Section Starter Question

Consider a family with 5 children. What is the probability of having all five children be boys? How many children must a couple have for at least a 0.95 probability of at least one girl? What is a proper and general mathematical framework for setting up the answer to these questions and similar questions?

Key Concepts

1. A **binomial random variable** S_n counts the number of successes in a sequence of n trials of an experiment.
2. A binomial random variable S_n takes only integer values between 0 and n inclusive and
$$\mathbb{P}\left[S_n = k\right] = \binom{n}{k} p^k (1-p)^{n-k}$$
for $k = 0, 1, 2, \ldots, n$.
3. The expectation of a binomial random variable with n trials and probability of success p on each trial is:
$$\mathbb{E}\left[S_n\right] = np$$
4. The variance of a binomial random variable with n trials and probability of success p on each trial is:
$$\operatorname{Var}\left[S_n\right] = npq = np(1-p)$$

Vocabulary

1. An **elementary experiment** is a physical experiment with two outcomes. An elementary experiment is also called a **Bernoulli trial**.
2. A **composite experiment** consists of repeating an elementary experiment n times.
3. The **sample space**, denoted Ω_n is the set of all possible sequences of n 0s and 1s representing all possible outcomes of the composite experiment.
4. A **random variable** is a function from the sample space Ω_n to the real numbers $\mathbb{R}$.

Mathematical Ideas

Sample Space for a Sequence of Experiments

An **elementary experiment** in this section consists of an experiment with two outcomes. An elementary experiment is also called a **Bernoulli trial**. Label the outcomes of the elementary experiment 1, occurring with probability p and 0, occurring with probability q, where $p + q = 1$. Often we name 1 as **success** and 0 as **failure**. For example, a coin toss would be a physical experiment with two outcomes, say with "heads" labeled as success, and "tails" as failure.

A **composite experiment** consists of repeating an elementary experiment n times. The **sample space**, denoted Ω_n is the set of all possible sequences of n 0's and 1's representing all possible outcomes of the composite experiment. We denote an element of Ω_n as $\omega = (\omega_1, \dots, \omega_n)$, where each $\omega_k = 0$ or 1. That is, $\Omega_n = \{0,1\}^n$. We assign a probability measure $\mathbb{P}_n[\cdot]$ on Ω_n by multiplying probabilities of each Bernoulli trial in the composite experiment according to the principle of independence. Thus, for $k = 1, \dots, n$,

$$\mathbb{P}[\omega_k = 0] = q \text{ and } \mathbb{P}[\omega_k = 1] = p$$

and inductively for each $(e_1, e_2, \dots, e_n) \in \{1, 0\}^n$

$$\mathbb{P}_{n+1}[\omega_{n+1} = 1 \text{ and } (\omega_1, \dots, \omega_k) = (e_1, \dots, e_n)] = \\ \mathbb{P}[\omega_{n+1} = 1] \times \mathbb{P}[(\omega_1, \dots, \omega_n) = (e_1, \dots, e_n)]$$

Additionally, let $S_n(\omega)$ be the number of 1's in $\omega \in \Omega_n$. Note that $S_n(\omega) = \sum_{k=1}^{n} \omega_k$. We also say $S_n(\omega)$ is the number of successes in the composite experiment. Then

$$\mathbb{P}_n[\omega] = p^{S_n(\omega)} q^{n-S_n(\omega)}.$$

We can also define a unified sample space Ω that is the set of all infinite sequences of 0's and 1's. We sometimes write $\Omega = \{0, 1\}^\infty$. Then Ω_n is the projection of the first n entries in Ω.

A **random variable** is a function from a set called the **sample space** to the real numbers $\mathbb{R}$. For example as a frequently used special case, for $\omega \in \Omega$ let

$$X_k(\omega) = \omega_k,$$

then X_k is an indicator random variable taking on the value 1 or 0. X_k (the dependence on the sequence ω is usually suppressed) indicates success or failure at trial k. Then as above,

$$S_n = \sum_{k=1}^{n} X_i = \sum_{k=1}^{n} \omega_i$$

is a random variable indicating the number of successes in a composite experiment.

Binomial Probabilities

Proposition 1. *The random variable S_n takes only integer values between 0 and n inclusive and*

$$\mathbb{P}_n[S_n = k] = \binom{n}{k} p^k q^{n-k}.$$

Remark. The notation $\mathbb{P}_n[\cdot]$ indicates that we are considering a family of probability measures indexed by n on the sample space Ω.

Proof. From the inductive definition

$$\mathbb{P}[\omega_i = 0] = q \text{ and } \mathbb{P}[\omega_i = 1] = p$$

and inductively for each $(e_1, e_2, \dots, e_n) \in \{1, 0\}^n$

$$\mathbb{P}_{n+1}[\omega_{n+1} = 1 \text{ and } (\omega_1, \dots, \omega_n) = (e_1, \dots, e_n)] = \mathbb{P}[\omega_{n+1} = 1] \times \mathbb{P}_n[(\omega_1, \dots, \omega_n) = (e_1, \dots, e_n)]$$

the probability assigned to an ω having k 1's and $n-k$ 0's is $p^k(1-p)^{n-k} = p^{S_n(\omega)}(1-p)^{n-S_n(\omega)}$. The sample space Ω_n has precisely $\binom{n}{k}$ such points. By the additive property of disjoint probabilities,

$$\mathbb{P}_n[S_n = k] = \binom{n}{k} p^k q^{n-k}.$$

and the proof is complete. □

Proposition 2. *If $X_1, X_2, \dots X_n$ are independent, identically distributed random variables with distribution $\mathbb{P}[X_i = 1] = p$ and $\mathbb{P}[X_i = 0] = q$, then the sum $X_1 + \dots + X_n$ has the distribution of a binomial random variable S_n with parameters n and p.*

Proposition 3. *1.*

$$\mathbb{E}[S_n] = np$$

2.

$$\operatorname{Var}[S_n] = npq = np(1-p)$$

Proof. First Proof: By the binomial expansion

$$(p+q)^n = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k}.$$

Differentiate with respect to p and multiply both sides of the derivative by p:

$$np(p+q)^{n-1} = \sum_{k=0}^{n} k \binom{n}{k} p^k q^{n-k}.$$

Now choosing $q = 1 - p$,

$$np = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} = \mathbb{E}[S_n].$$

For the variance, differentiate the binomial expansion with respect to p twice:

$$n(n-1)(p+q)^{n-2} = \sum_{k=0}^{n} k(k-1)\binom{n}{k} p^{k-2} q^{n-k}.$$

Multiply by p^2, substitute $q = 1-p$,and expand:

$$n(n-1)p^2 = \sum_{k=0}^{n} k^2 \binom{n}{k} p^k (1-p)^{n-k} - \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} = \mathbb{E}\left[S_n^2\right] - \mathbb{E}\left[S_n\right]$$

Therefore,

$$\operatorname{Var}\left[S_n\right] = \mathbb{E}\left[S_n^2\right] - \left(\mathbb{E}\left[S_n\right]\right)^2 = n(n-1)p^2 + np - n^2p^2 = np(1-p)$$

□

Proof. Second proof: Use that the sum of expectations is the expectation of the sum, and apply it to the corollary with $S_n = X_1 + \cdots + X_n$ with $\mathbb{E}\left[X_i\right] = p$.

Similarly, use that the sum of variances of independent random variables is the variance of the sum applied to $S_n = X_1 + \cdots + X_n$ with $\mathbb{E}\left[X_i\right] = p(1-p)$. □

Examples

Example. The following example appeared in the January 20, 2017 "Riddler" puzzler on the website `fivethirtyeight.com`.

You and I find ourselves indoors one rainy afternoon, with nothing but some loose change in the couch cushions to entertain us. We decide that well take turns flipping a coin, and that the winner will be whoever flips 10 heads first. The winner gets to keep all the change in the couch! Predictably, an enormous argument erupts: We both want to be the one to go first.

What is the first flippers advantage? In other words, what percentage of the time does the first flipper win this game?

First solve an easier version of the puzzle where the first person to flip a head will win. Let the person who flips first be A, and the probability that A wins by first obtaining a head is P_A. Then adding the probabilities for the disjoint events that the sequence of flips is H, or TTH, or TTTTH and so forth.

$$P_A = \frac{1}{2} + \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right) + \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right) + \dots$$
$$= \frac{1}{2}\left(1 + \frac{1}{4} + \left(\frac{1}{4}\right)^2 + \dots\right)$$
$$= \frac{1}{2} \cdot \frac{1}{1 - 1/4} = \frac{1}{2} \cdot \frac{4}{3} = \frac{2}{3}.$$

Another way to do this problem would be to use first-step analysis from Markov Chain theory. Then the probability of the first player winning P_A is the probability of winning on the first flip plus the probability of both players each losing their first flip at which point the game is essentially starting over,

$$P_A = \frac{1}{2} + \frac{1}{4}P_A.$$

Solving, $\frac{3}{4}P_A = \frac{1}{2}$ or

$$P_A = \frac{1}{2} \cdot \frac{4}{3} = \frac{2}{3}.$$

Now extend the same reasoning as in the first approach to the case of the first player to get 10 heads winning. The first case for A to win is to get 9 heads in flips $1, 3, 5, \dots, 17$ and the 10th head on flip 19 and for player B to get anywhere from 0 to 9 heads on flips $2, 4, 6, \dots, 18$. This is probability $\binom{9}{9}\left(\frac{1}{2}\right)^9 \cdot \frac{1}{2}$ and cumulative binomial probability $\sum_{k=0}^{9} \binom{9}{k}\left(\frac{1}{2}\right)^9$ respectively. The next disjoint probability case is for A to win is to get 9 heads in flips $1, 3, 5, \dots, 19$ and the 10th head on flip 21 and for player B to get anywhere from 0 to 9 heads on flips $2, 4, 6, \dots, 20$. This is probability $\binom{10}{9}\left(\frac{1}{2}\right)^{10} \cdot \frac{1}{2}$ and cumulative binomial probability $\sum_{k=0}^{9} \binom{10}{k}\left(\frac{1}{2}\right)^{10}$ respectively. In general, the disjoint probability case is for A to win is to get 9 heads in flips $1, 3, 5, \dots, 2j-1$ and the 10th head on flip $2j+1$ and for player B to get anywhere from 0 to 9 heads on flips $2, 4, 6, \dots, 2j$. This is probability $\binom{j}{9}\left(\frac{1}{2}\right)^j \cdot \frac{1}{2}$ and cumulative binomial probability $\sum_{k=0}^{9} \binom{j}{k}\left(\frac{1}{2}\right)^j$ respectively.

Then multiplying the independent probabilities for A and B in each case and adding all these disjoint probabilities

$$P_A = \sum_{j=9}^{\infty} \binom{j}{9} \left(\frac{1}{2}\right)^{j+1} \sum_{k=0}^{9} \binom{j}{k} \left(\frac{1}{2}\right)^{j}$$

There does not seem to be an exact analytic or closed form expression for this probability as in the case of winning with a single head, so we need to approximate it. In the case of winning with 10 heads, $P_A \approx 0.53278$.

Sources

This section is adapted from: *Heads or Tails*, by Emmanuel Lesigne, Student Mathematical Library Volume 28, American Mathematical Society, Providence, 2005, Sections 1.2 and Chapter 4 [3]. The example is heavily adapted from the weekly "Riddler" column of January 20, 2017 from the website `fivethirtyeight.com`.

Algorithms, Scripts, Simulations

Algorithm

The following Octave code in inefficient in the sense that it generates far more trials than it needs. However, writing the code that captures exactly the number of flips needed on each trial would probably take more lines, so it is easy to be inefficient here.

Scripts

```
p = 0.5;
n = 500;
trials = 2000;

victory = 10;
```

```
headsTails = ( rand(n,trials) <= p );
headsTailsA = headsTails(1:2:n, :);
headsTailsB = headsTails(2:2:n, :);
totalHeadsA = cumsum( headsTailsA);
totalHeadsB = cumsum( headsTailsB);

winsA = zeros(1,trials);

for j = 1:trials
        winsA(1,j) = ( min(find(totalHeadsA(:,j) == victory))
    <= min(find(totalHeadsB(:,j) == victory)) );
    endfor;
empirical = sum(winsA)/trials;

nRange = [9:40];
A = binopdf(9,nRange,1/2) * (1/2);
B = binocdf(9,nRange,1/2);
analytic = dot(A,B);

disp("The empirical probability is:")
disp(empirical)
disp("The approximation to the analytic probabily is:")
disp(analytic)
```

Problems to Work for Understanding

1. Solve the example problem for the cases of winning with $2, 3, 4, \ldots, 9$ heads.

2. Write a simulation to experimentally simulate the coin-flipping game of the example. Experimentally determine the probability of A winning in the cases of winning with $1, 2, 3, \ldots 10$ heads.