

## The Binomial distribution

Consider a two-child family. If we assume that for any one child,  $P(\text{male}) = .5$ , then we get the following result:

Outcome	Prob
-----	----
MM	.25
MF	.25
FM	.25
FF	.25

We are assuming a couple of things here: that the probability of a male child is the same for each child, and that the gender of each child is independent of that of the other child. We're also using that there are exactly two children in the family, and that one of only two genders is possible for each child, so that  $P(\text{female}) = 1 - P(\text{male})$ . Then, for example,  $P(\text{MF}) = P(\text{M})P(\text{F}) = .25$ . Consider a couple of refinements. It's well-known that (without any intervention) the actual proportion of male children born is .51, not .5. Further, a 1998 study in Nottingham, England, found that for children of mothers who did not eat meat or fish, the proportion of male children born was .46. This presumably reflects biochemical changes in the mother because of the diet (environmental factors known to be linked to higher proportions of boys include high levels of magnesium, potassium, and calcium, while smoking is associated with fewer boys being born). Using these values, we get

Outcome	Prob [P(M)=.51]	Prob [P(M)=.46]
-----	-----	-----
MM	.2601	.2116
MF	.2499	.2484
FM	.2499	.2484
FF	.2401	.2916

Consider now the random variable  $X =$  number of male children in a two-child family. The probability function is

X	Prob [P(M)=.5]	Prob [P(M)=.51]	Prob [P(M)=.46]
-	-----	-----	-----
0	.25	.2401	.2916
1	.50	.4998	.4968
2	.25	.2601	.2116

Consider now a three-child family:

Outcome	Prob [P(M)=.5]	Prob [P(M)=.51]	Prob [P(M)=.46]
-----	-----	-----	-----
MMM	.125	.132651	.097336
MMF	.125	.127449	.114264
MFM	.125	.127449	.114264
MFF	.125	.122451	.134136
FMM	.125	.127449	.114264
FMF	.125	.122451	.134136
FFM	.125	.122451	.134136
FFF	.125	.117649	.157464

The random variable  $X =$  number of male children in a three-child family has probability function

X	Prob [P(M)=.5]	Prob [P(M)=.51]	Prob [P(M)=.46]
-	-----	-----	-----
0	.125	.117649	.157464
1	.375	.367353	.402408
2	.375	.382347	.342792
3	.125	.132651	.097336

This type of process is called a *Binomial process*:

1. there is a fixed, finite number of trials,  $n$
2. each trial has two outcomes (success / failure)

3. the probability of success,  $p$ , is the same on all trials
4. the trials are independent

The random variable  $X =$  number of successes in the  $n$  trials is called a Binomial random variable ( $X \sim B(n, p)$ ). It has many applications, such as in gambling games, survival analysis and survey sampling. The probability function can be derived as follows. Consider the event  $X = k$ . That means we are interested in all outcomes with  $k$  successes and  $n - k$  failures. *Each* has probability  $p^k(1 - p)^{n-k}$ . Now we need to count up how many of them there are. This turns out to be the *Binomial coefficient*:

$$\binom{n}{k} = \frac{n!}{k!(n - k)!},$$

where  $n! = n \times (n - 1) \times (n - 2) \times \dots \times 1$  is read “ $n$  factorial,” and  $\binom{n}{k}$  is read “ $n$  choose  $k$ ,” because it represents the number of ways to choose  $k$  objects out of  $n$ , where order doesn’t matter. Thus, if  $X \sim B(n, p)$ ,

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad k = 0, 1, 2, \dots, n.$$

We would then use this formula to calculate any probabilities of interest; for example, the probability of having at least one boy in a three-child family, assuming that the probability of having a male child is .51, is

$$\begin{aligned} P(X \geq 1) &= P(X = 1) + P(X = 2) + P(X = 3) \\ &= .367353 + .382347 + .132651 \\ &= .882351. \end{aligned}$$

Note, by the way, that this could be done more easily by noting that

$$\begin{aligned} P(X \geq 1) &= 1 - P(X < 1) \\ &= 1 - P(X = 0) \\ &= 1 - .117649 \\ &= .882351. \end{aligned}$$

What are the properties of a Binomial random variable? We can guess what the mean is just by intuition. If we had a two-child family, how many boys would we expect there to be, on average, if a boy and a girl are equally likely? Half of them, or 1, of course. If we had a ten-child family? Again, half of them, or 5. That is, for  $X \sim B(n, p)$ ,

$$E(X) = np.$$

Remember what this number means: in the long run, the average number of successes in  $n$  trials is  $np$ . This does **not** mean that this specific number is very likely to occur! If we flipped a fair coin twice, the expected number of heads is 1, and the probability of that event occurring is

$$P(X = 1) = \binom{2}{1} (.5)^1 (.5)^1 = .5;$$

if we flipped it 100 times, the expected number of heads is 50, and the probability of that event occurring is

$$P(X = 50) = \binom{100}{50} (.5)^{50} (.5)^{50} \approx .08.$$

The reason for this is that with 100 flips, the probability of any one particular number of heads is smaller. Still, the number 50 *is* a “typical value,” in the sense that if lots of people flipped a fair coin 100 times, the average of all of their numbers of heads would be right around 50.

We don’t have any intuition to appeal to regarding the variance of the Binomial random variable, so I’ll just state it: for  $X \sim B(n, p)$ ,

$$V(X) = np(1 - p).$$

So, for the two-child family with  $p = .5$ ,  $V(X) = (2)(.5)(.5) = .5$ , while for the three-child family with  $p = .51$ ,  $V(X) = (3)(.51)(.49) = .7497$ .

The mean and variance of a Binomial random variable can be easily derived algebraically, since the binomial random variable  $X$  is just the sum of individual random variables  $X_i$ , where  $X_i$  refers to the  $i$ th trial, and equals 1 if the outcome of the trial is a success and 0 otherwise. It is easy to show that  $E(X_i) = p$  and  $V(X_i) = p(1 - p)$ ; thus,

$$E\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n E(X_i) = \sum_{i=1}^n p = np,$$

and since the  $X_i$  are independent of each other,

$$V\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n V(X_i) = \sum_{i=1}^n p(1 - p) = np(1 - p).$$

Note that these properties of the binomial help debunk the idea of the “law of averages.” Say you flip a fair coin 1000 times, and get 450 heads. What is the probability of heads on the next flip? Of course, it is .5, by independence. But what then of the “law of averages,” which says that you should end up with about half heads and half tails, so you

somehow have to “make up” the extra tails with extra heads? Let  $X$  be the number of heads in  $n$  flips of a fair coin, and let  $\bar{p} = X/n$  be the observed proportion of heads. We know that  $E(X) = (n)(.5) = n/2$ , and that  $V(X) = (n)(.5)(.5) = n/4$ , or  $SD(X) = \sqrt{n}/2$ . Further, we know that

$$E(\bar{p}) = E(X/n) = E(X)/n = .5$$

and

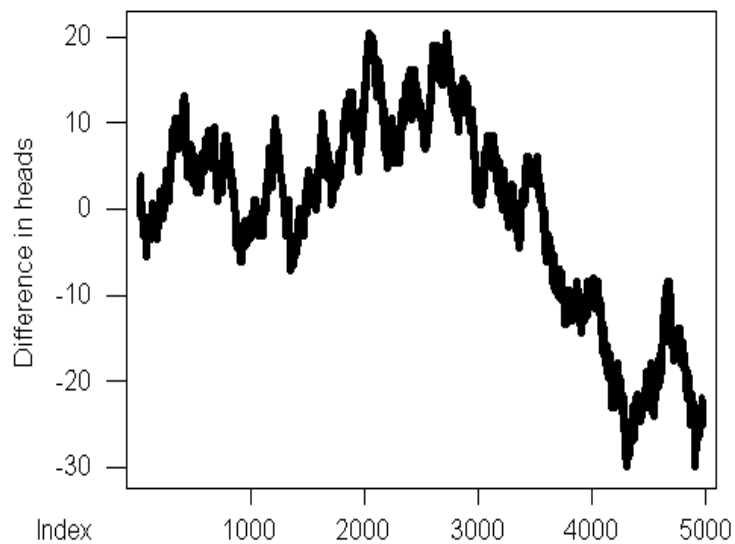
$$V(\bar{p}) = V(X/n) = V(X)/n^2 = 1/(4n),$$

or  $SD(\bar{p}) = 1/(2\sqrt{n})$ .

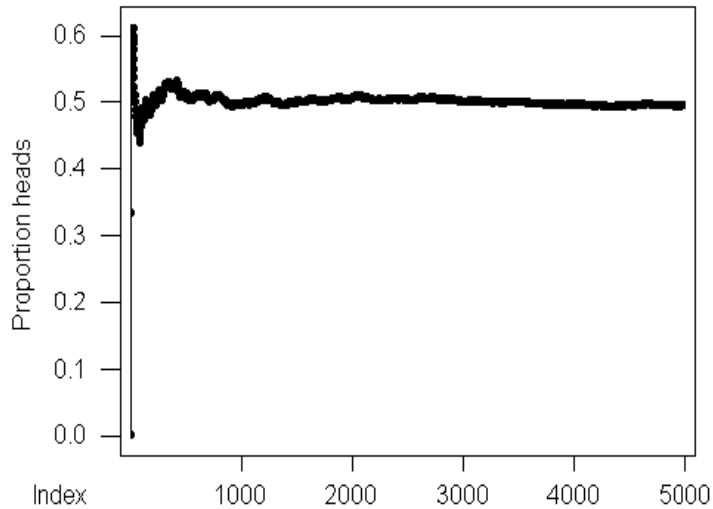
Now, what about those 1000 flips of the coin? It is certainly true that the expected number of heads is half, or 500, and the standard deviation of the number of heads is  $\sqrt{1000}/2 = 15.8$ . What if we flipped the coin 2000 times? The expected number of heads would be half, or 1000, but the standard deviation would be  $\sqrt{2000}/2 = 22.4$ . That is, *as the number of flips gets larger, the variability around the expected number of heads gets larger, not smaller*. There is no law of averages “pushing” the number of coins back towards half of the total, since the variability of the number of heads **increases** with more flips, not decreases.

All is not lost, however. What is the expected proportion of heads in 1000 flips of the coin? It’s  $E(\bar{p})$ , or .5. What is the standard deviation of that proportion? It’s  $SD(\bar{p})$ , or .0158. What if there are 2000 flips? The expected observed proportion of heads is still .5, but the standard deviation of the observed proportion is .0112. That is, the observed proportion of heads gets closer and closer to the expected proportion as the number of binomial trials gets larger, in the sense that the observed proportion is centered around the expected proportion with progressively smaller standard deviation. This is not the “law of averages,” but rather the *Law of Large Numbers*, and virtually any statistic satisfies a version of it: as the size of a random sample increases, the observed value of a statistic that can be written as an average of sample values gets closer and closer to the common expected value of each sample value in this probabilistic sense. If you’re interested, you can find proofs of two versions of the Law of Large Numbers at Wikipedia ([http://en.wikipedia.org/wiki/Law\\_of\\_large\\_numbers](http://en.wikipedia.org/wiki/Law_of_large_numbers)).

Here are two pictures that show what’s going on. Both of these are based on 5000 flips of a hypothetical coin (I used the pseudorandom number generator in **Minitab** to generate the coin flips). The first plot is a plot of the difference between the observed number of heads in  $n$  flips and the expected number of heads in  $n$  flips ( $n/2$ ), as the number of flips went from 1 to 5000:



As you can see, the difference between the observed and expected number of heads is “around” zero, but it is not getting closer to zero as the number of flips increases. On the other hand, this is a plot of the observed proportion of heads as the number of flips increases:



It is obvious that the observed proportion eventually settles down to its expected value (.5), and pretty much stays there.

What does this say about flipping a fair coin where the first 1000 flips have only 450 heads? It **doesn't** mean that the observed proportion of heads after 1000 flips must be a little above .5 to “make up” for the original 1000 flips. Say the first 1000 flips resulted in only 450 heads; the observed proportion of heads is .45. Now, say the next 1000 flips resulted in 500 heads; now the observed proportion of heads is .483. Say the next 10,000 flips resulted in 5000 heads; now the observed proportion of heads is .496. Even though the flips started out short of the expected number of heads, and even though the proportion of heads after the first 1000 flips is not above .5, the overall proportion of heads in all of the flips approaches the correct number, .5.