: $\int_{-\infty}^{\infty} f(x)\, dx = 1$

(b) Find the cumulative distribution function $F(x)$

(c) Find the survival function $S(x)$, with $S(x)$ defined for all $x \in (0, \infty)$.

(d) Plot the survival function using R.

(e) Find the hazard function $h(x)$