Probability, Proportions, and the Binomial Distribution

Probability

Let's review a few basic concepts of probability in order to ultimately better understand random binary variables and their distributions. A probability is the chance or likelihood that an event (or set of events) will occur (a single binary trial is called a *Bernoulli* trial). In mathematics and statistics, it is presented as a proportion out of 1.0, with a value of 0.0 meaning that there is absolutely no chance the event will occur and a value of 1.0 meaning there is a certainty the event will occur. In common use, probabilities are often presented as percents, with 0% and 100% representing no chance and certainty, respectively. Formally, an *elementary event* (sometimes a "sample point") is the simplest concept of an event, and the term *event* (or "event class") refers to a set of elementary events. The distinction is little hard to track at first, but consider drawing a single card from deck of cards. Drawing a queen of hearts from the deck is an elementary event in this context. Drawing a heart (any of 13 cards in the suit) is an event (event class), because it involves a single trial but a class of possible elementary events. The *sample space* is the set of all possible distinct outcomes. In the card example, the sample space is the deck of 52 cards, or 52 possible outcomes. A single opportunity for an elementary event or an event to occur is called a *trial*. Eventually (soon, I promise), we will be interested in many trials, which corresponds to multiple cases (e.g., participants) in a study.

The probability then is the chance that an event or event class will occur on a single trial. The probability has to be in relationship to all of the possible events that could occur on that trial, which is the sample space. In the card example, the probability, denoted $P(A)$, of drawing a queen of hearts from a (fair, shuffled, full) deck of cards will be 1/52 = .019, and the probability of drawing a heart of any value is 13/52 = .25. We actually got to the probability of event class of drawing any heart by intuitively combining all of the probabilities of the elementary events of any specific heart, such as the queen of hearts, *or* the jack of hearts, *or* the 2 of hearts, . . ., each with the probability of 1/52. This is a union of 13 elementary events, denoted with a $\cup$ symbol for union.

$$P(A \cup B \cup C \cup ... \cup M) = \frac{1}{52} + \frac{1}{52} + \frac{1}{52} \ldots \frac{1}{52} = \frac{13}{52} = .25$$

Thus, whenever we combine the probabilities of elementary (event classes) into an event class (or larger set of event classes), the "or" implies addition.

Alternatively, we may be interested in a joint probability, which involves the probability that one elementary event *and* another elementary event happen together. Two events would have to involve two consecutive trials (or perhaps an experiment conceived of as a single trial with two simultaneous events). If we draw one card each in two consecutive trials, we are interested in a joint event of one elementary event and another elementary event both occurring. For two consecutive draws from a deck, assuming we put the first card back in and reshuffle, the probability of drawing two aces, for example, would be:

$$P(A \cap B) = \frac{4}{52} \cdot \frac{4}{52} = \frac{16}{2,704} = .006$$

The symbol $\cap$ is for intersection, which corresponds to the joint events. Incidentally, the joint probability of drawing an ace and a 2 is the same, but it seems less unique for some reason. If we do not replace the first card, things get more complicated. The denominator must change on the second probability, because there will only be 51 cards remaining. And the numerator will depend on whether we drew an ace in the first trial. So, assuming we did, we would have a slightly lower probability of drawing two aces, 4/52 + 3/51 = .005. Now, imagine how complicated it is to count cards and estimate probabilities in your head during a real game with say four players when there are few cards available with each player's turn, you do not know which cards are being retained by other players, and the other players are retaining pairs or same suits (i.e., the cards they keep on each turn will be non-independent events).

One other definition, the probability that an event does not occur is usually denoted *P*(~*A*). This could be "not a queen of hearts" or "not heads" when flipping a coin. The probability rules above rely on the assumption that the possible events are mutually exclusive and mutually exhaustive, so that P(A) + *P*(~*A*) = 1.0, for a single trial. There is no way for one possible outcome (e.g., heads) to happen on the same trial if the other possible outcome (tails) happens instead, and there can be no other possible outcomes besides these two outcomes (heads and tails). This is known as the *partition rule*. If events are not independent, it will not make sense to simply add *P*(*A*) and *P*(*B*) any longer to determine *P*(*A* or *B*).

Proportions

Usually, we are interested in the expected probability of binary outcome in the long run, over many trials. We will call the number of trials *n* (e.g., five coin tosses) and the desired outcome (a "success) *k* (e.g., two heads). We can restate some of the above discussion by saying that a probability that a particular event class *k* (e.g., two heads out of five trials) occurs out of all of the possible ways any event class might occur (0 heads, 1 head, 2 heads, 3 heads, 4 heads, 5 heads). Two heads out of five tosses can happen in several different combinations—on the first two tosses, the first and third tosses, the first and fourth toss, and so on. The binomial coefficient (some long forgotten set of synapses just fired, didn't it?) is a convenient way to find the number of ways some *k* number of successes (some number of heads) can happen out of some *n* number of trials (some number of coin flips).

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

The exclamation symbol, !, is the factorial symbol, which is shorthand for multiplying a number by one less than the prior number, multiplied by one less than that number, and so on. $5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$. Plugging in our numbers, we find that the number of ways you can get two heads out of 5 tosses is 10.

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{5!}{2!(5-2)!} = \frac{120}{2(6)} = 10$$

This is just the *number* of possible ways two heads might occur, it does not give us the *probability* of two heads occurring. To find that, we need to take into account the probability that a single coin toss will turn up heads, which we will call π, as well as the probability that a single coin toss will be tails, which is $1 - \pi$ = .5. The probability that just one combination of *k* will happen is then $\pi^k = .5^2 = .25$ and the probability that the particular combination of k does not happen is $(1-\pi)^{n-k}$. Putting this together with all the possible combinations of *k*, we get the following equation, which will give us the probability of getting two heads out of five coin tosses.

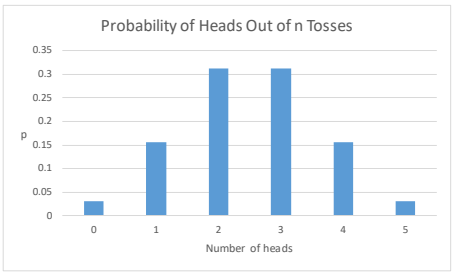
$$\binom{n}{k}\pi^k(1-\pi)^{n-k} = \frac{n!}{k!(n-k)!}\pi^k(1-\pi)^{n-k} = \frac{5!}{2!(5-2)!}(.5)^2(1-.5)^{5-2} = (10)(.25)(.125) = .3125$$

Binomial Distribution

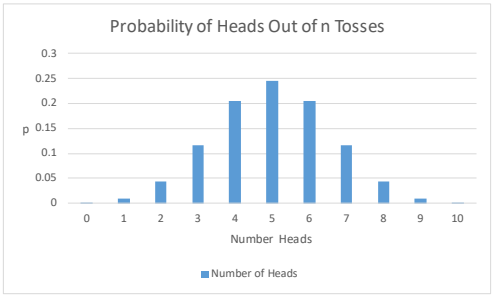
Now, back to the exciting stuff. This is all for the purpose of explaining the binomial probability of a certain sample value for some observed binary response. A common example would be any survey question with a dichotomous response format, such as a political poll in which the voter must choose between two candidates. The number of cases in the sample is the number of cases surveyed. In our result, we will observe just one proportion, but we want to know how unusual that is likely to be by comparing it to all the other possible outcomes we might get with the same sample size. Of course, a frequency plot of all of the possible outcomes is a distribution, and, for a binary outcome, the most common distribution is the binomial distribution based on the binomial probability. In general form, we can state that the probability of a particular outcome $Y = k$ is

$$P(Y=k;n,\pi)=\binom{n}{k}\pi^{k}(1-\pi)^{n-k}$$

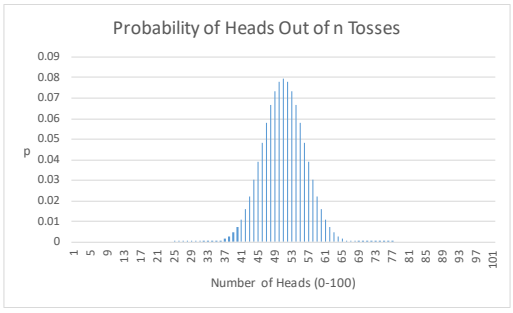
Which has a binomial distribution with two parameters n and π. If we create a frequency histogram using all possible outcome frequencies of *k* for some chosen sample size n, we obtain a figure for the density of the binomial distribution. Using the above equation, I used as spreadsheet to generate a frequency histogram for all possible outcomes (values of k), for π = .5 and n = 5.



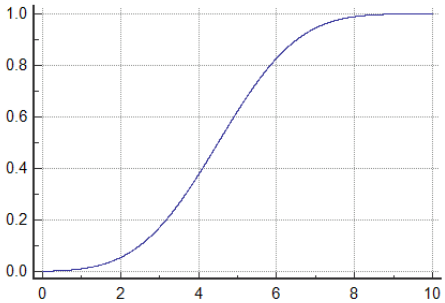
For, n = 10, it looks like this.



For n = 100, it looks like this.



Each of these distributions are referred to as a probability density function, or pdf. For statistical tests, we will often refer to the cumulative version of the probability function, called the cumulative density function, cdf. The cdf shows the proportion of values in the distribution at or a below each value.



The binomial distribution is based upon the known (or assumed) probability of the occurrence of an elemental or event class. For example, a toss of a coin has a probability of .5, and the distributions

shown above use π = .5, but other values of π could be used. Also note that, although the figures I used above have proportions on the y-axis, formally the distribution is given in terms of the number of trials (or observations), n, and thus its mean of the binomial distribution is $\mu = n\pi$ and the variance is $\sigma^2 = n\pi(1\text{-}\pi)$.

The binomial distribution is the most basic theoretical distribution, and there are several related distributions. The Bernoulli distribution is a binomial distribution for only one trial (i.e., n = 1). The Bernoulli distribution is most often discussed in connection with fundamental theoretical probabilities, such as the basis of the binomial distribution. The hypergeometric distribution is a binomial distribution in which each trial (or case) is sampled without replacement, which is used for inference with small, finite populations and small sample corrections like the Fisher's exact test. The multinomial distribution is a more general extension of the binomial distribution that applies to variables with more than two categories. The Poisson distribution, which is asymptotically related to the multinomial distribution, is used in regression modeling of count variables, estimation for small proportions, and some loglinear models.

Binomial Sample Descriptive Statistics

Now that we have discussed the probability distribution that will be integral to making inferences about the population, we should discussion some issues related to sample values. Presuming that we have coded the observed variable y_i as 0 or 1, we can compute the sample mean of the variable the same way that we compute the sample mean for a continuous variable. As long as we have used the 0,1 coding,[1] the mean is the same as the proportion of cases for which y = 1,

$$\bar{Y} = \sum Y_i / n = p\,.$$

In probability terms, the mean or proportion is also is the probability that the event occurs, $P(y = 1) = p$, whether it is a coin toss of heads or a survey response of "yes." The expected value of the binomial distribution would then be written as $E(Y) = np$. Because the binomial distribution is given in frequencies for a given sample size, we multiply by n.

Similarly, the variance and standard deviation can be computed in the same way as they are with continuous variables, where it can be shown that the variance, $\sum\left(Y_i - \bar{Y}\right)^2 / n$, is equal to $p(1\text{-}p)$. I will spare you the proof. The standard deviation is the square root of this number as usual, $\sqrt{p\left(1-p\right)}$. Notice that we do not use n - 1, the degrees of freedom in the binary case typically, and that is because the assumption of large samples and use of maximum likelihood that is conventional for most binomial-related statistics. Also notice that once we know the mean, we also know the variance. As with the mean, the variance and standard deviation of the binomial distribution are expressed in terms of a product of n, $\sigma^2 = np(1\text{-}p)$ and $\sigma = \sqrt{np\left(1-p\right)}$, respectively. Skewness and kurtosis have simplified, equivalent equations as well.

$$\text{skewness} = \frac{E\left(Y_i - \mu\right)^3}{\sigma^3} = \frac{\left(1-2p\right)}{\sqrt{np\left(1-p\right)}}$$

$$\text{kurtosis} = \frac{E\left[\left(Y_i - \mu\right)^4\right]}{\left(E\left[Y_i - \mu\right]^2\right)^2} = 3 - \frac{6}{n} + \frac{1}{np\left(1-p\right)}$$

[1] For some data sets, binary variables are coded 1,2. For this coding scheme, the mean will no longer be the proportion, but rather equal to 1 + p instead. The variance as usually computed is not changed, we cannot use the simplified formula unless the proportion (rather than the mean) is used.