Definition of Survival and hazard functions:

$$S(t) = \Pr\{T > t\} = 1 - F(t)$$

$$\lambda(t) = \lim_{u \to 0} \frac{\Pr\{t < T \le t + u \mid T > t\}\}}{u} = \frac{f(t)}{S(t)}$$

Relationship between Survival and hazard functions:

$$\frac{\partial \log S(t)}{\partial t} = \frac{\partial S(t) / \partial t}{S(t)} = -\frac{f(t)}{S(t)}$$

$$\lambda(t) = -\frac{\partial \log S(t)}{\partial t}$$

The cumulative hazard function

$$\Lambda(t) \equiv \int_0^t \lambda(v)dv = -\log S(t)$$

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

Exponential distribution

$$\Lambda(t) = \lambda t$$

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

$$\Pr\{T > t_0 + t \mid T > t_0\} = \Pr\{T > t\}$$

Weibull distribution

$$\lambda(t) = \alpha\gamma t^{\gamma-1}$$

$$\Lambda(t) = \alpha t^{\gamma}$$

$$S(t) = \exp(-\Lambda(t)) = \exp(-\alpha t^{\gamma})$$

With all deaths observed, can estimate nonparametrically: $\hat{S}(t) = prop(T_i > t)$

Or parametrically (method of moments)

Exponential: $E(T) = {}^{1}\!/_{\lambda} \Rightarrow \hat{\lambda}_m = {}^{1}\!/_{mean(T)}$

Weibull: $E(T) = {}^{\gamma}\!/_{\lambda}, Var(T) = \gamma / \lambda^2$

Plug in sample estimates and solve!

Types of Censorship:

Type 1: fixed censoring time (rare in medical applications, more common in engineering)
Type 2: censor after observe r failures (common in engineering)
RANDOM CENSORING: let C be a random censoring time, then for patient i

Observe
$$Y_i = \min(T_i, C_i)$$
$$\delta_i = I(T_i < C_i)$$

Convention: assume "death before censoring"!

ASSUME T and C are independent (nearly always false, but usefully so…weaker assumptions usually suffice)

Leads to non-parametric estimation such as KM, or the compactly named:
Altschuler-Nelson-Aalen-Fleming-Harrington estimator:

$$\hat{\Lambda}(t) = \sum_{i:t_i<t} \frac{d_i}{n_i}$$

$t_1, t_2, t_3, \ldots$ *are the ordered unique event times*
$d_1, d_2, d_3, \ldots$ *corresponding numbers of deaths*
$n_1, n_2, n_3, \ldots$ *numbers at risk*

$$\hat{S}_\Lambda(t) = \exp[-\Lambda(t)]$$

Compare to Kaplan-Meier PL estimator

$$\hat{S}_{KM}(t) = \prod_{i:t_i<t} \left(1 - \frac{d_i}{n_i}\right)$$

$$\hat{\Lambda}_{KM}(t) = -\log \hat{S}_{KM}(t) = -\sum_{i:t_i<t} \log\left(1 - \frac{d_i}{n_i}\right)$$

parametric estimation, for random censoring:

Exponential: $\hat{\lambda}_{ML} = \sum \delta_i \Big/ \sum Y_i$

(note, numerator is count of uncensored observations)

Weibull: not closed form…must iterate!

For now we will concentrate on non-(and semi-) parametric estimation and testing – leave parametrics (especially useful Weibull) to Prof. Olshen.

Mantel-Haenszel log-rank test

At each unique death, 2X2 table of vital status by group

	dead	alive	
group 1	a	b	n_1
group 2	c	d	n_2
	m_1	m_2	n

For example
Leukemia data, first death time=5

	dead	alive	
Maintained	0	11	11
Unmaintained	2	10	12
	2	21	23

leukemia data in full

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

>

Given fixed margins, null hypothesis (equal death rates)
A=upper left hand entry has hypergeometric distribution:

$$\Pr\{A=a\}=\frac{\binom{n_1}{a}\binom{n_2}{m_1-a}}{\binom{n}{m_1}}$$

$$E_0(A)=\frac{n_1 m_1}{n}$$

$$Var_0(A)=\frac{n_1 n_2 m_1 m_2}{n^2(n-1)}$$

Accumulate the differences between observed and expected over the set of 2X2 tables at each death time, and divide by the std error of the sum:

$$MH=\frac{\sum\left[a_i-E_0(A_i)\right]}{\sqrt{\sum Var_0(A_i)}}$$ …….refer to N(0,1)

Gehan modification of Wilcoxon
Let T and R be the true (possibly unobserved) times in Maintained and Unmaintained subjects, X and Y the corresponding censored values

$$U(X_i, Y_j) = U_{ij} = \begin{cases} +1 \text{ if we know } \mathrm{T_i} > \mathrm{R_j} \\ 0 \text{ otherwise} \\ -1 \text{ if we know } \mathrm{T_i} < \mathrm{R_j} \end{cases}$$

$$U = \sum_{i,j} U_{ij}$$

Reject null if U is large.
If no censorship:

$$Var_{0,P}(U) = \frac{mn(m+n+1)}{3}$$

but with censorship more complex (and larger)

……for leukemia data
MH = 1.8429, p= 0.0653
GW= 1.6671, p= 0.0955

Tarone-Ware class of tests:

$$MH = \frac{\sum w_{i=}\left[a_i - E_0(A_i)\right]}{\sqrt{\sum w_i^2 Var_0(A_i)}}$$

$w_i = 1$ gives MH

$w_i = n_i$ gives Gehan

$w_i = \sqrt{n_i}$ is TW suggestion

COX PH Model:

$$\lambda(t;\mathbf{X}) = \lambda_0(t)\exp(\mathbf{X}\beta)$$

single binary covariate *X*:

$$\frac{\lambda(t;X=1)}{\lambda(t;X=0)} = \frac{\lambda_0(t)\exp(1\beta)}{\lambda_0(t)\exp(0\beta)} = \exp(\beta)$$

Lehman Alternatives:

$$S(t;\mathbf{X}) = \exp\left[-\int_0^t \lambda_0(s)\exp(\mathbf{X}\beta)ds\right] = S_0(t)^\gamma$$

$$S_0(t) = \exp\left[-\int_0^t \lambda_0(s)ds\right]$$

$$\gamma = \gamma(\mathbf{X}) = \exp(\mathbf{X}\beta)$$

suppressing times, and taking logs…

$$\log(S) = \gamma \log(S_0) = -\gamma\Lambda_0$$
$$\log(-\log(S)) = \log(\gamma) + \log(\Lambda_0) = \mathbf{X}\beta + \log(\Lambda_0)$$

So estimates of survival for various subgroups should look parallel on the "log-minus-log" scale.

And – if the hazard is *constant:*

$$\log(\Lambda_0(t)) = \log(\lambda_0 t) = \log(\lambda_0) + \log(t)$$

so the survival estimates are all *straight lines* on the log-minus-log (survival) against log (time) plot.

How many subjects to enroll?

To detect a true log hazard ratio of $\theta = \log\left(\frac{\lambda_1}{\lambda_2}\right)$

(power $1-\beta$ using a 1-sided test at level α)

require D observed ***deaths***, where:

$$D = \frac{4\left(z_{1-\alpha} + z_{1-\beta}\right)^2}{\theta^2}$$

(for equal group sizes- if unequal replace 4 with 1/P(1-P) where P is proportion assigned to group 1)

The censored observations contribute nothing to the power of the test!

Sample size required for non-binary covariate X:

Deaths:

$$D = \frac{\left(z_{1-\alpha} + z_{1-\beta}\right)^2}{\sigma_X^2 \theta^2}$$

where is the variance of X and   is the log hazard ratio for a unit change in X

Note that "wider" X gives more power, as it should!

Epidemiology: non-binary exposure X (say, amount of smoking)

Adjust for confounders **Z** (age, sex, etc.), in the Cox model.

Adjust D above by "Variance Inflation Factor"

$VIF = {}^{1}\!\big/_{1-R^2}$ where R^2 = variance of X

explained by **Z**