POLYA'S URN AND THE BETA-BERNOULLI PROCESS

NORA HELFAND

Abstract. The Polya's Urn model is notable within statistics because it generalizes the binomial, hypergeometric, and beta-Bernoulli (beta-binomial) distributions through a single formula. In addition, Polya's Urn is a multivariate distribution whose variables are exchangeable but not independent. This paper introduces basic probability and Bayesian concepts in order to prove these properties.

Contents

1. Background

Modern statisticians generally ascribe to one of two philosophies: frequentist probability theory or Bayesian probability theory. Frequentist probability theory, or the traditional theory taught in probability courses, describes probability as a *fixed* measure on an event independent of previous observations of that event. Bayesian statistics teaches that an event's probability is inextricable from the fact of its observation – thus, probabilities are always *changing*. However, basic probability does motivate Bayesian methods. We will first introduce a probability measure and prove Bayes' theorem. Then we outline the basics of Bayesian statistics and introduce the binomial and hypergeometric densities conceptually, assuming knowledge of combinatorics and random variables. Finally, we define the beta function and distribution and explain its role as a conjugate prior to the binomial distribution. All of these results motivate our urn sampling model, since these distributions can all be modeled using urns. The final result is that the Polya's Urn process is identical to the beta-Bernoulli process under certain conditions, a surprising result. This result demonstrates the real-life significance of Bayesian methods.

Date: DEADLINES: Draft AUGUST 13 and Final version AUGUST 24, 2012.

2. Basic Probability

In this section we define the fundamentals of probability and prove Bayes' theorem for the countable case.

Definition 2.1. In probability theory we always deal with some sort of *experiment*, which is any well-defined procedure or chain of circumstances. The set of end results of an experiment is the *sample space* Ω whose elements ω represent individual *outcomes* of the experiment.

Definition 2.2. An *event space* $\mathcal{F}$ is a subset of the power set $\wp(\Omega)$ of Ω which satisfies the following:

(1) $\Omega \in \mathcal{F}$.

(2) if $A \in \mathcal{F}$ then $\Omega \setminus A \in \mathcal{F}$.

(3) if $A_j \in \mathcal{F}$ for $j \geq 1$, then $\bigcup_{j=1}^{\infty} A_j \in \mathcal{F}$ (or, any countable union of elements of $\mathcal{F}$ is in $\mathcal{F}$).

If $A \in \mathcal{F}$ we say A is an *event*.

Definition 2.3. Let A be an event. If some $\omega \in A$ is the outcome of our experiment, we say A *occurs*. The complement $\Omega \setminus A$ is written A^c. If some $\omega \in A^c$ is the outcome of our experiment, we say A *does not occur*.

Definition 2.4. We define *probability* as a function $\mathbf{P} : \mathcal{F} \to [0, 1]$ satisfying the following:

(1) $0 \leq \mathbf{P}(A) \leq 1$.

(2) $\mathbf{P}(\Omega) = 1$.

(3) $\mathbf{P}(A_1 \cup A_2 \cup \cdots) = \mathbf{P}(A_1) + \mathbf{P}(A_2) + \cdots$ whenever A_1, A_2, ... are disjoint events.

We can also call $\mathbf{P}$ a *probability distribution* on Ω.

Lemma 2.5. *For any event A, $\mathbf{P}(A) = 1 - \mathbf{P}(A^c)$.*

Proof. For every outcome $\omega \in \Omega$, either $\omega \in A$ or $\omega \in A^c$. Thus $\Omega \subseteq A \cup A^c$. Since $A \cup A^c \subseteq \Omega$ by definition, $A \cup A^c = \Omega$. We also have $A \cap A^c = \emptyset$ by definition, so by 2.4.3

$$\mathbf{P}(A \cup A^c) = \mathbf{P}(A) + \mathbf{P}(A^c) = \Omega.$$

Since $\mathbf{P}(\Omega) = 1$, $\mathbf{P}(A) = 1 - \mathbf{P}(A^c)$. □

We now define conditional probability and state and prove Bayes' Theorem.

Definition 2.6. Let A and B be events with $\mathbf{P}(B) > 0$. Given that B occurs, the *conditional probability* that A occurs is denoted by $\mathbf{P}(A|B)$ and defined as $\mathbf{P}(A|B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)}$.

Lemma 2.7. *Let $A \subseteq \bigcup_i B_i$ where $B_i \cap B_j = \emptyset$ for $i \neq j$. Then*

$$\mathbf{P}(A) = \sum_i \mathbf{P}(A|B_i)\mathbf{P}(B_i).$$

Proof. By Definition 2.6, $\sum_i \mathbf{P}(A|B_i)\mathbf{P}(B_i) = \sum_i \mathbf{P}(A \cap B_i)$. Since $B_i \cap B_j = \emptyset$ for $i \neq j$,

$$\sum_i \mathbf{P}(A \cap B_i) = \mathbf{P}\bigcup_i (A \cap B_i) = \mathbf{P}(A \cap \bigcup_i B_i),$$

which is equal to $\mathbf{P}(A)$ since $A \subseteq \bigcup_i B_i$. □

Theorem 2.8 (Bayes' Theorem)**.** *If $A \subseteq \bigcup_i B_i$, $\mathbf{P}(A) > 0$ and $B_i \cap B_j = \emptyset$ for $i \neq j$, then for all k,*

$$\mathbf{P}(B_k|A) = \frac{\mathbf{P}(A|B_k)\mathbf{P}(B_k)}{\sum_i \mathbf{P}(A|B_i)\mathbf{P}(B_i)}.$$

Proof. We have from Definition 2.6 that

$$\mathbf{P}(B_k|A) = \frac{\mathbf{P}(B_k \cap A)}{\mathbf{P}(A)}.$$

Applying 2.6 to the numerator, we have

$$\mathbf{P}(B_k|A) = \frac{\mathbf{P}(A|B_k)\mathbf{P}(B_k)}{\mathbf{P}(A)}.$$

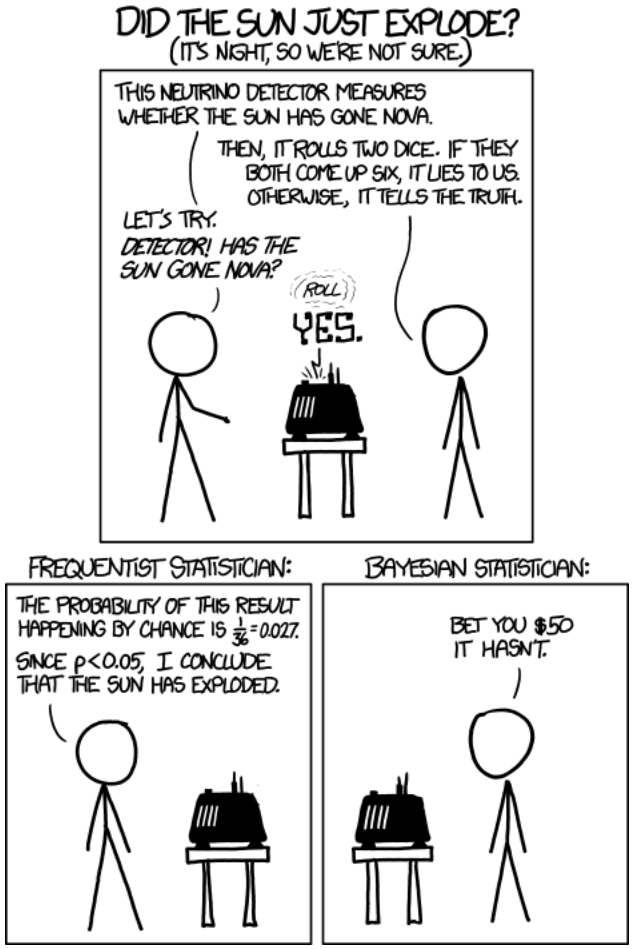
Applying Lemma 2.7 to the denominator,

$$\mathbf{P}(B_k|A) = \frac{\mathbf{P}(A|B_k)\mathbf{P}(B_k)}{\sum_i \mathbf{P}(A|B_i)\mathbf{P}(B_i)}.$$

□

Bayes' Theorem drives the logic of *Bayesian analysis*. Whereas a classic or *frequentist* approach to probability and statistics merely assigns fixed probabilities to events and uses these probabilities to deduce the "randomness" and significance of processes, Bayesian statistics asserts that probabilities should be re-interpreted and updated in light of all conditions on a process.

The distinction is well-illustrated by a comic strip from xkcd.com entitled "Frequentists vs. Bayesians."



What the frequentist fails to take into account is his belief that the sun would explode prior to asking the machine, and to what degree the machine's answer should "update" that belief. To see this, let X be the event that the sun explodes and let Y be the event that the machine answers "Yes." We can apply Bayes' Theorem to find the probability that the sun has exploded given that the machine says "Yes," or $\mathbf{P}(X|Y)$:

$$\mathbf{P}(X|Y) = \frac{\mathbf{P}(Y|X)\mathbf{P}(X)}{\mathbf{P}(Y|X)\mathbf{P}(X) + \mathbf{P}(Y|X^c)\mathbf{P}(X^c)}.$$

We know that $\mathbf{P}(Y|X) = \frac{35}{36}$. We do not know $\mathbf{P}(X)$ exactly but the Bayesian statistician knows that it is extremely small. Finally, we know that $\mathbf{P}(Y|X^c) = \frac{1}{36}$ and that $\mathbf{P}(X^c)$ is extremely close to 1 since $\mathbf{P}(X^c) = 1 - \mathbf{P}(X)$. Thus,

$$\mathbf{P}(X|Y) \approx \frac{\mathbf{P}(X)}{\mathbf{P}(X) + \frac{1}{36}}$$

We know $\mathbf{P}(X) << \frac{1}{36}$ so this is an extremely low probability (even though the machine has a much higher likelihood of saying "No" than "Yes," as the frequentist notes). The Bayesian will most likely win the bet.

3. Prior and Posterior Distributions

Bayesian statistics treats sought-after probabilities as random variables. The reasoning behind this is that, due to the variance in experimental results, experiments can only be interpreted through some lens of experience or knowledge, and thus a *prior distribution* is used to model current beliefs about the desired probability value. The result obtained from an experiment is used to describe how our beliefs should change due to the result using a *posterior distribution*. To see this, we will define some new sample spaces.

Definition 3.1. The *sample space* $\mathcal{Y}$ is the set of all possible datasets that could result from our experiment. The *parameter space* Θ is the set of possible parameter values, from which we hope to identify the value that best represents the true probability of an event. (Θ is analogous to the interval $[0, 1]$ into which all probability measures map.)

Definition 3.2.
(1) For each $\theta \in \Theta$, the *prior distribution* $p(\theta)$ describes our belief that θ represents the true probability.

(2) For each $\theta \in \Theta$ and $y \in \mathcal{Y}$, our *sampling model* $p(y|\theta)$ describes our belief that y would be the outcome of our study if we knew θ to be the true probability.

(3) For each $\theta \in \Theta$, our *posterior distribution* $p(\theta|y)$ describes our belief that θ is the true probability having observed our experiment's dataset y.

Remark 3.3. We have not yet stated Lemma 2.7 for the case in which the events are uncountable. If for some experiment, we have a random variable θ that takes values in Θ and an event space containing all events in which θ takes some value in Θ, then the events are uncountable. Suppose we also have some random variable y. We cannot sum over $\mathbf{P}(y|\theta_i)$, but we can integrate over these values which is the continuous analog to summing. Thus, if we use p to represent continuous densities of these variables:

$$p(y) = \int_0^1 p(y|\theta_i)p(\theta_i)d\theta_i.$$

Theorem 3.4. *Because θ can take any real value in $[0, 1]$, it behaves like a* continuous *random variable and thus Bayes' rule applies as follows for a fixed y and for every θ:*

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int_0^1 p(y|\theta_i)p(\theta_i)d\theta_i}.$$

(We use subscripts to distinguish the integration over all values of θ from the particular θ for which we desire a posterior probability.)

4. Sampling Models

We now outline the binomial and hypergeometric distributions whose distributions are derived intuitively. A basic knowledge of random variables, combinatorics, and probability densities is assumed.

Remark 4.1. For natural numbers n and r, we use the notation $n^{(r)}$ to refer to an ordered sample of r elements chosen from n elements – in other words, $n^{(r)} = r!\binom{n}{r}$.

Example 4.2. Let D be a set whose elements fall into one of two categories: successes and failures. Let $R \subseteq D$ be the set of all successes and let $|R| = r$ and $|D| = m$. In our experiment we take an ordered sample of n elements and let X_i be a random variable that takes the value 1 if the ith element is a success and 0 if the ith element is a failure. If Y is a random variable that gives the number of successes in a trial, we see that

$$Y = \sum_{i=1}^{n} X_i.$$

Theorem 4.3. *The probability density function of Y is given by*

$$\mathbf{P}(Y = y) = \frac{\binom{r}{y}\binom{m-r}{n-y}}{\binom{m}{n}},\ y \in \{\max(0, n-(m-r)), \ldots, \min(n, r)\}.$$

Proof. There are $\binom{r}{y}$ ways to choose y of the r successes, $\binom{m-r}{n-y}$ ways to choose from the $m-r$ failures, and $\binom{m}{n}$ ways to select from all the elements. □

Corollary 4.4. *The probability density function of Y is also given by*

$$\mathbf{P}(Y = y) = \binom{n}{y}\frac{r^{(y)}(m-r)^{(n-y)}}{m^{(n)}}. \tag{4.5}$$

Proof. By the definitions of combinations and permutations:

$$\begin{aligned}\frac{\binom{r}{y}\binom{m-r}{n-y}}{\binom{m}{n}} &= \frac{\frac{r^{(y)}}{y!} \cdot \frac{(m-r)^{(n-y)}}{(n-y)!}}{\frac{m^{(n)}}{n!}} \\ &= \binom{n}{y}\frac{r^{(y)}(m-r)^{(n-y)}}{m^{(n)}}.\end{aligned}$$

□

Note that this result also agrees with our intuition because, considering the ordered sample of elements, $\binom{n}{y}$ describes the number of ways to select the "positions" of the successes, $r^{(y)}$ is the number of ways to select an ordered sequence of y successes, $(m-r)^{(n-y)}$ is the number of ways to select an ordered sequence of $n-y$ failures, and $m^{(n)}$ describes all ordered sequences.

Definition 4.6. The density function given by (4.5) is known as the *hypergeometric distribution* with parameters m, r, and n.

Example 4.7. Consider another dichotomous population D and suppose we choose n of its elements *with* replacement – or, every element can be chosen in an infinite number of trials. Define X_i and Y as before. Since each trial is identical, $\mathbf{P}(\text{success}) = p$ and $\mathbf{P}(\text{failure}) = q = 1 - p$ remain fixed for each trial and trials are independent.

Theorem 4.8. *The probability density function of Y is given by*

$$\mathbf{P}(Y = y) = \binom{n}{y}p^y q^{n-y}. \tag{4.9}$$

Proof. If we choose a success y times and a failure $n-y$ times, there are then $\binom{n}{y}$ binary strings of length n with y 1's. □

Any such sampling model that has binary outcomes for independent trials with fixed probabilities is known as a *Bernoulli trial*.

Definition 4.10. The density function given by (4.9) is known as the *binomial distribution* with parameter p (q is always $1-p$ and thus is not listed as a parameter).

5. The Gamma Function and the Beta Function

In this section we introduce the gamma function and the beta function. The gamma function generalizes the factorial function to numbers that are not necessarily natural. It is related to the beta function which is central to the beta density.

Definition 5.1. The *Gamma function* (or Euler function of the second kind) $\Gamma(\cdot)$ is defined as follows for all x:

$$\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt.$$

Theorem 5.2. *For all $x > 1$, $\Gamma(x) = (x-1)\Gamma(x-1)$.*

Proof. The integration-by-parts rule states

$$\int udv = uv - \int vdu.$$

Letting $u(t) = t^{x-1}$ and $v(t) = -e^{-t}$, we have

$$\begin{aligned}\Gamma(x) &= -t^{x-1}e^{-t}|_0^\infty - \int_0^\infty -(x-1)t^{x-2}e^{-t}dt \\ &= 0 + (x-1)\int_0^\infty t^{x-2}e^{-t}dt \\ &= (x-1)\Gamma(x-1).\end{aligned}$$

□

Thus, the Gamma function has the following properties which are easily derived using induction:

Corollary 5.3. *For all natural numbers n, $\Gamma(n) = (n-1)!$.*

Corollary 5.4. *For all $a > 0$, the product $a\cdot(a+1)\cdot(a+2)\cdot\ldots\cdot(a+k-1) = \frac{\Gamma(a+k)}{\Gamma(a)}$ for all natural numbers k.*

We now define another important related function.

Definition 5.5. The *beta function* (or Euler function of the first kind) $B(\cdot)$ is defined as follows for all $0 < p < 1$ and all α and β:

$$B(\alpha,\beta) = \int_0^1 p^{\alpha-1}(1-p)^{\beta-1}dp.$$

Theorem 5.6. *For any α and β, $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$.*

Proof. The proof is a change-of-variables problem beyond the scope of this paper. □

6. Beta Density

Definition 6.1. A *conjugate prior* of a sampling model is a prior distribution $p(\theta)$ which, when applied to a sampling model $p(y|\theta)$, generates a posterior distribution of the same form as the prior. For example, if we use a binomial prior $p(\theta)$ on some sampling model $p(y|\theta)$ and the posterior $p(\theta|y)$ is *also* binomial, then we say that the binomial distribution is a conjugate prior of the sampling model.

The posterior of one experiment can be used as the prior for the next. Thus a conjugate prior is ideal for a given sampling model because the prior will be the same type for infinite experiments for which Bayes' rule is applied. We will now define the beta density, show that it is a conjugate prior for the binomial sampling model, and give a generalized formula for the beta-Bernoulli distribution.

Definition 6.2. The *beta density* $Be_{\alpha,\beta}(\theta)$ for $\alpha > 0$ and $\beta > 0$ is given by

$$Be_{\alpha,\beta}(\theta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}.$$

Theorem 6.3. *The beta density is a conjugate prior of the binomial distribution.*

Proof. Let $p(\theta) = Be_{\alpha,\beta}(\theta)$ for $\alpha, \beta > 0$ with parameter θ. Let $p(y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$ for some positive integer index n and all integers y such that $0 \leq y \leq n$ (note that we are simply treating y as a binomial random variable with parameter θ). By Bayes' rule, for each $y \in \mathcal{Y}$

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}.$$

Thus

$$p(\theta|y) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^y(1-\theta)^{n-y}}{p(y)B(\alpha,\beta)}.$$

However, $p(y)$ does not depend on θ and $B(\alpha,\beta)$ is a normalization constant, so we can write

$$p(\theta|y) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^y(1-\theta)^{n-y}$$

where $\propto$ means "is directly proportional to." Therefore,

$$p(\theta|y) \propto \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1}.$$

Normalizing our final result gives a beta distribution with parameters $y+\alpha$ and $n-y+\beta$ (recall that $y \geq 0$ and $n-y \geq 0$). □

Remark 6.4. When, as above, a binomial random variable (or random vector) y has a random parameter with the beta distribution, y is called a *Beta-Bernoulli* process.

Theorem 6.5. *Let $(X_1, X_2, \ldots, X_n)$ be a random vector such that each X_i takes the value 1 with probability p and 0 with probability $1-p$. Moreover let p be a beta random variable with parameters a and b that takes values in $[0,1]$. Then for all $x_i \in \{0,1\}$,*

$$\mathbf{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \frac{B(a+k, b+(n-k))}{B(a,b)}$$

where for all i, $x_i \in \{0,1\}$.

Proof. We know that $\mathbf{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n|p) = p^k(1-p)^{n-k}$. Thus by Definition 2.6:

$$\begin{aligned}\mathbf{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) &= \int_0^1 p^k(1-p)^{n-k}\frac{1}{B(a,b)}p^{a-1}(1-p)^{b-1}dp \\ &= \int_0^1 p^{k+a-1}(1-p)^{b+n-k-1}\frac{1}{B(a,b)}dp.\end{aligned}$$

Thus, by definition 5.5:

$$\mathbf{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \frac{B(a+k, b+(n-k))}{B(a,b)}.$$

□

Corollary 6.6. *Let X_i be as in Theorem 6.5 and let Y_n be a random variable such that $Y_n = \sum_{i=1}^n X_i$. Then*

$$\mathbf{P}(Y_n = y) = \binom{n}{y}\frac{B(a+k, b+(n-k))}{B(a,b)}. \tag{6.7}$$

Proof. There are $\binom{n}{y}$ bit strings of length n with exactly y 1's. Thus any combination of 0's and 1's has a $\binom{n}{y}\mathbf{P}(X_1 = x_i, X_2 = x_2, \ldots, X_n = x_n)$ chance of occurring. □

Definition 6.8. The density given by (6.7) is known as the *Beta-Bernoulli distribution.*

7. Polya's Urn

We now introduce the Polya's Urn sampling model. We will show that this sampling model concretely describes the binomial, hypergeometric, and beta-Bernoulli distributions under particular conditions.

Consider an urn that contains a azure and b balls. A ball is drawn from the urn, its color is noted, and it is returned to the urn. c balls of the same color as the ball that was just drawn are added to the urn, and the process repeats.

After n draws, if y is a random variable that gives the number of blue balls drawn, y has the binomial distribution for $c = 0$ and the hypergeometric distribution for $c = -1$ (assuming there are at least n balls to accommodate the n draws). This can be seen if we imagine that all azure balls represent successes and all blue balls represent failures. Then $c = 0$ represents the case in which we sample with replacement, and $c = -1$ represents sampling without replacement (leaving out the ball we just drew). We prove this result in Theorem 7.6 and 7.8.

Definition 7.1. For any r and s and any natural number j, define

$$r^{(s,j)} = r(r+s)(r+2s)\cdots[r+(j-1)s].$$

This is the generalized permutation formula.

Definition 7.2. Let $\mathcal{C}$ be a collection of random variables for an experiment. This collection is said to be *exchangeable* if for any $\{X_1, X_2, \ldots, X_n\} \subseteq \mathcal{C}$, the distribution of the random vector $(X_1, X_2, \ldots, X_n)$ depends only on n.

Theorem 7.3. *Consider an urn containing a azure and b blue balls governed by the Polya's Urn process for some c. Let $(X_1, X_2, \ldots, X_n)$ be a random vector where each random variable X_i takes the value 1 if the ith ball is azure and 0 if the ith ball is blue. Then*

$$\mathbf{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \frac{a^{(c,y)} b^{(c,n-y)}}{(a+b)^{(c,n)}} \text{ where } y = \sum_{i=1}^{n} x_i.$$

Proof. If one draws an azure ball y times, at the drawing of the first azure ball there are a azure balls, at the second drawing there are $a + c$ azure balls, and so on up to $a + (y-1)$ – this is represented by $a^{(c,y)}$. This is similarly true for blue balls ($b^{(c,n-y)}$ possibilities). The denominator $(a+b)^{(c,n)}$ denotes the change in the total number of balls; at the first drawing there are $a + b$, then $a + b + c$, and so on. □

As a result of this theorem, we can say that the random vector X which takes values in all ordered binary strings of length n is exchangeable. Thus, Polya's Urn has the surprising result that the variables $X_1, X_2, \ldots, X_n$ are exchangeable but not *independent.*

Theorem 7.4. *Let Y_n be a random variable that gives $\sum_{i=1}^{n} X_i$ for a trial of n drawings of Polya's Urn. Then*

$$\mathbf{P}(Y_n = y) = \binom{n}{y} \frac{a^{(c,y)} b^{(c,n-y)}}{(a+b)^{(c,n)}}.$$

Proof. As before, for any y there are $\binom{n}{y}$ random vectors with y successes or 1's. □

Now let's verify that this result agrees with our intuition about binomial and hypergeometric sampling models.

Lemma 7.5. *For any r and positive integer j:*

$$r^{(0,j)} = r^j.$$

Proof. By definition:

$$r^{(0,j)} = r(r + 1 \cdot 0)(r + 2 \cdot 0) \cdots (r + (j-1) \cdot 0) = r^j.$$

□

Theorem 7.6. *Polya's urn has the binomial distribution for $c = 0$.*

Proof. By Lemma 7.5:

$$\begin{aligned}
\binom{n}{y} \frac{a^{(c,y)} b^{(c,n-y)}}{(a+b)^{(c,n)}} &= \binom{n}{y} \frac{a^y b^{n-y}}{(a+b)^n} \\
&= \binom{n}{y} \frac{a^y}{(a+b)^y} \frac{b^{n-y}}{(a+b)^{n-y}} \\
&= \binom{n}{y} \left(\frac{a}{a+b}\right)^y \left(\frac{b}{a+b}\right)^{n-y}.
\end{aligned}$$

Since $\mathbf{P}(\text{ball is azure}) = \frac{a}{a+b}$ for all trials and similarly for $\mathbf{P}(\text{ball is blue})$, this is a binomial distribution. □

Lemma 7.7. *For all real r and natural numbers j, $r^{(-1,j)} = r^{(j)}$.*

Proof. By definition:

$$r^{(-1,j)} = r(r-1)(r-2)\dots(r-j+1) = r^{(j)}.$$

□

Theorem 7.8. *Polya's urn has the hypergeometric distribution for $c = -1$.*

Proof. By Lemma 7.7:

$$\binom{n}{y}\frac{a^{(c,y)}b^{(c,n-y)}}{(a+b)^{(c,n)}} = \binom{n}{y}\frac{a^{(y)}b^{(n-y)}}{(a+b)^{(n)}}.$$

Of course, this only makes sense if $n \leq a+b$. □

Theorem 7.9. *The density $\mathbf{P}(Y_n = y)$ of Polya's Urn can also be written as follows:*

$$\mathbf{P}(Y_n = y) = \binom{n}{y}\frac{B(\frac{a}{c}+y, \frac{b}{c}+n-y)}{B(\frac{a}{c}, \frac{b}{c})}.$$

Proof. By the definition of the generalized permutation formula,

$$\begin{aligned}\frac{a^{(c,y)}b^{(c,n-y)}}{(a+b)^{(c,n)}} &= \frac{\prod_{i=1}^{y}[a+(i-1)c]\prod_{i=1}^{n-y}[b+(i-1)c]}{\prod_{i=1}^{n}[a+b+(i-1)c]} \\ &= \frac{\prod_{i=1}^{y}[\frac{a}{c}+i-1]\prod_{i=1}^{n-y}[\frac{b}{c}+i-1]}{\prod_{i=1}^{n}[\frac{a+b}{c}+i-1]}.\end{aligned}$$

Then, by Corollary 5.4, we can write this in terms of the gamma function:

$$\begin{aligned}\frac{\prod_{i=1}^{y}[\frac{a}{c}+i-1]\prod_{i=1}^{n-y}[\frac{b}{c}+i-1]}{\prod_{i=1}^{n}[\frac{a+b}{c}+i-1]} &= \frac{\frac{\Gamma(\frac{a}{c}+y)\Gamma(\frac{b}{c}+n-y)}{\Gamma(\frac{a}{c})\Gamma(\frac{b}{c})}}{\frac{\Gamma(\frac{a+b}{c}+n)}{\Gamma(\frac{a+b}{c})}} \\ &= \frac{\Gamma(\frac{a}{c}+y)\Gamma(\frac{b}{c}+n-y)}{\Gamma(\frac{a}{c}+\frac{b}{c}+n)}\frac{\Gamma(\frac{a}{c}+\frac{b}{c})}{\Gamma(\frac{a}{c})\Gamma(\frac{b}{c})} \\ &= \frac{B(\frac{a}{c}+y, \frac{b}{c}+n-y)}{B(\frac{a}{c}, \frac{b}{c})} \text{ by Theorem 5.6.}\end{aligned}$$

Thus $\mathbf{P}(Y_n = y) = \binom{n}{y}\frac{B(\frac{a}{c}+y, \frac{b}{c}+n-y)}{B(\frac{a}{c}, \frac{b}{c})}$. □

Thus, when $c = 1$, Polya's Urn generates the beta-Bernoulli distribution ($p(y|\theta)$) with parameters a and b. The significance of this result lies in its usefulness in determining the values of $\mathbf{P}$(ball is azure) and $\mathbf{P}$(ball is blue). As in section six, we can assign a beta prior $p(\theta)$ with parameters a and b and generate a beta posterior $p(\theta|y)$ that more closely models our beliefs about the value of θ with each trial of n draws. This result is surprising given that each drawing of a ball is not an identical Bernoulli trial – each drawing affects the probability of successive drawings. Yet the probabilities follow a binomial distribution with a beta parameter.

Acknowledgments. I would like to thank Peter May for his devotion to giving every student an enjoyable experience, my father for his enthusiasm when I would talk about what I was learning, and my mentor, Olga, for her encouragement and flexibility.

References

[1] Saad Mneimneh. Class Notes for Intro to Bayesian Statistics. http://www.cs.hunter.cuny.edu/ saad/courses/bayes/notes/.
[2] Kyle Siegrist. Virtual Laboratories in Probability and Statistics. University of Alabama in Huntsville, 1997-2013. http://www.math.uah.edu/stat/.
[3] David Stirzaker. Elementary Probability. http://carlossicoli.free.fr/S/Stirzaker-D.-Elementary-probability-Cambridge-University-Press(2003).pdf.
[4] Peter D. Hoff. A First Course in Bayesian Statistical Methods. http://link.springer.com.proxy.uchicago.edu/book/10.1007/978-0-387-92407-6/page/1.