Supporting Information for

The Communication Distance of Intermittent Streams

Ken Aho[1], Dewayne Derryberry[2], Sarah E. Godsey[3], Rob Ramos[4], Sara Warix[3†], Sam Zipper[5]

[1]Department of Biological Sciences, Idaho State University, Pocatello, ID, USA, 83209-8007, [2]Department of Mathematics and Statistics, Idaho State University, Pocatello, ID, USA, 83209-8085, [3]Department of Geosciences, Idaho State University, Pocatello, ID, USA, [4]Division of Biology, Kansas State University, Manhattan, KS 66506, [5]Kansas Geological Survey, The University of Kansas, Lawrence, KS 66047-3724

Corresponding author: Ken Aho (kenaho@isu.edu)

[†]Current address: Department of Geology and Geological Engineering, Colorado School of Mines, Golden, CO, USA.

Contents of this file

Supplementary Materials S1 – Derivation of the Inverse Beta Distribution and Its Moments

S1-1 The inverse beta probability density function

If a random variable X follows a beta distribution with parameters $\alpha, \beta > 0$, then it will have the probability density function (PDF):

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}.$$

where $\Gamma(.)$ refers to the gamma function, $0 < y < 1$, $\alpha, \beta > 0$.

To derive the inverse-beta PDF, we have the one-to-one transformation: $s(Y) = 1/Y = X$, with Jacobian $\frac{d}{dy}s(y)$. Thus, we have:

$$\begin{aligned} g(y) &= f[s(y)]\left|\frac{d}{dy}s(y)\right| \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\left(\frac{1}{y}\right)^{\alpha-1}\left(1-\frac{1}{y}\right)^{\beta-1}\left|\frac{1}{y^2}\right|. \end{aligned}$$

We refer to the result of this transformation as the inverse-beta probability density function. Specifically, if a random variable Y follows an inverse beta distribution, then we denote this as: $Y \sim BETA^{-1}(\alpha, \beta)$ for $1 < y < \infty$, $\alpha, \beta > 0$. Several examples of the inverse-beta PDF are shown in Fig. S1-1

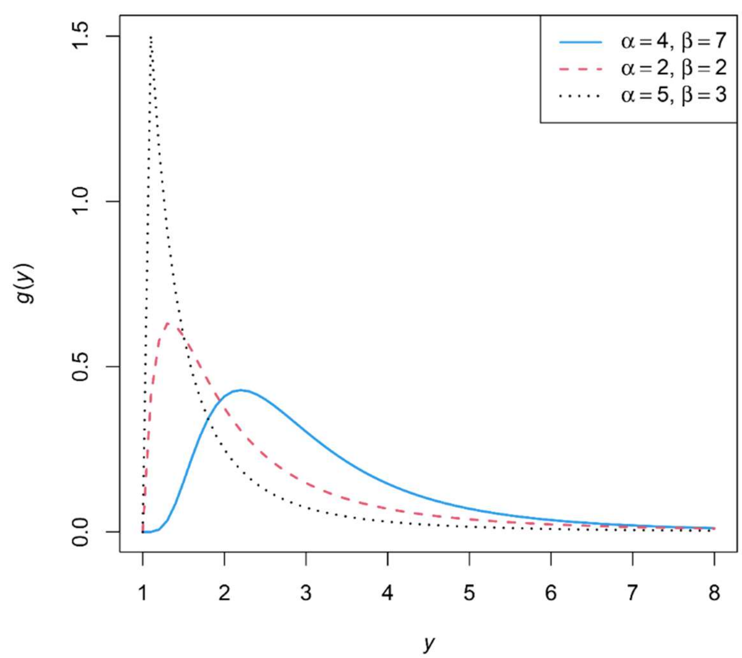


Figure S1-1. Distributions for $BETA^{-1}(\alpha, \beta)$.

S1-2 Moment generating function

$$E(Y^k) = \int_1^\infty \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^k \left(\frac{1}{y}\right)^{\alpha+1} \left(1-\frac{1}{y}\right)^{\beta-1} dy$$

Note: $y^k = \left(\frac{1}{y}\right)^{-k}$

$$= \int_1^\infty \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \left(\frac{1}{y}\right)^{\alpha-k+1} \left(1-\frac{1}{y}\right)^{\beta-1} dy$$

Let $u = \frac{1}{y}$, so that $-u^{-2}du = dy$

$$= \int_1^0 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} u^{\alpha-k+1}(1-u)^{\beta-1}(-u^{-2}du)$$

$$= -\int_1^0 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} u^{\alpha-k+1}(1-u)^{\beta-1}du$$

$$= \int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} u^{\alpha-k-1}(1-u)^{\beta-1}du$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 u^{\alpha-k-1}(1-u)^{\beta-1}du$$

Note, below we are integrating a beta PDF, when $(\alpha + \beta) - k > 0$, over its entire support.

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \frac{\Gamma(\alpha-k)\Gamma(\beta)}{\Gamma(\alpha-k+\beta)} \int_0^1 \frac{\Gamma(\alpha-k+\beta)}{\Gamma(\alpha-k)\Gamma(\beta)} u^{\alpha-k-1}(1-u)^{\beta-1}du$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \frac{\Gamma(\alpha-k)\Gamma(\beta)}{\Gamma(\alpha-k+\beta)}(1)$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta-k)} \cdot \frac{\Gamma(\alpha-k)}{\alpha}.$$

Note, the moment integral assumes $\alpha > k$, suggesting that the variance and mode do not exist for $\alpha \leq 2$.

S1-2.1 First moment (mean)

$$E(Y) = E(Y^1) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta-1)} \cdot \frac{\Gamma(\alpha-1)}{\Gamma(\alpha)}$$

Note, $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$ and $\Gamma(\alpha+\beta) = (\alpha+\beta-1)\Gamma(\alpha+\beta-1)$

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

S1-2.2 Second moment, and variance

$$\begin{aligned}
E(Y^2) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta-2)} \cdot \frac{\Gamma(\alpha-2)}{\Gamma(\alpha)} \\
&= \frac{(\alpha+\beta-1)(\alpha+\beta-2)\Gamma(\alpha+\beta-2)}{\Gamma(\alpha+\beta-2)} \cdot \frac{\Gamma(\alpha-2)}{(\alpha-1)(\alpha-2)\Gamma(\alpha-2)} \\
&= \frac{(\alpha+\beta-1)(\alpha+\beta-2)}{(\alpha-1)(\alpha-2)}
\end{aligned}$$

$$\begin{aligned}
Var(Y) &= E(Y^2) - E(Y^2) \\
&= \frac{(\alpha+\beta-1)(\alpha+\beta-2)}{(\alpha-1)(\alpha-2)} - \left(\frac{\alpha+\beta-1}{\alpha-1}\right)^2 \\
&= \frac{(\alpha+\beta-1)(\alpha+\beta-2)}{(\alpha-1)(\alpha-2)} - \frac{(\alpha+\beta-1)^2}{(\alpha-1)^2} \\
&= \frac{(\alpha-1)(\alpha+\beta-1)(\alpha+\beta-2) - (\alpha-2)(\alpha+\beta-1)^2}{(\alpha-1)^2(\alpha-2)} \\
&= \frac{(\alpha+\beta-1)(\alpha-1)(\alpha+\beta-2) - (\alpha-2)(\alpha+\beta-1)}{(\alpha-1)^2(\alpha-2)} \\
&= \frac{(\alpha+\beta-1) \cdot \alpha \left[(\beta-2) - (\beta-1) + 1 + \frac{\beta}{\alpha}\right]}{(\alpha-1)^2(\alpha-2)} \\
&= \frac{(\alpha+\beta-1)(\beta)}{(\alpha-1)^2(\alpha-2)} \\
&= \frac{(\alpha+\beta-1)}{\alpha-1} \cdot \frac{\beta}{(\alpha-1)(\alpha-2)}.
\end{aligned}$$

Note: The form of the variance indicates the mode $= \frac{\alpha+\beta-2}{\alpha-1}$. The median cannot be found in closed form.

Supplementary Materials S2 – Prior Weights for Obtaining Reasonable Inverse-Beta Posteriors

If $Y \sim BETA^{-1}(\alpha, \beta)$, then $E(Y) = \frac{\alpha+\beta-1}{\alpha-1}$. Thus, for a finite mean to exist, $\alpha > 1$, (see Supplementary Materials S1).

For our application, the α parameter in the posterior inverse beta distribution had the form,

$$\alpha = w \cdot n \cdot \hat{p}_{k(prosper)} + \sum \boldsymbol{x}_k$$

where *w* is the weight given to the prior and *n* denotes the sample size. The $\hat{p}_{k(prosper)}$ outcomes were based on the Probability of Streamflow Permanence model (PROSPER; Jaeger et al. 2019), as reported for Murphy Creek stream segments by the United States Geological Survey (USGS) StreamStats web-based application (USGS, 2016).

Assume that for the *k*th arc, PROSPER assigns the minimum probability given to any arc in the Murphy Creek network, $\hat{p}_{k(prosper)} = 0.21$. Assume further that no stream presence outcomes were recorded at the *k*th segment (i.e., $\sum \boldsymbol{x}_k = 0$), over *n* = 10 trials. To obtain finite posterior means, we have the constraint:

$$2.1(w) + 0 > 1$$

$$w > 0.47$$

Weighting constraints under *n* = 5, *n* = 10, and *n* = 20 are shown in Fig S2-1.

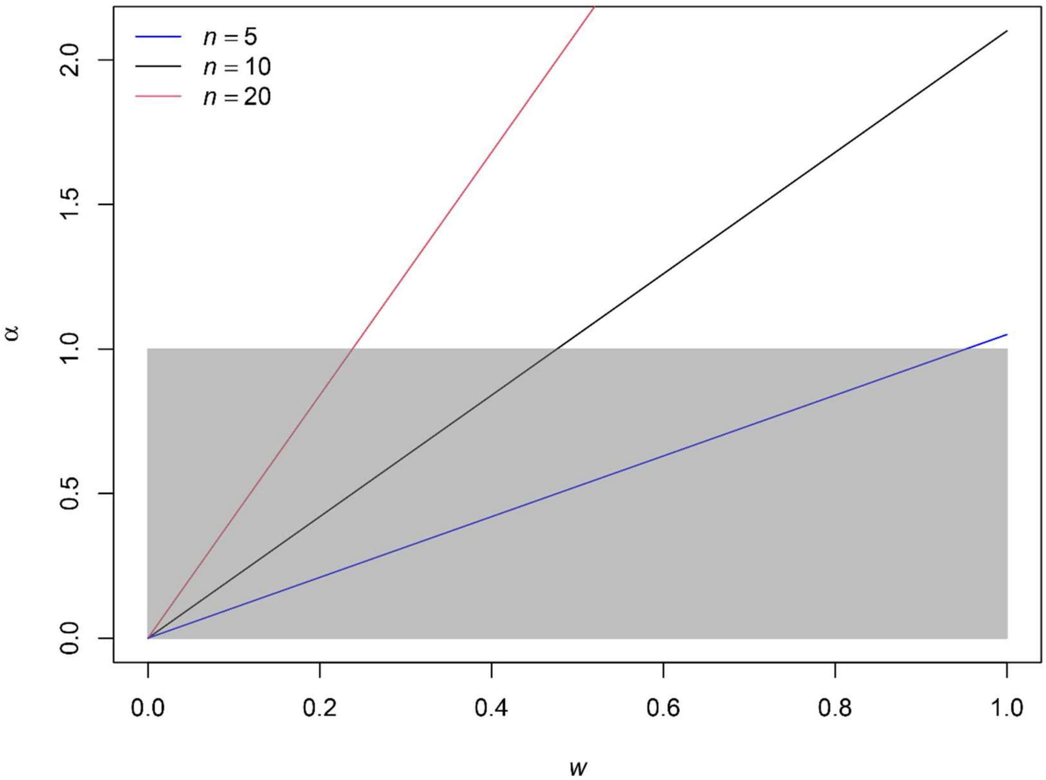


Fig S2-1. Effect of conceptual weights on finite values for the inverse beta mean for the *k*th arc with a low prior probability of stream presence ($\hat{p}_{k(prosper)} = 0.21$), and no observations of stream presence (i.e., $\sum \boldsymbol{x}_k = 0$) . Graph regions in which an inverse beta posterior mean can become infinitely large are shown in grey. Note that these regions include weights ≤ 0.47 when n = 10.

We note that for our application at Murphy Creek, the issue of infinitely large posterior inverse beta means was only important in the driest season at the driest segments. In general, for a stream segment to actually exist over some segment of time, we assume that $\sum \boldsymbol{x}_k \geq 1$ as n grows large and, thus $\hat{p}_k > 0$ over some large number of trials. This outcome would allow any prior weight ≥ 0. For the inverse beta distributional variance to exist, larger prior weights will be needed, since this requires $\alpha > 2$. The issue of infinitely large posterior beta distribution means will not be an issue unless $\alpha = \beta = 0$.

References

Jaeger, K., Sando, R., McShane, R. R., Dunham, J. B., Hockman-Wert, D., Kaiser, K. E., Hafen, K., Risley, J., & Blasch, K. (2019), Probability of streamflow permanence model (PROSPER): A spatially continuous model of annual streamflow permanence throughout the pacific northwest. *Journal of Hydrology X*, *2*, 100005.

USGS. (2016), *StreamStats*. http://streamstats.usgs.gov.

Supplementary Materials S3 – A Suggested Step by Step Guide to Analyses, with Examples

S3-1 Suggested steps in analyses

1. Record stream presence data from logically located nodal locations in a stream network.
2. Obtain stream arc (segment) flow presence data for arc bounding nodes. For the kth arc with bounding nodes u and v, for the ith time frame, $i = 1,2,3, \ldots, n$, there are three possibilities:

$$x_{k,i} = \begin{cases} 1.0, & \text{both } u \text{ and } v \text{ wet} \\ 0.0, & \text{both } u \text{ and } v \text{ dry} \\ 0.5, & \text{only one of } u \text{ or } v \text{ wet} \end{cases}$$

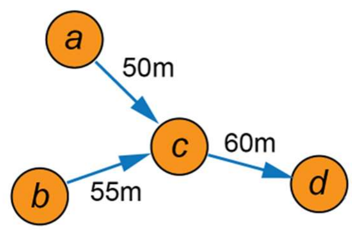
   For outlet and source segments, information from only node may be available resulting in only two possible values for $x_{k,i}$: 0 (dry) and 1 (wet). Implementation of this rule is facilitated by the function `arc.pa.from.nodes` in the **R** package *streamDAG* (Aho 2022).
3. Estimate the segment marginal probability of arc surface water presence by finding the sample arithmetic mean of presence/absence data from step two[1].
4. Estimate the variance/covariance structure of arcs to each other using conventional method of moments estimators (i.e., `var()` in **R**) using data from step 2. Segment correlations (standardized covariances) are constrained by the bounds given in Eq 6 and 7 in the manuscript. These bounds can be checked using the function `R.bounds` from *streamDAG*.
5. An estimate for average stream length can now be obtained by taking the sum of dot product of the marginal probabilities of arc presence and their respective marginal arc lengths. An estimate for the average communication distance can be obtained by taking the sum of the dot product of the inverse marginal probabilities of arc presence and the respective marginal arc lengths. These computations are provided by the function `bern.length` in *streamDAG*.
6. For Bayesian applications, one will need to either acquire prior information for the probability of arc surface water presence, independent of current data, or specify naïve priors. We suggest defining prior distributions as beta distributions to allow conjugacy with binomial likelihood functions. An informative beta prior distribution with a mean equal to a previously observed probability of a surface water presence can be defined using the approach shown in the manuscript in Eq 8. Weights for the priors (compared to current data) can be assigned to further constrain the prior distribution using the approach described in Section 2.2. Resulting beta posteriors for the probability of arc surface water presence, and

[1] Note that node outcomes here are treated as independent observations when calculating probabilities for the presence their respective (conjoining) arc. However, this outcome is more realistically a correlated bivariate Bernoulli process. We will consider this issue in future work.

the inverse beta inverse probability of arc surface water presence are given in Eq. 19-21, and Eq 22-24, respectively. These calculations are provided by the function `beta.posterior` in *streamDAG*.

S3-2 Example

Consider a network of four nodes: *a*, *b*, *c*, *d*, and three arcs: $\overrightarrow{ac}$, $\overrightarrow{bc}, and\ \overrightarrow{cd}$. Stream flow presence is observed at the four nodes at three time frames, with wet = 1, dry = 0 (Fig S3-1).



Time	*a*	b	c	d
1	1	1	1	1
2	0	1	1	1
3	0	0	1	1

Figure S3-1. Example stream network with four nodes and three arcs, observed at three time frames.

In practice, we recommend using relatively large time series datasets for the simulation approaches described in this section. For instance, the Murphy Creek dataset in this paper was based on sensor readings at 15-minute intervals over a four-month period, resulting in over 11,600 observations for each of the 24 measured nodes.

From the data in Fig S3-1, we obtain the arc presence outcomes shown in Table S3-1, using the rule from step 2.

Table S3-1. Arc presence/absence outcomes based on the rules given in Step 2 above.

	$\overrightarrow{ac}$	$\overrightarrow{bc}$	$\overrightarrow{cd}$
Time 1	1.0	1.0	1.0
Time 2	0.5	1.0	1.0
Time 3	0.5	0.5	1.0

Implementation of this rule is facilitated by the function arc.pa.from.nodes in the **R** package *streamDAG* (Aho 2022).

```
library(streamDAG)
G <- graph_from_literal(a --+ c, b --+ c --+ d)
p.a <- matrix(nrow = 3, data = c(1,0,0,1,1,0,1,1,1,1,1,1))
colnames(p.a) <- c("a","b","c","d")
m <- arc.pa.from.nodes(G, p.a)[,c(1,3,2)]
m
     a -> c c -> d b -> c
[1,]    1.0      1    1.0
[2,]    0.5      1    1.0
[3,]    0.5      1    0.5
```

Thus, the estimated marginal probabilities for surface water presence are $\hat{p}_{ac} = 0.67, \hat{p}_{bc} = 0.83, \hat{p}_{cd} = 1.0$.

The estimated mean stream length is: $0.67(50\text{m}) + 0.83(55\text{m}) + 1.0(60\text{m}) = 139.17\text{m}$. The estimated mean communication distance is: $(1/0.67)(50\text{m}) + (1/0.83)(55\text{m}) + 1.0(60\text{m}) = 200.89\text{m}$. These calculations can be made using the function `bern.length` from *streamDAG*.

```
length <- c(50,55,60)
bern.length(colMeans(m), length)
  a -> c   b -> c   c -> d
33.33333 45.83333 60.00000

bern.length(colMeans(m), length, "global")
[1] 139.1667
```

We used the marginal arc probabilities of surface water presence and stream network covariance structures to obtain multivariate Bernoulli outcomes using functions from the **R** library *mipfp* (Barthélemy & Suesse 2018). The estimated variance-covariance matrix is shown in Table S3-2.

Table S3-2. Estimated variance-covariance matrix for the data in Table S3-1.

	$\overrightarrow{ac}$	$\overrightarrow{bc}$	$\overrightarrow{cd}$
$\overrightarrow{ac}$	0.83	0.04	0.00
$\overrightarrow{bc}$	0.04	0.83	0.00
$\overrightarrow{cd}$	0.00	0.00	0.00

Here we obtain the matrix using **R**:

```
var(m)
           a -> c     b -> c c -> d
a -> c 0.08333333 0.04166667      0
b -> c 0.04166667 0.08333333      0
c -> d 0.00000000 0.00000000      0
```

Note that the marginal variance of arc $\overrightarrow{cd}$ is zero. Because this will prevent calculation of correlations of arcs to $\overrightarrow{cd}$ (and because this will save computational time), we will only simulate multivariate Bernoulli outcomes for arc $\overrightarrow{ac}$ and arc $\overrightarrow{bc}$. We will assume that, because only ones were observed for $\overrightarrow{cd}$, that $\overrightarrow{cd}$ is perennial, and that its Bernoulli outcomes will always be 1.

We confirm that the estimated correlations for the matrix $\boldsymbol{R}$ are within the possible bounds given in the manuscript in Eqs. 6 and 7, using the function `R.bounds` from *streamDAG*.

```
R <- cor(m[,1:2]) # correlation matrix for segments ac and bc
marg.probs <- colMeans(m)[1:2] # marginal probs for ac and bc
R <- R.bounds(marg.probs, R)
```

To generate 10 random multivariate Bernoulli outcomes from this framework we have:

```
library(mipfp)
p.joint.all <- ObtainMultBinaryDist(corr = R, marg.probs = marg.probs,
tol = 0.001, tol.margins = 0.001, iter = 100)

set.seed(2) # to get same results as those shown below
out <- RMultBinary(n = 10, p.joint.all, target.values =
NULL)$binary.sequences

# bind c to d arc outcomes (ones) to outcomes from other two arcs
out.w.cd <- cbind(out, rep(1,10))
colnames(out.w.cd)[3] <- "c -> d"

out.w.cd
      a -> c b -> c c -> d
 [1,]      1      1      1
 [2,]      0      0      1
 [3,]      0      0      1
 [4,]      0      1      1
 [5,]      1      1      1
 [6,]      1      1      1
 [7,]      1      1      1
 [8,]      0      1      1
 [9,]      1      1      1
[10,]      0      0      1
```

The resulting simulation-based marginal probability estimates are:

```
colMeans(out.w.cd)
 a -> c  b -> c c --> d
```

```
    0.5      0.7      1.0
```

Basing estimates on many iterations will stabilize them.

S3-3 Bayesian extensions

Assume now that we wish to apply a naïve Bayesian prior for the probability of surface water presence to a stream segment. For instance, $\theta_k \sim BETA(1,1)$. This distribution is equivalent to a continuous uniform distribution in 0,1, and will have the mean, $E(\theta_k) = 0.5$. Assume further that we wish to give the priors 1/3 of the weight of the data (outcomes in our single ten observation simulation shown above).

According to Eq. 19 in the manuscript, the posteriors for the probability of surface water presence for $\overrightarrow{ac}$, $\overrightarrow{bc}$ and $\overrightarrow{cd}$ will have the form: $BETA(w \cdot n \cdot \hat{p}_k + \sum \boldsymbol{x}_k, w \cdot n(1-\hat{p}_k) + n - \sum \boldsymbol{x}_k)$. Thus, the posteriors for $\overrightarrow{ac}$, $\overrightarrow{bc}$ and $\overrightarrow{cd}$ are:

$$BETA(1/3 \cdot 10 \cdot 0.5 + \ 5, 1/3 \cdot 5 + 10 - 5) \ = BETA(6.67, 6.67)$$

$$BETA(1/3 \cdot 10 \cdot 0.5 + \ 7, 1/3 \cdot 5 + 10 - 7) \ = BETA(8.67, 4.67)$$

$$BETA(1/3 \cdot 10 \cdot 0.5 + \ 10, 1/3 \cdot 5 + 10 - 10) \ = BETA(11.67, 1.67)$$

The means of these distributions are $\frac{\alpha}{\alpha+\beta}$ = 0.5, 0.65, and 0.875, respectively.

Linearly transforming the stream segment posteriors by multiplying them by their respective lengths provides a posterior for stream segment length.

The posteriors for the inverse probability of surface water presence for $\overrightarrow{ac}$, $\overrightarrow{bc}$ and $\overrightarrow{cd}$ are the inverse beta distributions: $BETA^{-1}(6.67, 6.67)$, $BETA^{-1}(8.67, 4.67)$, and $BETA^{-1}(11.67, 1.67)$. The means of these distributions are: $\frac{\alpha+\beta-1}{\alpha-1}$ = 2.18, 1.61, and 1.15, respectively.

Linearly transforming the stream segment posteriors by multiplying them by their respective lengths provides a posterior for stream segment communication distance.

Derivation of these distributions and their expectations is provided by the function `beta.posterior` in *streamDAG*.

```
beta.posterior(p.prior = 0.5, dat = out.w.cd, length = length, w = 1/3)

$alpha
   a -> c    b -> c   c --> d
 6.666667  8.666667 11.666667

$beta
```

```
  a -> c   b -> c  c --> d 
6.666667 4.666667 1.666667 

$mean
 a -> c  b -> c c --> d 
  0.500   0.650   0.875 

$var
     a -> c      b -> c     c --> d 
0.017441860 0.015872093 0.007630814 

$mean.inv
  a -> c   b -> c  c --> d 
2.176471 1.608696 1.156250 

$var.inv
    a -> c     b -> c    c --> d 
0.54869006 0.14688091 0.01868939 

$Com.dist
[1] 266.6768

$Length
[1] 113.25

$x
 a -> c  b -> c c --> d 
      5       7      10 
```

References

Aho, K. (2022), *StreamDAG*: Descriptors and Methods for Stream DAGs. R package version 0.1.2. https://github.com/moondog1969/streamDAG

Barthélemy, J., & Suesse, T. (2018), mipfp: An R package for multidimensional array fitting and simulating multivariate Bernoulli distributions. *Journal of Statistical Software, Code Snippets*, *86*(2), 1–20. https://doi.org/10.18637/jss.v086.c02