BCS TTE User's Guide

John D. Cook

August 15, 2006

1 Introduction

This software computes the Bayesian chi square test of Valen Johnson[1] for right-censored time-to-event data. It tests the goodness of fit of the best fit to the data from the following distribution families:

- exponential
- gamma
- inverse gamma
- Weibull
- log normal
- log logistic
- log odds rate

At least one observation must be uncensored.

2 Input

2.1 Software arguments

The only required argument is a data file name. This file must be a plain text file with two numbers per line. The first number is a (possibly censored) survival time, a real number, and the second is either 0 or 1. The two numbers must be separated by white space. One may include comments in the data file if they begin with a # character in the first column.

[1] "A Bayesian χ^2 test for goodness-of-fit", *Annals of Statistics*, 2004, Vol 32, No. 6, 2361-2384

By default, the second number on each line is the indicator of death. If this number is 1, the survival time is uncensored because the terminal event has occurred. A value of 0 indicates the survival time is right-censored.

The software produces HTML output. If a second file name is provided as an argument, the output will be written to that file. Otherwise the output will be directed to the command line standard output.

The `--censor` option reverses the meaning of the second number on each line of the input. If this option is specified, the second number is the indicator of censoring: 0 for uncensored and 1 for censored data.

The `--discrete` option indicates that the survival times have been rounded to integer values. If the survival times have indeed been rounded to integer values, it is best to set this option as it will improve the goodness of fit.

By default, the number of bins used for the Bayesian chi square test is the nearest integer to $n^{2/5}$ where n is the number of observations. One may specify the number of bins with the option `--numbins=`b where b is the number of bins.

By default, the number of posterior samples used for the Bayesian chi square test is 1000. One may specify a different number of samples with the option `--numreps=`r where r is the number of samples.

The seed for random number generation is obtaining from the system time by default. Hence results will differ each time the analysis is run. One may specify the random number generation seed by using the `--seed=`s option where s is a positive integer.

Arguments may be supplied in any order.

2.2 Example 1

The command

```
bcstte sample1.txt sample1.html --discrete
```

will analyze the observations in the text file `sample1.txt`, writing its output to the file `sample1.html` which can be viewed in a web browser. Survival times are assumed to be rounded to integer values, and so each time t will be interpreted as a random value in the interval $(t - 1/2, t + 1/2)$. The number of bins used for the Bayesian chi square test will be the 2/5 power of the number of observations in `sample1.txt`.

2.3 Example 2

The command

```
bcstte sample2.txt --numbins=5 --censor --seed=12345 > sample2.html
```

will analyze the observations in the text file `sample2.txt`. Since no output file is specified, the results will be dumped to standard output, which is then redirected to the file `sample2.htm`. The test will use five bins, corresponding to a chi square test with four degrees of freedom. The software will interpret a 1 in the second column as indicating a death and will interpret a 0 as indicating censoring. The seed for the random number generation will be set to 12345.

3 Output

The output consists of three tables. The first echoes the input arguments. It is important to review this table. If, for example, the notation conventions for censoring do not match your intentions, the results will be meaningless. If there were comments in the data file, they will be echoed below this table.

The second table gives the Bayesian chi square results for each distribution family. The table is sorted in increasing order of Bayesian chi square value. If there is an asterisk by a value, this indicates that there was a numerical problem in the computation and an alternative prior was used to recover. The meaning of the output columns are documented in the output itself. The upper bound on the p-value is determined by applying an inequality of Rychlik[2].

The third table gives the parameters used for each distribution family. See the following section for the meaning of the parameters.

4 Distribution parameterizations

4.1 Exponential

The exponential distribution with parameter $\beta > 0$ has mean β, variance β^2, and PDF

$$\frac{e^{-x/\beta}}{\beta}.$$

for all positive x.

This is the same as a Gamma distribution with shape $\alpha = 1$ and scale β.

4.2 Gamma

Let $\alpha > 0$ be the shape parameter and $\beta > 0$ the scale parameter. Then the gamma distribution has mean $\alpha\beta$, variance $\alpha\beta^2$ and PDF

$$\frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}$$

[2] "Bounds for order statistics based on dependent variables with nonidentical distributions", Statistics & Probability Letters 23 (1995) 351–358

for all positive x.

The software lists the parameters in the order (α, β).

4.3 Inverse gamma

Let $\alpha > 0$ be the shape parameter and $\beta > 0$ the scale parameter. Then the inverse gamma distribution has PDF

$$\left(\frac{\beta^\alpha}{x^{\alpha+1}\Gamma(\alpha)}\right) e^{-\beta/x}$$

for all positive x.

If $\alpha > 1$ the mean is

$$\frac{\beta}{\alpha - 1}.$$

If $\alpha > 2$ the variance is

$$\frac{\beta^2}{(\alpha-1)^2(\alpha-2)}.$$

If X is distributed as a gamma distribution with parameters (α, β) then $1/X$ is distributed as an inverse gamma with parameters $(\alpha, 1/\beta)$.

NB: The β in our parameterization of the inverse gamma corresponds to $1/\beta$ in another common convention.

The software lists the parameters in the order (α, β).

4.4 Log logistic

The log logistic distribution has parameters k and θ. The PDF is

$$\frac{k e^\theta t^{k-1}}{(1 + e^\theta t^k)^2}.$$

For $k > 1$, the mean is given by

$$\frac{\pi e^{-\theta/k}}{k \sin\left(\frac{\pi}{k}\right)}$$

and for $k > 2$ the variance is given by

$$\frac{2\pi e^{-2\theta/k}}{k \sin\left(\frac{2\pi}{k}\right)}.$$

The log logistic is most simply expressed in terms of its survival function

$$\frac{1}{1 + e^\theta t^k}.$$

The software lists the parameters in the order (k, θ).

4.5 Log normal

The log normal distribution is parameterized by μ and $\sigma > 0$. If X is log normal with these parameters, $\log X$ is $N(\mu, \sigma)$. Note that μ and σ are *not* the mean and standard deviation of X but rather of $\log X$.

X has mean $\exp(\mu + \frac{1}{2}\sigma^2)$ and variance $\exp(2\mu + \sigma^2)(\exp(\sigma^2) - 1)$. The PDF is

$$\frac{1}{x\sigma\sqrt{2\pi}} \exp\left(\frac{-(\log(x) - \mu)^2}{2\sigma^2}\right)$$

for all positive x.

The software lists the parameters in the order (μ, σ).

4.6 Log odds rate

The log odds rate distribution has three parameters: φ (shape), λ (scale), and c. If $c = 1$, this distribution reduces to the log logistic. As $c \to 0$, the limiting distribution is Weibull.

The distribution is most simply characterized by its survival function

$$S(t) = \left(1 + \left(\frac{t}{\lambda}\right)^{\varphi}\right)^{-1/c}.$$

The PDF if $f(t) = -S'(t)$.

If $\varphi > c$, the mean and second moment are given by

$$\mu = \frac{\lambda c^{-1/\varphi}}{\varphi} \, B\left(\frac{1}{\varphi}, \frac{1}{c} - \frac{1}{\varphi}\right)$$

and

$$\mu_2' = \frac{2\lambda^2 c^{-2/\varphi}}{\varphi} \, B\left(\frac{2}{\varphi}, \frac{1}{c} - \frac{2}{\varphi}\right)$$

respectively. The variance is $\mu_2' - \mu^2$.

The software lists the parameters in the order (φ, λ, c).

4.7 Weibull

The Weibull distribution has a shape parameter $\beta > 0$ and a scale parameter $\eta > 0$. It has mean

$$\eta\Gamma((\beta + 1)/\beta),$$

variance

$$\eta^2 \left(\Gamma((\beta + 2)/\beta) - \Gamma((\beta + 1)/\beta)^2\right),$$

and PDF

$$\frac{\beta x^{\beta-1}}{\eta^{\beta}} \exp(-(x/\eta)^{\beta})$$

for positive x.

The PDF is the product of the hazard function

$$\frac{\beta x^{\beta-1}}{\eta^{\beta}}$$

and the complementary CDF

$$\exp(-(x/\eta)^{\beta}).$$

A shape parameter $\beta = 1$ is an exponential with mean η, which has constant hazard.

The software lists the parameters in the order (β, η).