Beta Product Confidence Procedure Confidence Intervals and Discrete Data

Michael P. Fay and Erica H. Brittain

May 3, 2016

Summary

Fay, Brittain, and Proschan (2013) developed the beta product confidence procedure (BPCP) to create confidence intervals for a survival distribution for right censored data. Here we detail how the `bpcp` R package handles discrete failure times. Prior to version 1.3.0, the `bpcp` function had an awkward convention for defining the confidence interval exactly at the failure time. These notes explain that convention and detail the new one described in Fay and Brittain (2016) which is used in versions 1.3.0 or greater. Then the notes give the details of how the `bpcp` handles discrete failure times in terms of defining a grouping interval for all "observed" failures.

For users not interested in details who only want to know the recommended confidence intervals on right censored data when ties are allowed, we recommend the `bpcp` function version 1.3.0 or greater using the default Delta=0 argument. That recommendation will give pointwise confidence intervals that treats ties similarly to the way that the Kaplan-Meier estimator treats ties, and hence will give confidence intervals that enclose the Kaplan-Meier estimate.

1 Changing Conventions for Confidence Intervals Exactly at a Failure Time

1.1 Example of the Problem with Old Convention

The beta product confidence procedure (BPCP) of Fay, Brittain and Proschan (2013) is based on the assumption that the data are continuous. For continuous failure times $X_1, \ldots, X_n$, then $S(t) = Pr[X_i > t]$ is the same as $\bar{S}(t) = Pr[X_i \geq t]$. So for continuous data, a confidence interval procedure for $S(t)$ has the same probability of covering $S(t)$ as it does of covering $\bar{S}(t)$. Of course real data are not continuous, since they can be represented only with a finite number of digits (or as rational numbers, since technically 1/3 does not have a finite number of digits). For example, consider the fake data with 4 failure times at $t = 3, 7, 8$ and 14 and no censoring. Even though there are no ties, the data are not continuous because each failure time is an integer. If we run the default BPCP for versions 1.2.6 and earlier, then we get

```
> ## Actually run on bpcp Version 1.2.6. Not run on current version of package
> library(bpcp)
> packageVersion("bpcp")

[1]    '1.2.6'

> b1<-bpcp(c(3,7,8,14),c(1,1,1,1))
> summary(b1)
```

```
  time interval survival lower 95% CL upper 95% CL
1         (0,3)     1.00  0.397635364    1.0000000
2         [3,3]     0.75  0.397635364    0.9936905
3         (3,7)     0.75  0.194120450    0.9936905
4         [7,7]     0.50  0.194120450    0.9324140
5         (7,8)     0.50  0.067585986    0.9324140
6         [8,8]     0.25  0.067585986    0.8058796
7        (8,14)     0.25  0.006309463    0.8058796
8       [14,14]     0.00  0.006309463    0.6023646
9      (14,Inf)     0.00  0.000000000    0.6023646
```

At the last failure time, $t = 14$, the Kaplan-Meier estimate of survival is 0, but the 95% BPCP confidence interval excludes 0, since it is the middle 95% of the $Beta(1, 4)$ distribution,

```
>  qbeta(c(.025,.975),1,4)

[1] 0.006309463 0.602364636
```

The BPCP is derived from the probability integral transformation, which says that for continuous data, the survival distribution at a random failure time is uniformly distributed. Thus, the survival at the j failure time out of n uncensored failure times is the jth order statistic of n independent uniform random variables, and it is distributed $Beta(n-j+1, j)$. Let T_4 be the random variable for the 4th failure time. For continuous data the distribution of $S(T_4)$ and $\bar{S}(T_4)$ are identical. But the convention of using the distribution $Beta(n - j + 1, j)$ for $S(T_j)$ is not as useful as using it for $\bar{S}(t)$, which will translate better to discrete data. So it is better to assume $\bar{S}(T_j) \sim Beta(n-j+1, j)$. Then we use the fact that for any t, $S(t) = \lim_{\epsilon \to 0} \bar{S}(t+\epsilon) \equiv \bar{S}(t+)$, to get the confidence interval on $S(t)$ exactly at $t = T_j$. For example, by the new convention the lower limit for $S(14)$ from the previous example will be 0. This is clearly a better convention since the confidence interval now includes the Kaplan-Meier estimator.

So in `bpcp` Version 1.3.0 or greater, we use this convention.

```
> library(bpcp)
> packageVersion("bpcp")

[1] '1.3.3'

> b2<-bpcp(c(3,7,8,14),c(1,1,1,1))
> summary(b2)

  time interval survival lower 95% CL upper 95% CL
1         [0,3)     1.00  0.397635364    1.0000000
2         [3,7)     0.75  0.194120450    0.9936905
3         [7,8)     0.50  0.067585986    0.9324140
4        [8,14)     0.25  0.006309463    0.8058796
5      [14,Inf)     0.00  0.000000000    0.6023646
```

1.2 Definition of BPCP under Both Conventions with Continuous Failure Times

Let $Y(t)$ be the number of subjects at risk just before time t. Suppose the failure times are continuous, and let $T_1 < T_2 < \cdots < T_k$ be the ordered observed failure times. Because the failure times are continuous the probability that any two failure times are equal is zero. Additionally, the probability that any failure time falls on the same time as a censoring time is

also zero. For convenience define $T_0 = 0$ and $T_{k+1} = \infty$. Let $B(a, b)$ be a beta random variable with parameters $a > 0$ and $b > 0$, and define $B(0, b) = \lim_{a \to 0} B(a, b)$ as a point mass at 0. Let

$$W(t) = \begin{cases} 1 & \text{if } t = 0 = T_0 \\ \prod_{i=1}^{j} B\{Y(T_i), 1\} & \text{if } t = T_j, \text{ for } j = 1, \ldots, k \\ B\{Y(t), 1\} \prod_{i=1}^{j} B\{Y(T_i), 1\} & \text{if } T_j < t < T_{j+1}, \text{ for } j = 1, \ldots, k \end{cases} \quad (1)$$

where all the random variables are independent. For t larger than the largest censored observation, then $Y(t) = 0$ and $W(t)$ is a point mass at 0.

Let $W(T_j-) = \lim_{\epsilon \to 0} W(T_j - \epsilon)$ and $W(T_j+) = \lim_{\epsilon \to 0} W(T_j + \epsilon)$ with $\epsilon > 0$. Define the $100(1-\alpha)\%$ BPCP confidence interval for $S(t)$ as

$$\left[q\left\{\alpha/2, W^+(t)\right\}, q\left\{1-\alpha/2, W^-(t)\right\}\right]$$

The definition of $W^+(t)$ and $W^-(t)$ differs based on the convention. For $T_{j-1} < t < T_j$ then $W^-(t) = W(T_{j-1})$ and $W^+(t) = W(t)$ for both conventions. The conventions differ at $t = T_j$. Here are both conventions:

	Old Convention		New Convention	
t	$W^+(t)$	$W^-(t)$	$W^+(t)$	$W^-(t)$
T_j-	$W(T_j)$	$W(T_{j-1})$	$W(T_j)$	$W(T_{j-1})$
T_j	$W(T_j)$	$W(T_j)$	$W(T_j+)$	$W(T_j)$
T_j+	$W(T_j+)$	$W(T_j)$	$W(T_j+)$	$W(T_j)$

2 Discrete Data and the Beta Product Confidence Procedure

We now give the details of how discrete data is handled under the new convention of Fay and Brittain (2016). As in Fay, Brittain, and Proschan (2013), under the new convention we assume that the underlying data generating process produces continuous time failures, but we can only assess whether those failures have occurred or not at a finite number of assessment times.

2.1 Discrete Data in Continuous Notation

Assume that the failures occur in continuous time, so that there are no ties. As previously, let $X_1, \ldots, X_n$ denote the n failure times. Now suppose that all individuals have a potential censoring time, and let C_i^* be that potential censoring time for the ith individual. If we could observe the data in continuous time, we would be able to assign each individual with the indicator, $\delta_i = I(X_i \leq C_i^*)$, where $\delta_i = 1$ would denote an observed failure, and $\delta_i = 0$ would denote a right censored observation. As before let $T_1 < \cdots < T_k$ be the observed failure times assuming continuous observation, and now let $C_1 \leq \cdots \leq C_{n-k}$ be the ordered censoring times (i.e., the ordered values of C_i^* with $\delta_i = 0$). In other words, if the ith subject has a censored value at C_j, then $Pr[X_i > C_j] = 1$.

Now suppose that we do not observe the data in continuous time, but can only make assessments at a finite number of assessment times, $g_0 \equiv 0 < g_1 < g_2 < \cdots < g_m < \infty \equiv g_{m+1}$. At the jth assessment time we determine how many individuals are known to have failed since the last assessment time (d_{j-1}), and how many individuals are still under observation and at risk for failure sometime in the future (n_j). Define $n_0 = n$ and $n_{m+1} = 0$. We relate the d_j and n_j to previous notation. Let

$n_j =$ the number at risk for failure just after g_{j-1}.

$d_j = \#T_i \in (g_{j-1}, g_j]$, the number of failures known to have occurred in $(g_{j-1}, g_j]$. Note $Pr[T_i = g_j] = 0$ for all i, j, so we need not worry about inclusion or exclusion the boundary of the interval.

Using simular notation for censoring, let

$c_j = \#C_i \in (g_{j-1}, g_j]$, the number of censored values known to have occurred in $(g_{j-1}, g_j]$.

For tractability, we make the conventional assumption that within each interval, $(g_{j-1}, g_j]$, all the failures occur before all the censored individuals, and no failure occurs exactly at a boundary point. In other words, if $d_j > 0$ and $c_j > 0$ that all the censored values occur after the failure values. Thus, $n_{j+1} = n_j - d_j - c_j$. So although we do not observe the $T_1, \ldots, T_k$ and $C_1, \ldots, C_{n-k}$ values, by the assumption about the order of the values within an interval, we can know $W^-(g_j)$ and $W^+(g_j)$ for $j = 0, 1, 2, \ldots, m+1$.

To motivate the expression for $W^-(g_j)$ generally using d_j and n_j notation, first consider $W^-(g_1)$. Note that $W^-(t)$ is defined as $W(T_j)$, where T_j is the largest failure time less than or equal to t. So by assumption, the largest failure time less than or equal to g_1 is T_{d_1}. If $d_1 = 0$ then trivially, $T_0 \equiv 0$ and $W^-(0) = 1$. If $d_1 > 0$ then because we assume that all the failures occur before the censored values within $(g_0, g_1]$,

$$\begin{aligned} W^-(g_1) &= \prod_{i=1}^{d_1} B(Y(T_i), 1) \\ &= \prod_{i=1}^{d_1} B(n - i + 1, 1) \\ &= B(n_1 - d_1 + 1, d_1), \end{aligned} \tag{2}$$

where expression 2 comes from Fay, et al 2013, equation 2.1 (see also Casella and Berger, 2002, p. 158) who show that for $a > j > 0$,

$$\prod_{i=1}^{j} B(a - i + 1, 1) = B(a - j + 1, j).$$

Since $B(a, 0)$ is a point mass at 1 for $a > 0$, we can define $W^-(g_j)$ iteratively as

$$W^-(g_j) = W^-(g_{j-1}) B(n_j - d_j + 1, d_j).$$

Now consider the expression for $W^+(g_j)$. Using the new convention, $W^+(g_j) = W(g_j+)$. From equation 1, and using the continuity assumption which ensures that the probability of a failure at g_j is 0, we see that

$$\begin{aligned} W^+(g_j) &= W^-(g_j) B(Y(g_j+), 1) \\ &= W^-(g_j) B(n_{j+1}, 1). \end{aligned}$$

For the BPCP confidence interval, we use conservative assumptions. We also use the new convention that the confidence interval is right continuous for each interval. So we define the $100(1-\alpha)\%$ BPCP confidence interval for $S(t)$ for any $t \in [g_{j-1}, g_j)$ as:

$$q\left\{\alpha/2, W^+(g_j)\right\}, q\left\{1 - \alpha/2, W^-(g_{j-1})\right\} \tag{3}$$

where $q(a, X)$ is the ath quantile of the random variable X.

2.2 Discrete Data as Handled in bpcp R package: Delta= 0

Suppose you have data that are nearly continuous failure times, but where ties are allowed. The default way that the `bpcp` function (version 1.3.0 or greater) handles this is to assume discrete data with Delta=0. The notation and details are given in the next section, but roughly speaking the Delta=0 means that ties are allowed and the width of the intervals for the failures approach

0 in the limit. This will give confidence intervals that make sense with the way the Kaplan-Meier estimator treats tied values. So for most situations this will be a reasonable default method for obtaining confidence intervals for any right censored data (even with ties).

The details with Delta>0 (rarely needed) and Delta$=0$ (with precise notation) are given in the next section.

2.3 Discrete Data as Handled in bpcp R package: Delta>0

In this section, we detail how the data are input into the `bpcp` function. We then translate that into the notation of the Section 2.1 in order to get the BPCP confidence intervals.

We input the data as times for each of n individuals, $t_1^*, \ldots, t_n^*$, and associated with each time is a status indicator, $\delta_1^*, \ldots, \delta_n^*$, where $\delta_i^* = 1$ if t_i^* represents a failure time, and $\delta_i^* = 0$ if t_i^* represents a censoring time. Let X_i be the unobserved continuous failure time associated with the ith individual. Let C_i^* represent an unobserved right censored observation associated with the ith observation in continuous time.

Assume that the time is grouped into discrete intervals. Let Δ be the size for the time intervals. Then we use the convention that the data for the ith individual, (t_i^*, δ_i^*) represents the following:

$(t_i^*, 1)$ means that $X_i \in (t_i^* - \Delta, t_i^*]$

$(t_i^*, 0)$ means that the right censored observation associated with i, say C_i^*, has $C_i^* \in (t_i^* - \Delta, t_i^*]$, and $X_i > t_i^*$.

As in Section 2.1, we assume that if there are ties, then all the failures come before all the censored observations within an interval.

We require that Δ is less than or equal to the smallest absolute difference $|t_i - t_j|$, for any i, j to avoid overlapping failure time intervals (i.e., the first type of intervals).

Let $u_1 < u_2 < \cdots < u_h$ be the ordered unique values of $t_1^*, \ldots, t_n^*$. For convenience define $u_0 = 0$ and $u_{h+1} = \infty$. First, suppose that Δ is less than the smallest difference $u_j - u_{j-1}$ for $j = 1, \ldots, h$. (We will deal with the case when Δ equals that smallest difference later.) Then we can partition the positive real line into $2h + 1$ intervals:

$$(0, u_1 - \Delta], \quad (u_1 - \Delta, u_1], \quad (u_1, u_2 - \Delta], \quad (u_2 - \Delta, u_2], \quad \ldots, \quad (u_h - \Delta, u_h], \quad (u_h, \infty).$$

or

$$(g_0, g_1], \quad (g_1, g_2], \quad (g_2, g_3], \quad (g_3, g_4], \quad \ldots, \quad (g_{2h-1}, g_{2h}], \quad (g_{2h}, g_{2h+1}).$$

in the notation of Section 2.1.

So because the BPCP confidence interval is right continuous, the BPCP interval for $S(t)$ for $t \in [g_{j-1}, g_j)$ is given by expression 3. Thus, we just need to define d_j and n_j in this context. Let

$d_j = \sum_{i=1}^n I(t_i^* = g_j)\delta_i^*$.

$c_j = \sum_{i=1}^n I(t_i^* = g_j)(1 - \delta_i^*)$.

and n_j is defined iteratively, with $n_0 = n$ and $n_{j+1} = n_j - d_j - c_j$. Note that when $g_j = u_i - \Delta$ then $d_j = c_j = 0$ and $n_{j+1} = n_j$. In this case, you can show that $W^-(g_j) = W^-(g_{j-1})$ and $W^+(g_j) = W^+(g_{j-1})$. This can save some computation time.

This defines the BPCP confidence interval for each of the $2h + 1$ intervals, $[g_0, g_1), \ldots, [g_{2h}, g_{2h+1})$. Now consider the case where $\Delta = u_j - u_{j-1}$ for some j. If $u_j - \Delta = u_{j-1}$ then we remove the interval $[u_{j-1}, u_j - \Delta)$, and the other intervals remain as previously defined.

When $\Delta = 0$ then the even intervals (e.g., $[u_j - \Delta, u_j)$ for $j = 1, \ldots, h$) disappear. leaving the $h + 1$ intervals

$$[0, u_1), [u_1, u_2), \ldots, [u_{h-1}, u_h), [u_h, \infty).$$

References

Casella, G, and Berger, RL (2002). *Statistical Inference, second edition* Duxbury: Pacific Grove, Ca.

Fay, MP, Brittain, EH, and Proschan, MA (2013). "Pointwise Confidence Intervals for a Survival Distribution with Small Samples or Heavy Censoring." *Biostatistics* **14**(4): 723-736.

Fay, MP, and Brittain, EH (2016). "Finite Sample Pointwise Confidence Intervals for a Survival Distribution with Right-Censored Data." *Statistics in Medicine.* DOI: `10.1002/sim.6905`