The Number of Heads in Coin Tosses

YAP Von Bing (stayapvb@nus.edu.sg)
Department of Statistics and Applied Probability
National University of Singapore
November 17, 2015

1 Introduction

This is a lesson plan for thinking about coin tossing and more generally finite discrete random variables. It begins with a series of elementary questions that lean on students' intuition and prime them towards definitions which are then presented in a more didactic fashion. Interpretation of expectation and standard deviation (SD) in terms of data from repeated experiments is an important component in the narrative. Rudimentary knowledge of summary statistics and probability topics in the Singapore secondary school mathematics curriculum is assumed. The material evolved from a talk given at Mathematics Seminar 2015 organised by the Ministry of Education for about 250 upper secondary school students.

The Dialogue is a fairly accurate reproduction of the first part of the talk. The teacher's questions can be used directly to engage students. The stylised responses can be improved in light of actual responses. The Lecture links up a set of ideas logically. Its first subsection calculates the average and SD of data from coin-tossing experiments with one and two tosses, which motivates the general formulae for the expectation and SD of the number of heads. At the end, the term "expectation" supplants "expected number of heads". The second subsection carries the narrative onto random variables and associated concepts like distributions, expectation and SD. Like coin tosses, the expectation and SD are interpreted in terms of repeated observations of the random variable. The third subsection touches on the sum of independent random variables, and states the addition rules which justify the formulae in the first subsection. The algebra in the Lecture can be quite dense, so the teacher may need to elaborate on some materials, perhaps substantially.

2 Dialogue

Teacher: Toss a coin 30 times. Roughly how many heads will you get?
Pupil: 15, because each toss produces a head or a tail.
Teacher: Suppose the coin has a chance 0.5 of producing a head.
Roughly how many heads will you get in 30 tosses?
Pupil: 15. This is the meaning of 0.5: about half of the tosses will produce heads.
Teacher: Would you revise your response to the first question?
Pupil: Yes, I will ask "What is the chance of head?", since this will affect the answer.
For example, if it is 0.6, then I will get about $40 \times 0.6 = 18$ heads.
Teacher: My coin has been run over by a car, causing the head to be less likely.
Roughly how many heads will you get in 30 tosses, if the chance of head is 0.45?
Pupil: $30 \times 0.45 = 13.5$, so I will say 13 or 14.
Teacher: Say we use 13. Do the same if the chance of head is 13/30.
Pupil: Then it will be $30 \times 13/30 = 13$, but $13/30 \approx 0.43 < 0.45$.
Perhaps the previous answer should be larger than 13.
Teacher: 14?
Pupil: But that is the answer for a coin with chance of head $14/30 \approx 0.47$.
Teacher: Are you forced to say 13.5?
Pupil: Yes. But it doesn't make sense to get 13.5 heads!
Teacher: Fair enough. Let us make a definition.
13.5 is the **expected number of heads** when a coin with chance 0.45 is tossed 30 times.
Pupil: All right, I will go along with you.
Teacher: To interpret 13.5, let us generate many outcomes from the experiment.
We will get integers between 0 and 30 inclusive, appearing in no particular order.
But if the number of repetitions is large, their average will be around 13.5.
Pupil: Really? Can you show some evidence?
Teacher: It is very time-consuming to toss a coin so many times,
but a computer can do a very good imitation quickly.
I use a computer to generate 1,000 outcomes, and the first 5 are 10, 12, 11, 15, 18.
The average of my numbers is 13.4, which is very close to the prediction.
Pupil: I see. What if we accumulate more outcomes, say 10,000?
Will their average be closer to 13.5?

Teacher: That is very likely.
To consolidate our current understanding, let p be a number between 0 and 1.
What is the expected number of heads from n tosses of a coin with chance of head p?
Pupil: That is $n \times p = np$.
Teacher: How can this quantity be interpreted?
Pupil: Get a list of outcomes from the experiment, by repeating it.
Their average will be roughly np.
The larger number of outcomes, the closer the average will be to np.
Teacher: Excellent.

3 Lecture

3.1 Deriving the expected number of heads

In the dialogue, the expected number of heads np is gently hoisted on intuitive notions about coin tossing, and it extends them to other cases where intuition does not work so well. The interpretation involves the consideration of a large number of outcomes, which is a very effective approach to get acquainted with a random variable.

By a p-coin, we mean a coin with chance of head equal to p. For example, a 0.5-coin is a fair coin. Suppose n is so large that $np > 1$. Here is an intuitive justification for the interpretation of np. Because in n tosses we should get more or less np heads, it seems reasonable that np should be the average number of heads from n tosses, if the experiment is repeated many times. This is a rough argument, since np need not be an integer. Now we demonstrate the cases $n = 1, 2$ completely.

For the first case, each experiment produces 0 head (a tail) or 1 head. Repeating the experiment r times gives around $r(1-p)$ 0's and rp 1's. The average and variance are roughly

$$\begin{aligned} \frac{r(1-p) \times 0 + rp \times 1}{r} &= p \\ \frac{r(1-p) \times (0-p)^2 + rp \times (1-p)^2}{r} &= p(1-p) \end{aligned}$$

For the second case, each experiment produces 0, 1 or 2 heads. 0 occurs when both tosses turn up tails, which happens with chance $(1-p)^2$; here we are implicitly assuming that the tosses are independent. Similarly, the chance of 2 heads is p^2, hence the chance of 1 head is $1-(1-p)^2-p^2 = 2p(1-p)$. Therefore r repetitions give around $r(1-p)^2$ 0's, $r2p(1-p)$ 1's and rp^2 2's, with average and variance roughly

$$\begin{aligned} \frac{r(1-p)^2 \times 0 + r2p(1-p) \times 1 + rp^2 \times 2}{r} &= 2p \\ \frac{r(1-p)^2 \times (0-2p)^2 + r2p(1-p) \times (1-2p)^2 + rp^2 \times (2-2p)^2}{r} &= 2p(1-p) \end{aligned}$$

Thus, the formula np checks out for $n = 1, 2$, and there seems to be a formula for the variance as well. A similar treatment for larger n gets progressively more difficult: the number of terms in the sums is $n + 1$, and it is necessary to know the chances of all possibilities. Even for a fair coin, the number of heads in two tosses does not have equally likely outcomes. Techniques for overcoming the difficulty will be introduced later, to justify

Rule 1: *Toss a p-coin n times independently and under the same conditions; record the number of heads observed. Repeat the experiment to obtain a large number of outcomes. The average will be around np, and the SD will be around $\sqrt{np(1-p)}$.*

A more concise statement of Rule 1 is

Toss a p-coin n times independently and under the same conditions. The number of heads will be around np, give or take $\sqrt{np(1-p)}$ or so.

Rule 1 holds for real coin tosses, provided the conditions are approximately satisfied. Independence, for instance, means tosses do not influence each other. We illustrate Rule 1 with two examples. The number of heads in 100 tosses of a fair coin will be around $100 \times 0.5 = 50$, give or take $\sqrt{100 \times 0.5 \times (1 - 0.5)} = 5$ or so. The mathematician Kerrich performed an extensive coin-tossing experiment ([K], [FPP]). Part of his data consisted of 100 numbers of heads from tossing a coin 100 times: a total of 10,000 tosses! These outcomes have average 50.7 and SD 5.6, which provide empirical support for Rule 1. It is quite remarkable that the outcomes, all integers between 0 and 100, have such a small SD. Next, consider the 1,000 competer-simulated outcomes from tossing a 0.45-coin 30 times. We saw in the Dialogue that their average was 13.4, quite close to $30 \times 0.45 = 13.5$. Their SD was 2.6, also quite close to $\sqrt{30 \times 0.45 \times 0.55} \approx 2.7$. The computer imitates coin tosses satisfactorily.

The number of heads in n tosses of a p-coin is an example of a **random variable**. The next two subsections will explain why np and $\sqrt{np(1-p)}$ are respectively its **expectation** and **standard deviation (SD)**.

3.2 Random variables

By a **random variable**, we mean a mechanism that generates random numbers. We now describe a **finite discrete** random variable, i.e., it generates only a finite number of values with positive probability. Let X take k distinct values $x_1, \ldots, x_k$, with chances $p_1, \ldots, p_k$ summing to 1; we write $\Pr(X = x_i) = p_i$, $1 \le i \le k$. The vectors x and p are called the **distribution** of X. Like the number of heads in one toss last section, we want to calculate the average and SD of a large number, r, of outcomes from repeated observations of X. For each i, about rp_i outcomes are x_i, so the average is roughly

$$\frac{rp_1 \times x_1 + \cdots + rp_k \times x_k}{r} = \sum_{i=1}^{k} p_i x_i$$

We define the **expectation** of X as $\mathrm{E}X = \sum_{i=1}^{k} p_i x_i$. The variance of the outcomes is roughly

$$\frac{rp_1 \times (x_1 - \mathrm{E}X)^2 + \cdots + rp_k \times (x_k - \mathrm{E}X)^2}{r} = \sum_{i=1}^{k} p_i (x_i - \mathrm{E}X)^2$$

We define the **variance** of X as $\mathrm{var}X = \sum_{i=1}^{k} p_i (x_i - \mathrm{E}X)^2$. The **standard deviation (SD)** of X is $\sqrt{\mathrm{var}X}$, denoted by $\mathrm{SD}(X)$. We have

Rule 2: *X will be around EX, give or take $SD(X)$ or so.*

Rule 2 generalises Rule 1, and has an analogous meaning. *Collect a long series of outcomes from the random variable X. Then the average and SD of the data will be around EX and $SD(X)$.* It is tacitly assumed that repeated observations of X are obtained independently and under the same conditions. Here are some examples.

1. $k = 2, x_1 = 0, x_2 = 1, p_1 = 1 - p, p_2 = p$. X represents the number of heads in one toss of a p-coin. Then $\mathrm{E}X = p$ and $\mathrm{SD}(X) = \sqrt{p(1-p)}$, as seen before. In particular, the number of heads in one toss of a fair coin will be around 0.5, give or take 0.5 or so.

2. $k = 3, x_1 = 0, x_2 = 1, x_3 = 2, p_1 = (1-p)^2, p_2 = 2p(1-p), p_3 = p^2$. X represents the number of heads in two tosses of a p-coin, so $\mathrm{E}X = 2p$ and $\mathrm{SD}(X) = \sqrt{2p(1-p)}$, also seen before. The number of heads in two tosses of a fair coin will be around 1, give or take 0.7 or so. Unlike the previous example, if $p = 0.5$, X does not have a uniform distribution.

3. $k = 6, x_i = i, p_i = 1/6, 1 \le i \le 6$. X represents the number of spots from rolling a fair die, with $\mathrm{E}X = 3.5$ and $\mathrm{SD}(X) \approx 1.7$. X will be around 3.5, give or take 1.7 or so. If you roll a fair die many times, the average and SD of the outcomes will be around 3.5 and 1.7.

4. Let X represent the number of heads in 100 tosses of a fair coin. Then the possible values are $0, 1, \ldots, 100$, but the chances are not so simple to know. By Rule 1, $\mathrm{E}X = 50$ and $\mathrm{SD}(X) = 5$. The next section presents a derivation of these numbers, without having to know the chances, i.e. the distribution of X.

3.3 Sum of random variables

Let X and Y be finite discrete random variables. They are **independent** if for any possible values x of X and y of Y, $\Pr(X = x, Y = y) = \Pr(X = x)\Pr(Y = y)$. We usually assume that two tosses of a p-coin are independent. For example, the chance of two heads equals $p \times p = p^2$, etc. In many board games, the number of moves is determined by the total number of spots from rolling two dice. We assume the dice are independent, so that the chance of getting twelve spots is $1/36 = 1/6 \times 1/6$.

An example of dependent random variables is as follows. A box has two physically identical tickets, each bearing two numbers, on the left and on the right. One ticket has 0

in both positions, while another has 1 in both positions. Let X and Y be the left and right number in one random draw. Then $\Pr(X = 0, Y = 0) = \Pr(X = 0) = \Pr(Y = 0) = 0.5$, so X and Y are not independent.

Random variables can be added together. Generate x from X and y from Y. Then $x+y$ can be viewed as coming from a new random variable, denoted by $X + Y$. An elementary proof of the following can be found in [R].

Theorem. *For finite discrete random variables X and Y,*

$$E(X + Y) = EX + EY$$

If X and Y are independent, then

$$var(X + Y) = var\,X + var\,Y$$

We illustrate the Theorem with some examples.

5. Let H_2 be the number of heads from 2 tosses of a p-coin. Since H_2 is the sum of two independent random variables, each being the number of heads in one toss, $\mathrm{E}H_2 = 2p$ and $\mathrm{var}H_2 = 2p(1-p)$, so $\mathrm{SD}(H_2) = \sqrt{2p(1-p)}$. Compare with Example 2.

6. Roll two fair dice. The expectation and SD of the total number of spots are $2 \times 3.5 = 7$ and $\sqrt{2} \times 1.7 \approx 2.4$, using results from Example 3.

7. Let H_n be the number of heads in n tosses of a p-coin. An induction argument shows that $\mathrm{E}H_n = np$ and $\mathrm{SD}(H_n) = \sqrt{np(1-p)}$. This completes the derivation of the formulae in Rule 1.

8. For the dependent X and Y described before the Theorem, $X + Y$ has equal chance of being 0 or 2, so $\mathrm{var}(X + Y) = 1$. But $\mathrm{var}X + \mathrm{var}\,Y = 0.25 + 0.25 = 0.5$, illustrating a need of some restriction for the variance formula to hold.

Reference

[FPP] Freedman, Pisani and Purves, *Statistics*. Norton (2007).
[K] Kerrich, *An Experimental Introduction to the Theory of Probability*. Munkskaard (1946).
[R] Ross, *A First Course in Probability*. Pearson (2012).