The Beta-Binomial Distribution

Florian H. Hodel* James G. Booth †

Contents

*Michigan State University, hodelflo@msu.edu
†Cornell University, jim.booth@cornell.edu

Sequence of Bernoulli trials

Pr will denote probability or a probability mass function (PMF), f will denote a probability density function (PDF) and E an expectation value.

Let $\mathbf{X} = (X_1, X_2, ...)$ be a sequence of identically and independently distributed (i.i.d.) indicator random variables, i.e., random variables that can take the values 0 and 1. The probability that one of the random variables, X_i, takes the value 1 is $\Pr(X_i = 1) = p$. We say this random variable follows a Bernoulli distribution with parameter p. The probability to obtain a sequence of Bernoulli trials, $(X_1 = x_1, X_2 = x_2, ..., X_n = x_n)$; e.g., a sequence of $(1, 0, 0, 1, ..., 1)$, of length n with k successes (i.e., 1s) is $\Pr(x_1, x_2, ..., x_n) = p^k (1-p)^{n-k}$.

Since we stated above that we work with sequences of independent random variables, it is clear that this probability is the product of the probabilities of each Bernoulli trial. Therefore, the probability of a sequence is completely determined by the length n, and number of successes k, and does not depend on the actual sequence of 1s and 0s. This means that all sequences of the same length with the same number of 1s and same number of 0s have the same probability (and thus, are *exchangeable*). Because of this, if we want to model the number of successes in a sequence of Bernoulli trials, we choose a sequence that contains the desired number of 1s and 0s and simply multiply its probability by the number of different sequences that contain that same number of 1s and 0s.

Let's assume that $\mathbf{x}$ is one of those sequences. There will always be $\frac{n!}{k!(n-k)!} = \binom{n}{k}$ different sequences of length n with k successes, and the probability of that outcome will therefore be

$$\Pr(k, n) = \frac{n!}{k!\,(n-k)!} \prod_{i=1}^{n} \Pr(X_i = x_i) = \binom{n}{k} p^k (1-p)^{n-k} \,. \tag{1}$$

This is called a binomial distribution.

Sequence of beta-Bernoulli trials: an introduction

So far, we have treated the parameter p as a constant (i.e., we have operated within a frequentist paradigm). Now we will give it a statistical distribution (say, a beta distribution), and we will treat p as a realization of a random variable, denoted by P, following that distribution. We call this the Bayesian paradigm.

A beta distribution has two parameters, α and β. These parameters can take any positive real value, but to make intuition easier, we will mostly use integers larger than or equal to 1. The larger the integer value of α and the larger the integer value of β, the taller and narrower the shape of the distribution will be. The larger the integer value of α, the more the shape of the distribution will be shifted to the right. If $\alpha = \beta = 1$, the shape of the distribution will be uniform (i.e., flat). If, in general, $\alpha = \beta$, then the shape of the distribution will be symmetric around a mean of 0.5.

The probability that one of the random variables, X_i, takes on the value of 1 (i.e., is a success) given that P takes a certain value p is $\Pr(X_i = 1|p) = p$. Akin to the logic shown above, conditional on $P = p$, X_i follows a Bernoulli distribution with parameter p.

To get the marginal distribution, $\Pr(X_i = x_i)$, we need to multiply the PDF of P at p (i.e., $f(p)$) with the probability $\Pr(X_i = x_i|p)$, then integrate over all possible values of P. This process will allow us to obtain the PMF of a beta Bernoulli distribution

$$\Pr(X_i = x_i) = \int_0^1 \Pr(X_i = x_i|p)\, f(p) dp \,. \tag{2}$$

If we assume that all random variables comprising the sequence $\mathbf{X} = (X_1, X_2, ...)$ are independent, we can write

$$\Pr(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} \int_0^1 \Pr(x_i|p)\, f(p) dp\,, \tag{3}$$

which is a product of beta Bernoulli PMFs. We obtain the first entry of a sequence of data, $\mathbf{X}$, by drawing a value of p_1 from $f(p)$ and then drawing the value of 0 or 1 from a Bernoulli distribution with parameter p_1. We obtain the second entry of $\mathbf{X}$ by drawing a new value of p_2 from $f(p)$ and then drawing the value of 0 or 1 from a Bernoulli distribution with parameter p_2. We repeat this process n times to obtain a sequence, $\mathbf{X}$, of length n. This is could be called a "sequence of independent beta Bernoulli trials". However, we will show below that we can also get representative samples from that distribution by simply drawing repeatedly from regular Bernoulli distributions with a probability of success of $\mathrm{E}(P)$ (see equation (17)). In other words, a beta Bernoulli distribution is just a Bernoulli distribution with a probability of success of $\mathrm{E}(P)$ and the "sequence of independent beta Bernoulli trials" is actually just a sequence of independent Bernoulli trials.

Now, let $\mathbf{Y} = (Y_1, Y_2, ...)$ be a sequence similar to $\mathbf{X}$ but without the assumption that the random variables in this sequence are independent (the reason will become clear below). Assume that $\mathbf{Y} = (y_1, y_2, ..., y_n)$ is of length n with k successes (again denoted as a 1). The PMF of the corresponding beta-binomial distribution is obtained by substituting the conditional Bernoulli distribution in the PMF of the beta Bernoulli distribution (equation (2)) with the conditional binomial distribution

$$\begin{aligned}
\Pr(k, n) &= \int_0^1 \Pr(k, n|p)\, f(p)\, dp \\
&= \binom{n}{k} \int_0^1 \left(\prod_{i=1}^{n} [\Pr(Y_i = y_i|p)]\, f(p) \right) dp \\
&= \binom{n}{k} \Pr(y_1, y_2, \ldots, y_n)\ .
\end{aligned} \tag{4}$$

We obtain the first entry of a sequence, $\mathbf{Y}$, by drawing a value of p from $f(p)$ and then drawing a value of 0 or 1 from a Bernoulli distribution with parameter p. Unlike the process described above, we obtain the second entry in $\mathbf{Y}$ by using the *same value* of p as the parameter in the Bernoulli distribution to draw the value of 0 or 1. We repeat this process n times to obtain a sequence, $\mathbf{Y}$, of length n. This is usually called a "beta Bernoulli process".

While the random variables making up $\mathbf{X} = (X_1, X_2, \ldots)$ are independent, the random variables in $\mathbf{Y} = (Y_1, Y_2, \ldots)$ are not. The random variables in $\mathbf{Y}$ are connected by the fact that we used the Bernoulli distribution with the same parameter, p, for every entry in the sequence.

Note that even when sequences $\mathbf{y} = \mathbf{x}$ (i.e., instance $\mathbf{y}$ of a sequence of dependent random variables $\mathbf{Y}$ has the same values as instance $\mathbf{x}$ of a sequence of independent random variables $\mathbf{X}$),

$$\Pr(y_1, y_2, ..., y_n) = \int_0^1 \left(\prod_{i=1}^{n} [\Pr(y_i|p)]\, f(p) \right) dp \neq \prod_{i=1}^{n} \left[\int_0^1 (\Pr(x_i|p) f(p)) dp \right] = \Pr(x_1, x_2, ..., x_n)\,. \tag{5}$$

Therefore, the PMF of the beta-binomial is *not equal* to the product of *identical* independent beta Bernoulli PMFs multiplied by the binomial coefficient $\binom{n}{k}$.

How can we write $\Pr(y_1, y_2, ..., y_n)$ as a product of *non-identical* independent beta Bernoulli PMFs? As we already established, there is a dependence and correlation between the elements of the sequence $\mathbf{Y}$. Therefore, every beta Bernoulli PMF will have a different distribution of the parameter p depending on the previous elements of the sequence. The second random variable, Y_2, will use $f(p|y_1)$, the third random variable, Y_3, will use $f(p|y_1, y_2)$, and so on, such that the PDF of P is conditional on the previous values, which is necessary to capture the dependence structure and

$$\begin{aligned}
\Pr(y_1, y_2, \ldots, y_n) &= \int_0^1 \left(\prod_{i=1}^{n} [\Pr(y_i|p)]\, f(p) \right) dp \\
&= \int_0^1 (\Pr(y_1|p) f(p))\, dp \prod_{i=2}^{n} \left[\int_0^1 (\Pr(y_i|p)\, f(p|y_{i-1}, \ldots, y_1))\, dp \right] \qquad (6) \\
&= \Pr(y_1) \Pr(y_2|y_1) \Pr(y_3|y_2, y_1) \prod_{i=4}^{n} \Pr(y_i|y_{i-1}, y_{i-2}, \ldots, y_1)\,.
\end{aligned}$$

This is the chain rule of probability: A joint probability (density) can always be expressed as a product of conditional probabilities (or probability densities). These conditional random variables are necessarily independent and, in our case, non-identically distributed.

In the special case where the random variables of a sequence $\mathbf{X} = (X_1, X_2, \ldots)$ are independent, the conditional distributions are equal to the marginal distributions (e.g., $\Pr(x_3|x_2, x_1) = \Pr(x_3)$) and the joint PMF is just the product of the marginal PMFs ($\Pr(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} \Pr(x_i)$).

In the following sections, we will mostly concern ourselves with finding probabilities of the type $\Pr(n, k)$. Since $\mathbf{Y}$ is exchangeable (i.e., the joint distribution does not change when the positions of random variables in the sequence are altered), this is equivalent (except for the binomial coefficient) to finding $\Pr(y_1, y_2, \ldots, y_n)$ where $\sum_{i=1}^{n} y_i = k$.

Real-life uses of the beta-Binomial distribution

Families

Why or in what situations would we want to use a beta-binomial distribution? A frequently invoked example is the number of male and female children born. Let's assume the sex ratio in the entire population is $1:1$. However, the number of male children, k, in a family with n children will not follow a binomial distribution, because there is a tendency for some families to have more children of one sex. In other words, the sex ratio in individual families tends to be lower or higher than $1:1$ while remaining $1:1$ for the overall population. We can capture this with a beta-binomial distribution with parameters α and β, where $\alpha = \beta$ (recall that α and β must be the same in order for the mean to be 0.5).

By changing the probability of a male birth, p, from a single number (frequentist paradigm) to a random variable with a PDF $f(p)$ (Bayesian paradigm), we have introduced an additional degree of freedom and increased the variance of the overall distribution relative to the binomial distribution (with $p = 0.5$). This technique is called *overdispersion*, and it allows us to model dependency in our sequence of successes.

In the case of children born to families, there appears to be some correlation between sexes in a birth sequence, such that it is more probable for a family to have another boy given that a boy has already been born to that family. Suppose we want to simulate the sequence of births in a family with n children. We draw a value p from $f(p)$, say $p = 0.6$ and then draw the number of boys born in that family from a binomial distribution with parameter $p = 0.6$. This would be a family where there is a tendency to have more boys born since $p > 0.5$. Another way to simulate a family would be to draw the first birth from a Bernoulli distribution with $p = \frac{\alpha}{\alpha+\beta} = 0.5$. If the first child is a boy, we can assume that this is again a family with a tendency to have boys, but the evidence is pretty low, as the sample size so far is just 1. For the next draw from a Bernoulli distribution, we increase the parameter p slightly to $p = \frac{\alpha+1}{\alpha+\beta+1}$ (i.e., the probability to have a second boy is higher than it was to have the first boy). If the beta distribution is narrow (α and β are large), then there is only a weak tendency to have more boys or more girls in a family than predicted by a binomial distribution, and p will change only very slightly from the first to the second birth. In this case, using Bayesian language, we remark that the prior is very strong and the inclusion of additional data

only weakly influences the posterior. If the second child is a girl, despite the family having a slightly higher probability of having a second boy, there is some evidence that the sex ratio in the family is $1:1$. We therefore set $p = \frac{\alpha+1}{\alpha+\beta+2} = 0.5$ for the third birth, then record the outcome. We continue this same process n times until we have a sequence of n children in the family.

Both procedures produce a valid sample from the beta-binomial distribution, and if you repeat these procedures many times, the average number of boys over all those families (with n children) will be the same as the average number of girls, meaning that the overall ratio will be $1:1$.

Let's look at the probability that a single family will have four boys. You can think of the beta-binomial as an infinite mixture of binomial distributions, weighted by $f(p)$. In short, you could calculate the probability to have four boys according to a binomial distribution with parameter p, then multiply or weigh this by $f(p)$, and integrate over all values of p. Alternatively (and because we are dealing with *exchangeable* sequences), you could select any sequence of births that contains exactly four boys. You would initiate the system by setting the probability of the first child being male to 0.5, then condition the probability of the next child on the gender of the first, and so on. Overall, you take the product over non-identical Bernoulli PMFs, which is now the probability of our representative sequence containing exactly four boys, and multiply this by the number of possible sequences containing exactly four boys, thus arriving at the same result as before.

Coin tosses

We will now use coin tosses to illustrate the beta Bernoulli process.

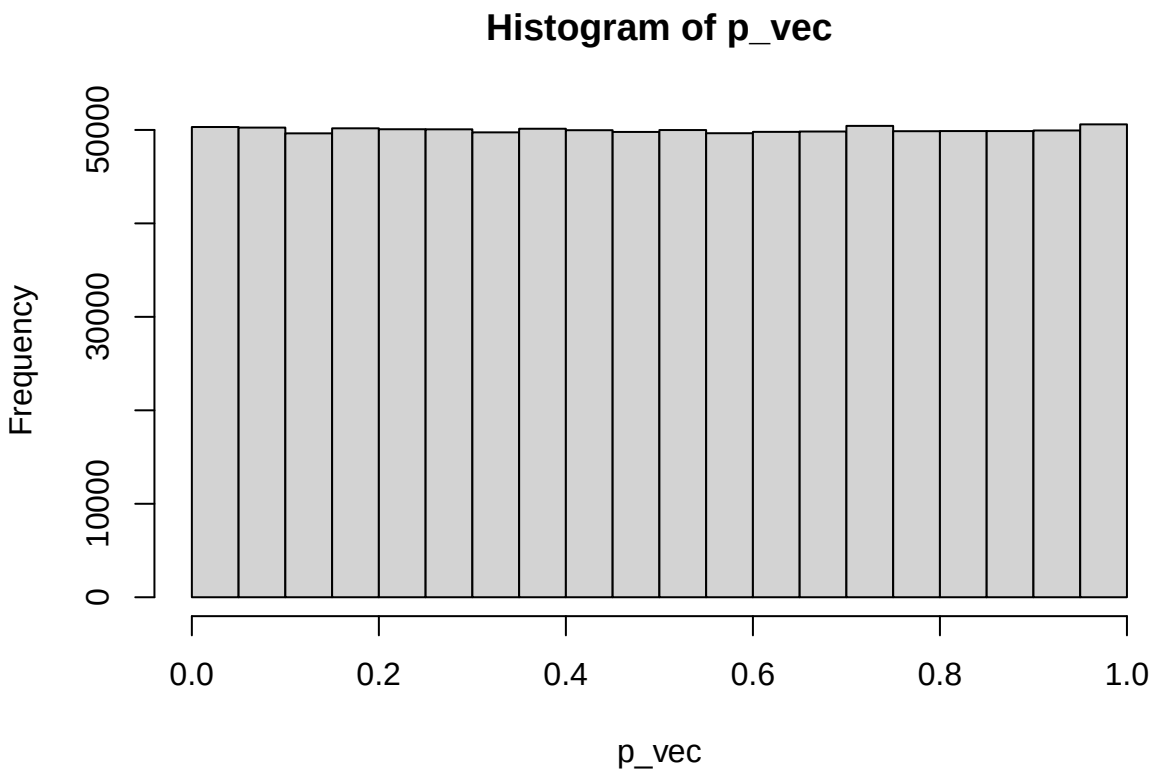
First, we draw a value p from a beta distribution. We will then flip n identical coins, all with a probability of heads of p. We will repeat this process t times to produce t different sequences of length n, or equivalently, t instances of a beta Bernoulli process.

Next, we will again produce t sequences, but this time we will not choose a static probability, p, of heads for all coins in a sequence, but instead we will draw from n beta Bernoulli distributions, meaning we will use a different probability of heads for each coin.

Let's look at a specific example: Assume we are drawing p from a beta distribution with parameters $\alpha = 1$ and $\beta = 1$ (i.e., from a uniform distribution), and we wish to generate $t = 1,000,000$ sequences of length $n = 3$.

First, let's draw $t = 1,000,000$ values from the uniform distribution

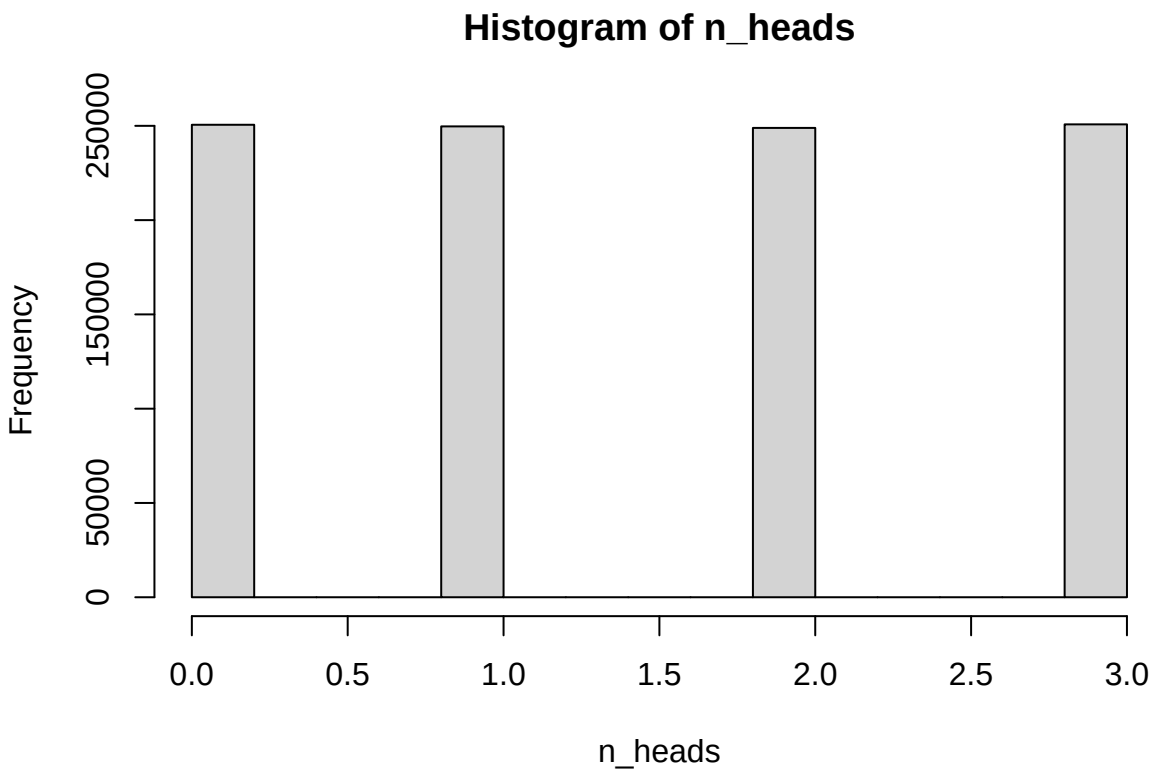
```
t <- 1000000
p_vec <- rbeta(t,1,1)
hist(p_vec)
```



Next, we use these values to generate $t = 1,000,000$ sequences of length $n = 3$ (where each of the 3 coin tosses in a given sequence are done with the same probability of heads, p).

When we calculate the number of heads in each sequence

```
n <- 3
n_heads <- rep(0,t)
for (i in 1:t) {
n_heads[i] <- sum(rbern(n,prob=p_vec[i]))
}
hist(n_heads)
```

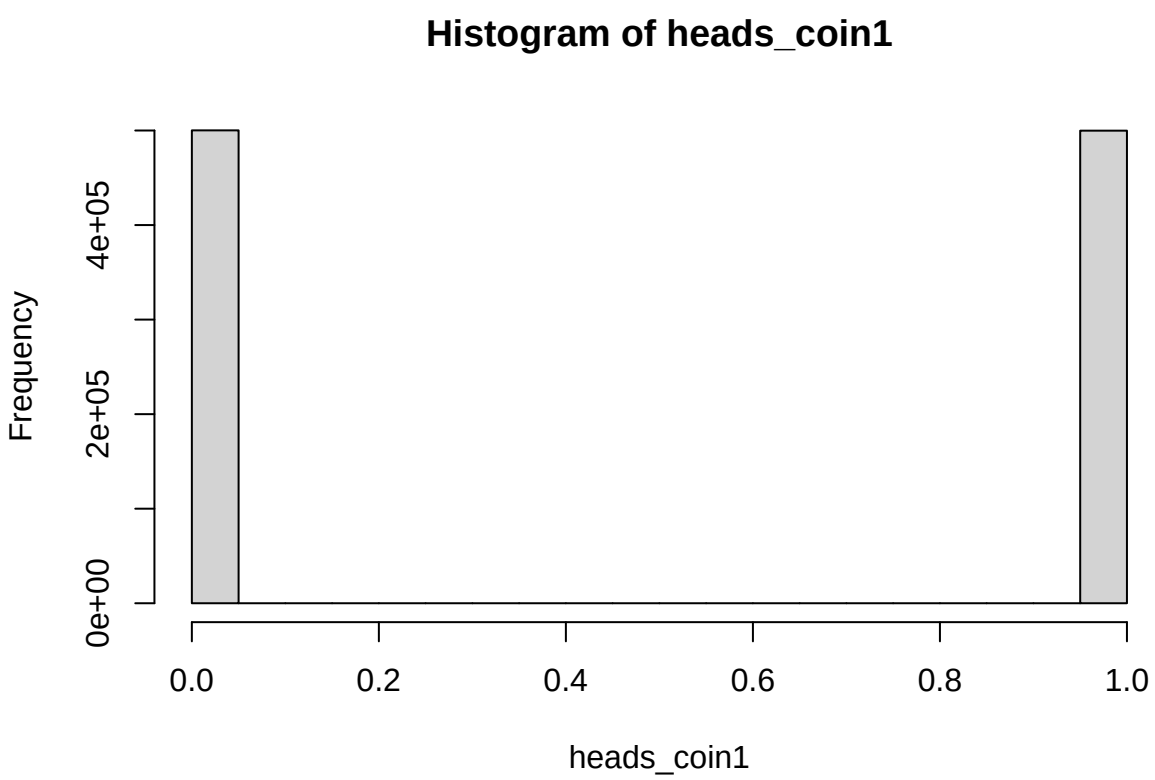


the distribution of the number of heads is uniform. This is indeed the case when the probability of heads, P, follows a uniform distribution, as we will show later.

We will now draw from a beta Bernoulli distribution. That is, we will flip $t = 1,000,000$ different coins, each coin with a different probability of success, p, and those probabilities are drawn from a Beta$(1, 1)$ (i.e., uniform) distribution. We repeat this $n = 3$ times.

We will start with the first random variable (i.e the first entry in each of the t sequences):

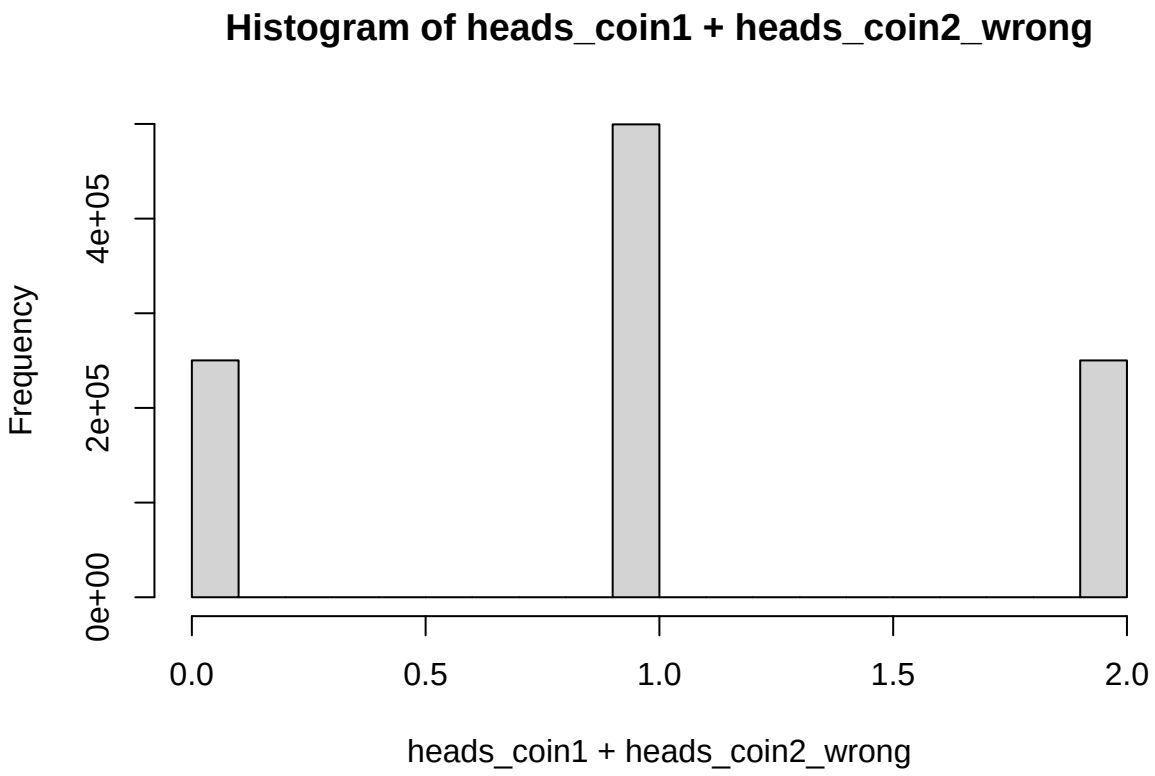
```
heads_coin1 <- rbern(t, rbeta(t,1,1))
hist(heads_coin1)
```



When we go to repeat this with the second coin, we run into a problem. In the first sampling scheme, we sampled each coin toss of a sequence with the same probability. Therefore, the coin tosses of a sequence are not independent anymore. If we now (in the second sampling scheme) simply toss the second coin (and again draw the value of p from a Beta$(1, 1)$ distribution), we failed to account for the correlation of outcomes in the sequence. On the one hand, we could just "remember" the values of p we used for the tosses of the first coin, and then use those same probabilities for the second coin, and this process would give us exactly the same situation as in the first sampling scheme. But let's instead assume that is not possible for us to remember, and we have to draw new values of p for each coin. How should we draw those new values so that the same overall situation is achieved? That is, to converge on the situation where we had "remembered", and thus, reused the appropriate values of p?

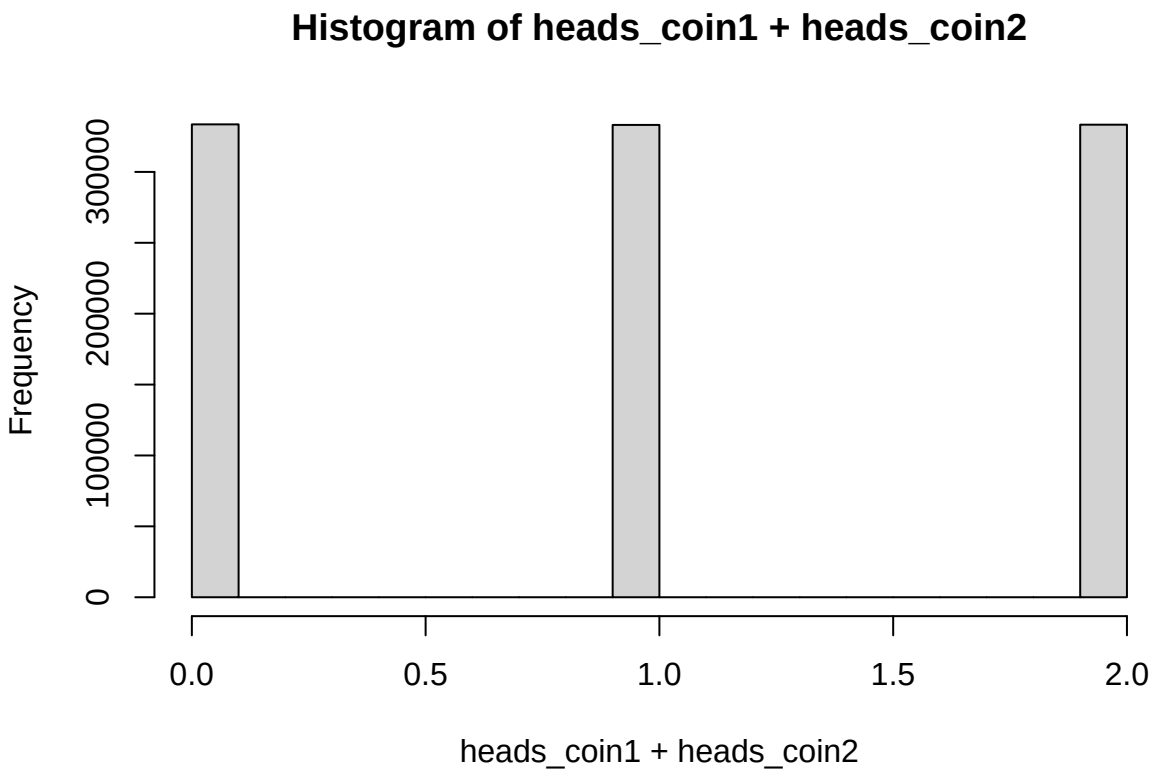
If we did it "wrong" and inappropriately drew all values of p from a Beta$(1, 1)$ distribution, we would end up with a solution of approximately 25% $k = 0$, 50% $k = 1$, and 25% $k = 2$. In other words, 50% of the density would be "transferred" from $k = 0$ to $k = 1$ (or in words, 50% of the coins that showed tails in the first toss would show tails in the second toss). Similarly, 50% of the density would be transferred from $k = 1$ to $k = 2$ (or in words, 50% of the coins that showed heads in the first toss would show heads in the second toss). In this incorrect method, the density is equal to the product of beta-Bernoulli densities with parameters α and β, which is just the product of Bernoulli densities with $p = \frac{\alpha}{\alpha+\beta}$, and certainly not a uniform distribution.

```
heads_coin2_wrong <- rbern(t, rbeta(t,1,1))
hist(heads_coin1+heads_coin2_wrong)
```



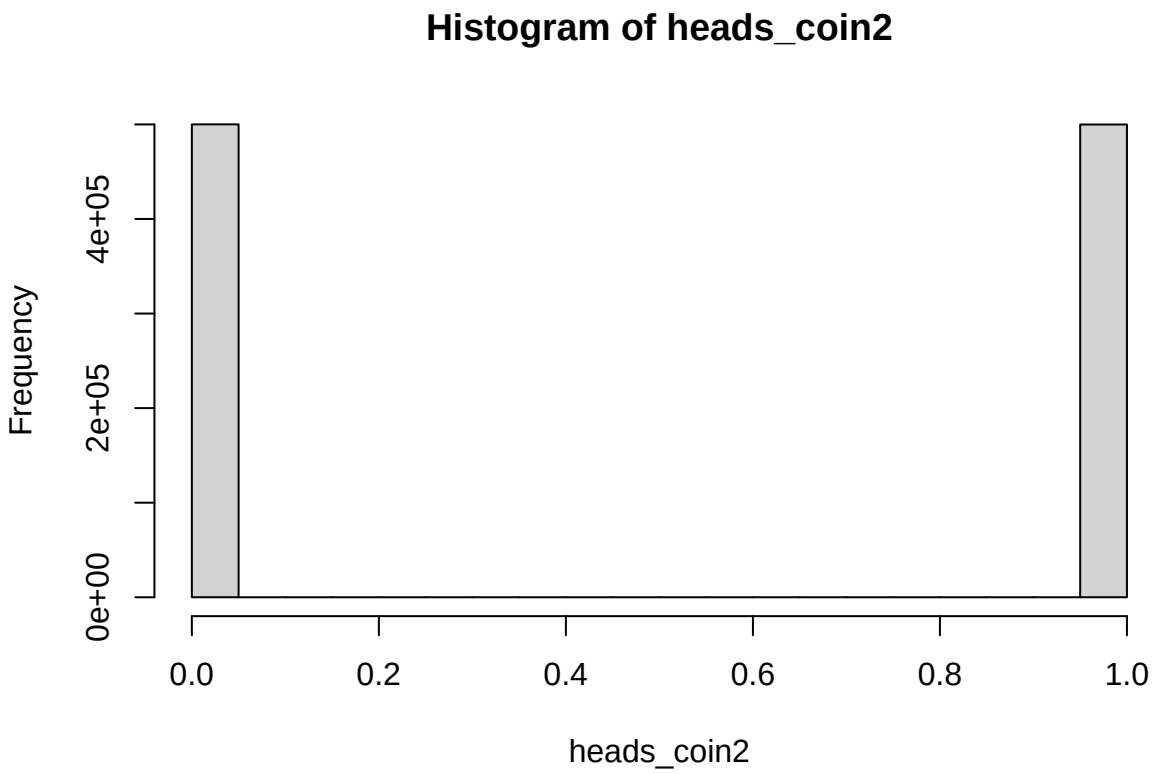
To achieve an approximately uniform distribution, we would instead need to transfer one-third of the density from $k = 0$ to $k = 1$ and two-thirds of the density from $k = 1$ to $k = 2$. Equivalently, we need about 66% of the coins that showed tails in the first toss to show tails in the second toss, and about 66% of the coins that showed heads in the first toss to show heads in the second toss. In this example there is an obvious correlation between the first and the second toss in each of the sequences. We will next show that in order to have, on average, 66% of coins show heads, we would need to draw p from a Beta $(2, 1)$ distribution, and in order to have, on average, 66% of coins show tails, we would need to draw p from a Beta $(1, 2)$ distribution.

```
heads_coin2 <- rbern(t, rbeta(t,1+heads_coin1,1+(1-heads_coin1)))
hist(heads_coin1+heads_coin2)
```



Note that the number of heads in the second toss has exactly the same distribution as the number of heads in the first (or any other) toss. About 50% of the coins show heads in the first toss, and of those, about 66% show heads in the second toss. Similarly, about 50% of the coins show tails in the first toss, and of those, about 33% show heads in the second toss. In summary, on average, about 50% of the coins in the second toss show heads, so the marginal probability of heads remains constant for every toss. Only the conditional probabilities change with every toss.

```
hist(heads_coin2)
```



The beta distribution

In the coin tosses problem, we looked at the question, "What is the probability of k heads out of n tosses given the probability of heads is p?". Now we will look at the *inverse* of that question, "What is the probability of $P = p$ given we observed k heads out of n tosses?".

In this case, *inverse* actually means *posterior distribution*. Following Bayes' theorem, we want to find the probability density given by

$$f\left(p|k,n\right) = \frac{\Pr\left(k,n|p\right) f\left(p\right)}{\Pr\left(k,n\right)}\,. \tag{7}$$

From our demonstration above, we know that $\Pr\left(k,n|p\right) = \binom{n}{k} p^k \left(1-p\right)^{n-k}$ follows a binomial distribution. If we assume that we have no prior knowledge of P, we can choose a uniform prior and $f\left(p\right) = 1$. The marginal probability in the denominator is then simply

$$\Pr\left(k,n\right) = \int_0^1 \Pr\left(k,n|p\right) f\left(p\right) dp = \binom{n}{k} \int_0^1 p^k \left(1-p\right)^{n-k} dp = \binom{n}{k} \mathrm{B}\left(k+1,(n-k)+1\right)\,, \tag{8}$$

where B is called the *beta function*. Formally, the *beta function* is defined as

$$\mathrm{B}\left(a+1,b+1\right) = \int_0^1 p^a \left(1-p\right)^b dp\,. \tag{9}$$

The posterior distribution is thus called a beta distribution and can be calculated as

$$f\left(p|k,n\right)=\frac{p^{k}\left(1-p\right)^{n-k}}{\mathrm{B}\left(k+1,(n-k)+1\right)}\,. \tag{10}$$

Statisticians usually transform the parameters of this distribution to $\alpha=k+1$ and $\beta=(n-k)+1$, which makes the formula for the posterior density

$$f\left(p|\alpha,\beta\right)=\frac{p^{\alpha-1}\left(1-p\right)^{\beta-1}}{\mathrm{B}\left(\alpha,\beta\right)}=\mathrm{Beta}\left(\alpha,\beta\right)\,. \tag{11}$$

To reiterate, the density of a $\mathrm{Beta}\left(\alpha,\beta\right)$ distribution is the probability density of P given we observed $\alpha-1$ heads and $\beta-1$ tails and have no other prior knowledge of the coin.

Beta-binomial probability

We will now look again at the beta-binomial distribution. Recall we drew a value p from a $\mathrm{Beta}\left(\alpha,\beta\right)$ distribution and then flipped n identical coins, each with a probability of heads of p. The probability of observing k heads in those n tosses can be calculated simply by integrating over the possible values of p (this is equation (4), which we will write here again with the particular choice of $f(p)=\mathrm{beta}(\alpha,\beta)$)

$$\begin{aligned}\Pr\left(n,k\right)&=\int_{0}^{1}\Pr\left(n,k|p\right)f\left(p\right)dp\\&=\binom{n}{k}\int_{0}^{1}p^{k}\left(1-p\right)^{n-k}\mathrm{Beta}\left(\alpha,\beta\right)dp\\&=\binom{n}{k}\frac{1}{\mathrm{B}\left(\alpha,\beta\right)}\int_{0}^{1}p^{k+\alpha+1}\left(1-p\right)^{n-k+\beta+1}dp\\&=\binom{n}{k}\frac{\mathrm{B}\left(k+\alpha,n-k+\beta\right)}{\mathrm{B}\left(\alpha,\beta\right)}\,.\end{aligned} \tag{12}$$

The beta function can be written as $\mathrm{B}\left(m,n\right)=\frac{(m-1)!(n-1)!}{(m+n-1)!}$, which allows us to rewrite the previous equation as

$$\begin{aligned}\Pr\left(n,k\right)&=\binom{n}{k}\frac{\mathrm{B}\left(k+\alpha,n-k+\beta\right)}{\mathrm{B}\left(\alpha,\beta\right)}\\&=\binom{n}{k}\frac{\left(\alpha\times\left(\alpha+1\right)\times\cdots\times\left(\alpha+k-1\right)\right)\left(\beta\times\left(\beta+1\right)\times\cdots\times\left(\beta+n-k-1\right)\right)}{\left(\alpha+\beta\right)\times\left(\alpha+\beta+1\right)\times\cdots\times\left(\alpha+\beta+n-1\right)}\,.\end{aligned} \tag{13}$$

We can simplify this expression by defining $x^{[t]}=x\left(x+1\right)\left(x+2\right)\ldots\left(x+t-1\right)$, which gives

$$\Pr(n,k)=\binom{n}{k}\frac{\alpha^{[k]}\beta^{[n-k]}}{(\alpha+\beta)^{[n]}}\,. \tag{14}$$

We will now find a product of non-identical beta Bernoulli densities (times the binomial coefficient) that equals $\Pr\left(n,k\right)$. To this end (and the details of why will become clear below), it makes sense to split the equation into multiple terms:

$$\begin{aligned}\Pr(n,k) = \binom{n}{k} &\frac{\alpha}{\alpha+\beta}\frac{\alpha+1}{\alpha+\beta+1}\cdots\frac{\alpha+k-1}{\alpha+\beta+k-1} \\ &\times \frac{\beta}{\alpha+\beta+k}\frac{\beta+1}{\alpha+\beta+k+1}\cdots\frac{\beta+n-k-1}{\alpha+\beta+n-1}\,.\end{aligned} \tag{15}$$

Product of beta Bernoulli densities

Direct calculation of posterior

In the narrative above, we found that the outcome of each coin toss depends on the previous outcome in a given sequence, i.e. the distribution of the probability to show heads given the previous tosses changes from toss to toss. We again define Y_i to be the number of heads (either 1 or 0) in the ith toss. The probability $\Pr(y_{n+1}|y_1, y_2, \ldots, y_n) = \Pr(n+1, k+y_{n+1}|n,k)$ of seeing y_{n+1} heads in the $n+1$-th toss given the previous tosses is the same as the probability of observing $k+y_{n+1}$ heads in $n+1$ tosses given we observed k heads in n tosses. It depends on the distribution of P given we observed k heads in n tosses (i.e., given all previous tosses, $f(p|n,k) = f(p|y_1, y_2, ..., y_n)$), and

$$\begin{aligned}\Pr(n+1, k+y_{n+1}|n,k) &= \int_0^1 \Pr(n+1, k+y_{n+1}|n,k,p) f(p|n,k)\,dp \\ &= \int_0^1 p^{y_{n+1}}(1-p)^{1-y_{n+1}} f(p|n,k)\,dp\,.\end{aligned} \tag{16}$$

If $f(p|n,k)$ is the PDF of a beta distribution, then $\Pr(n, k+y_{n+1}|n,k)$ is the PMF of a beta Bernoulli distribution.

Below, we will show that this is the case and we will also show that a product of PMFs of non-identical beta Bernoulli distributions (times a binomial coefficient) is equal to the PMF of a beta Binomial distribution. First, however, note that we can rewrite the beta Bernoulli PMF as

$$\Pr(n+1, k+y_{n+1}|n,k) = \mathrm{E}(P|n,k)^{y_{n+1}}(1-\mathrm{E}(P|n,k))^{1-y_{n+1}} = \mathrm{Bern}(\mathrm{E}(P|n,k))\,. \tag{17}$$

In words, the probability of a coin showing heads (meaning k increasing by 1) given the previously observed number of heads in the total number of tosses, is equal to the *expectation* of the distribution of P conditional on those previous tosses. Alternatively, the probability of a coin showing tails, i.e. k staying the same, given the previously observed number of heads in the total number of tosses is equal to 1 minus the *expectation* of the distribution of P conditional on those previous tosses. In other words, the outcome of a coin toss given the previous tosses is a Bernoulli random variable with parameter $p = \mathrm{E}(P|n,k)$.

We now need an expression for $f(p|n,k)$, which is the posterior density of P. This expression can be calculated according to Bayes' theorem, which states that

$$f(p|n,k) = \frac{\Pr(n,k|p)\,f(p)}{\int_0^1 \Pr(n,k|p)\,f(p)\,dp}\,. \tag{18}$$

We use the Beta density as prior

$$f(p|\alpha,\beta) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}\,, \tag{19}$$

and when inserted in the above equation, we get the direct posterior

$$f\left(p|n,k,\alpha,\beta\right)=\frac{\frac{p^{\alpha+k-1}(1-p)^{\beta+n-k-1}}{\mathrm{B}(\alpha,\beta)}}{\frac{\mathrm{B}(\alpha+k,\beta+n-k)}{\mathrm{B}(\alpha,\beta)}}=\frac{p^{\alpha+k-1}\left(1-p\right)^{\beta+n-k-1}}{\mathrm{B}\left(\alpha+k,\beta+n-k\right)}=\mathrm{Beta}\left(\alpha+k,\beta+n-k\right). \tag{20}$$

The posterior distribution is again a beta distribution. This is the case because the beta distribution is what we call a *conjugate prior* to the Binomial and the Bernoulli distribution.

Posterior through Bayesian updating

Note that we can also use a Bayesian updating scheme to obtain the result in equation (20). To do this, we toss a coin, then calculate the posterior distribution of P. Then, we use the outcome as the prior to calculate the posterior after the second toss

$$\begin{aligned} f\left(p|n=1,k=y_1,\alpha,\beta\right)=f\left(p|y_1,\alpha,\beta\right)&=\frac{\Pr\left(y_1|p\right)f\left(p|\alpha,\beta\right)}{\Pr\left(y_1|\alpha,\beta\right)}\\ &=\frac{p^{\alpha+y_1-1}\left(1-p\right)^{\beta-y_1}}{\mathrm{B}\left(\alpha+y_1,\beta+1-y_1\right)}\\ &=\mathrm{Beta}\left(\alpha+y_1,\beta+1-y_1\right) \end{aligned} \tag{21}$$

and

$$\begin{aligned} f\left(p|n=2,k=y_1+y_2,\alpha,\beta\right)&=f\left(p|y_1,y_2,\alpha,\beta\right)\\ &=\frac{\Pr\left(y_2|p\right)f\left(p|y_1,\alpha,\beta\right)}{\Pr\left(y_2|y_1,\alpha,\beta\right)}\\ &=\frac{p^{y_2}\left(1-p\right)^{1-y_2}\frac{p^{\alpha+y_1-1}(1-p)^{\beta-y_1}}{\mathrm{B}(\alpha+y_1,\beta+1-y_1)}}{\int_0^1 p^{y_2}(1-p)^{1-y_2}\frac{p^{\alpha+y_1-1}(1-p)^{\beta-y_1}}{\mathrm{B}(\alpha+y_1,\beta+1-y_1)}dp}\\ &=\frac{p^{y_1+y_2+\alpha-1}\left(1-p\right)^{\beta-(y_1+y_2)+1}}{\mathrm{B}\left(\alpha+\mathrm{y}_1+\mathrm{y}_2,\beta-(\mathrm{y}_1+\mathrm{y}_2)+1\right)}\\ &=\mathrm{Beta}\left(\alpha+y_1+y_2,\beta+2-(y_1+y_2)\right). \end{aligned} \tag{22}$$

The first step in both equations is because the sequence is *exchangeable*.

We start with a Beta (α,β) density, the PDF of P given we observed $\alpha-1$ heads and $\beta-1$ tails, and from this probability distribution we draw a value p_1 and toss the first coin. It comes up heads. The distribution from which we draw p_2 is now Beta $(\alpha+1,\beta)$, the PDF of P given we observed α heads and $\beta-1$ tails. We toss the coin again, and it comes up heads again. The distribution from which we draw p_3 is now Beta $(\alpha+2,\beta)$, the PDF of P given we observed $\alpha+1$ heads and $\beta-1$ tails. We toss the coin a third time, it comes up tails. The distribution from which we draw p_4 is now Beta $(\alpha+2,\beta+1)$, the PDF of P given we observed $\alpha+1$ heads and β tails. Repeated many times, this procedure will produce the same proportion for each possible sequence relative to the scenario where we draw a value p from Beta (α,β), then toss all coins with that same value of p. Repeated many times, i.e. generating many different sequences of a certain length, this procedure will produce the same proportion for each possible sequence relative to the scenario where we draw a value p from Beta (α,β), then toss all coins of a given sequence with that same value of p and repeat this many times.

Joint distribution

We can obtain a *joint density* by multiplying the conditional PMFs (non-identical Bernoulli PMFs)

$$\begin{aligned}
\Pr(y_1, y_2 \ldots, y_n) &= \Pr(y_n|y_1, y_2, \ldots, y_{n-1}) \Pr(y_1, y_2, \ldots, y_{n-1}) \\
&= \Pr(y_n|y_1, y_2, \ldots, y_{n-1}) \Pr(y_{n-1}|y_1, y_2, \ldots, y_{n-2}) \Pr(y_1, y_2, \ldots, y_{n-2}) \\
&= \Pr(y_1) \prod_{i=2}^{n} \Pr(y_i|y_1, \ldots, y_{i-1}) \\
&= \mathrm{E}(P|\alpha, \beta)^{y_1} (1 - \mathrm{E}(P|\alpha, \beta))^{1-y_1} \prod_{i=2}^{n} \mathrm{E}(P|y_1, \ldots, y_{i-1})^{y_i} (1 - \mathrm{E}(P|y_1, \ldots, y_{i-1}))^{1-y_i} .
\end{aligned} \tag{23}$$

For simplicity with our notation, we have dropped the explicit conditioning on α and β above.

Since the sequences are *exchangeable*, we can "group" all the heads and all the tails together into a new sequence $\mathbf{Z}$ where $(z_1, z_2, \ldots, z_k)$ are all 1 and $(z_{k+1}, z_{k+2}, \ldots, z_n)$ are all 0. Thus, $\mathbf{Z}$ and $\mathbf{Y}$ have the same number of 1s and 0s, just in a different arrangement $\sum_{i=1}^{n} z_i = \sum_{i=1}^{n} y_i = k$

$$\begin{aligned}
\Pr(n, k) &= \binom{n}{k} \Pr(y_1, y_2, \ldots, y_n) \\
&= \binom{n}{k} \mathrm{E}(P|\alpha, \beta) \prod_{i=2}^{k} \mathrm{E}(P|z_1, \ldots, z_{i-1}) \prod_{i=k+1}^{n} 1 - \mathrm{E}(P|z_1, \ldots, z_{i-1}) .
\end{aligned} \tag{24}$$

The expectation of a $\mathrm{Beta}(\alpha, \beta)$ distribution is $\frac{\alpha}{\alpha+\beta}$ and we can insert this in the equation above using equation (20)

$$\begin{aligned}
\Pr(n, k) &= \binom{n}{k} \frac{\alpha}{\alpha+\beta} \prod_{i=2}^{k} \frac{\alpha+i-1}{\alpha+\beta+i-1} \prod_{i=k+1}^{n} 1 - \frac{\alpha+k-1}{\alpha+\beta+i-1} \\
&= \binom{n}{k} \frac{\alpha}{\alpha+\beta} \prod_{i=1}^{k-1} \frac{\alpha+i}{\alpha+\beta+i} \prod_{i=k}^{n-1} \frac{\beta+i-k}{\alpha+\beta+i} \\
&= \binom{n}{k} \frac{\mathrm{B}(k+\alpha, n-k+\beta)}{\mathrm{B}(\alpha, \beta)} .
\end{aligned} \tag{25}$$

This is the same as the product of beta Bernoulli densities equation (15).

Example

Assume we tossed a coin 3 times and obtained the result {1,0,1} (i.e., {heads, tails, heads}) and the probability p of the coin to show heads is drawn from a beta distribution with $\alpha = 3$, $\beta = 5$.

The prior

Imagine first there is a hypothetical coin toss experiment, where the coin is tossed six times and comes up heads $\alpha - 1 = 2$ and tails $\beta - 1 = 4$ times. We know nothing about the prior probability of that hypothetical coin to show heads, so we say the prior distribution is uniform. Equivalently, we say the prior distribution is $\mathrm{Beta}(1, 1)$. The posterior probability density after six hypothetical tosses is

$$
\begin{aligned}
f\left(p|\alpha=3,\beta=5\right) &= \frac{\Pr\left(n=6,k=2|p\right)f\left(p\right)}{\int_0^1 \Pr\left(n=6,k=2|p\right)f\left(p\right)dp} \\
&= \frac{\left(p^{3-1}\left(1-p\right)^{5-1}\right)\times\left(p^{1-1}\left(1-p\right)^{1-1}\right)}{\int_0^1\left(p^{3-1}\left(1-p\right)^{5-1}\right)\times\left(p^{1-1}\left(1-p\right)^{1-1}\right)dp} \\
&= \frac{p^{2+0}\left(1-p\right)^{4+0}}{\mathrm{B}\left(2+0+1,4+0+1\right)} \\
&= \frac{p^2\left(1-p\right)^4}{\mathrm{B}\left(3,5\right)} = \mathrm{Beta}\left(3,5\right)\,.
\end{aligned}
\tag{26}
$$

We will now use this probability as the prior probability of the first actual coin toss in our example. By saying that we toss a coin with p following a beta distribution with $\alpha=3$, $\beta=5$, we are saying that before the first toss of the experiment, we have already observed that coin to come up heads $\alpha-1=2$ and tails $\beta-1=4$ times. That is, we have calculated the prior probability for the coin to come up heads from those preliminary tosses.

The first real toss

We want to calculate the probability that the coin came up heads in the first actual toss. As we showed above, this probability is the *expected value* of P. For the Beta$(3,5)$ distribution, this *expectation* is equal to $\frac{3}{3+5}$. This means that

$$
\Pr\left(n=1,k=1|\alpha,\beta\right)=\frac{\alpha}{\alpha+\beta}=\frac{3}{8}\,. \tag{27}
$$

Next, we need to calculate the probability distribution of P conditional on the first toss being heads. In other words, we need to calculate the posterior density of P after the first toss came up heads and with the prior probability following a Beta$(3,5)$ distribution.

$$
\begin{aligned}
f\left(p|n=1,k=1\right) &= \frac{\Pr\left(n=1,k=1|p\right)\mathrm{Beta}(3,5)}{\int_0^1 \Pr(1,1|p)\mathrm{Beta}\left(3,5\right)dp} \\
&= \frac{p\frac{p^2(1-p)^4}{\mathrm{B}(3,5)}}{\frac{1}{\mathrm{B}(3,5)}\int_0^1 pp^2\left(1-p\right)^4dp} = \frac{\frac{p^3(1-p)^4}{\mathrm{B}(3,5)}}{\frac{\mathrm{B}(4,5)}{\mathrm{B}(3,5)}} \\
&= \frac{p^3\left(1-p\right)^4}{\mathrm{B}\left(4,5\right)} = \mathrm{Beta}\left(4,5\right)\,.
\end{aligned}
\tag{28}
$$

The posterior distribution is again a beta distribution, because the beta distribution is the *conjugate prior* to the Bernoulli distribution.

We started with the prior distribution Beta$(3,5)$, the probability distribution of P after hypothetically observing 2 heads and 4 tails. After one additional heads, the posterior distribution of p is Beta$(4,5)$, or the probability distribution of P obtained when observing 3 heads and 4 tails. We will now use $f\left(p|n=1,k=1\right)$ as the prior for the next coin toss.

The second real toss

Now we calculate the probability of observing tails in the second toss, given the outcome of the first toss.

$$\begin{aligned}\Pr\left(n=2,k=1|k_{prev}=1\right) &= 1-\Pr\left(n=2,k=2|k_{prev}=1\right)\\ &= 1-\mathrm{E}\left(P|n=1,k=1\right)=1-\mathrm{E}\left(\mathrm{Beta}\left(4,5\right)\right)=1-\frac{4}{4+5}=\frac{5}{9}.\end{aligned} \tag{29}$$

Now we will find the probability distribution of P conditional on the first two tosses being heads and tails. We will use the posterior of P after the first toss as the prior probability for the second toss, $f\left(p|n=1,k=1\right)$.

$$\begin{aligned}f\left(p|n=2,k=1\right) &= \frac{\Pr\left(n=1,k=0|p\right)f\left(p|n=1,k=1\right)}{\int_0^1 \Pr\left(1,0|p\right)f\left(p|n=1,k=1\right)dp}\\ &= \frac{\left(1-p\right)\frac{p^3(1-p)^4}{\mathrm{B}(4,5)}}{\frac{1}{\mathrm{B}(4,5)}\int_0^1\left(1-p\right)p^3\left(1-p\right)^4dp}\\ &= \frac{p^3\left(1-p\right)^5}{\mathrm{B}\left(4,6\right)}=\mathrm{Beta}\left(4,6\right)\ .\end{aligned} \tag{30}$$

The third real toss

The probability of observing heads in the third toss given the previous tosses is

$$\Pr\left(n=3,k=2|k_{prev}=1\right)=\mathrm{E}\left(P|n=2,k=1\right)=\mathrm{E}\left(\mathrm{Beta}\left(4,6\right)\right)=\frac{4}{4+6}=\frac{4}{10}. \tag{31}$$

The posterior distribution of P after the first three tosses can be calculated similarly as above, which gives us

$$\begin{aligned}f\left(p|n=3,k=2\right) &= \frac{\Pr\left(n=1,k=1|p\right)f\left(p|n=2,k=1\right)}{\int_0^1 \Pr\left(1,1|p\right)f\left(p|n=2,k=1\right)dp}\\ &= \frac{p^4\left(1-p\right)^5}{\mathrm{B}\left(5,6\right)}=\mathrm{Beta}\left(5,6\right)\ .\end{aligned} \tag{32}$$

The entire sequence

Now let's calculate the probability of observing heads twice and tails once. First, note that as with the binomial distribution, there are $\binom{3}{2}$ different sequences (with the same probability) that will lead to an outcome of three tosses with exactly two heads and one tails. Therefore, we will choose one representative sequence, say, the one described above ($\{1,0,1\}$), then multiply its resulting probability by $\binom{3}{2}$.

$$\begin{aligned}\Pr(n=3,k=2) &= \binom{3}{2}\Pr\left(n=1,k=1|\alpha=3,\beta=5\right)\\ &\quad\times\Pr\left(n=2,k=1|k_{prev}=1\right)\Pr\left(n=3,k=2|k_{prev}=1\right)\\ &= \binom{3}{2}\frac{3}{3+5}\frac{5}{4+5}\frac{4}{4+6}.\end{aligned} \tag{33}$$

We could have also obtained this value using the beta-binomial distribution instead of multiple Bernoulli distributions

$$\begin{aligned}\Pr(n=3,k=2) &= \int_0^1 \Pr(n=3,k=2|p)\, f(p)\, dp \\ &= \int_0^1 \text{Binom}(3,2|p)\,\text{Beta}(3,5)\, dp \\ &= \binom{3}{2}\frac{\text{B}(3+2,5+1)}{\text{B}(3,5)}\,.\end{aligned} \tag{34}$$

By definition of the beta function, this value is

$$\Pr(n=3,k=2) = \binom{3}{2}\frac{\frac{4!5!}{10!}}{\frac{2!4!}{7!}} = \binom{3}{2}\frac{3\times 4\times 5}{8\times 9\times 10} = \binom{3}{2}\frac{3}{8}\frac{5}{9}\frac{4}{10}\,. \tag{35}$$

Note that the marginal probability of any coin showing heads, irrespective of the previous tosses, is constant. This is because the joint probability distribution of a sequence of tosses only depends on the number of tosses and the number of heads, not on the actual sequence. As an example, imagine we toss a coin 3 times. The probability of it showing heads in the third toss is

$$\begin{aligned}\Pr(x_3=1) &= \Pr(x_1=1,x_2=1,x_3=1) + \Pr(x_1=0,x_2=1,x_3=1) \\ &\quad + \Pr(x_1=1,x_2=0,x_3=1) + \Pr(x_1=0,x_2=0,x_3=1) \\ &= \frac{\alpha^{[3]}}{(\alpha+\beta)^{[3]}} + 2\frac{\alpha^{[2]}\beta}{(\alpha+\beta)^{[3]}} + \frac{\alpha\beta^{[2]}}{(\alpha+\beta)^{[3]}} = \frac{\alpha}{\alpha+\beta}.\end{aligned} \tag{36}$$

Summary and comments

A Beta (α,β) distribution describes the distribution of P, the probability of a coin to come up heads, given we observed tossing heads $\alpha-1$ times and tails $\beta-1$ times. Every time the coin is tossed and comes up heads, the parameter α of the posterior distribution of P increases, but every time the coin is tossed and it comes up tails, the parameter β increases.

The probability of observing heads given the previous tosses is the *expectation* of the corresponding posterior of P.

When sampling from a beta-binomial distribution, a coin is tossed several times, each time with the same probability drawn from a Beta (α,β) distribution. Therefore, the *marginal* probability to show heads must be the same for all coins and by the properties of the Bernoulli distribution, it must be the expectation of the Beta distribution, $\frac{\alpha}{\alpha+\beta}$

The relationship

$$\Pr(n=n_1,k=0)\Pr(n=n_2,k=0) = \Pr(n=n_1+n_2,k=0)\,, \tag{37}$$

holds for binomial random variables, but does not hold in the case of a beta-binomial distribution. Since we cannot split a sequence into two independent sequences, the probability to have 0 heads in n_1+n_2 tosses is not equal to the probability to have 0 heads in n_1 tosses times the probability to have 0 heads in n_2 tosses. If we sample from a beta-binomial distribution, every toss of a sequence is carried out with the same probability p drawn from a beta distribution, and when we split a sequence in two, they will not both be generated with the same parameter p. These tosses are only independent conditional on the parameter p. Another way to look at this is that the probability of a toss in a sequence depends on the number of heads observed in the previous tosses of that sequence, and that the tosses of a sequence are correlated.

Correlation

The random variables Y_i and Y_k (i.e., the outcome of the toss of coin i and coin k in the same sequence) are correlated. In fact, they are identically distributed, dependent random variables. $Y_i|Y_1, Y_2, \ldots, Y_{i-1}$ and $Y_k|Y_1, Y_2, \ldots, Y_{k-1}$, on the other hand, are independent but non-identically distributed random variables.

To calculate the correlation, we need the variance of Y_i. We know that $\Pr(Y_i) = \frac{\alpha}{\alpha+\beta} = \mathrm{E}(Y_i) = \mathrm{E}(Y_k)$ and the variance is

$$\mathrm{var}(Y_i) = \Pr(Y_i)(1 - \Pr(Y_i)) = \frac{\alpha}{\alpha+\beta}\frac{\beta}{\alpha+\beta}. \tag{38}$$

Next, we note that

$$\Pr(Y_1 = 1, Y_2 = 1) = \frac{\alpha}{\alpha+\beta}\frac{\alpha+1}{\alpha+\beta+1} = \Pr(Y_i = 1, Y_j = 1), \tag{39}$$

where the last equality holds because the sequence is *exchangeable*. We can now calculate the covariance

$$\begin{aligned}\mathrm{cov}(Y_i, Y_k) &= \mathrm{E}(Y_i Y_k) - \mathrm{E}(Y_i)\mathrm{E}(Y_k) \\ &= \frac{\alpha}{\alpha+\beta}\frac{\alpha+1}{\alpha+\beta+1} - \left(\frac{\alpha}{\alpha+\beta}\right)^2 \\ &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)},\end{aligned} \tag{40}$$

and the correlation

$$\mathrm{corr}(Y_i, Y_k) = \frac{\mathrm{cov}(Y_i, Y_k)}{\mathrm{var}(Y_i)\,\mathrm{var}(Y_k)} = \frac{1}{\alpha+\beta+1}. \tag{41}$$

We can use the correlation $\mathrm{corr}(Y_i, Y_k) = \frac{1}{\alpha+\beta+1}$ and expected value $\mathrm{E}(Y_i) = \frac{\alpha}{\alpha+\beta}$ as an alternative parameterization of the beta distribution, and through that, the beta-binomial distribution.

Example cases and limits

For illustration, we have limited our examples to integer values of α and β. This is not necessary as either parameter can take any positive value. We will now explore some beta-binomial distributions with different parameter values.

Uniform distribution: $\alpha = 1$ and $\beta = 1$

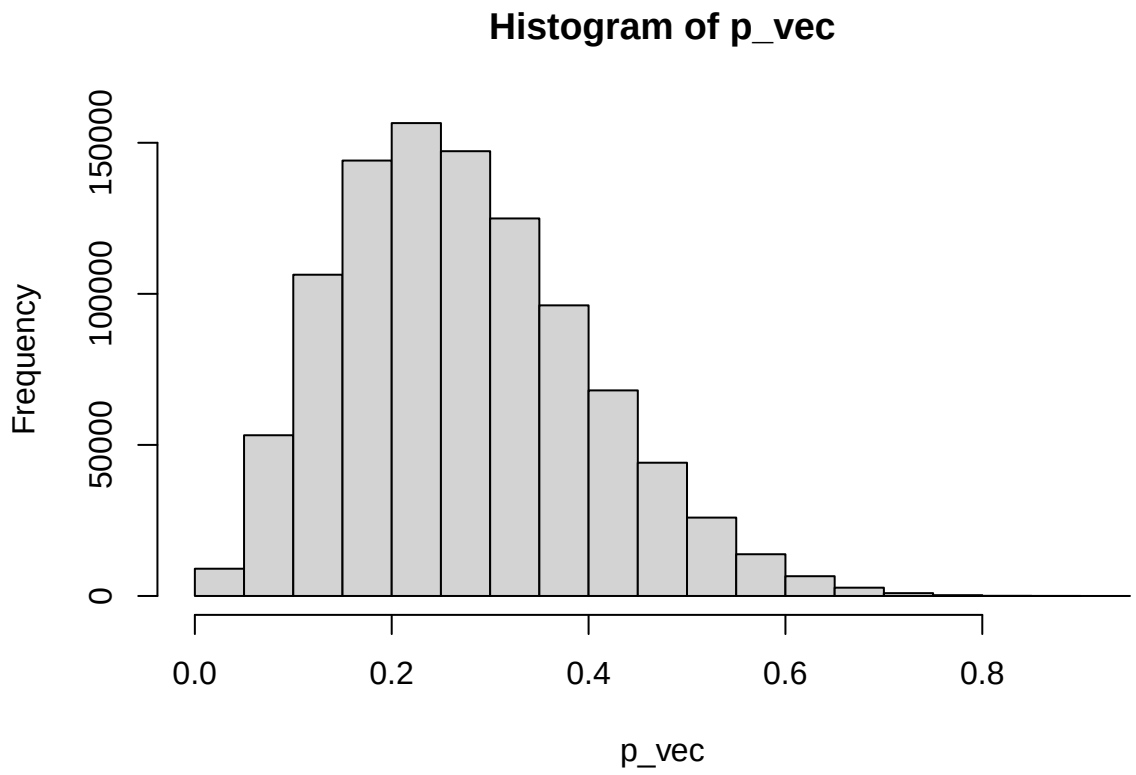
We have already seen that a $\mathrm{Beta}(1, 1)$ distribution is equivalent to a uniform distribution. The correlation is $\mathrm{corr}(Y_i, Y_k) = \frac{1}{3} = \rho$, the one-third of the density that had to be "transferred" from $k = 0$ to $k = 1$ in our example above. The mean is $\mathrm{E}(Y_i) = 0.5 = \pi$. The beta-binomial distribution can therefore also be parameterized as $\mathrm{BetaB}\left(n, \pi = 0.5, \rho = \frac{1}{3}\right)$. The probability mass function is uniformly distributed

$$\begin{aligned}\Pr(n,k) &= \int_0^1 \text{Binom}(n,k|p)\,\text{Beta}(1,1)\,dp \\ &= \binom{n}{k}\frac{\text{B}(k+1,n-k+1)}{\text{B}(1,1)} \\ &= \frac{n!k!(n-k)!}{k!(n-k)!(n+1)!} = \frac{n!}{(n+1)!} = \frac{1}{n}.\end{aligned} \tag{42}$$

Non-integer parameter values: $\alpha = 3.2$ and $\beta = 8.5$

We can also use non-integer parameters. For example, when we consider a distribution with the values $\alpha = 3.2$ and $\beta = 8.5$, the mean (or marginal probability of a coin to show heads) is $\text{E}(X_i) = \frac{3.2}{8.5+3.2} = 0.2745 = \pi$. The correlation is $\text{corr}(Y_i, Y_k) = \frac{1}{8.5+3.2+1} = 0.0787 = \rho$. If we plot the $\text{Beta}(3.2, 8.5)$ distribution, we see:

```
t <- 1000000
p_vec <- rbeta(t,3.2,8.5)
hist(p_vec)
```

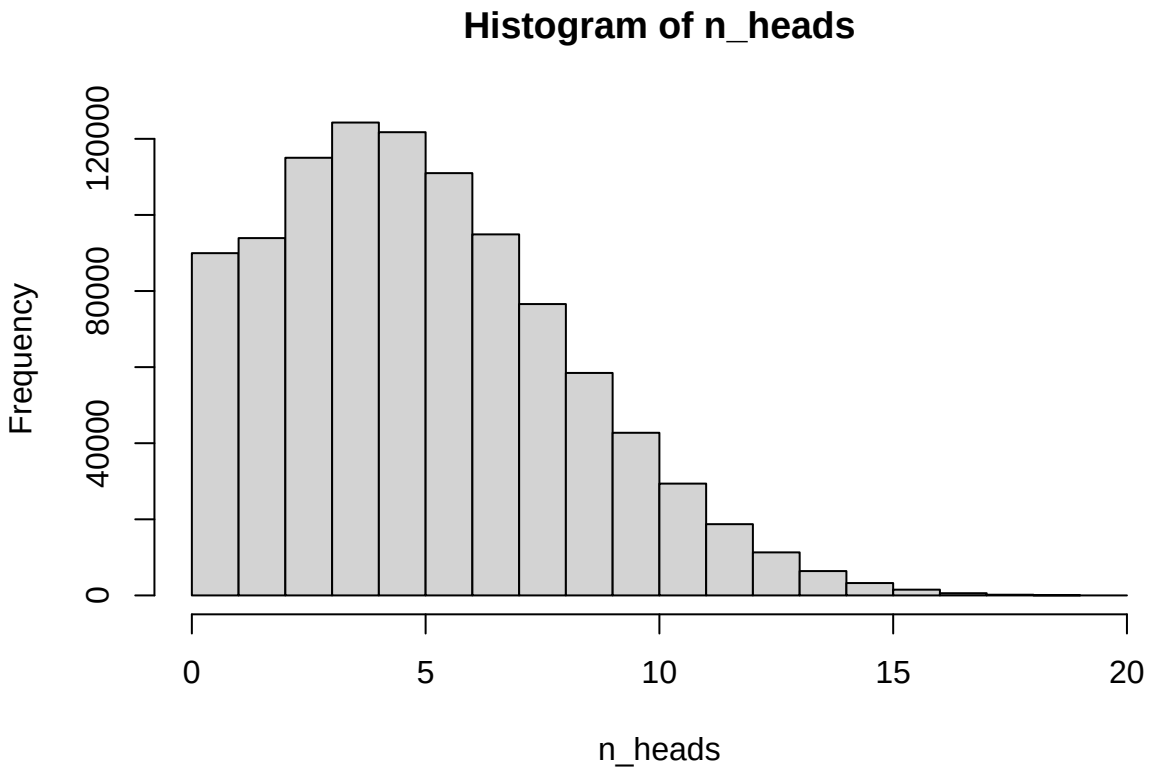


The probability mass function is

$$\Pr(n,k) = \int_0^1 \text{Binom}(n,k|p)\,\text{Beta}(3.2,8.5)\,dp = \binom{n}{k}\frac{\text{B}(k+1,n-k+1)}{\text{B}(3.2,8.5)} \tag{43}$$

and we plot it here for $n = 20$

```
n <- 20
n_heads <- rep(0,t)
for (i in 1:t) {
n_heads[i] <- sum(rbern(n,prob=p_vec[i]))
}
hist(n_heads)
```

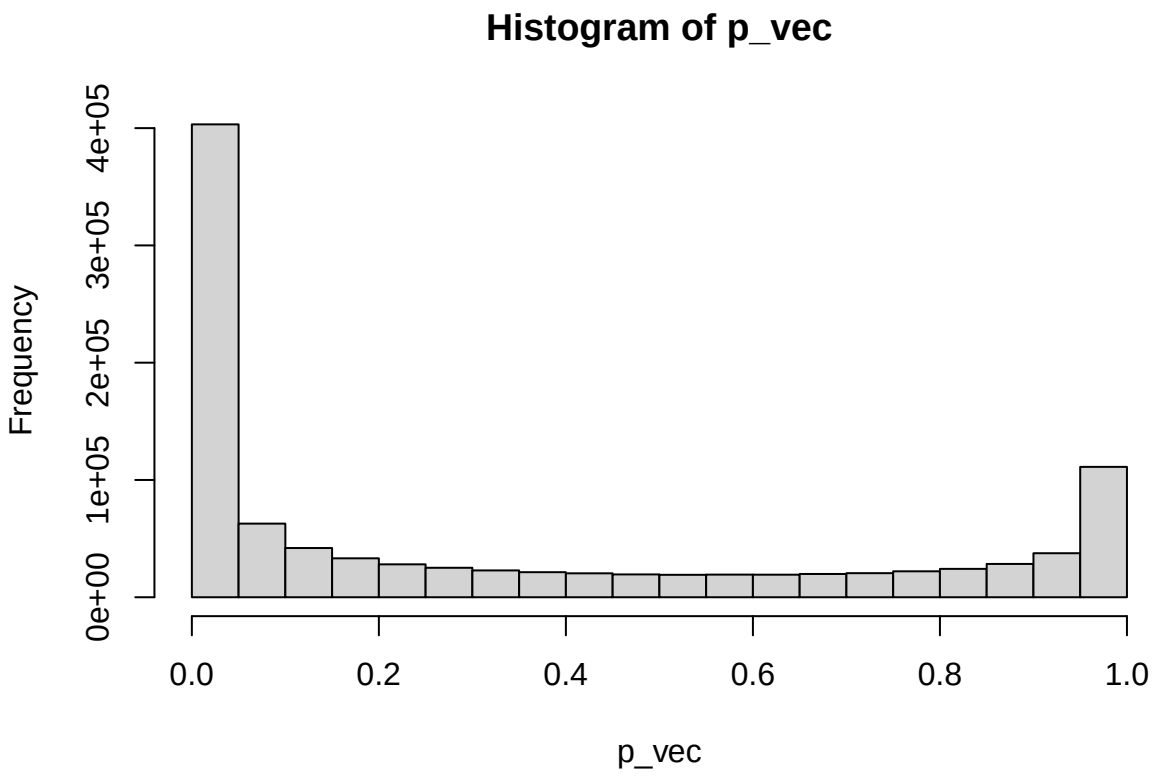


Parameters smaller than 1: $\alpha = 0.4$ and $\beta = 0.2$

In this distribution, it is marginally twice as probable to see heads than to see tails, $\mathrm{E}(Y_i) = \frac{0.4}{0.4+0.2} = 0.6667 = \pi$. We find that the correlation is very strong, $\mathrm{corr}(Y_i, Y_k) = \frac{1}{0.4+0.2+1} = 0.625 = \rho$. In fact, it is so strong that the beta distribution is bimodal, meaning, in this instance, that it is more probable to get "extreme sequences" with many more 1s than 0s or many more 0s than 1s than it is to get sequences with an approximately $1:1$ ratio.

To see this, we will first plot the Beta $(0.2, 0.4)$ distribution:

```
t <- 1000000
p_vec <- rbeta(t,0.2,0.4)
hist(p_vec)
```

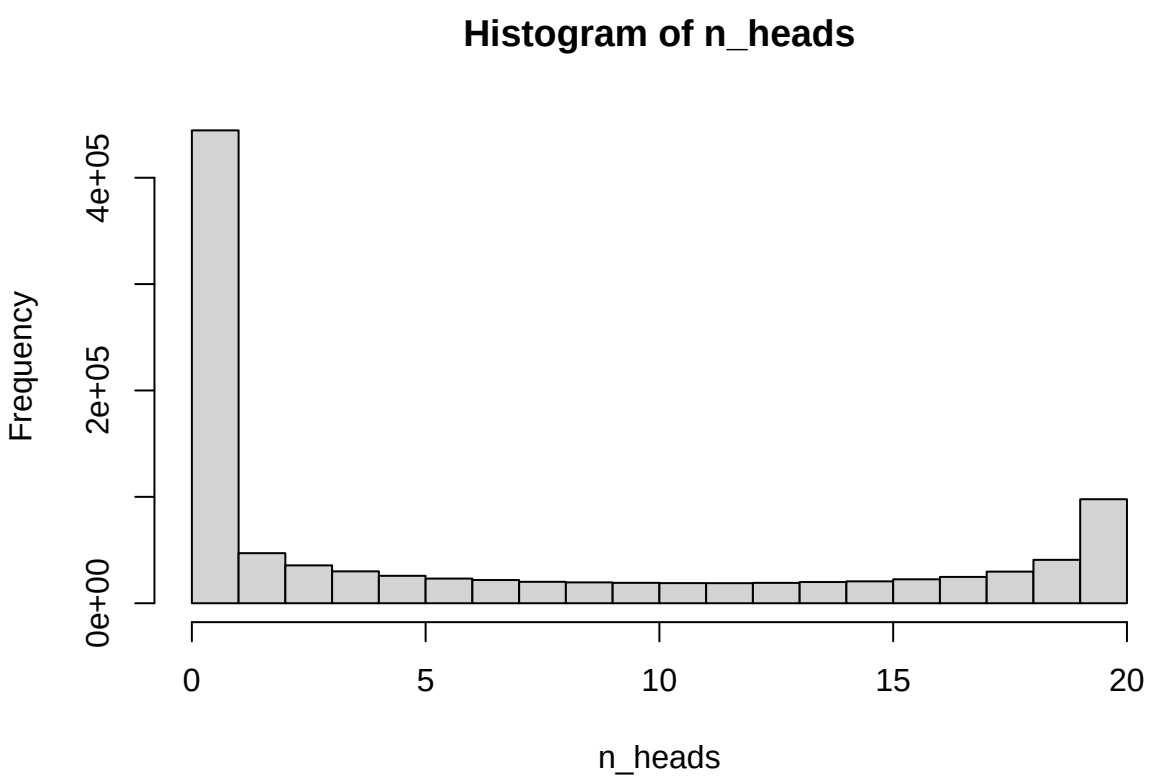


The probability mass function is

$$\Pr(n,k) = \int_0^1 \text{Binom}(n,k|p)\,\text{Beta}(0.2,0.4)\,dp = \binom{n}{k}\frac{\text{B}(k+1,n-k+1)}{\text{B}(0.2,0.4)} \tag{44}$$

and we plot it here for $n = 20$

```
n <- 20
n_heads <- rep(0,t)
for (i in 1:t) {
n_heads[i] <- sum(rbern(n,prob=p_vec[i]))
}
hist(n_heads)
```



Limits

As α and β approach 0 ($\alpha \to 0$ and $\beta \to 0$), the Beta PDF approaches a mixture of 2 Bernoulli PMFs with equal probability 0.5 at 0 and 1. This means that any sampled sequence will be either all 1s or all 0s and the correlation approaches 1 ($\rho \to 1$), while the mean is still $\pi = 0.5$. If $\alpha \to \infty$ or $\beta \to \infty$, the beta PDF will approach a Dirac delta function centered at 1 or 0, respectively. A Dirac delta function is a function whose value is 0 everywhere except at a single point. Its integral over the entire real line is equal to one, creating a cumulative distribution function that is the unit step function. Thus, in the case of $\alpha \to \infty$, sampled sequences will consist only of 1s and in $\beta \to \infty$ only of 0s. If both $\alpha \to \infty$ and $\beta \to \infty$, such that $\frac{\alpha}{\beta}$ and $\frac{\alpha}{\alpha+\beta}$ are constant, the beta PDF will approach a Dirac delta function centered at $\frac{\alpha}{\alpha+\beta}$. The beta-binomial distribution will approach a binomial distribution with parameter $p = \frac{\alpha}{\alpha+\beta}$. The correlation approaches 0 ($\rho \to 0$).

Summary

One can think of the beta-binomial distribution as a binomial distribution where the probability of success is not fixed but randomly drawn from a beta distribution. Conditionally on a certain value of the probability of success, the PMF of the binomial distribution is the product of the PMFs of identical Bernoulli distributions (times the binomial coefficient). Marginally (i.e. marginalizing over the probability of success), this is not true and the PMF of the beta-binomial distribution is not proportional to the product of the PMFs of identical beta-Bernoulli distributions.

In other words, if we draw a probability of success, p, from a beta distribution, generate a sequence of n indicator random variables from identical Bernoulli distributions, all with parameter p and repeat this

many times, we will obtain a distribution of sequences of length n. If we draw a value of p_1 from a beta distribution and then simulate an outcome of an indicator random variables from a Bernoulli distribution with parameter p_1, then draw a different value of p_2 from the beta distribution and another outcome of an indicator random variables from a Bernoulli distribution with parameter p_2, and so on until we have n draws, and then repeat this entire process many times, the distribution of the resulting sequences will not be the same as the distribution of the sequences obtained with the first sampling scheme. This is because in the first case, members of a sequence are correlated, whereas in the second they are not.

To see this better, remember that we have shown that a beta-Bernoulli distribution is just a Bernoulli distribution where the probability of success is the expectation of the beta distribution. Thus, we could rewrite the second sampling scheme above as: we generate a sequence of n indicator random variables, where each value is drawn from an identical Bernoulli distribution with p equal to the expectation of the beta-distribution.

However, we can generate sequences from *non-identical* Bernoulli distributions with different probabilities of success, i.e. from beta-Bernoulli distributions with different beta distributions. The beta distributions and the probabilities of success will depend on previous outcomes in the sequence, hence capturing the correlation. One way to view this is that in a Bayesian updating scheme, every toss in a sequence updates our knowledge about the probability of success.