Lecture seventeen: Common Distributions for Survival Data

In additon to exponential, Weibull and Gompertz distributions mentioned in chapter 5, there are other probability distributions for survival data, following is a summary of them. Notice that each one of the four quantities, namely, density function $f(t)$, survival function $S(t)$, hazard function $h(t)$ and cumulative hazard function $H(t)$, uniquely determines the underlying distribution, hence the other three quantities.

1. **Derivative, max, min, range, change rate**

2. **Common Distributions**

   (a) Exponential: $Exp(\lambda)$:

   $$f(t) = \lambda exp(-\lambda t), \qquad t \geq 0$$

   $$S(t) = exp(-\lambda t), \qquad t \geq 0$$

   Constant hazard; linear cumulative hazard in time t.

   Piece-wise exponential:

   (b) Weibull: $W(\lambda, \gamma)$:

   $$f(t) = \lambda\gamma t^{\gamma-1} exp(-\lambda t^{\gamma}), \qquad t \geq 0$$

   $$S(t) = exp(-\lambda t^{\gamma}), \qquad t \geq 0$$

   $$h(t) = \lambda\gamma t^{\gamma-1}, \qquad t \geq 0$$

   i. $\gamma < 1$, decreasing hazard over time
   ii. $\gamma = 1$, Exponential distribution: $Exp(\lambda)$
   iii. $\gamma > 1$, increasing hazard over time

   (c) The log-logistic: $log - logistic(\theta, \kappa)$

   $$f(t) = \frac{e^{\theta}\kappa t^{\kappa-1}}{(1 + e^{\theta}t^{\kappa})^2}, \qquad t \geq 0$$

   $$S(t) = [1 + e^{\theta}t^{\kappa}]^{-1}, \qquad t \geq 0$$

   $$h(t) = \frac{e^{\theta}\kappa t^{\kappa-1}}{1 + e^{\theta}t^{\kappa}}, \qquad t \geq 0$$

i. $\kappa < 1$, hazard decreases from $+\infty$
ii. $\kappa = 1$, hazard decreases from e^{θ} to 0
iii. $\kappa > 1$, hazard increases from 0 to a maximum, and then decreases to 0

log(T) has a logistic distribution, whose density function is similar to that of normal distribution. The pth percentile is

$$t(p) = [pe^{-\theta}/(100 - p)]^{1/k},$$

(d) Gamma: $Gamma(\lambda, \rho)$ using scale and shape parameters

$$f(t) = \frac{\lambda^{\rho} t^{\rho-1} e^{-\lambda t}}{\Gamma(\rho)}, \qquad t \geq 0$$

There are no closed form for $S(t), h(t)$, for example,

$$h(t) = \frac{\lambda^{\rho} t^{\rho-1} e^{-\lambda t}}{\Gamma(\rho)\{1 - \Gamma_{\lambda t}(\rho)\}}, \qquad t \geq 0$$

where $\Gamma(\rho)$ is a gamma function and $\Gamma_{\lambda t}(\rho)$ is the imcomplete gamma function given by

$$\Gamma_{\lambda t}(\rho) = \frac{1}{\Gamma(\rho)} \int_0^{\lambda t} u^{\rho-1} e^{-u} du,$$

which is the cumulative distribution function.

i. when $\rho < 1$, hazard decreases
ii. when $\rho = 1$, exponential distribution: $Exp(\lambda)$
iii. when $\rho > 1$, hazard increases
iv. If T_i are k independent random variables with $Gamma(\lambda, \rho_i)$, $i = 1, ..., k$, then $T = \sum_{i=1}^{k} T_i$ has $Gamma(\lambda, \sum_{i=1}^{k} \rho_i)$.

(e) Log-normal: $log - normal(\mu, \sigma^2)$

$$f(t) = \frac{1}{\sigma\sqrt{(2\pi)}} t^{-1} exp\{-(logt - \mu)^2/2\sigma^2\}, \qquad t \geq 0$$

There are no closed form for $S(t), h(t)$, hazard non-monotonic, increasing from 0 to a maximum, and then decreasing to 0. For example,

$$S(t) = 1 - \Phi(\frac{log\, t - \mu}{\sigma}) = \int_{-\infty}^{\frac{log\, t-\mu}{\sigma}} \frac{1}{\sqrt{(2\pi)}} exp\{-u^2/2\} du,$$

The log-normal model will tend to be similar to log-logistic model; the Weibull and Gamma distributions will generally lead to very similar results.

(f) *Generalized gamma distribution and inverse Gaussian distribution (page 224)*: Ref: The Inverse Gaussian Distribution, Seshadri, 1999, Spinger.

(g) Gompertz distribution: $Gompertz(\beta, \gamma)$

$$f(t) = \beta e^{\gamma t} exp[\frac{\beta}{\gamma}(1 - e^{\gamma t})],$$

where $\beta, \gamma > 0, t \geq 0$.

$$S(t) = exp[\frac{\beta}{\gamma}(1 - e^{\gamma t})],$$

what if $\gamma < 0$?(Cure rate, Ref.: *Survival Analysis with Long-Term Survivors, Maller; Zhou, 1996*)

$$h(t) = \beta e^{\gamma t},$$

(h) Mixture (Maller, Zhou, 1996) and non-mixture (Tsodikov et al: JASA, 2003; 98: 1063-1078) cure models: See also section 6.10.

(i) More general form of Gompertz distribution above is the hazard has following forms

$$h(t) = \theta - \beta e^{-\gamma t}, \qquad t \geq 0$$

and

$$h(t) = \theta + \beta e^{-\gamma t},$$

where $\theta, \beta, \gamma > 0$. There are other constraints on the parameters.

The hazard of death is to increase or decrease with time in the short term, and then become constant.

(j) the 'bathtub' hazard:

$$h(t) = \alpha t + \frac{\beta}{1 + \gamma t},$$

which decreases to a single minimum and increases thereafter.

3. Choose a distribution: graphic tools

(a) Quantitle-Quantile Plot (QQ plot): without Censoring

i. Quantile/Percentile: t_q is the q^{th} percentile if $P(T < t_q) = q$.

ii. If a theoritical distribution approximates data reasonably well.

A. Theoretical quantitles (from distribution) should be comparable with emperical quantile (based on data).

B. Plot of theoritical quantitles versus empirical quantile is close to a straighline.

iii. If two samples of data follow the same distribution

A. Empirical quantiles of sampe 1 should be similar to the empirical quantile of sample 2.

B. Plot of quantiles from sample 1 vs quantiles of sample 2 is roughly a straightline.

iv. Splus Implementation

A. qqnorm(y): normal quantiles vs quantiles of sample y.

B. qqplot(x,y): quantiles of sample x vs quantiles of sample y, $length(x) = length(y)$ is NOT required.

C. plot(qdist(ppoints(y)), sort(y)): quantiles of theoretical 'dist' vs quantiles of sample y.

D. 'qdist' can be one of 'qexp', 'qgamma', 'qlogis','qlnorm','qunif', 'qweibull', and many more.

E. 'ppoints': a sample of probability points corresponding to the sample y.

F. 'sort': re-arrange y based on ranking.

v. Using formal tests to compare a sample with a theoretical distribution

A. Examples: Kolmogorov-Smirnov test, Chisq test, etc.

B. **Warning**: many common t-tests, Wilcoxon etc. are not appropriate because they test only difference in mean/median, not the distribution.

vi. censoring cannot be dealt with by 'qqplot' (how to rank censored observations?)

(b) Hazard and other plots

i. The estimates of hazard function mentioned in chapter 2 are very unstable, which depend heavily on time interval between two consecutive distinct failures/ time intervals chosen.

ii. plotting $\hat{h}(t_j)$ vs t_j: unstable because of substantial fluctuation. Instead, use cumulative hazard plot.

(c) Cumulative Hazard Plot

i. Estimating H(t) by using $\hat{S}(t)$ (e.g. KM).

ii. Appropriate scale to plotting H(t) vs t: determined by the analytical behavior of a distribution.

A. Exponential: $H(t) = \lambda t$, $log(S(t))$ linear in t.

B. Weibull: $log(H(t)) = log(-log S(t))$ linear in $log\, t$.

C. log-logistic: $log\{\frac{S(t)}{1-S(t)}\} = -\theta - \kappa log\, t$. Illustration (Example 6.1):

D. log-normal: $\Phi^{-1}\{1 - S(t)\} = \frac{log\, t - \mu}{\sigma}$

E. Not trivial for most of other distributions.

iii. Example 6.1: Time to discontinuation (IUD data)

Figure 1:

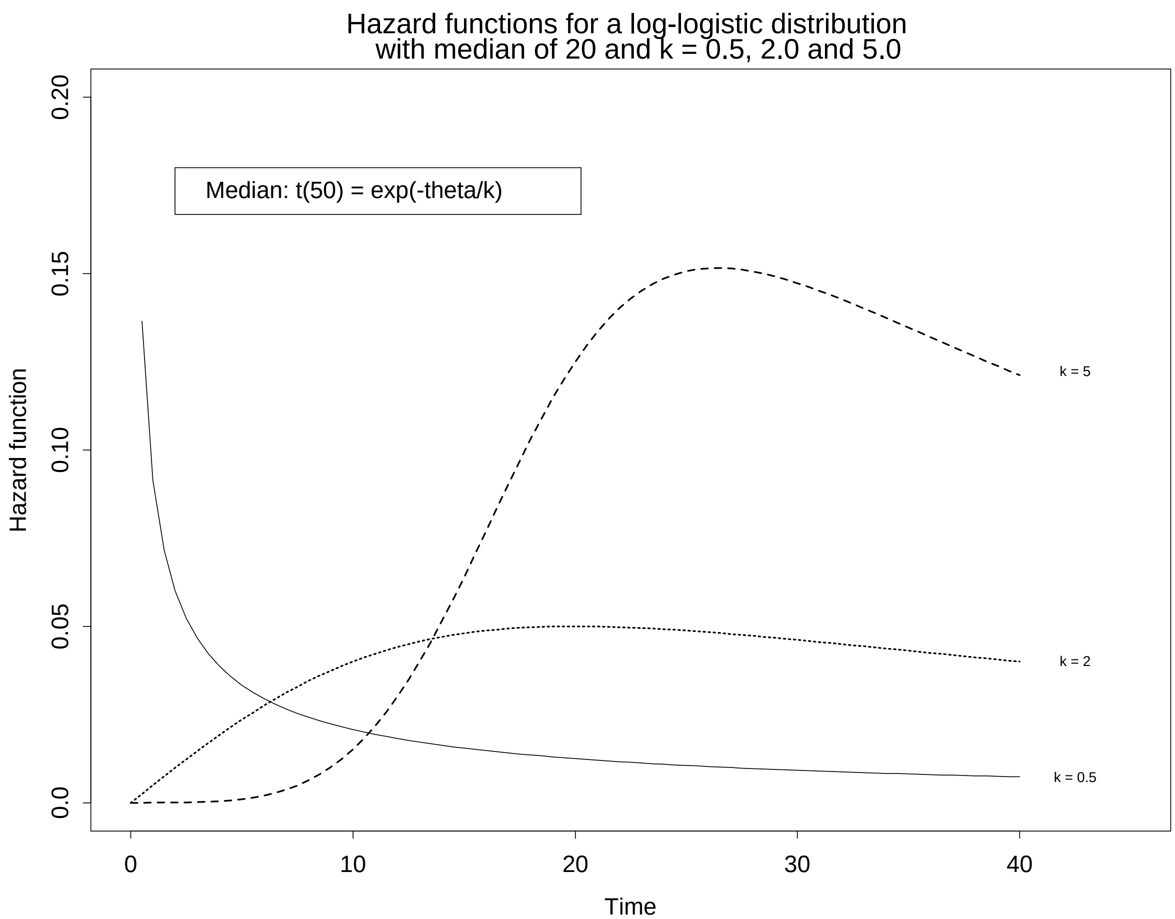
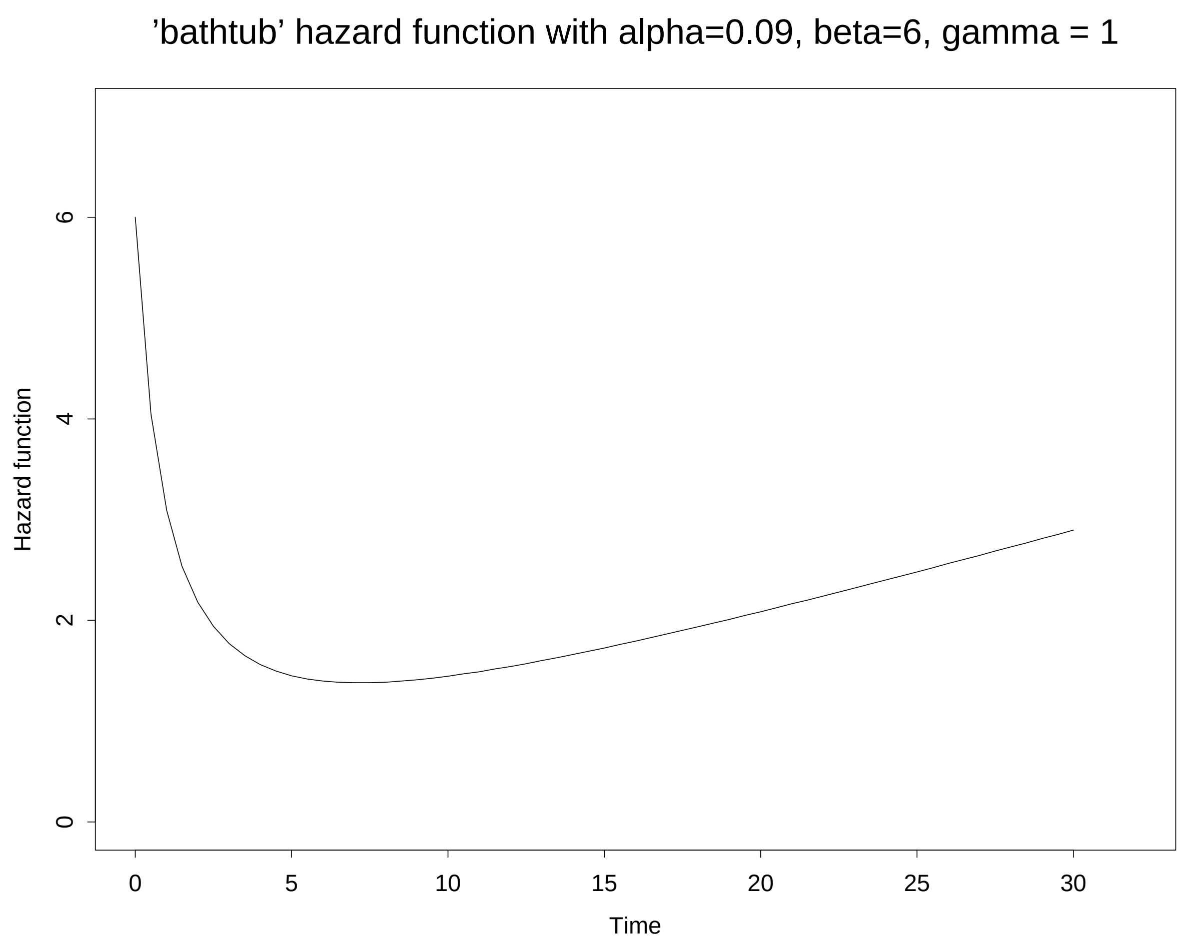
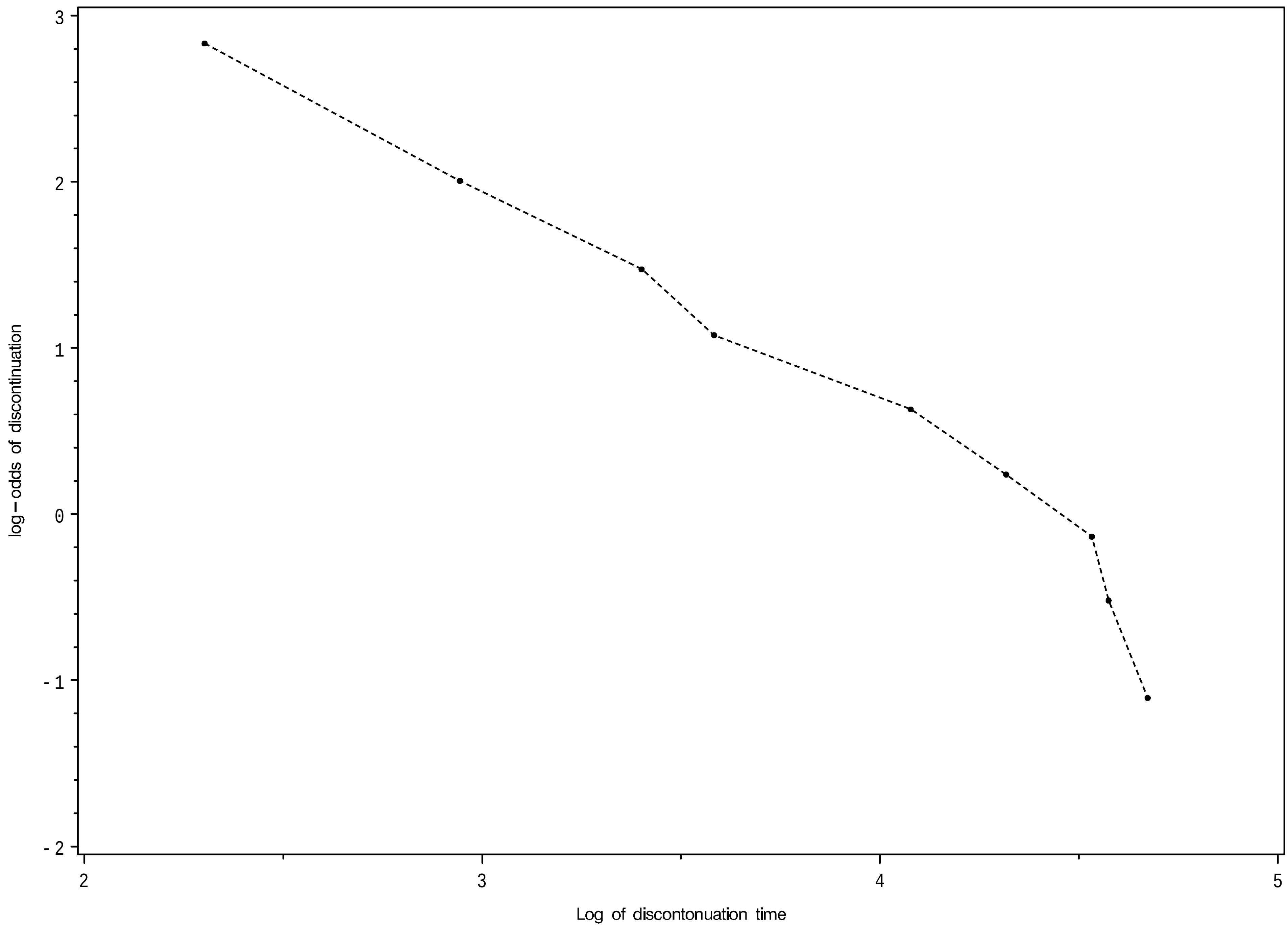

Hazard functions for a log-logistic distribution
with median of 20 and k = 0.5, 2.0 and 5.0
Median: t(50) = exp(-theta/k)
Hazard function
0.0
0.05
0.10
0.15
0.20
k = 5
k = 2
k = 0.5
0
10
20
30
40
Time

Figure 2:


'bathtub' hazard function with alpha=0.09, beta=6, gamma = 1
Hazard function
0
2
4
6
0
5
10
15
20
25
30
Time



3
2
1
0
-1
-2
log-odds of discontinuation
2
3
4
5
Log of discontonuation time