Beta random fractions of a size-biased sample

Geurt Jongbloed
Delft Institute of Applied Mathematics
Delft University of Technology
The Netherlands

November 15, 2012

1 Introduction

A well known problem in applied probability is the 'waiting time paradox' or 'inspection paradox' ([6], section 7.7). The original statement of this problem is concerned with a homogeneous Poisson process according to which buses arrive at a bus stop. At a randomly chosen point, a person arrives at the bus stop and at that moment his 'waiting period' starts. Upon arrival of the next bus, this waiting period ends and his waiting time is the time elapsed in between arrival and departure. In this idealistic setting, the 'paradox' is that the distribution of this remaining waiting time is exactly the same as the (exponential) distribution of the inter-arrival time between two consecutive buses. At first sight this might look strange since one can expect only having to wait *half* the time between two consecutive arrivals. The solution to this paradox is that the length of the interval of arrival does not have the same distribution as a typical inter arrival time. The customer is more likely to arrive in a longer interval than a short one. In fact, the inter arrival interval that is 'hit' by the customer can be viewed as a draw from a biased sample corresponding to the inter arrival distribution. Moreover, the remaining waiting time is a (uniform) random fraction of the length of the sampled interval. In the setting of the classical waiting time paradox with exponentially distributed inter-arrival distribution, the

two opposing mechanisms (selected interval tends to be *longer* than usual, but only *part of its length* is the actual waiting time) cancel. This cancelation is specific for the Poisson process; more general renewal processes do not share this property.

In section 2, the relation between the distribution of a random fraction from a biased draw from a distribution will be derived. The relation is specialized to the situation where the biasing happens proportional to a specific measure of size and the random fractions have a Beta distribution. Section 3 contains classical as well as more recent examples of the general problem.

2 The general problem

Consider a distribution function F on $(0,\infty)$ and a *weight function* $w : [0,\infty) \to [0,\infty)$ such that

$$0 < \int w(x)\, dF(x) < \infty.$$

The w-biased distribution associated to F has distribution function

$$F^{(w)}(x) = \frac{1}{w_F} \int_0^x w(y)\, dF(y), \quad \text{where } w_F = \int_0^\infty w(y)\, dF(y).$$

For weight function $w(x) = x$, the distribution w-biased distribution is known as the *length biased distribution*.

Now consider the second mechanism. Suppose $X \geq 0$ has distribution function G, $Y \geq 0$ has probability density k and X and Y are independent. Then, defining for $z > 0$ the set $C_z = \{(x,y) \in \mathbb{R}_+^2 : y \leq z/x\}$ the random variable $Z = XY$ has distribution function

$$\begin{aligned} G_Z(z) &= P(XY \leq z) = P((X,Y) \in C_z) = \int_{y=0}^\infty \int_{x=0}^{z/y} k(x) dG(y)\, dx \\ &= \int_{y=0}^\infty \int_{x=0}^{z} \frac{1}{y} k\left(\frac{x}{y}\right) dx\, dG(y) = \int_{x=0}^{z} \int_{y=0}^\infty \frac{1}{y} k\left(\frac{x}{y}\right) dG(y)\, dx \end{aligned}$$

with density

$$g_Z(z) = \int_{y=0}^\infty \frac{1}{y} k\left(\frac{z}{y}\right) dG(y)$$

If the random variable Y takes values in $(0,1)$, this boils down to

$$g_Z(z) = \int_{y=z}^\infty \frac{1}{y} k\left(\frac{z}{y}\right) dG(y)$$

Specializing even more, to Y having a $B(\alpha,\beta)$ distribution, so

$$k(y) = \frac{y^{\alpha-1}(1-y)^{\beta-1}}{B(\alpha,\beta)} 1_{(0,1)}(y),$$

we obtain the following density of the product Z:

$$g_Z(z) = \frac{z^{\alpha-1}}{B(\alpha,\beta)} \int_{y=z}^{\infty} y^{\alpha-\beta-1}(y-z)^{\beta-1}\, dG(y).$$

In combining the two mechanisms, we assume G to be a biased version of distribution function F. More specifically, for $\gamma \in \mathbb{R}$ we assume that

$$G(z) = \frac{1}{w_{F,\gamma}} \int_0^z y^\gamma\, dF(y) \text{ with } w_{F,\gamma} = \int_0^\infty y^\gamma\, dF(y).$$

Then the distribution of the observable Z can be written as

$$g_Z(z) = \frac{z^{\alpha-1}}{B(\alpha,\beta) w_{F,\gamma}} \int_{y=z}^{\infty} y^{\alpha-\beta+\gamma-1}(y-z)^{\beta-1}\, dF(y).$$

In section 3 we will see that some well known statistical models fit within the framework of estimating a distribution function F based on a sample of Beta random fractions of a 'size-biased' sample from F, where 'size' is measured as a power of the variable of interest: X^γ. Moreover, the Beta distributions all come from the branch with $\alpha = 1$.

3 Specific instances of the model

In this section we consider three models that fit within the framework of section 2. For the first two examples, the basic setting is identical. In an opaque medium, ball centers are distributed according to a (low intensity) homogeneous Poisson process and the corresponding squared ball radii all have the same distribution function F. Moreover, these are independent. One is interested in the distribution of the (squared) ball radii, but direct observations on the radii are not obtainable. In example 3.1, data are obtained by intersecting the medium with a line and observing the traces of the balls on the line. In example 3.2, the medium is cut by a plane and (circular) profiles of the balls on the intersecting plane are observed.

Example 3.1 *(Linear probe problem).* This model was described in [8]. Let us derive the distribution of the squared radius of a ball *given* that it is hit by the line. Think of a large block of size $M \times M \times M$ and assume for simplicity that the radii of the balls do not exceed the value $M/2$. Figure 1 gives the top-view of this block. Suppose the line intersects the block perpendicular to the plane in Figure 1, right at the middle (see the dot). Now, given the squared radius of a ball equals x, what is the probability that the (fixed) line intersects this ball when its center is placed uniformly at random within the block? This event occurs if and only if the center of the ball is placed in the circular cylinder with the intersecting line as axis, with radius $\sqrt{x}$. Denote this cylinder by C_x. Because the location of the center is taken uniformly over the opaque block, we have

$$P(\text{ball hits line}|X = x) = \frac{\text{volume of cylinder } C_x}{\text{volume of block}} = \frac{M\pi x}{M^3} = \frac{\pi x}{M^2}.$$

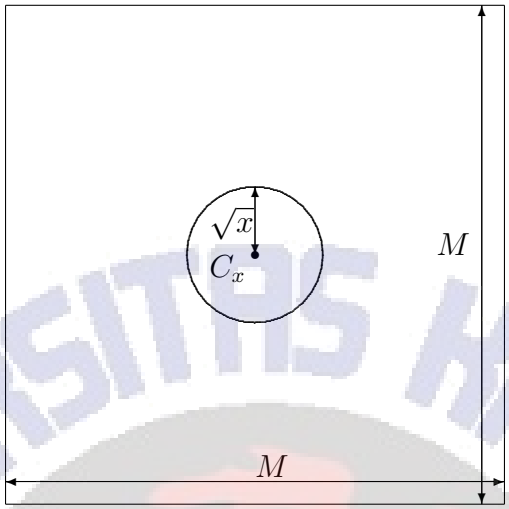

Figure 1: Top view of the $M \times M \times M$-block, perpendicular to the intersecting line (appearing as dot in the middle). Also the top view of cylinder C_x is given.

This leads to

$$P(X \leq x | \text{ball hits line}) = \frac{P(X \leq x \text{ and ball hits line})}{P(\text{ball hits line})} = \frac{\int_0^x y \, dF(y)}{\int_0^\infty y \, dF(y)}.$$

We recognize this distribution as the *length biased* distribution associated with F.

Now, given that a ball with squared radius X was hit, the distribution of the squared half-length of the trace of this ball on the line can be derived from the picture in Figure 2. The top view shows the contour of the ball of radius $\sqrt{X}$. Uniformly at random on this contour, the line intersects the ball. The dot indicates this position. The (perpendicular) distance of the ball center to the intersecting line is given by $V\sqrt{X}$, where for $0 \leq v \leq 1$

$$P(V \leq v) = \frac{\text{area of circle with radius } v}{\text{area of circle with radius one}} = v^2.$$

From the right picture in Figure 2 we see that the observed squared half-length of the line segment intersecting the ball is given by

$$Z = X(1 - V^2) = ZY$$

where Y is a random variable, independent of X, and for $0 \leq y \leq 1$

$$P(Y \leq y) = P(V^2 \geq 1 - y) = P(V \geq \sqrt{1 - y}) = y.$$

Hence, we see that a typical ball that is hit by the line has the length-biased distribution corresponding to F and that the actual observation is a (standard) uniform random fraction

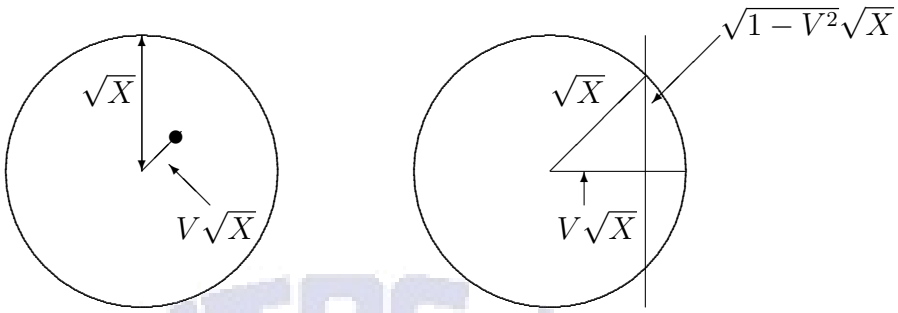

Figure 2: The left picture shows the top-view of an intersected ball of radius $\sqrt{X}$. The point indicates the (uniformly at random chosen) position where the line transects the ball. The distance of the ball center to the intersecting line is $V\sqrt{X}$. The right picture shows the same ball, but now from the direction perpendicular to the plane spanned by the ball center and the intersecting line. The 'squared half-length' of the observable line segment is given by $(1-V^2)X$.

of this observation. This leads to the following sampling density for the observed squared half-length of the trace:

$$g(z) = \frac{1}{B(1,1)w_{F,1}} \int_{y=z}^{\infty} dF(y) = \frac{1-F(z)}{\int_0^\infty (1-F(y))\,dy}.$$

Note that the inverse relation can be easily obtained from this expression. Indeed,

$$1 - F(z) = \frac{g(z)}{g(0)}.$$

This means that estimating F at a point z entails estimating the bounded decreasing density g at z as well as at 0. Estimation of g is usually done using the maximum likelihood estimator, also known as Grenander estimator ([1]). This estimator has good asymptotic properties on intervals $[\epsilon, \infty)$, but is inconsistent at zero. See e.g. [10] for a discussion on this inconsistency as well as a penalization approach that can be adopted to consistently estimate $g(0)$.

It is clear that the linear probe problem is related to the waiting time paradox described in the introduction. Note that an exponential distribution function F corresponds to the same exponential sampling density, as remarked in the introduction. Indeed, for $x \geq 0$

$$F(x) = 1 - e^{-\theta x} \iff g(z) = \theta e^{-\theta x}.$$

Example 3.2 *(Wicksell problem).* The setting of the classical Wicksell problem (see e.g. [9], [2], [7]) is the same as in the linear probe problem. However, instead of viewing traces of the spheres on a line (as in the linear probe problem), traces are observed on a plane that randomly intersects the medium. The first question in this model is again: what is the distribution of the squared radius of a ball, given the ball is intersected by the plane?

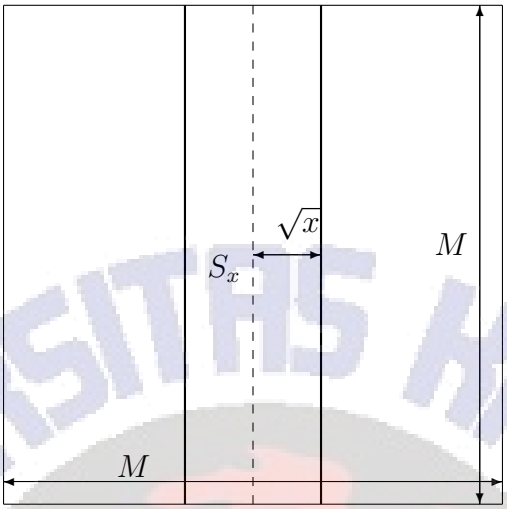

Figure 3: Top view of the $M \times M \times M$-block, perpendicular to the intersecting line (appearing as dashed line in the middle). Also the top view of slice S_x is given.

Putting the ball center in the block will now only give a nonempty intersection with the cutting plane if the center is put in the slice S_x depicted in Figure 3. We therefore get

$$P(\text{ball hits plane}|X = x) = \frac{\text{volume of slice } S_x}{\text{volume of block}} = \frac{2M^2\sqrt{x}}{M^3} = \frac{2\sqrt{x}}{M}.$$

This leads to

$$P(X \leq x|\text{ball hits plane}) = \frac{P(X \leq x \text{ and ball hits plane})}{P(\text{ball hits plane})} = \frac{\int_0^x \sqrt{y}\, dF(y)}{\int_0^\infty \sqrt{y}\, dF(y)}.$$

We recognize this distribution as the weighted distribution associated with F with weight function $w(x) = \sqrt{x}$.

As illustrated in Figure 4, given a sphere with squared radius X is cut, the observable squared radius of the observable circle can be expressed as $(1 - U^2)X$ for U standard uniformly distributed and independent of X. This means that Wicksell's problem fits within the framework of section 2 with $\gamma = \beta = 1/2$ and $\alpha = 1$. Indeed, the sampling density g can be expressed in terms of F:

$$g(z) = \frac{\int_z^\infty (y - z)^{-1/2}\, dF(y)}{2\int_0^\infty \sqrt{y}\, dF(y)}$$

The inverse relation is given by

$$1 - F(z) = \frac{\int_z^\infty (y - z)^{-1/2} g(y)\, dy}{\int_0^\infty y^{-1/2} g(y)\, dy}$$

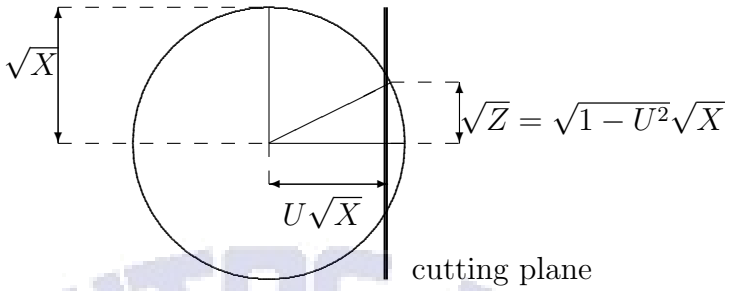

Figure 4: View on the sphere in the direction of the vertical cutting plane. The sphere radius is $\sqrt{X}$, the position of the cut is uniformly distributed on the right half of the sphere.

Example 3.3 *(Hampel problem).* Consider a population of birds of a certain type that cross the desert individually and stop over at an oasis. Ornithologists are interested in the time spent by a generic bird at this oasis. The following model can be used to investigate these sojourn times. This problem was first described in [4].

To each bird, a positive random variable X with distribution function F is attached, denoting the time spent at the oasis. This quantity cannot be observed. Also, independent of X, the bird has a homogeneous Poisson process $(N(t))_{t\geq 0}$ attached to it, with intensity λ, assumed to be small. The times the bird is caught at the oasis are then those jump points of the Poisson process that occur during the interval $[0, X]$. Note that $N(X)$ is the number of times the bird is caught.

The data for one bird consists of those catching times that occurred before X. Of course, conditional on the fact that $N(X) \geq 1$. In Hampel's model only those observations are used that correspond to birds that have been caught exactly twice. We will derive the distribution of the difference in time between the two catches in terms of the unknown distribution function X of the sojourn time.

The first question is: what is the distribution of the sojourn time of a bird given it is caught exactly twice?

$$\begin{aligned} P(X \leq x | N(X) = 2) &= \frac{P(X \leq x \wedge N(X) = 2)}{P(N(X) = 2)} = \frac{\int_{y\in[0,x]} P(N(X) = 2 | X = y) dF(y)}{\int_{y\in[0,\infty)} P(N(X) = 2 | X = y) dF(y)} \\ &= \frac{\int_{y\in[0,x]} \frac{1}{2}(\lambda y)^2 e^{-\lambda y} dF(y)}{\int_{y\in[0,\infty)} \frac{1}{2}(\lambda y)^2 e^{-\lambda y} dF(y)} \approx \frac{\int_{y\in[0,x]} y^2 dF(y)}{\int_{y\in[0,\infty)} y^2 dF(y)} \end{aligned}$$

for small λ (and, for example, if F has bounded support). In other words, the conditional distribution is (for small λ) the weighted distribution associated with F with weight $w(x) = x^2$.

The second question is: what is the distribution of Z (time between the two catches) given that $X = x$ and there are exactly two catches? Given X and $N(X) = 2$, the jumps of N in $[0, x]$ are uniformly distributed on this interval. Therefore, writing $U_{(1)}$ and $U_{(2)}$

for the order statistics of two uniformly distributed random variables in $(0, 1)$

$$P(Z > z|X = x) = P(U_{(2)} - U_{(1)} > z/x) = (1 - z/x)^2 \quad \text{for } 0 < z < x.$$

Consequently,

$$P(Z > z) = \int_0^\infty P(Z > z|X = x)\, d\tilde{F}(x) = \frac{\int_z^\infty (x - z)^2\, dF(x)}{\int_0^\infty x^2\, dF(x)}. \tag{1}$$

We see that we explicitly have to require the second moment of F to exist. However, if we take F to have bounded support, this assumption certainly holds.

Differentiating (1) with respect to z, we get for the density g of Z that

$$g(z) = \frac{2\int_z^\infty (x - z)\, dF(x)}{\int_0^\infty x^2\, dF(x)}$$

and differentiating once again gives

$$g'(z) = -\frac{2(1 - F(z))}{\int_0^\infty x^2\, dF(x)}$$

This shows that the density g is necessarily convex and decreasing on $(0, \infty)$. Moreover, the inverse relation expressing F in terms of g is given by

$$F(z) = 1 - \frac{g'(z)}{g'(0)}$$

The problem of estimating F boils down to estimating the derivative of a convex decreasing density g. In particular, estimating $F(z)$ for $z > 0$ entails estimation of the derivative of g at z and 0. This problem was studied thoroughly in [3].

4 Discussion

In this note we derive the relation between an underlying distribution function F and the density of a random variable that is obtained from a two-step procedure. This procedure entails drawing a biased observation from F and taking an (independent) random fraction of this draw, where the fraction has a beta distribution. This gives a whole family of inverse problems, where one can study the problem of estimating F based on a sample from density g.

Based on the two-step approach, for three particular inverse problems the relation between a distribution function of interest, F, and a sampling density g is derived. These problems are more or less well studied in the literature. An interesting question is whether also other problems from the specified family of inverse problems are relevant in practice.

References

[1] Grenander, U. (1957). On the theory of mortality measurement. II. *Skand. Aktuarietidskr.*, **39**, p. 125-153.

[2] Groeneboom, P. and Jongbloed, G. (1995). Isotonic estimation and rates of convergence in Wicksell's problem. *The Annals of Statistics* **23**, p. 1518-1542.

[3] Groeneboom, P., Jongbloed, G. and Wellner, J.A. (2001). Estimation of a convex function: characterizations and asymptotic theory. *The Annals of Statistics* **29**, p. 1653-1698.

[4] Hampel, F.R. (1987). Design, modelling and analysis of some biological datasets. In C.L. Mallows, editor, *Design, data and analysis, by some friends of Cuthbert Daniel*, p. 111-115, Wiley, New York.

[5] McGarrity, K.S., Sietsma, J. and Jongbloed, G. (2012). Nonparametric inference in a stereological model with oriented cylinders applied to dual phase steel. Submitted.

[6] Ross, S.M. (2010). *Introduction to Probability Models,* 10*-th edition*. Elsevier, Amsterdam.

[7] Sen, B. and Woodroofe, M. (2012). Bootstrap Confidence Intervals for Isotonic Estimators in a Stereological Problem. To appear in Bernoulli.

[8] Watson, G.S. (1971). Estimating functionals of particle size distributions. *Biometrika* **58**, p. 483-490.

[9] Wicksell, S.D. (1925). The corpuscle problem. *Biometrika* **17**, p. 84-99.

[10] Woodroofe, M. and Sun, J. (1993). A penalized maximum likelihood estimate of $f(0+)$ when f is non-increasing. *Statistica Sinica* **3**, p. 501-515.