Bayesian analysis in finite-population models*

Ryan Martin
www.math.uic.edu/~rgmartin

February 4, 2014

1 Introduction

Bayesian analysis is an increasingly important tool in all areas and applications of statistics. The problem of sampling from finite populations is no exception. This is especially true of one takes the (somewhat extreme) view that all populations in real applications are finite. In any case, Bayesian methods are useful tools for statisticians to have available in their toolbox, so it is good for students to have some exposure to Bayesian ideas, in addition to the classical ones. In these notes, we will explore a few relatively basic ideas in the Bayesian analysis of data obtained via sampling from a finite population.

To fix notation, consider a population with N units, where N is assumed to be known. The sampling frame, call it $\{1, 2, \ldots, N\}$, is also known. A sample is a subset s of the frame, of size $n(s)$, and this particular sample is selected with probability $p(s)$. There would be some feature of the population of interest, i.e., a measurement that can be taken[1] on each unit. We shall denote the characteristic of interest by Y, and its measurement on the sampled units, s, as $Y(s)$. After sampling, one has the measured characteristics $Y(s)$ on the sampled units, but those characteristics $Y(s^c)$ on the not-sampled units remains unknown. To goal is to make use of the observed $Y(s)$ to learn something about $Y(s^c)$ so that inference about the population $\{Y(s), Y(s^c)\}$ can be made. For example, we might be interested in saying something (e.g., point or interval estimates) about the population mean of the characteristic Y.

2 Bayesian analysis

2.1 High-level overview

Classical statistics focuses on the construction of procedures, such as point or interval estimates for unknown parameters, based on observed data, that perform well in terms

*These notes are meant to supplement the few lectures given by the author in Prof. Samad Hedayat's STAT 532 (Sampling Theory II) course at UIC in Spring 2014. The author makes no guarantee that these notes are free of typos or other more serious errors.

[1]Here, assume this measurement is without error, though this is not necessary in general.

of repeated sampling. An example is the construction of an estimator that is unbiased or consistent. Students are assumed to be familiar with this approach.

The Bayesian approach is different in a number of ways. First, the Bayesian approach is based on the idea that only probability can be used to describe uncertainties. So, to the Bayesian, the parameter being unknown means that it's uncertain and he/she must use probability to describe the what is believed to be true about the parameter. This amounts to assigning a probability distribution to the parameter, which is called the *prior distribution*. Given this prior distribution for the parameter, and the conditional distribution of data given parameter encoded by the likelihood, one obtains a posterior distribution of the parameter, given observe data, via Bayes theorem. This posterior distribution is understood as the logical update of the prior uncertainties about the parameter in light of the new information from the observed data. So, to the Bayesian, the posterior distribution is everything. Once this posterior distribution is available, all kinds of things can be done. There are many references that describe the mechanics of Bayesian analysis; see, for example, Gelman et al. (2004) and Ghosh et al. (2006). Some details about Bayesian analysis in the finite-population problem will be given below; students can also see my Stat 411 and Stat 511 notes[2] on Bayesian analysis.

As we shall see below, there are also a number of reasons to care about Bayesian ideas, even if you're not technically a Bayesian. For example, admissibility questions are often resolved based on Bayesian techniques. Also, there are general theorems[3] which give conclusions like the following: *if a method is good, then it must be at least approximately Bayes*. Therefore, Bayesian ideas and techniques ought to be of interest to all.

2.2 Details in finite-population models

The explanation of the Bayesian approach described above and the mechanics of implementing the Bayesian approach in the references above are general enough to cover both the finite- and infinite-population models simultaneously. However, there are some special features of the finite-population problem. Here, we flesh out some of these details, and finish with some remarks to highlight those special features.

First, observe that the full population feature $Y = (Y_1, \ldots, Y_N)$ is the unknown parameter. Let $\pi(y)$ be a prior distribution for Y; selecting a reasonable prior in this likely very-high-dimensional case is a non-trivial task, but discussion of this important point is postponed till later in this section. After the sample s is selected, a portion $Y(s)$ of the unknown parameter Y is observed. The goal is to update prior beliefs about Y, encoded in $\pi(y)$, in light of the observed $Y(s)$. Bayes theorem is the tool. A slight difference in the finite-population sampling context is that the likelihood is sort of a trivial one—it is constant in Y except that it identifies those Y values which agree with the observed $Y(s)$. Specifically,

$$L(y \mid s, Y(s)) = \begin{cases} p(s) & \text{if } y(s) = Y(s) \\ 0 & \text{otherwise.} \end{cases}$$

Since the likelihood is essentially constant, the posterior is proportional to the prior,

[2] Go to: `www.math.uic.edu/~rgmartin/teaching.html`

[3] These are called "complete-class theorems." A summary is given in my Stat 511 lecture notes on statistical decision theory.

restricted to those y values that agree with the observed $Y(s)$, i.e.,

$$\pi(y \mid s, Y(s)) \propto \begin{cases} \pi(y) & \text{if } y(s) = Y(s), \\ 0 & \text{otherwise.} \end{cases}$$

Given that the posterior takes essentially the same form as the prior, it should be clear that the choice of the prior is important (and challenging). Some aspects of the prior choice will become clear next; a "reasonable" choice of prior is presented in Section 4.

Suppose we are interested in the mean value μ_Y of Y in the population. Clearly, the posterior mean is

$$\mathsf{E}(\mu_Y \mid s, Y(s)) = \frac{1}{N}\Big(\sum_{i \in s} Y_i + \sum_{j \notin s} \mathsf{E}(Y_j \mid s, Y(s))\Big);$$

that is, just impute the expected values of $Y(s^c)$, given observed data $Y(s)$, in the usual formula for the population mean. This is a very intuitively appealing formula. However, it does reveal an important feature that the chosen prior must satisfy for the results to be meaningful. If the prior models the components of Y as independent, then $Y(s^c)$ is independent of $Y(s)$, so nothing about the observed Y's can be learned. So, it is essential that the prior introduce some dependence between the components of Y. It turns out, as in many other Bayesian settings, a very natural way to do this is via *exchangeability*.

Finally, there is one somewhat philosophical point that is worth mentioning. In general, if the prior is fixed and does not depend on the specified model, then the Bayesian approach satisfies what is called the *likelihood principle*. That is, any two experiments[4] leading to likelihood functions which are proportional will produce the same posterior distribution, hence the same Bayesian solution. That this is a fundamental property for logical inference is debatable (Martin and Liu 2014; Mayo 2014), but this has a particularly important implication in the finite-population setting. That is, *the sampling design is irrelevant* or, in other words, *all the matters is the $Y(s)$ values observed, not how the particular units were obtained*. This is a striking conclusion that may be difficult for some to swallow. But, after a little reflection, students should see that the sampling mechanism has nothing to do with the population features and, therefore, does not contribute any information about the unknown $Y(s^c)$ values after seeing $Y(s)$.

To follow up on the previous point, students should understand that the motivation for selecting a good sampling design is to justify the belief that the obtained sample is "representative." In the classical setting, for example, one cannot justify the use of the sample mean as an estimate of the population mean if the sample is not a representative one. Arguments based on unbiasedness—which do rely on the sampling design—attempt to justify the observed-to-unobserved reasoning by saying that the results balance out with respect to repeating sampling from the given design. In the Bayesian setting, this connection between observed and unobserved is made through the choice of a suitable prior and the regular rules of probability. More on this in Section 4.

[4]Here, "experiment" means the triplet $(p, s, Y(s))$.

3 Admissibility of sample mean

3.1 Preliminaries

In an estimation problem, for example, when considering what kind of estimator one should use, the first question is *what estimators are admissible?* Specifically, one surely should not use an estimator if there is another estimator that is always better in terms of risk (mean square error). More specifically, if there exists an estimator $\delta'(X)$ with mean square error everywhere smaller than the mean square error of of an estimator $\delta(X)$, as a function of the unknown parameter, then δ is not worth considering. In this case, δ is said to be *inadmissible*; an estimator that is not inadmissible is called *admissible*. One should only consider admissible estimators (or decision rules in general) that are admissible.

In general, and especially in finite-population settings, the sample mean is an attractive estimator for the population mean. However, it is not clear that the sample mean is admissible; that is, how do we know that there isn't some other estimator that's always better? It turns out that Bayesian analysis provides a very useful tool for proving admissibility of estimators. To understand this approach, and apply it to the finite-population problem, we need some notation and a few basic facts.

Let θ denote the population mean and, for an estimator $\delta(X)$, write

$$R(\theta, \delta) = \mathsf{E}_\theta\{(\delta(X) - \theta)^2\},$$

for the *risk* or mean square error. Admissibility of an estimator δ means that for any other estimator δ', there exists a θ, which depends on δ and δ', such that $R(\theta, \delta) < R(\theta, \delta')$. Now consider a Bayesian setup where θ has a prior Π. The *Bayes risk* of δ is

$$r(\Pi, \delta) = \int R(\theta, \delta)\, \Pi(d\theta),$$

the expectation of $R(\theta, \delta)$ with respect to the prior. The Bayes rule δ_Π is the δ that minimizes the Bayes risk, i.e.,

$$r(\Pi, \delta_\Pi) = \inf_\delta r(\Pi, \delta),$$

where the infimum is over all estimators δ (measurable functions from sample space to parameter space). Since we are considering squared error loss, it follows that the Bayes rule is just the posterior mean of θ, i.e., $\delta_\Pi(X) = \mathsf{E}(\theta \mid X)$; see Exercise 1.

A useful fact connecting admissibility questions with Bayes rules is the following fact:

for a proper prior Π, the Bayes rule $\delta_\Pi(X)$ is admissible.

So, if one can demonstrate that the estimator in question is a Bayes rule with respect to some proper prior, then the admissibility issue is automatically resolved. However, in many cases, including the one discussed in the following subsection, the estimator of interest is not a Bayes rule with respect to a proper prior. For such cases, the above fact is not useful. Fortunately, the above fact can be extended to the case where the estimator is, in a certain sense, a limit of a sequence of prior-prior Bayes rules. We summarize this general fact as a theorem.

Theorem 1. *Assume the risk function $R(\theta, \delta)$ is a continuous function of θ for all δ. Suppose there exists a sequence of finite measures $\{\pi_t : t \geq 1\}$ such that $\liminf_{t\to\infty} \Pi_t(B) > 0$ for all open balls/intervals $B \subset \Theta$. If*

$$\lim_{t\to\infty} \{r(\Pi_t, \delta) - r(\Pi_t, \delta_{\Pi_t})\} = 0,$$

then δ is admissible.

Proof. See Keener (2010), Theorem 11.9, or Schervish (1995, p. 158–159). □

We apply this theorem below to prove admissibility of the sample mean in the case where the finite population is binary, i.e., contains only zeros and ones. Then we discuss how to extend the result to general (not necessarily binary) populations.

3.2 Binary case

Let the population consist of only zeros and ones, with θ the population mean. Here the goal is to apply Theorem 1 to prove that the sample mean is an admissible estimator of θ. To connect the notation from above, write $X = Y(s)$ for the observed data, so that the estimator in question is $\delta(X) = X/n$, where $n = n(s)$ is the sample size. Here, we are assuming that either the sampling is done with replacement or that the population size N is sufficiently large that a binomial model for X is appropriate.

To prove admissibility of δ using Theorem 1, we need a sequence of proper priors $\{\Pi_t : t \geq 1\}$ for θ, though these need not be probability measures (i.e., they don't need to integrate to 1). Since beta priors are conjugate for the binomial model (Exercise 2), a reasonable starting point is to consider $\theta \sim \mathsf{Beta}(t^{-1}, t^{-1})$. For such a model, the Bayes rule is

$$\mathsf{E}(\theta \mid X) = \frac{X + t^{-1}}{n + 2t^{-1}}.$$

It is clear that, as $t \to \infty$, the Bayes rule $\delta_{\Pi_t}(X)$ converges to the sample mean $\delta(X) = X/n$. But the limit of the beta priors is not a proper prior for θ; in fact, the limit of the priors has a density that is proportional to $\{\theta(1-\theta)\}^{-1}$, which is improper. Though this limiting prior is improper, it does have some nice properties; see Section 4.

The beta priors themselves do not satisfy the conditions of Theorem 1 (Exercise 3). Fortunately, there is a simple modification of the beta prior that does work. Take Π_t to have density π_t which is just the $\mathsf{Beta}(t^{-1}, t^{-1})$ without the normalizing constant:

$$\pi_t(\theta) = \{\theta(1-\theta)\}^{1/t-1}$$

or, in other words,

$$\pi_t(\theta) = \lambda(t)\mathsf{Beta}(\theta \mid t^{-1}, t^{-1}), \quad \text{where} \quad \lambda(t) = \frac{\Gamma(t^{-1})^2}{\Gamma(2t^{-1})}.$$

Since Π_t is just a rescaling of the beta prior from before, it is clear that the Bayes rule $\delta_{\Pi_t}(X)$ is the same as for the beta prior above. Then the Bayes risk calculations are relatively simple (Exercise 4). With the sequence of priors Π_t, which are proper but not probability measures, it follows from Theorem 1 that the sample mean $\delta(X) = X/n$ is an admissible estimator of θ.

3.3 Towards the general case

A proof of admissibility of the sample mean as an estimator of the population mean, for the general case of arbitrary finite population and arbitrary sampling designs p, is given in Ghosh and Meeden (1997). Though the technical details are a bit different, the idea is quite similar to that presented above for the binomial sampling case. That is, construct a sequence of priors and corresponding Bayes rules that have a sort of limit which gives, as the Bayes rule, the sample mean.

4 The Polya posterior

4.1 "Non-informative" priors

As discussed previously, the choice of prior is a crucial one in the Bayesian context in general. For the finite-population problem, this is especially true since the unknown parameter Y is often N-dimensional with N very large. In this case, there is no large-sample theory to say that the effect of the prior is washed out by a lots of observations, so one must be careful in choosing a good prior.

The one thing we know the prior needs is dependence between the unknown Y values; otherwise, the observed Y's provide no information about the unobserved Y's. One way this can be done is by modeling the Y's as *exchangeable*. This strategy appears in all kinds of Bayesian problems, especially in high-dimensional problems (e.g., Ghosh et al. 2006). Exchangeable simply means that the order of the labels is irrelevant, i.e., the joint distribution is invariant to permutations of the coordinates. Such a "symmetry" assumption implies a very interesting dependence structure, which is (essentially) equivalent to conditionally iid. That is, an exchangeable prior would have a density of the form

$$\pi(y) = \pi(y_1, \ldots, y_N) = \int \prod_{i=1}^{N} h(y_i \mid \theta)\, \Phi(d\theta),$$

where $h(y \mid \theta)$ is some density function and Φ is a probability measure, like a prior for θ. That conditionally iid random variable are exchangeable is easy to verify; that (essentially) the converse is true is not at all obvious, and is the conclusion of de Finetti's famous theorem. The above formula provides a very convenient tool for specifying an exchangeable model for the Y's.

The exchangeable format above requires specification of h and Φ. One would likely want to make these choices in such a way that the posterior distribution has reasonable properties. One reasonable property, which is somewhat vague, is that the posterior should be driven primarily by data, having only minimal dependence on choices made by the statistician. If a prior admits a posterior which is effectively data-driven, then we call that prior "non-informative." We will not attempt to give a formal definition of what it means for a prior to be non-informative; see Ghosh et al. (2006) for considerable discussion on this point. Here, instead, we will focus on a particularly nice result in the binary population case, and a brief discussion of the extension to the non-binary case.

4.2 Binary case

Consider the same finite population consisting of only zeros and ones in Section 3.2. For this problem, we wanted to use the formal prior "$\mathsf{Beta}(0,0)$" for inference on the population mean θ, i.e., for this formal prior, the Bayes rule is the sample mean. Since this formal prior admits a Bayes rule which is a good estimator of θ, one might wonder if this posterior is reasonable more generally. Here we explore some other properties of this posterior in the binary case; some remarks about the general case follow.

To start, let's consider the prior on $\{Y_1, \ldots, Y_N\}$ induced by the prior "$\theta \sim \mathsf{Beta}(0,0)$" on the mean. Given θ, the Y's are conditionally iid $\mathsf{Ber}(\theta)$, so their joint prior looks like

$$\pi(y_1, \ldots, y_N) = \int_0^1 \Big[\prod_{i=1}^N \theta^{y_i}(1-\theta)^{1-y_i}\Big] \frac{c}{\theta(1-\theta)}\, d\theta = c \int_0^1 \theta^{\sum y_i - 1}(1-\theta)^{N - \sum y_i - 1}\, d\theta.$$

(The density of "$\mathsf{Beta}(0,0)$" is defined only up to a proportionality constant $c > 0$.) The right-hand side is just a beta integral, so we get

$$\pi(y_1, \ldots, y_N) = \frac{c\Gamma(\sum_{i=1}^N y_i)\Gamma(N - \sum_{i=1}^N y_i)}{\Gamma(N)}.$$

This prior is exchangeable due to the conditionally iid structure. Hence, this prior has the necessary dependence between the Y's for reasonable sharing of information. As mentioned before, this dependence facilitates learning about unobserved from observed.

In general, conditional distributions are proportional to joint distribution, so we get for the posterior distribution of $Y(s^c)$, given $Y(s)$,

$$\pi(y(s^c) \mid Y(s)) \propto \Big(\sum_{i\in s} Y_i + \sum_{i\notin s} y_i - 1\Big)!\Big(n(s) - \sum_{i\in s} Y_i + N - n(s) - \sum_{i\notin s} y_i\Big)!,$$

where, in this expression, the generic values $y(s^c)$ of $Y(s^c)$ range over $\{0,1\}^{N-n(s)}$. Though this expression looks rather messy, it is a distribution for a relatively simple kind of process, called the *Polya urn model*.[5] (Exercise 5 outlines a proof of this claim.) The simple Polya urn model starts with a bag that consists of r red balls and g green balls. At step t, $t \geq 1$, a ball is sampled at random from the bag; if the ball is red, set $X_t = 1$, replace the red ball, and put another red ball in the bag; otherwise, set $X_t = 0$, replace the green ball, and put another green ball in the bag. In the present context, we have a bag that consists of $\sum_{i\in s} Y_i$ red balls and $n(s) - \sum_{i\in s} Y_i$ green balls. Then the posterior distribution of $Y(s^c)$ is the same as that of the corresponding Polya urn process (after $N - n(s)$ steps). Therefore, the posterior distribution for $Y(s^c)$, given $Y(s)$, based on the formal prior "$\theta \sim \mathsf{Beta}(0,0)$" for the the population mean is called the *Polya posterior*.

Besides being a relatively simple (exchangeable) model for $Y(s^c)$ after seeing $Y(s)$, the Polya posterior has some nice intuition and properties:

- It is easy to simulate (see the next subsection).
- It is completely data-driven, i.e., it requires no input from the analyst besides the observations $(s, Y(s))$. So, in this sense, we can say that the Polya posterior corresponds to a "non-informative" prior for Y.

[5] `http://en.wikipedia.org/wiki/Polya_urn_model`

- It encodes the belief that the obtained sample s is a representative one, e.g., if $Y(s)$ has lots of ones, then $Y(s^c)$ is also likely to have lots of ones. So, even though the Bayesian posterior does not depend on the sampling design p, there is an implicit dependence here because if the sampling design were bad, then the sample s may not be representative, and the Polya posterior would not be justifiable.

4.3 General case

The extension to the general (not necessarily binary) case is complicated, and the reader is referred to Ghosh and Meeden (1997) for the details. Here I want to explain more the intuition of the move from the binary case to the general case.

The idea is to consider a bag containing balls labeled by the values of $Y(s)$; each value gets its own ball, not just the unique values, so there could be several identical balls in the bag. Then one follows the same recipe as before: pick a ball from the bag at random, replace the ball, and add another ball identical to the selected one. The bag starts with $n(s)$ balls in it, and continue this process till there are N balls in the bag. Now this bag makes up a single sample from the posterior distribution of Y, given $Y(s)$. Repeat this process M times to get M (independent) samples from the posterior distribution of Y. R code to implement sampling from the Polya posterior is given in Figure 1.

Any feature of the Y population of interest, call it $\varphi(Y)$, has a posterior distribution, and its posterior can be obtained from the Polya posterior sample discussed above. That is, get samples $Y^{(1)}, \ldots, Y^{(M)}$ from the Polya posterior, and compute $\varphi(Y^{(m)})$, $m = 1, \ldots, M$. A histogram of these φ values can be plotted to visualize its posterior. For a concrete example, suppose the mean of the Y population is of interest. Given a sample

```
20.89  19.01  6.88  1.61  1.34  4.57  9.13  8.83  29.98  1.55
```

of size $n = 10$, if we know that the population size is $N = 50$, then we get a posterior distribution for the Y population mean, based on sampling from the Polya posterior, is displayed in Figure 2. The figure shows the mean of this posterior as well as the sample mean. It is a consequence of the results presented above that the Bayes rule corresponding to the Polya posterior is the sample mean, so we are not surprised that these two values agree in this simulated data example.

5 Exercises

1. If the loss function is squared error, i.e., $L(\theta, a) = (a - \theta)^2$, then show that the Bayes rule (minimizer of the the Bayes risk) is the posterior mean.

2. Suppose that the conditional distribution of X, given θ, is $\mathsf{Bin}(n, \theta)$. Show that if the the marginal (prior) distribution of θ is $\mathsf{Beta}(a, b)$, then the conditional (posterior) distribution of θ, given X, is $\mathsf{Beta}(a + X, b + n - X)$.

3. Consider $\theta \sim \mathsf{Beta}(a, a)$, and let $B = (b_1, b_2)$ where b_1 and b_2 are bounded away from 0 and 1, respectively. Show that $\mathsf{P}(\theta \in B) \to 0$ as $a \to 0$. Hint: Use the fact that $\Gamma(x) = \Gamma(x + 1)/x$.

4. Consider admissibility of the sample mean as discussed in Section 3.2.

```
rpolyapost <- function(M, y, N) {

  n <- length(y)
  out <- matrix(0, nrow=M, ncol=N)
  for(m in 1:M) {

    out[m, 1:n] <- y
    for(j in (n + 1):N) {

      I <- sample(1:(j - 1), 1)
      out[m, j] <- out[m, I]

    }

  }
  return(out)

}

y <- c(20.89, 19.01, 6.88, 1.61, 1.34, 4.57, 9.13, 8.83, 29.98, 1.55)
set.seed(7)
yy <- rpolyapost(5000, y, 50)
avg <- apply(yy, 1, mean)
hist(avg, freq=FALSE, xlab="Population mean", col="gray", border="white", main="")
points(mean(y), 0, pch="X")
abline(v=mean(avg))
```

Figure 1: R code for sampling from the Polya posterior, given the observed data y; N is the known population size and M is the desired Monte Carlo sample size.

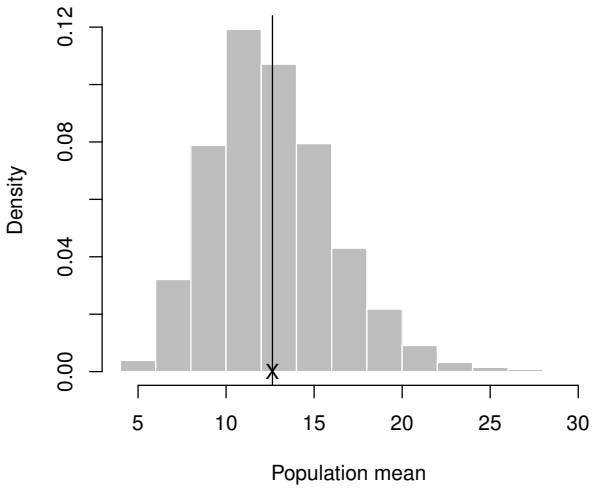

Figure 2: Plot of the simulated Polya posterior distribution of the population mean. X marks the sample mean and the vertical line marks the mean of the posterior sample, the Bayes estimator. That X and the line agree demonstrates that the sample mean is the Bayes rule under the model that admits the Polya posterior.

(a) Show that the Bayes risk of the sample mean $\delta(X) = X/n$ with respect to Π_t is

$$r(\Pi_t, \delta) = \frac{1}{4n}\Big(3 - \frac{1}{2t^{-1}+1}\Big).$$

(b) Show that the Bayes risk of the posterior mean $\delta_{\Pi_t}(X) = \mathsf{E}(\theta \mid X)$ with respect to the prior Π_t is

$$r(\Pi_t, \delta_{\Pi_t}) = \Big(\frac{1}{2t^{-1}+1}\Big)^2\Big[\frac{3n}{4} - \frac{n}{4(2t^{-1}+1} + \frac{t^{-2}}{2t^{-1}+1}\Big].$$

(c) Show that $\{r(\Pi_t, \delta) - r(\Pi_t, \delta_{\Pi_t}\} \to 0$ as $t \to \infty$. Hint: You'll probably need the property of the gamma function from Exercise 3.

5. (a) Let $\{X_t : t = 1, \ldots, n\}$ be the first n iterations of the simple Polya urn process, with bag set to have r red and g green balls at time $t = 0$. Show that the joint distribution of $(X_1, \ldots, X_n)$ is given by

$$\mathsf{P}(X_1 = x_1, \ldots, X_n = x_n) = \frac{r^{[k]}g^{[n-k]}}{(r+g)^{[n]}},$$

where $k = x_1 + \cdots + x_n$ and $a^{[k]} = a(a+1)\cdots(a+k-1)$ is the rising factorial.

(b) Use the calculation above to justify the claim that, in the binary case, the posterior distribution of $Y(s^c)$, given $Y(s)$, is a Polya urn process with $r = \sum_{i\in s} Y_i$ and $g = n(s) - \sum_{i\in s} Y_i$.

References

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman & Hall/CRC, Boca Raton, FL, second edition.

Ghosh, J. K., Delampady, M., and Samanta, T. (2006). *An Introduction to Bayesian Analysis*. Springer, New York.

Ghosh, M. and Meeden, G. (1997). *Bayesian Methods for Finite Population Sampling*, volume 79 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.

Keener, R. W. (2010). *Theoretical Statistics*. Springer Texts in Statistics. Springer, New York.

Martin, R. and Liu, C. (2014). Comment: Foundations of statistical inference, revisited. *Statist. Sci.*, to appear, `arXiv:1312.7183`.

Mayo, D. (2014). On the Birnbaum argument for the strong likelihood principle. *Statist. Sci.*, to appear.

Schervish, M. J. (1995). *Theory of Statistics*. Springer-Verlag, New York.