TUTORIAL ON LARGE DEVIATIONS FOR THE BINOMIAL DISTRIBUTION

R. Arratia* and L. Gordon†
Department of Mathematics,
University of Southern California,
Los Angeles, CA 90089, U.S.A.

We present, in an easy to use form, the large deviation theory of the binomial distribution: how to approximate the probability of \( k \) or more successes in \( n \) independent trials, each with success probability \( p \), when the specified fraction of successes, \( a = k/n \), satisfies \( 0 < p < a < 1 \).

1. Introduction. Large deviation estimates for the binomial distribution are quite useful for assessing attained statistical significance, or \( p \)-values, for many situations in macromolecular sequence analysis. The binomial distribution governs the total number of outcomes of one type ("success") in a group of independent random trials where there is the same probability of success at each trial. An observation in which some property occurs somewhere more often than it would on average might be just a chance fluctuation, so the analyst needs to estimate the very small probability of such "large deviations" due to chance alone.

Large deviation theory, rather than the central limit theorem, is relevant because we are interested in extreme events, in which the observed average is not close to its expected value. For observations which are several standard deviations away from the mean, the normal approximation to the binomial may be very poor, while the large deviation approximation is good. We present a numerical example of this in Section 2.

In biology, the property labeled as "success" might be hydrophobicity or positive charge for amino acids in a protein, or the property might be "A" or purine in a DNA sequence. In sequence alignment of unrelated sequences, without gaps inserted to increase matching, the number of matches can be modeled by a binomial random variable; see Arratia et al. (1989) for situations which use large deviation estimates in this context. An analysis of the statistical significance of clusters of mildly similar proteins, using large deviation estimates, appears in Arratia and Lander (1988). Other more complex situations

* Supported by NIH grant GM 36230 and NSF grant DMS 8601986.
† Supported by NIH grant GM 36230 and a grant from the System Development Foundation.
arise, as in Waterman et al. (1984) where a binomial model is used in consensus sequence alignment.

The binomial distribution is the distribution of the total number of successes $S_n$ in a collection of $n$ independent trials each with the same individual success probability $p$, for $j = 0, 1, 2, \ldots, n-1, n$:

$$P(\text{binomial}(n, p) = j) = P(S_n = j) = \frac{n!}{j!(n-j)!} p^j (1-p)^{n-j},$$

where the factorial symbol is the product $n! = n(n-1)(n-2)\ldots(3)(2)(1)$. (The symbol $\equiv$ means "defined to equal".) To assess the significance of an observation with an unusually large number $k$ of successes, we need to estimate the chance of $k$ or more successes. The probability of $k$ or more successes is given by summing the above for $j = k, k+1, k+2, \ldots, n-1, n$. This sum could be approximated by brute force with a computer, but a more elegant and insightful approach is possible with large deviation theory.

Large deviation theory in its general form is one of the most active frontiers of research in probability. This theory is usually presented in a general version which may be hard for the non-specialist, see for example Ellis (1985), Varadhan (1984), or Bahadur (1971). The large deviation theory is much simpler when specialized to the binomial distribution. This important and simple case is not presented in Feller, Vol. 1 (1968), which is still the best probability text for scientists. In this paper, which is expository and self-contained, we present only the large deviation theory of the binomial distribution: how to approximate the probability of $k$ or more successes in $n$ independent trials, each with success probability $p$, when the specified fraction of successes, $a = k/n$, satisfies $0 < p < a < 1$.

Here is a survey of the results we present. For tests of significance, the inequality in Theorem 1:

$$P(S_n \geq an) \leq e^{-nH},$$

is often the most useful; it is an upper bound which is valid even before taking the limit $n \to \infty$. The exponential decay rate in this upper bound, $H$, is sharp, as shown by the first statement in Theorem 2. The second statement in Theorem 2:

$$P(S_n \geq an) \sim \frac{1}{1-r} \left( \frac{1}{\sqrt{2\pi a(1-a)n}} \right) e^{-nH},$$

gives two refinements to the estimate in Theorem 1: the factor in brackets which corresponds to the central limit theorem for $n$ independent trials with success probability $a$, in which the variance is $a(1-a)n$, and a factor $1/(1-r)$,
which makes the ratio of the two sides tend to 1 as the number \( n \) of trials increases to infinity. The third statement in Theorem 2:

\[
P(S_n = an + i) \sim \frac{1}{\sqrt{2\pi a(1-a)n}} r^i e^{-nh}, \text{ for } i = 0, 1, 2, \ldots
\]

can be viewed in two ways. First, it says that for a large number of trials, when asking for the fraction of successes to be a large value \( a \) instead of its typical value \( p \), the probability of a particular large number \( k \) of successes goes down by a factor of \( r \) each time \( k \) is increased by one. Second, it explains the correction factor \( (1/r) \): the relative contributions to \( P(S_n \geq k) \) made by each of the possibilities \( k, k+1, k+2, \ldots \) for the actual number of successes; these contributions are proportional to \( 1, r, r^2, r^3, \ldots \), and the sum of this geometric series is \( (1/r) \).

2. A Numerical Example. What is the probability \( p^* \) of 16 or more successes in 20 independent trials, when the probability of success at each trial is 0.25? Formally, with \( n = 20, \ p = 0.25, \ a = 0.8 \), what is \( P(S_n \geq an) \)? We express all small probabilities as multiples of one millionth.

Exact calculation with the binomial density gives: \( p^* = P(\text{binomial}(20, 0.25) \geq 16) = 0.3865 \times 10^{-6} \). The main contribution here is \( P(\text{binomial}(20, 0.25) = 16) = 0.3569 \times 10^{-6} \).

Approximation with the normal distribution would be done as follows. The variance in \( n \) tosses of a \( p \)-coin is \( \sigma^2 = np(1-p) \), which in this example is \( \sigma^2 = 20(0.25)(0.75) = 3.75 \). The mean is \( np \), which in this example is 5. The number of standard deviations above the mean represented by \( k \) is \( (k - np)/\sigma \), which in this example is \( (16 - 5)/\sqrt{3.75} = 5.68 \). The chance that a standard normal exceeds this is 0.0069 \times 10^{-6}. Thus the normal approximation is way too small, by a factor of 56. If we were more careful and included the continuity correction, we would obtain the approximation \( 1 - \Phi((15.5 - 5)/\sqrt{3.75}) = 0.03038 \times 10^{-6} \), which is still too small by a factor of 12.7.

Approximation with the large deviation theory below would be done as follows. The relative entropy is \( H = a \log(a/p) + (1-a) \log((1-a)/(1-p)) \), which in this example is \( 0.8 \log(3.2) + 0.2 \log(4/15) = 0.93 - 0.26 = 0.666 \). The upper bound from Theorem 1 is \( e^{-nh} \) which in this example is \( 1.64 \times 10^{-6} \), too big by a factor of 4.24. The central limit factor is \( \sqrt{2\pi na(1-a)} = \sqrt{2\pi(20)(0.8)(0.2)} = \sqrt{20.11} = 4.48 \). The combination of these two factors is the approximation given by formula (6) for \( P(\text{binomial}(20, 0.25) = 16) \), in this case \( 1/(2\pi na(1-a))^{1/2} e^{-nh} = 1.64 \times 10^{-6}/4.48 = 0.366 \times 10^{-6} \). The odds ratio in this example is \( 3/1 = 1/12 \), so the correction factor corresponding to "\( k \) or more heads" instead of "exactly \( k \) heads" is \( 1/(1-r) = 12/11 \), which is not far
from 1. The approximation specified by equation (5) is $0.398 \times 10^{-6}$, which is only 3% above the exact value.

Typically, an observation of data involves a large number of groups of trials, so the overall significance of finding at least one group with a high proportion of successes must be assessed by multiplying the large number of groups observed by the small probability of seeing a large deviation in any particular group.

For example, suppose that in a sequence of one million base pairs of DNA, in which the fraction of “A” is one quarter, a search finds a window of twenty consecutive letters in which the number of places with “A” is sixteen. The overall statistical significance of this observation is reasonably bounded by multiplying the probability $p^*$ considered above by 1,000,000 (more precisely, by 999,981, which is the number of ways to position a window of length 20.) Using the normal approximation we would incorrectly assign an overall significance of 0.0069 to our observation. Using the large deviation theory or the exact binomial probability gives us an overall significance of approximately 0.4, so we conclude that even in purely random data, the chance of some window of length 20 with at least 16 heads is not small, and our observation could easily be due to chance alone.

In the example above, the effect of multiplying by 999,981 is to ignore the possibility that two or more windows of length 20 may each show at least 16 successes. We have overestimated the chance of at least one window by using the expected number of such windows, which is easier to compute. If the multiplication had produced a value greater than 1, then clearly we have an overestimate of the probability of at least one window, but the large expected number of windows is a clear indication that the probability of at least one window is not small, i.e. the overall observation is not statistically significant. The correction for this overestimate is complicated to work out, since the windows can overlap each other. For example, a region with 18 consecutive successes surrounded by several failures corresponds to 7 windows of length 20, each with at least 16 heads. This correction is carried out in Arratia et al. (1989).

3. Main Results. Formally, we let $C_1, \ldots, C_n$ be $p$-coins, i.e. independent, identically distributed Bernoulli variables with $p = P(C_i = 1) = 1 - P(C_i = 0)$. Let $S_n = C_1 + \cdots + C_n$, so that $S_n$ represents the total sum of successes in $n$ independent trials. The results we state are valid uniformly as $k$ and $n$ tend to infinity in such a way that their ratio:

$$a = k/n,$$

satisfies $p < a < 1$, and $a$ is bounded away from the endpoints $p$ and 1.

Let $H(a, p)$ be the relative entropy between a $p$-coin and an $a$-coin, i.e.
\[ H \equiv H(a, p) \equiv (a)\log \left( \frac{a}{p} \right) + (1-a)\log \left( \frac{1-a}{1-p} \right). \] \hfill (1)

We observe that \( H(a, p) \) increases from 0 to \( \log(1/p) \) as \( a \) increases from \( p \) to 1. This value \( H \) is also called the Kullback–Liebler distance; it measures the distance from the binomial \((n, p)\) distribution under which the data are generated to an alternative, the binomial \((n, a)\) distribution. The key concept and difficulty in understanding large deviations is dealing simultaneously with two probability measures on the same space of possible outcomes. Theorem 1 gives an upper bound valid for all \( n, p, \) and \( a \).

**Theorem 1.** For \( p < a < 1 \), and for \( n = 1, 2, 3, \ldots \), with \( H \) the relative entropy defined in equation (1) above:

\[ P(S_n \geq an) \leq e^{-nH} \]

In order to state Theorem 2, we let:

\[ r \equiv \frac{p}{1-p} \left/ \frac{a}{1-a} \right. = \frac{p}{a} \frac{1-a}{1-p}, \] \hfill (2)

denote the “odds ratio” between \( p \)-coins and \( a \)-coins. Observe that \( 0 < r < 1 \) since \( p < a < 1 \), and that all of the quantities \( a, H, \) and \( r \) are allowed to vary depending on \( k \) and \( n \). Observe also that \( r \) and \( H \) are related: \( H'(a, p) = -\log(r) \), where the derivative is taken with respect to \( a \), for fixed \( p \). We write \( \sim \) to denote asymptotic equality, i.e. that the ratio of two quantities tends to the limit 1, and \( \asymp \) to denote exponential equivalence, i.e. that the logarithms of two quantities are asymptotic.

Equation (3) below, which is the result of dividing equation (5) into equation (6), says that conditional on the number of successes being at least \( k = an \), the excess over \( k \) has asymptotically a geometric distribution with parameter \( 1-r \). This geometric distribution is the distribution of the number of heads before the first tail when tossing an \( r \)-coin, and has mean \( r/(1-r) \).

\[ P(S_n = an + i | S_n \geq an) \rightarrow r^i(1-r) \text{ for } i = 0, 1, 2, \ldots \text{ as } n \rightarrow \infty. \] \hfill (3)

**Theorem 2.** For \( p < a < 1 \), and \( r \) the odds ratio defined in equation (2) above, as \( n \rightarrow \infty \),

\[ P(S_n \geq an) \approx e^{-nH}, \text{ i.e. } \log P(S_n \geq an) \sim -nH, \] \hfill (4)

\[ P(S_n \geq an) \sim \frac{1}{1-r} \frac{1}{\sqrt{2\pi a(1-a)n}} e^{-nH}, \] \hfill (5)
\[ P(S_n = an + i) \sim \frac{1}{\sqrt{2\pi a(1-a)n}} r^i e^{-nH}, \text{for } i = 0, 1, 2, \ldots \]

**Proof 1.** To prove Theorem 1, observe that for any positive value \( \beta \),
\[ P(S_n \geq an) = P(\exp(\beta S_n) \geq e^{\beta an}) \leq E \exp(\beta S_n)/e^{\beta an} = \left[e^{-a\beta}(1 - p + p e^\beta)\right]^n. \]

The inequality above is known as Chebychev’s or as Markov’s inequality; multiplied out it simply says \( E \exp(\beta S_n) \geq e^{\beta an} P(\exp(\beta S_n) \geq e^{\beta an}). \) When the expression in brackets is minimized by the optimal choice of \( \beta \), its value is \( e^{-H} \), proving Theorem 1.

To prove Theorem 2, combine the binomial density, \( P(S_n = j) = \binom{n}{j} p^j (1-p)^{n-j} \), and the detailed form of Stirling’s formula, as presented in Section II.9 of Feller, Vol. 1, and reproduced below:
\[ \sqrt{2\pi n} \ n^n e^{-n} e^{1/(12n + 1)} < n! < \sqrt{2\pi n} n^n e^{-n} e^{1/(12n)}. \]

The proof above, although elementary, has three drawbacks: it provides little insight, the manipulations are complicated and hard to manage without error, and you must take something on faith, namely Stirling’s formula. The second proof we present, due to Cramer (1938), shares none of these drawbacks. It is a beautiful proof, and well worth studying even by non-specialists in probability.

**Proof 2.** Begin by considering a second probability model, \( Q \), in which the event \( \{S_n \geq an\} \) is not rare but rather typical. More precisely, let \( Q \) be the model in which \( C_1, C_2, \ldots \) are \( a \)-coins, so that under \( Q \), the probability of the event \( \{S_n = j\} \) is \( Q(S_n = j) = n!/s!(n-j)!a^j(1-a)^{n-j} \). By considering the ratio of the \( P \) and \( Q \) probabilities of the event \( \{S_n = j\} \), we cause the factorials to cancel. We denote this likelihood ratio by \( \frac{dP}{dQ}(j) \):
\[ \frac{dP}{dQ}(j) = \frac{p^j(1-p)^{n-j}}{a^j(1-a)^{n-j}}. \]

Observe that \( nH = k \log(a/p) + (n-k) \log((1-a)/(1-p)) \), so that \( e^{-nH} = (p/a)^k((1-p)/(1-a))^{n-k} \), hence the likelihood ratio above can be expressed in terms of the relative entropy \( H \) and the odds ratio \( r \):
\[ \frac{dP}{dQ}(j) = \frac{p^j(1-p)^{n-j}}{a^j(1-a)^{n-j}} = e^{-nH}r^{j-k}. \]

Equation (6) follows from the above by writing:
\[ P(S_n = an + i) = Q(S_n = an + i) \frac{dP}{dQ}(i) = Q(S_n = an + i)r^i e^{-nH}, \]
and quoting the local central limit theorem for tossing an \( a \)-coin: 

\[
Q(S_n = an + i) = \frac{1}{(2\pi a(1-a)n)^{1/2}}.
\]

Equations (4) and (5) are proved in the same way: sum the above over \( i = 0, 1, 2, \ldots \), and take out the common factor \( e^{-nH} \) to get:

\[
P(S_n \geq an = k) = \sum_{0 \leq i \leq n-k} \frac{dP}{dQ}(k+i)Q(S_n = an + i)
= e^{-nH} \sum_{0 \leq i \leq n-k} r^i Q(S_n = an + i).
\]

Use \( 0 < r < 1 \) and \( \sum Q(S_n = an + i) \leq 1 \) to prove Theorem 1. Use a version of the local central limit theorem with a uniformity estimate to prove equation (5).

We thank Ann Barber, for showing us work that motivated this paper.

LITERATURE


Received for publication 1 July 1988