Median Bounds and their Application*

Alan Siegel

Department of Computer Science, Courant Institute, New York University, NYC, NY 10012-1185
E-mail: siegel@cs.nyu.edu

Basic methods are given to evaluate or estimate the median for various probability distributions. These methods are then applied to determine the precise median of several nontrivial distributions, including weighted selection, and the sum of heterogeneous Bernoulli Trials conditioned to lie within any centered interval about the mean. These bounds are then used to give simple analyses of algorithms such as interpolation search and some aspects of PRAM emulation.

Key Words: median, Bernoulli Trials, hypergeometric distribution, weighted selection, conditioned Bernoulli Trials, conditioned hypergeometric distribution, interpolation search, PRAM emulation

1. INTRODUCTION

While tail estimates have received significant attention in the probability, statistics, discrete mathematics, and computer science literature, the same cannot be said for medians of probability distributions, and for good reason. First, the number of known results in this area seems to be fairly meager, and even they are not at all well known. Second, there seems to have been very little development of mathematical machinery to establish median estimates for probability distributions. Third (and consequently), median estimates have not been commonly used in the analysis of algorithms, apart from the kinds of analyses frequently used for Quicksort, and the provably good median approximation schemes typified by efficient selection algorithms.

This paper addresses these issues in the following ways. First, a framework (Theorems 2.1 and 2.4) is presented for establishing median estimates. It is strong enough to prove, as simple corollaries, the two or three non-trivial median bounds (not so readily identified) in the literature. Second, several new median results are presented, which are all, apart from one, derived via this framework. Third, median estimates are shown to simplify the analysis of some probabilistic algorithms and processes. Applications include both divide-and-conquer calculations and tail bound estimates for monotone functions of weakly dependent random variables. In particular, a simple analysis is given for the $\log_2 \log_2 n + O(1)$ probe cost for both successful and unsuccessful Interpolation Search, which is less than two probes worse than the best bound but much simpler. Median bounds are also used, for example, to attain a tail bound to show that n random numbers can be sorted in linear time with probability $1 - 2^{-cn}$, for any fixed constant c. This result supports the design of a pipelined version of Ranade's Common PRAM emulation algorithm on an $n \times 2^n$ butterfly network with only one column of 2^n processors, by showing that each processor can perform a sorting step that was previously distributed among n switches.

The tenor of the majority of median estimates established in Section 2 is that whereas it may be difficult to prove that some explicit integral (or discrete sum) exceeds $\frac{1}{2}$, by some tiny amount, it is often much easier to establish global shape-based characteristics of a function – such as the number of zeros in some interval, or an inequality of the form $f < g$ – by taking a mix of derivatives and logarithmic derivatives to show that, say, f and g both begin at zero, but g grows faster than f. This theme characterizes Theorems 2.1 and 2.4, plus all of their applications to discrete random variables.

* Supported, in part, by NSF grant CCR-9503793.

1.1. Notation, conventions, and background

By supposition, all random variables are taken to be real valued. A median of a random variable X is defined to be any value x where Prob$\{X \geq x\} \geq \frac{1}{2}$, and Prob$\{X \leq x\} \geq \frac{1}{2}$. Many random variables—including all of the applications in this paper—will have essentially unique medians.

The functions X, Y, and Z will be random variables. Cumulative distribution functions will be represented by the letters F or G, and will have associated density functions $f(x) = F^{'}(x)$, and $g(x) = G^{'}(x)$. The mean of a random variable X will be represented as E$[X]$, and the variable μ will be used to denote the mean of the random variable of current interest. The variance of X is defined to be E$[X^2] - E[X]^2$, and will sometimes be denoted by the expression σ^2. The characteristic function $\chi\{event\}$ is defined to be 1 when the Boolean variable $event$ is true, and 0 otherwise. If X and Y are random variables, the conditional expectation E$[X|Y]$ is a new random variable that is defined on the range of Y. In particular, if Y is discrete, then E$[X|Y]$ is also discrete, and for any x where Prob$\{Y = x\} \neq 0$, E$[X|Y](x) = \frac{\text{E}[X \cdot \chi\{Y=x\}]}{\text{Prob}\{Y=x\}}$ with probability Prob$\{Y = x\}$. Thus, E$[X|Y](x)$ is just the average value of X as restricted to the domain where $Y = x$. Conditional expectations preserve the mean: E$[E[X|Y]] = E[X]$. However, they reduce variances: E$[E[X|Y]^2] \leq E[X^2]$. Intuitively, averaging reduces uncertainty.

We also would like to define the random variable Z as X conditioned on the event $Y = k$. More formally, Z is distributed according to the conditional probability: Prob$\{Z = s\} \equiv$ Prob$\{X = s|Y = k\} = \frac{\text{Prob}\{X=s \wedge Y=k\}}{\text{Prob}\{Y=k\}}$. Sometimes Z will be defined with the wording X given Y, and sometimes it will be formulated as $Z = [X|Y = k]$. Here the intention is that k is fixed and X is confined to a portion of its domain. The the underlying measure is rescaled to be 1 on the subdomain of X where $Y = k$. According to the conventions of probability, this conditioning should be stated in terms of the underlying probability measure for the r.v. pair (X, Y) as opposed to the random variables themselves. Thus, our definition in terms of X and Y and the notation $[X|Y = k]$ are conveniences that lie outside of the formal standards.

These definitions all have natural extensions to more general random variables. Indeed, modern probability supports the mathematical development of distributions without having to distinguish between discrete and continuous random variables. While all of the median bounds in this paper concern discrete random variables, almost all of the proofs will be for continuous distributions. When a result also applies to discrete formulations, we will say so without elaboration, since a point mass (or delta function) can be defined as a limit of continuous distributions. In such a circumstance, the only question to resolve would be how to interpret an evaluation at the end of an interval where a point mass is located. For this paper, the issue will always be uneventful.

If Z is a nonnegative integer valued random variable, its generating function is often defined as $G_Z(x) = E[x^Z] \equiv \sum_j x^j$Prob$\{Z = j\}$. For more general random variables, the usual variant is $G_Z(\lambda) = E[e^{\lambda Z}] \equiv \int e^{\lambda x}$Prob$\{Z \in [x, x+dx)\}$. This notation is intended to be meaningful regardless of whether the underlying density for Z comprises point masses, a density function, etc. One of the advantages of these transformations is that if X and Y are independent, then $G_{X+Y} = G_X \cdot G_Y$ (where all G's use the same formulation). Hoeffding-Chernoff tail estimates are based on generating functions. The basic procedure is as follows. For $\lambda > 0$: Prob$\{X > a\} = $Prob$\{\lambda X > \lambda a\} = $Prob$\{e^{-\lambda a} e^{\lambda(X)} > 1\} = E[\chi\{e^{-\lambda a} e^{\lambda(X)} > 1\}] \leq E[e^{-\lambda a} e^{\lambda(X)}] = e^{-\lambda a}E[e^{\lambda(X)}]$, because $\chi\{e^{-\lambda a} e^{\lambda(X)} > 1\} < e^{-\lambda a} e^{\lambda(X)}$. This procedure frequently yields a specific algebraic expression $e^{-a\lambda} f(\lambda)$, where by construction Prob$\{X > a\} < e^{-a\lambda} f(\lambda)$ for any $\lambda > 0$. The task is then to find a good estimate for $\min_{\lambda > 0} e^{-a\lambda} f(\lambda)$ as a function of a, which can often be done. If $X = X_1 + X_2 + \ldots + X_n$ is the sum of n independent random variables X_i, then the independence can be used to represent $f(\lambda)$ as a product of n individual generating functions $G_{X_i}(\lambda)$. If the random variables have some kind of dependencies, then this procedure does not apply, and an alternative analysis must be sought.

Some of the basic random variables that will be used to build more complex distributions are as follows.

A Boolean variable X with mean p satisfies:

$$X = \begin{cases} 1 & \text{with probability } p; \\ 0 & \text{with probability } 1 - p. \end{cases}$$

A random variable X that is exponentially distributed with (rate) parameter λ satisfies:

$$\text{Prob}\{X \leq t\} = 1 - e^{\lambda t}, \ \ t \geq 0.$$

The density function is $f(t) = \lambda e^{-\lambda t}$, and the mean is $\int_0^\infty t\lambda e^{-\lambda t} dt = \frac{1}{\lambda}$.

A Poisson random variable P with (rate) parameter λ satisfies:

$$P = j \text{ with probability } e^{-\lambda}\frac{\lambda^j}{j!}, \text{ for } j = 0, 1, 2, \dots.$$

A straightforward summation gives

$$\mathrm{E}[P] = e^{-\lambda}\sum_{j>0}\frac{\lambda^j}{(j-1)!} = \lambda e^{-\lambda}\sum_{j\geq 0}\frac{\lambda^j}{j!} = \lambda.$$

The hypergeometric distribution corresponds to the selection (without replacement) of n balls from an urn containing r red balls and g green balls, for $n \leq r+g$. If R is the number of red balls chosen, then

$$\text{Prob}\{R = k\} = \frac{\binom{r}{k}\binom{g}{n-k}}{\binom{r+g}{n}}.$$

Informally, a stochastic process is a random variable that evolves over time. The processes of interest remain consistent with their past but do not know their future changes. One of the simplest such processes is a 0–1 switch $e(t)$ with an exponential waiting time. A very informal definition might be

$$e(t) = \begin{cases} 0, & 0 \leq t < P; \\ 1, & P \leq t, \end{cases}$$

where P is an exponentially distributed random variable. P is the waiting time for e to become 1. The switch $e(t)$ is consistent with its past in the sense that if $e(s) = 1$ for some $s < t$, then $e(t) = 1$. We will soon review just how poorly it can predict future changes in the sense that if $e(t) = 0$, the future value $e(t+s)$ will be random, for $s > 0$.

The difficulty with our formulation in terms of P, of course, is that the value of P, is unknown as long as $e(t) = 0$.

The probability distribution for $e(t)$ satisfies

$$e(t) = \begin{cases} 0 & \text{with probability } e^{-\lambda t}; \\ 1 & \text{with probability } 1 - e^{-\lambda t}. \end{cases}$$

If we have k independent identically distributed copies of such a random variable, then the probability that all k are zero at time t is $e^{-k\lambda t}$. Thus, the waiting time until the first switching event occurs is exponentially distributed with rate parameter $k\lambda$. Evidently, this additivity of rates for the arrival of the first switching event is additive; the individual rate parameters need not be the same.

The random variable $e(t)$ exhibits a memorylessness that is formalized by observing that for $s, t \geq 0$, $\text{Prob}\{e(t+s) = 0 | e(t) = 0\} = \text{Prob}\{\hat{e}(s) = 0\}$, where $\hat{e}$ is a statistically identical switching function that is independent of e. The calculations are:

$$\begin{aligned} \text{Prob}\{e(t+s) = 0 | e(t) = 0\} &= \frac{\text{Prob}\{e(t) = 0 \wedge e(t+s) = 0\}}{\text{Prob}\{e(t) = 0\}} \\ &= \frac{\text{Prob}\{e(t+s) = 0\}}{\text{Prob}\{e(t) = 0\}} = \frac{e^{-\lambda(t+s)}}{e^{-\lambda t}} = e^{-\lambda s}. \end{aligned}$$

This memorylessness can be combined with the additivity of rate parameters to conclude that if we have k such switches that are 0 at time t, the probability that the first switching of a variable to 1 occurs at time $t+s$ is exponentially distributed (in s) with rate parameter $k\lambda$.

This random variable is also convenient for turning Bernoulli Trials into a random variable that evolves over continuous time. If $e(t)$ has an exponentially distributed waiting time with parameter λ, then $\text{Prob}\{e(t) = 0\} = e^{-\lambda t}$, so $e(t_0)$ is statistically equivalent to a Bernoulli Trial with probability of success $1 - e^{-\lambda t_0}$.

A Poisson process $P(t)$ with rate parameter λ is a stochastic process where

$$\text{Prob}\{P(t)=j\}=e^{-\lambda t}\frac{(\lambda t)^j}{j!},\ \text{for } j=0,1,2,\cdots.$$

We informally summarize the other key properties by noting that $P(t)-P(s)$, for $t\geq s$ is a stochastic process that is independent of $P(s)$ (including the history $P(h)$, for $h\leq s$), and is statistically identical to an independent copy of $P(t-s)$. Clearly, $\text{E}[P(t)]=\lambda t$.

The waiting time during which a Poisson process remains at a fixed value k is exponentially distributed with parameter λ. The usual term for λ is interarrival rate, since these processes are often used to model arrivals of discrete events (such as packets, parts failures, gamma rays, etc) in many physical systems. In plain English, $P(t)$, for $t\geq 0$, defines a random function that is piecewise constant and has jumps from 0 to 1 to 2 to 3 to .... Each piecewise constant portion has a random length that is exponentially distributed with mean (waiting time) $\frac{1}{\lambda}$ and is independent of the other lengths. Poisson processes give a natural way to define a Poisson random variable that evolves over time.

Lastly, we use a convention to simplify the discussion of inequalities that we seek to prove. Target inequalities that are not yet established might be expressed with the relation $<?$ to facilitate reformulation without confusion. As long as every step in a derivation is reversible, we can work in the more intuitive direction to reduce such a target inequality to a known result, and implicitly recognize that a formal proof would reverse all of the steps.

2. COMPUTING MEDIANS

Some median bounds are straightforward. Perhaps the most elementary is the fact that for any random variable with mean μ and standard deviation σ, the median must lie in the interval $[\mu-\sigma,\ \mu+\sigma]$.

There is nothing new about this observation.

The bound follows from considering the extreme case, which must comprise a two-point distribution. To be specific, let X be a random variable with a bounded mean μ and median m where, say, $m>\mu$. We seek to change X to have a smaller variance while keeping the mean at μ and without diminishing the median. The way to do this is with conditional expectations. Let $Z=\text{E}[X|\chi(X\geq m)]$. By definition, $Z(1)=\frac{\text{E}[X\cdot\chi\{X\geq m\}]}{\text{Prob}\{X\geq m\}}$ with probability $\text{Prob}\{X\geq m\}$, and similarly, $Z(0)=\frac{\text{E}[X\cdot\chi\{X<m\}]}{\text{Prob}\{X<m\}}$ with probability $\text{Prob}\{X<m\}$. The median of Z is $Z(1)$, which is the average of all X-values that are at least m.

Observe that Z has the same mean as X, a variance that is no larger, and a median that is no smaller. Thus this projection of X shows that some two-point discrete random variable (or limit of two-point random variables) has the largest median for a given mean and variance. So let Z be a random variable where $Z=M+\mu$ with probability $p>\frac{1}{2}$, and let $Z=-\frac{Mp}{1-p}+\mu$ with probability $1-p$. The median is $M+\mu$, the mean is μ, the variance is $\sigma^2=\frac{M^2p}{1-p}$, which is minimized by decreasing p to $\frac{1}{2}$, whence σ becomes M. It follows that even for this smallest possible variance, $M+\mu\in[\mu-\sigma,\mu+\sigma]$.

Even this bound can be useful. For example, if μ and σ are known, a randomized binary search procedure might, at each iteration, select one of the two endpoints $\mu\pm\sigma$ probabilistically.

In this section, we will determine the precise value of the median for a variety of discrete probability distributions.[1] The estimation process begins fairly innocuously; Lemma 2.1 simply states that the cumulative distribution function of a nonnegative random variable X has an average value of at least $\frac{1}{2}$, over the interval $[0,2\text{E}[X]]$.

LEMMA 2.1. *Let X be a nonnegative random variable with distribution function F and mean $\mu=\text{E}[X]$.*

Then
$$\int_0^{2\mu}F(t)dt=\mu+\int_{2\mu}^{\infty}(t-2\mu)f(t)dt\geq\mu.$$

[1] Some of our theorems will require that the random variable of interest—such as a sum of heterogeneous Bernoulli Trials—have a mean μ that is an integer. In these cases, including instances where the original random variable is subsequently studied subject to some conditioning, the median will also be μ. If the mean were not an integer, the theory would still be applicable even though we will not say so explicitly, and the median will turn out to be one of the two consecutive integers $\lfloor\mu\rfloor$, $\lceil\mu\rceil$.

Proof. Suppose, for simplicity, that the density $f(x)$ is a function, so that there are no discrete point masses. We can also assume that $\mu < \infty$. Integrating by parts gives:

$$\begin{aligned}\int_0^{2\mu} F(t)dt &= (t-2\mu)F(t)\Big|_{t=0}^{2\mu} - \int_0^{2\mu}(t-2\mu)f(t)dt \\ &= 0 - \int_0^{\infty}(t-2\mu)f(t)dt + \int_{2\mu}^{\infty}(t-2\mu)f(t)dt \\ &= -(\mathrm{E}[X]-2\mu) + \int_{2\mu}^{\infty}(t-2\mu)f(t)dt \\ &= \mu + \int_{2\mu}^{\infty}(t-2\mu)f(t)dt \geq \mu.\end{aligned}$$

▮

This simple fact suffices to establish the following.

THEOREM 2.1. *Let X be a nonnegative random variable, and suppose, for simplicity, that its density is defined by a function f, so that there are no discrete point masses, and suppose that f is also continuous. In addition, suppose that*

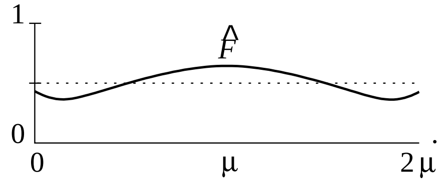
1) $\mathrm{E}[X] = \mu < \infty$,
2) $f(x) - f(2\mu - x)$ *has at most one zero for x restricted to the open interval $(0, \mu)$.*
3) $f(0) < f(2\mu)$.

Then
$$\mathrm{Prob}\{X \leq \mu\} > \frac{1}{2}.$$

Requirement 2) can be replaced by the more applicable albeit slightly less general formulation

2′) *The density f is continuously differentiable and $\frac{f'(x)}{f(x)} + \frac{f'(2\mu-x)}{f(2\mu-x)}$ has at most two zeros in $(0, 2\mu)$.*

Proof. Let $\widehat{F}(x) \equiv \frac{F(x)+F(2\mu-x)}{2}$. By supposition, $\widehat{F}$ is moustache-shaped on $[0, 2\mu]$:



It is symmetric about μ by construction. Since $f(0) < f(2\mu)$, it is decreasing at the origin from the value $\widehat{F}(0) = \frac{F(2\mu)}{2} \leq \frac{1}{2}$. Since $\frac{d}{dx}\widehat{F}$ has at most one zero inside $(0, \mu)$, this zero must be a minimum for $\widehat{F}$, which must subsequently increase to exceed, at $x = \mu$, the overall average, which by Lemma 2.1 is at least $\frac{1}{2}$.

As for the alternative criterion, suppose that f were continuously differentiable. The original requirement for F is that $f(x) = f(2\mu - x)$ (i.e., $\widehat{F}'(x) = 0$) should hold for at most three points in $[0, 2\mu]$. If $\log f(x)$ were to equal $\log f(2\mu - x)$ at four or more locations in the interval, then there would be at least three internal points where the derivative of $\log f(x) - \log f(2\mu - x)$ would be zero, which is precisely what is prohibited by the alternative $2'$). ▮

This formulation also extends to the (actually occurring) trivial cases where $f(0) \geq f(2\mu)$, and $f(x) - f(2\mu - x)$ has no zero (or $F(x) + F(2\mu - x)$ has no local maximum) for x restricted to the open interval $(0, \mu)$.

2.1. Applications of Theorem 2.1

2.1.1. Bernoulli Trials.

We can now give a generic proof of the following result by Jogdeo and Samuels [6]:

THEOREM 2.2 (Jogdeo and Samuels [6]). *Let $X_n = x_1 + x_2 + \cdots + x_n$ be the sum of n independent heterogeneous Bernoulli Trials. Suppose that* $\mathrm{E}[X_n]$ *is an integer.*

Then $$\mathrm{Prob}\{X_n > \mathrm{E}[X_n]\} < \frac{1}{2} < \mathrm{Prob}\{X_n \geq \mathrm{E}[X_n]\}.$$

Proof. Let $\mathrm{Prob}\{x_i = 1\} = 1 - e^{-\lambda_i}$, and let $\mathrm{E}[X_n] = np$. It suffices to show that $\frac{1}{2} < \mathrm{Prob}\{X_n \geq np\}$, since the complement of this bound, when the notions of success and failure are reversed, shows that $\mathrm{Prob}\{X_n > np\} < \frac{1}{2}$.

Suppose, for the moment, that the λ_i are all equal to a fixed value λ. Let $X_n(t)$ be the sum of n independent 0–1 random switching functions with exponentially distributed waiting times and rate parameter λ. Define the waiting time $T = \min_s\{s : X_n(s) \geq np\}$. Then T will have the cumulative distribution function:

$$F(t) \equiv \mathrm{Prob}\{T \leq t\} = \sum_{j=np}^{n} \binom{n}{j} (1 - e^{-\lambda t})^j e^{-\lambda t(n-j)}.$$

We want to show that $\mathrm{Prob}\{T \leq 1\} > \frac{1}{2}$, since $X_n(1)$ is distributed according to the Bernoulli Trails, and the event $T \leq 1$ is equivalent to the event $X_n(1) \geq np$.

Straightforward differentiation of F or an understanding of exponentially distributed waiting times shows that the density function for T is

$$f(t) = \binom{n}{np-1} (1 - e^{-\lambda t})^{np-1} e^{-\lambda t(n+1-np)} \lambda(n+1-np).$$

The stopping time T and its density function f are suitable for Theorem 2.1 where we set X to be T.

As is standard, we use the memorylessness of switches with exponentially distributed waiting times and the additivity of the rate parameters to conclude that $X_n(t)$ will switch from j (to $j+1$) according to a waiting time that is exponentially distributed with rate parameter $(n-j)\lambda$. Let T_{n-j} be the waiting time during which $X_n(t) = j$. Then $T = T_n + T_{n-1} + \ldots + T_{n-np+1}$, and $\mathrm{E}[T] = \mathrm{E}[T_n] + \mathrm{E}[T_{n-1}] + \cdots + \mathrm{E}[T_{n-np+1}]$. We conclude that

$$\begin{aligned}
\mathrm{E}[T] &= \frac{1}{\lambda}\left(\frac{1}{n} + \frac{1}{n-1} + \cdots + \frac{1}{n+1-np}\right) \\
&< \frac{1}{\lambda}\int_{n-np}^{n} \frac{1}{x}dx = \frac{1}{\lambda}(\ln n - \ln(n-np)) = \frac{1}{\lambda}\left(\ln\frac{1}{1-p}\right) \\
&\leq 1, \quad \text{because } p = 1 - e^{-\lambda}, \text{ and hence } \ln\tfrac{1}{1-p} = \lambda.
\end{aligned}$$

Alternatively, $\mathrm{E}[T]$ can also be evaluated directly via integration by parts.

Let $\mu = \mathrm{E}[T]$. We use Theorem 2.1 to show that $\mathrm{Prob}\{T \leq \mu\} > \frac{1}{2}$, which is sufficient to establish the application, since $\mu < 1$. We need only verify that $f(t)$ satisfies conditions 1), $2'$) and 3). The first and last are trivial. As for $2'$), let $\zeta = e^{-\lambda\mu}$, and note that

$$\frac{f'(t)}{f(t)} + \frac{f'(2\mu - t)}{f(2\mu - t)} = \lambda\left[\frac{(np-1)e^{-\lambda t}}{1 - e^{-\lambda t}} - (n+1-np)\right] + \lambda\left[\frac{(np-1)\zeta^2 e^{\lambda t}}{1 - \zeta^2 e^{\lambda t}} - (n+1-np)\right].$$

Multiply the numerator and denominator of the second fraction by $e^{-\lambda t}$ so that each term is a rational expression in $e^{-\lambda t}$, and substitute $x = e^{-\lambda t}$ everywhere. The resulting expression has the form $\frac{ax}{bx+c} + \frac{d}{ex+f} + g$. Expressing this over a common denominator yields a numerator that is quadratic in x, and which therefore has at most two zeros. Hence $\mathrm{Prob}\{X_n(1) \geq np\} > \frac{1}{2}$ when all Trials are identically distributed.

The general case follows immediately from a very simple but clever convexity analysis that is due to Hoeffding. Let n and $\mathrm{E}[X_n] = \mu$ be fixed. Let S be the the set of every random variable that is the sum of n Bernoulli Trials with mean μ, and let I be any fixed subset of the integers. Hoeffding showed that Prob$\{X \in I\}$, for $X \in S$, is minimized by some extremal distribution (and is likewise minimized by another such extremal distribution) where some of the n Bernoulli Trials are stuck at 1 (with individual probability of success equal to 1), some are stuck at zero (with probability of success equal to 0), and all other probabilities of success are set to some identical constant p [4].[2] Thus, the special cases where all probabilities are identical is also the worst case. The constant Trials just have the effect of adjusting n and p. ∎

For completeness, we note that Jogdeo and Samuels gave a subtle but very elegant argument – specific to Bernoulli Trials – to attain a slightly but remarkably stronger result. Whereas the mean is the median says that $|\text{Prob}\{X \leq np\} - \text{Prob}\{X \geq np\}| < \text{Prob}\{X = np\}$, Jogdeo and Samuels showed that the difference is in fact bounded by $\frac{1}{3}\text{Prob}\{X = np\}$, and is positive (without the absolute value) when $p < 1/2$. They also characterized the behavior of the difference as $n \to \infty$ with np fixed [6].

Before presenting some additional probability applications, we give an easy algorithmic application.

2.1.2. Interpolation Search.

Interpolation Search was introduced by Peterson [13], and its performance has been analyzed extensively [19], [3], [11]. Moreover, the underlying scheme has been extended to support search for data generated from unknown probability distributions [18] and more general data access systems [9]. The basic probing strategy, however, has yet to be analyzed by elementary means. Indeed, it has been open whether a simple analysis could establish a performance bound of $\log_2 \log_2 n + O(1)$ [1], [2]. Perl and Reingold, in their presentation of a less efficient (but easily analyzed) search strategy, speculate that the difficulty of the analysis is responsible for the absence of this remarkable search procedure from algorithms texts [12].

The basic problem is essentially as follows. Let D comprise n real values randomly (and independently) selected according to the uniform distribution on the interval $[l_0, u_0]$. Suppose D is stored in sorted order in the array (table) $T[1..n]$. We seek to determine if the search key (number) x is in T.

Preliminarily suppose, for expositional simplicity, that x actually resides in the table. Let x be in table location s_*, so that $T[s_*] = x$. We initially use an Interpolation Search procedure that is optimized for successful search, although alternative strategies are equally amenable to analysis. The search will proceed in rounds indexed by the subscript i. For the i-th round, let the search interval comprise the index locations $\{l_i+1, l_i+2, \ldots, u_i-1\}$, so that the keys in locations l_i and u_i are known, no interior values have been probed and s_* is within this index range. Let $n_i = u_i - l_i - 1$, and let p_i be the expected fraction of the $n_i - 1$ keys (excluding x) that are less than x, so that $p_i = \frac{x - T[l_i]}{T[u_i] - T[l_i]}$. Thus, the number of keys in $T[l_i + 1..u_i - 1]$ that are less than x is statistically equivalent to the sum of $n_i - 1$ Bernoulli Trials with probability of success p_i. Let s_{i+1} be the interpolated search location computed at round i, and probed at the next round, so that $s_{i+1} = l_i + p_i(n_i - 1) + 1$. (For unsuccessful search, comparable considerations suggest the formulation $s_{i+1} = l_i + p_i n_i + \frac{1}{2}$.) Assume, for the moment, that $(n_i - 1)p_i$ is an integer, so that s_{i+1} needs no rounding. At round $i + 1$, location s_{i+1} is probed, and the value is used to establish the updated index values l_{i+1}, u_{i+1} for the round $i + 1$.

We present two analyses that are reasonably elementary, provided we support the premise that a better knowledge of Bernoulli Trials – including mean-median results – is worthy of dissemination. The more precise of the two, which comes second, uses some probability that is only infrequently seen in the computer science literature, but the technical difficulty is otherwise less.

Analysis 1. An easy $O(\log \log n)$ analysis results from 2 facts:

[2] Hoeffding's wonderful proof technique is emulated in the more complicated setting of Lemma 4.1 of Section A.4. The interested reader might prefer to preview the approach as applied to the simpler context of unconditioned Bernoulli Trials [4]. Hoeffding's proof is easier to read because he does not state his lemma in full generality. Incidentally, this simple proof ignited more than a decade of related work in probability and statistics. See, for example, [8], which comprises some 500 pages of subsequent developments.

1) A standard Hoeffding tail bound shows that s_{i+1} will be within, say, $\sqrt{n_i \log n_i}$ locations of s_* with a probability that exceeds $1-\frac{2}{n_i^2} \geq \frac{1}{2}$. In particular, the Hoeffding bound [5] states: suppose X_m is the sum of m independent Bernoulli Trials with mean $\mu \equiv \mathrm{E}[X_m]$. Then Prob$\{|X_m - \mu| \geq am\} \leq 2e^{-2a^2m}$. We use this inequality with $a = \sqrt{\frac{\log_e m}{m}}$.

2) If x has not yet been found by the conclusion of round i, it will have more than a 50% chance of being in the smaller subinterval comprising $\min[s_{i+1} - l_i, u_i - s_{i+1}]$ items, since the median is the mean in this case of implicit Bernoulli Trials.

Now, two consecutive probes s_{2i} and s_{2i+1} will both satisfy closeness criterion 1), and s_{2i+1} will also satisfy 2) with a joint probability exceeding $\frac{1}{8}$. In this case, the resulting n_{2i+1} will satisfy

$$n_{2i+1} < |s_{2i+1} - s_{2i}| \leq |s_{2i} - s_*| + |s_{2i+1} - s_*| < 2\sqrt{n_{2i-1} \log n_{2i-1}},$$

and we call the probe pair very close. It is easy to show that the search interval will comprise at most $8 \log n$ indices after $\log\log n$ such pairs of very close probes:
A simple way to formalize this fact is by weakening closeness criterion 1) to be within, say, $\sqrt{n_i \log n}$ locations, when $n_i > 8 \log n$. It follows that $\log\log n$ such very close pairs will reduce the size of the index range from (the overestimate) $4n \log n$ down to $8 \log n$ or less, because a close pair is certain to reduce an overcount of the form $4(2^{2^h}) \log n$ down to the overcount $4(2^{2^{h-1}}) \log n$, and $h < \log_2 \log_2 n$.

Since each subsequent probe has at least a 50% chance of producing a new search range that is no bigger than half the size of the previous range, termination occurs within $\log_2 \log_e n + 4$ additional probes that satisfy criterion 2).

Since very close probe pairs will occur, and later probes will satisfy 2) with respective probabilities exceeding $\frac{1}{8}$ and $\frac{1}{2}$, the expected probe count must be $O(\log\log n)$:
This remark can be formalized by viewing disjoint pairs of consecutive probes $(s_{2i},\ s_{2i+1})$ as individual Bernoulli Trials with probability of success $\frac{1}{8}$ when $n_i > 8 \log n$, and subsequent individual probes as Trials with probability of success $\frac{1}{2}$. As is readily calculated by recurrences as well as several other methods, the expected number of Bernoulli Trials needed to get r successes is just $\frac{r}{p}$, where p is the probability of success.

Analysis 2. More interesting is a $\log_2 \log_2 n + O(1)$ bound for the probe count. The proof has two parts. The first half originates in an elegant, short (but sophisticated) analysis by Perl, Itai, and Avni [11]. These authors presented a three-part analysis of Interpolation Search that was technically incomplete. The first part uses notions no more advanced than Jensen's Inequality in the restricted form $\mathrm{E}[X]^2 \leq \mathrm{E}[X^2]$, conditional expectations for discrete random processes, plus a few standard facts about Bernoulli Trials. The second and third parts use the Optional Sampling Theorem for supermartingales, plus some mild technicalities to show that $\mathrm{E}[P] = \log_2 \log_2 n + O(1)$, where P is the number of rounds necessary to reach an index that is exactly two locations away from the actual search key x. The technicalities use interpolation to formalize P as real valued, since there may be no probe that lands exactly two locations away from the search key.

The authors note that sequential search can complete the process in the desired performance bound, but do not analyze the expected number of additional probes that would be used by pure Interpolation Search. Since, as Section 2.1.4 shows, the expected size of the remaining search interval is about $O(\frac{n}{\log n})$, some additional analysis would seem to be necessary to bound the performance of the pure algorithm.

We replace the second and third parts, at a cost of about one probe, by a triviality, thanks to the fact that the mean is the median for Bernoulli Trials. With the parameters as defined, Perl, Itai and Avni showed, in the first part of their proof, that

$$\mathrm{E}[|s_{t+1} - s_t| \;\Big|\; s_t, s_{t-1}] < \sqrt{\mathrm{E}[|s_t - s_{t-1}|]}.$$

Section A.1 of the Appendix presents their derivation of this inequality with enhancements to include the effects of rounding, and the differences between successful and unsuccessful search; an alternative derivation based upon direct calculation is given in Section A.1.3.

In any case, the expected distance between consecutive probes turns out to be less than 2 after $i_0 \equiv \log_2 \log_2 n+1$ probes. This fact does <u>not</u> guarantee that search range size n_{i_0} is small on average, but does prove that $\min\{n_{i_0} p_{i_0}, n_{i_0}(1-p_{i_0})\}$

has a small expectation, since, $|s_{t+1} - s_t|$ is, if termination has not yet occurred, either $\hat{n}_t p_t + 1$ or $\hat{n}_t(1 - p_t) + 1$, where $\hat{n}_i \equiv n_i - 1$. We now show that the bound $\mathrm{E}[|s_{i_0} - s_{i_0-1}|] < 2$ is indeed sufficient to prove that the expected number of probes is no more than $\log_2 \log_2 n + 3$.

It is easy to show that $|s_{i+1} - s_i|$ is nonincreasing in i, and since Interpolation Search, in the case of successful search, has less than a 50% chance of selecting the larger subinterval at each step, the desired bound follows trivially. To formalize these observations, let ρ_i be the largest number of consecutive probes $s_t \in \{s_{i+1}, \ldots\}$, beginning with s_{i+1}, where the index sequence $s_{i-1}, s_i, s_{i+1}, \ldots, s_{i+\rho(i)}$ is strictly monotone (increasing or decreasing). Once x has been located, we may view all subsequent probes as stuck at s_*. Thus, $\rho_i = 0$ if x has already been found within the first i probes, or if s_{i+1} lies between s_{i-1} and s_i. In any case, our definition of $\rho(i)$ will ensure that $s_{i+\rho(i)-1}, s_{i+\rho(i)}, s_{i+\rho(i)+1}$ is not strictly monotone.

Let $\mathcal{P}$ be the random variable that equals number of probes used in the search. We claim that in the case of successful search,

$$\mathcal{P} \leq i - 1 + \rho_i + |s_i - s_{i-1}|, \tag{1}$$

for any $i > 1$. The idea behind the definition of ρ_i is as follows. The probe index s_t splits the search interval $[l_{t-1}+1, u_{t-1}-1]$ into the two pieces $[l_{t-1}+1, s_t-1]$ and $[s_t+1, u_{t-1}-1]$. One of these pieces has the probabilistically small size $|s_t - s_{t-1} - 1|$, and the other might well be much larger. If the value $T[s_t]$ determines that x lies in the (probabilistically) larger piece for $i < t \leq i + \rho_i$, then s_t will belong to a monotone sequence of probes that march monotonically into this "larger" interval. However, $s_{i+\rho_i+1}$ will reverse this direction and lie between $s_{i+\rho_i-1}$ and $s_{i+\rho_i}$.

Thus, the bound for $\mathcal{P}$ counts the first $i - 1$ probes as definite. If there is an i-th probe, then $|s_i - s_{i-1}| \neq 0$, and the cost of this prove is counted in the probabilistic expression $|s_i - s_{i-1}|$. The subsequent ρ_i probes are viewed as a useless increase of $\mathcal{P}$. Probes s_k, for $k > i + \rho_i$, will be to different index locations that lie between $s_{i+\rho_i-1}$ and $s_{i+\rho_i}$. This count is bounded by $|s_{i+\rho_i} - s_{i+\rho_i-1}| - 1 \leq |s_i - s_{i-1}| - 1$, which corresponds to a worst-case exhaustive search.[3]

The claim $\mathrm{E}[\mathcal{P}] < \lceil \log_2 \log_2 n \rceil + 3$ follows by setting $i = \lceil \log_2 \log_2 n \rceil + 1$, applying the expectation to Display (1) and noting that $\mathrm{E}[\rho_i] < 1$, for the following reason. Our median bound for Bernoulli Trials says that the key found at s_t will exceed x with a probability that is below $\frac{1}{2}$, and likewise $T[s_t]$ will be less than x with a probability that is less than $\frac{1}{2}$. Thus, each probe s_{i+1}, s_{i+2}, …, has less than a 50% chance of continuing a streak of wrongly directed probes as counted by ρ_i.

Currently, this argument has a minor technical flaw: the analysis assumes that the interpolated location is always an integer. Section 2.1.3 gives a very simple extension of our median bound, which enables the reasoning about $\rho(i)$ to apply exactly as presented above. Section A.1 extends the analysis of Perl, Itai, and Avni to include rounding.

For completeness, we note that the tight upper and lower bound of $\log_2 \log_2 n \pm O(1)$ for the expected number of probes in both successful and unsuccessful Interpolation Search originates with Yao and Yao [19]. Gonnet, Rogers and George also analyze Interpolation Search [3]. Their study is by no means elementary, and uses extensive asymptotics. Their bound of $\log_2 \log_2 \frac{\pi n}{8} + 1.3581 + O(\frac{\log_2 \log_2 n}{n})$ for successful search is intended to be a very strong asymptotic formulation, and the derivation as presented does not prove that a bound of $\log_2 \log_2 n + 1.3581$, or so, must hold unconditionally (i.e., for all n). For unsuccessful search, Section A.1.2 gives a fairly elementary median-based approach to establish an unconditional upper bound of $\lceil \log_2 \log_2 n \rceil + 4$ for the expected number of probes; Gonnet et al. report a bound of $\log_2 \log_2 \frac{\pi n}{8} + 3.6181$ in this case.

2.1.3. Interpolated medians.

While the issue of rounding computed indices for Interpolation Search cannot possibly change any of the results by more than one probe, or so, the algorithmic implications of rounding are nevertheless worthy of consideration.

One possibility might be to round s_i to the median, but we would not wish to compute, at each iteration, which value is the true median. The work of Jogdeo and Samuels (both implicitly and explicitly) establishes classification results to

[3]We chose to avoid writing the inequality (1) as $\mathcal{P} \leq i + \rho_i + |s_i - s_{i-1}| - 1$ because the random variable $|s_i - s_{i-1}| - 1$ will be negative if x is found in the first $i - 1$ probes. Of course, the undercount in this incorrect interpretation is corrected by the overcount in the unconditional count of i initial probes, but the truth of this observation is, hopefully more evident because of the formulation we used.

determine some cases where the exact median can be distinguished from the candidates $\lfloor np \rfloor$ and $\lceil np \rceil$, but a complete formulation would seem to be unattainable as long as estimation procedures are employed. An algorithmic perspective allows us to simplify the objective so substantially that a complete answer becomes easy.

DEFINITION 2.1. Let p-rnd(x) be the probabilistic rounding

$$\text{p-rnd}(x) = \begin{cases} \lfloor x \rfloor & \text{with probability } 1 - \{x\}, \\ \lceil x \rceil & \text{with probability } \{x\}, \end{cases}$$

where $\{x\}$ is the fractional part $x - \lfloor x \rfloor$.

A very convenient way to round an interpolated probe location s_{i+1} is by replacing the value with p-rnd(s_{i+1}). This assignment preserves the mean, and it is straightforward to verify that

$$\text{E}[|\text{p-rnd}(s_i) - s_{i+1}|] = \text{E}[|\text{p-rnd}(s_i) - \text{p-rnd}(s_{i+1})|].$$

As Corollary 2.1 shows, this rounding scheme is sufficient to guarantee that the probing will be in a probabilistically balanced way. Consequently, an analysis of Interpolation Search with probabilistic rounding turns out to require only a little adaptation of the preceding performance argument, which was somewhat inexact.

COROLLARY 2.1. *Let $X_n = x_1 + x_2 + \cdots + x_n$ be the sum of n independent Bernoulli Trials with mean $\text{E}[X_n] = \mu$. Let $\tilde{\mu} =$ p-rnd(μ).*

Then $$\text{Prob}\{X_n < \tilde{\mu}\} < \frac{1}{2}, \textit{ and } \text{Prob}\{X_n > \tilde{\mu}\} < \frac{1}{2}.$$

Proof. Let $\epsilon = \{\mu\}$. Let y be a Bernoulli Trial that is independent of X_n and has probability of success $1 - \epsilon$, so that $X_n + y$ has an expected outcome of $\lfloor \mu \rfloor + 1$. Then

$$\begin{aligned} \text{Prob}\{X_n \geq \tilde{\mu}\} &= (1-\epsilon)\text{Prob}\{X_n \geq \lfloor \mu \rfloor\} + \epsilon \text{Prob}\{X_n \geq \lfloor \mu \rfloor + 1\} \\ &= \text{Prob}\{X_n + y \geq \lfloor \mu \rfloor + 1\} \\ &> \frac{1}{2}. \end{aligned}$$

The complement event satisfies $\text{Prob}\{X_n < \tilde{\mu}\} < \frac{1}{2}$, and the analogous formulation for $\text{Prob}\{X_n > \tilde{\mu}\}$ follows by symmetry. ∎

2.1.4. The expected size of the Interpolation Search range.

While Interpolation Search converges rapidly toward the search key, it is easy to see that the expected size of the search interval behaves quite poorly. After all, each probe has a remarkably even chance of discovering that the true key location s_* lies to the left or to the right of the current point. Thus, there is a $\Theta(\frac{1}{2^k})$ chance that each of the first k probes will all be to the same side of s_*, so that $\text{E}[n_k] \geq \Theta(\min(\frac{n}{2^k},\ s_1,\ n - s_1))$, which is $\Theta(\frac{n}{\log n})$ for most keys. We avoid a more formal analysis, and choose to rest on the insight provided by this example of elementary median tracking. Evidentially, the left-right distribution of probes does not exhibit asymptotic smoothness because of the overall number of probes is tiny when compared to n.

For completeness, we note that the above comments about the expected size of the search interval do not apply to search keys that should be very near locations 1 or n. For these keys, there is a high probability of termination within a small number of rounds because the number of index locations in the smaller interval of size $\min(s_*, n - s_*)$ will probably be insufficient to sustain a large number of unsuccessful probes.

2.1.5. The Poisson distribution.

Since the Poisson distribution can be attained as a limit of sums of Bernoulli Trials, it follows that Theorem 2.1 also applies to this case. We omit the details.

For completeness, we note that Ramanujan conjectured (in different wording) the following very precise median bound, and used messy asymptotics to attain a partial proof [14].

THEOREM 2.3 (Ramanujan [14]). *Let P_n, for $n > 0$, be a Poisson distributed random variable so that* $\mathrm{Prob}\{P_n = j\} = e^{-n}\frac{n^j}{j!}$, *and* $\mathrm{E}[P_n] = n$, *and suppose that n is an integer. Let* $\frac{1}{2} = \mathrm{Prob}\{P_n < n\} + \alpha_n \mathrm{Prob}\{P_n = n\}$.

Then
$$\frac{1}{3} < \alpha_n < \frac{1}{2}.$$

His original formulation was as an analysis of how well the first n Taylor Series for e^n approximate $\frac{e^n}{2}$. The bound was subsequently viewed in terms of probability, and later served to motivate the exactly analogous generalization to Bernoulli Trials by Jogdeo and Samuels.

2.2. Additional median estimators

The characterization of Theorem 2.1 has natural extensions. While the following formulation may seem somewhat abstract, it has, as the subsequent examples show, natural applications. Loosely stated, the intuition is this. Suppose that Rhoda and George engage in a stochastic race. To make the game fair, there are handicaps. Rhoda wins if she achieves her target score before George, and likewise George wins if he meets his objective first. In the discrete interpretation, which has a relaxed notion of time, both sides win if they meet their objectives by the time the game is over. We characterize some realistic the games where both George and Rhoda win at least half the time.

As an example, let an urn have r red balls and g green balls. Suppose the green balls weigh one ounce, and the red balls weigh gram. We draw 100 balls without replacement. Each ball is selected with a probability that is in proportion to its weight among the balls that remain in the urn. George gets the green balls and Rhoda the red. How can we specify targets for George and Rhoda that sum to 100, and give each a chance of winning that exceeds 50%? The question can be answered fairly well because of the continuous version of this ball game. On the other hand, if each ball is selected based on the outcome of a heterogeneous Bernoulli Trial, and the conditioning is based on a fixed number of balls being selected, different methods seem to be needed.

THEOREM 2.4. *Let X and Y be independent nonnegative random variables with respective probability density functions f and g. Suppose that*

$a1)$ $\mathrm{E}[X] = \mu < \infty$.
$a2)$ $g(t)$ *is maximized at* $\rho \geq \mu$.
$a3)$ g *satisfies the (simplified) right skew condition* $g(t) \leq g(2\rho - t)$, *for* $t \in [0, \rho]$, *and* $g(t)$ *is nondecreasing on* $[0, \rho]$.

Furthermore, suppose that f is continuous and

$b1)$ $f(x) - f(2\mu - x)$, *has at most one zero for t in the open interval* $(0, \mu)$.
$b2)$ $f(0) < f(2\mu)$.

Then
$$\int_0^\infty F(x)g(x)dx > \frac{1}{2}.$$

Requirement $a3)$ can be replaced by the more applicable albeit slightly less general formulation

$a3')$ g *is differentiable and* $\frac{g'(t)}{g(t)} \geq -\frac{g'(2\rho-t)}{g(2\rho-t)} \geq 0$, *for* $t \in (0, \rho)$.

Another alternative to $a2)$ and $a3)$ is

$a2'')$ $\int_0^{2\mu}(1 - G(t) - G(2\mu - t))dt > 0$.

$a3''$) $g(t) + g(2\mu - t)$ is increasing for $t \in (0, \mu)$.
$a4''$) For $t \geq 2\mu$, $g(t) \leq \frac{G(2\mu)}{\mu}$.

Requirement $b1$) can be replaced by the more applicable albeit slightly less general

$b1'$) The density f' is continuously differentiable and $\frac{f'(x)}{f(x)} + \frac{f'(2\mu-x)}{f(2\mu-x)}$ has at most two zeros in $(0, 2\mu)$.

Remarks. The conditions $a2''$), $a3''$) and $a4''$) are intended to capture some cases where g might have a maximum slightly to the left of μ. There are many simple properties that are sufficient to establish condition $a2''$); the integrand, for example, might be nonnegative. Condition $a4''$ is quite mild, has many alternative formulations, and is unlikely to be an impediment in most applications.

Proof (Proof of Theorem 2.4 (Discussion)). The basic strategy for $a1$–$a3$ and $b1, 2$ is similar to that of Theorem 2.1. The idea is to create a unimodal symmetrization of g on $[0, 2\mu]$ via leftward shiftings of area, which is lossy in terms of the integral. Then F can be symmetrized into a moustache $\widehat{F}$ as in Theorem 2.1. Now $\widehat{F}$ can be flattened to its average on $[0, 2\mu]$, which is lossy for the integral as symmetrized. The result is at least $\int_0^\infty \frac{g(x)}{2} dx = \frac{1}{2}$. The exact details plus extensions to the other cases can be found in Section A.2. ∎

2.2.1. Conditional Poisson distributions.

The following is an illustrative application of Theorem 2.4, although the bound can be established by alternative means as explained below.

COROLLARY 2.2. *Let P_r and P_s be independent Poisson distributions with integer means* $\mathrm{E}[P_r] = r$, $\mathrm{E}[P_s] = s$. *Let Q_r be the conditional random variable P_r given that $P_r + P_s = r + s$.*

Then
$$\text{Prob}\{Q_r < r\} < \frac{1}{2} < \text{Prob}\{Q_r \leq r\}.$$

Proof. By symmetry, we need only show that, say, $\text{Prob}\{Q_r \geq r\} > \frac{1}{2}$. Let $P_r(t)$ be a Poisson process with parameter r, so that $\text{Prob}\{P_r(t) = j\} = e^{-rt}\frac{(rt)^j}{j!}$, and $\mathrm{E}[P_r(1)] = r$. Let $P_s(t)$ be an independent Poisson process with parameter s. Let T_r be the stopping time defined by $P_r(T)$ becoming r:
$$T_r = \min_t\{t : P_r(t) = r\},$$
and let
$$T_s^+ = \min_t\{t : P_s(t) = s + 1\}.$$
The intuition behind these definitions is based on a stochastic game. If we play the continuous version of the game via Poisson processes, then the r side wins iff it has already won by the time $P_s(t)$ first reaches the crushing value of $s + 1$.

Since this special (crushing) win time is T_s^+, the statement that the r side wins can be formulated as:

$$\text{Prob}\{Q_r \geq r\} = \text{Prob}\{T_r \leq T_s^+\}.$$

Define the cumulative distribution function $F_r(t) = \text{Prob}\{T_r \leq t\}$, which is the probability that the r side has already achieved its goal by time t. Let $g_s^+(t)$ be the density function $\frac{d}{dt}\text{Prob}\{T_s^+ \leq t\}$, which gives the instantaneous probability that the s side achieves the crushing value of $s + 1$ at time t exactly.

Then the probability that the r side wins is

$$\text{Prob}\{Q_r \geq r\} = \int_0^\infty F_r(t) g_s^+(t) dt,$$

where $F_r(t) = \sum_{j=r}^\infty e^{-rt}\frac{(rt)^j}{j!}$, and $g_s^+(t) = \frac{d}{dt}\sum_{j=s+1}^\infty e^{-st}\frac{(st)^j}{j!} = e^{-st}\frac{(st)^s}{(s-1)!}$.

To apply Theorem 2.4, let X be the stopping time T_r, and Y the stopping time T_s^+.

We calculate that

$$\mathrm{E}[T_r] = \int_0^\infty t dF_r(t) = \int_0^\infty e^{-rt}\frac{(rt)^r}{(r-1)!}dt = 1.$$

The maximum of $g_s^+(t)$ can be found by setting its derivative to zero, which yields the equation $-s + \frac{s}{t} = 0$, or $t = 1$. So $a1$ and $a2$ are satisfied. We now check $a3'$: $\frac{g'(t)}{g(t)} \geq -\frac{g'(2-t)}{g(2-t)} \geq 0$, for $t \in (0,1)$. The first part of the candidate inequality reads: $-s + \frac{s}{t} \geq ?\, s - \frac{s}{2-t}$, or $\frac{1}{t} + \frac{1}{2-t} = \frac{2}{t(2-t)} \geq 2$, which is true for $t \in (0,1)$, since $t(2-t)$ is maximized at $t = 1$, where equality holds. Similarly, it is clear that $s - \frac{s}{2-t} \geq 0$ for $t \in (0,1)$.

The b conditions are straightforward to verify. Theorem 2.4 now guarantees that the median results are as claimed. ∎

For completeness, we note that the memorylessness of Poisson Processes can be used with the Jogdeo-Samuels bound to establish this bound more directly and with slightly more precision by adapting the following viewpoint. The joint process $< P_r(t), P_s(t) >$ is statistically equivalent to arrivals for the single process $P_{r+s}(t)$ with a Bernoulli Trial to type each arrival as being due to P_r or P_s.

For additional completeness, we reiterate that the same perspectives hold for Poisson Processes.

2.2.2. Weighted Selection.

For the following distribution, alternative approaches (apart from a direct attack via fairly messy asymptotics) are by no means evident.

COROLLARY 2.3. *Let an urn contain R red balls and B black balls. Suppose each red ball has weight $w_\circ$, and each black has weight $w_\bullet$. Suppose that the balls are selected one-by-one without replacement where each as yet unselected ball is given a probability of being selected at the next round that equals its current fraction of the total weight of all unselected balls. Suppose r and b are integers that satisfy $r = R(1 - e^{-w_\circ \rho})$, and $b = B(1 - e^{-w_\bullet \rho})$, for some fixed $\rho > 0$.*

Let $r + b$ balls be drawn from the urn as prescribed. Let $X_\circ$ be the number of red balls selected by this random process, and let $X_\bullet$ be the number of black, so that $X_\circ + X_\bullet = r + b$. Then r and and b are the medians of $X_\bullet$ and $X_\circ$:

$$\text{Prob}\{X_\circ > r\} < \frac{1}{2} < \text{Prob}\{X_\circ \geq r\}.$$

Proof (Discussion). It is straightforward to turn this selection process into a time evolution process; each ball is selected according to an independent exponentially distributed waiting time. Thus, the probability that a given red ball is selected by time t is set to $1 - e^{-w_\circ t}$, with density function $w_\circ e^{-w_\circ t}$. Black balls enjoy analogous distributions with weighting rate $w_\bullet$. It is easy to verify that each waiting ball will be selected next with a probability equal to its fraction of the weight in the pool of unselected balls.

Let $X_\circ(t)$ be the number of red balls so selected by time t, and $X_\bullet(t)$ be the number of black. Let $T_\bullet^+$ be the random stopping time when $X_\bullet$ becomes $b+1$: $T_\bullet^+(t) = \min\{t : X_\bullet(t) = b+1\}$.

Section A.3 completes the proof via Theorem 2.4 in a manner very similar to that for Corollary 2.2. ∎

Remarks. When $w_\circ = w_\bullet$, $X_\circ$ will be distributed hypergeometrically, which characterizes selection.

2.2.3. Conditional Bernoulli Trials and Prohibited Outliers: A different approach.

Up to now, we have modeled discrete problems with continuous versions, and have used global formulations that say if g is suitably bell shaped (with some lopsidedness to the right), and F is suitably Error Function-like, and if g is not centered to the left of F, then the integral of gF is at least half the integral of g. This approach uses symmetrization to replace the function $\widehat{F}(t)$ by the value $\frac{1}{2}$, which represents as naive an asymptotic expansion as possible. The next two median derivations deviate from this convenient scenario, and there is indeed a concomitant penalty (which is paid in the Appendix).

THEOREM 2.5. *Let $x_1, x_2, \ldots, x_{l+n}$ be $l + n$ independent Bernoulli Trials where* $\text{Prob}\{x_i = 1\} = p_i$. *Let* $X_n = x_1 + x_2 + \cdots + x_n$, *and let* $Y_l = x_{n+1} + x_{n+2} + \ldots + x_{n+l}$. *Let* $y = \text{E}[Y_l]$, $x = \text{E}[X_n]$, *and suppose that both x and y are integers. Let ζ_n be X_n conditioned on the event $X_n + Y_l = x + y$.*

Then
$$\text{Prob}\{\zeta_n < x\} < \frac{1}{2} < \text{Prob}\{\zeta_n \le x\}.$$

Remarks. When $p_1, p_2, \ldots, p_{l+n}$ are all equal, ζ_n will also be distributed hypergeometrically. Before outlining portions of the proof, we take the liberty of formulating a key step in the proof, which is of interest in its own right.

THEOREM 2.6. *Let $X_n = x_1 + x_2 + \cdots + x_n$ be the sum of n independent (possibly heterogeneously distributed) Bernoulli Trials. Let* $\text{E}[X_n] = \mu$, *and suppose that μ is an integer. Then for any j:*

$$\big|\ \text{Prob}\{-j < X_n - \mu \le 0\} - \text{Prob}\{0 \le X_n - \mu < j\}\ \big| < \frac{1}{2}\text{Prob}\{X_n = \mu\}.$$

Remarks. This bound, which does not appear to be implied by any previous work, also applies to the Poisson distribution, provided equality is allowed. The bound is tight (achieved with equality) for the Poisson distribution with means 1 and 2, and the maximum difference converges, as the mean goes to infinity, to $\frac{e^{-3/2}+1}{3}\text{Prob}\{X_n = \mu\}$, where $\frac{e^{-3/2}+1}{3} \approx .482$. Theorem 2.6 is similar in content (but not proof) to the strong median estimates of Ramanujan, and Jogdeo and Samuels, which can be written as

$$\big|\ \text{Prob}\{-\infty < X_n - \mu \le 0\} - \text{Prob}\{0 \le X_n - \mu < \infty\}\ \big| < \frac{1}{3}\text{Prob}\{X_n = \mu\}.$$

A potential application is as follows. Suppose a probabilistic algorithm starts afresh if the underlying Bernoulli Trials yield a sum that is more than r from the mean, and suppose that this mean is an integer. Then the actual random variable $[X_n | r \ge |X_n - \mu|]$ is no longer Bernoulli, but an analysis can use the median of the unconditioned Trials without error. For completeness, we remark that such a symmetric conditioning about the mean cannot shift the mean, for a sum of Bernoulli Trials, by more than $\frac{1}{5}$, and this size shift occurs only for the Poisson distribution with mean 1 and restriction to the symmetric range comprising 0, 1, and 2.

Proof (Proof of Theorem 2.5 (Discussion)). The first step is to show that the probability of any event for X_n is extremal when p_1 through p_n are restricted to the three values 0, 1, and some fixed α in $(0, 1)$, and similarly p_{n+1} through p_{n+l} are restricted to the three values 0, 1, and some fixed β in $(0, 1)$. The next step is to show that Theorem 2.6 implies Theorem 2.5. See Section A.4.

∎

Proof (Proof of Theorem 2.6 (Discussion)). The chief step is to show that the extreme case, among all Bernoulli Trials with a fixed expectation μ, is always for the Poisson distribution, which reduces the parameters that must be considered to one. The proof is completed by hand calculations when $\mu = 1, 2$, asymptotics for μ big, say $\mu > 20,000$, and a computer program for the rest. A program was written to verify Theorem 2.6 for integer $\mu \le 1000000$. It succeeded because of the gap between the limit value $.482\ldots$ and the bound $\frac{1}{2}$. See Section A.5.

∎

3. ELIMINATING WEAK DEPENDENCIES

The purpose of this paper is, in part, to outline additional easy-to-use estimation procedures for some probabilistic events where the underlying random variables are constrained by weak dependencies.

When a sequence of random variables $x_1, x_2, \ldots, x_n$ is somehow inconvenient to analyze, the desired tail bound or (related) expectation estimate is often computed by replacing the x-s with a simpler probabilistic sequence $y_1, y_2, \ldots, y_n$ that dominates, somehow, the x-s. For example, Hoeffding established the first peakedness properties to show that for a sum of Bernoulli Trials, deviations from the mean are maximum when the Trials (apart from some stuck at one or zero) all have the same probability of success [4]. Hoeffding also showed $\mathrm{E}[f(X)] \le \mathrm{E}[f(Y)]$, for convex f, where X is a sum of values randomly selected without replacement, and Y is the same with replacement [5]. These formulations have been widely generalized (principally for symmetric functions that exhibit some kind of convexity) via a variety of majorization concepts (cf. [8]). Panconesi and Srinivasan gave a related bound for Chernoff-style deviation formulations [10] of the form $\mathrm{E}[e^{tX}] < \lambda \mathrm{E}[e^{tY}]$, where X is a sum of random variables having (essentially negative) correlations that are, up to a fixed factor λ, bounded by the comparable correlations for the independent random variable comprising Y: for integer $\alpha_1, \ldots \ge 0$, $\mathrm{E}[\prod_i x_i^{\alpha_i}] \le \lambda \mathrm{E}[\prod_i y_i^{\alpha_i}]$.

We now suppose that $x_1, x_2, \ldots, x_n$ is a sequence of random variables such as Bernoulli Trials, and suppose that Z is another random variable that is dependent on and is increasing in the x_i-s. Let $f(x_1, \ldots, x_n)$ be a non-negative n-ary function that is non-decreasing in each coordinate, and let F_k be defined as the conditional random variable $f(x_1, \ldots, x_n)$ given that $Z = k$. The value of interest is $\mathrm{E}[F_{k_0}]$, where k_0 has some fixed value that need not equal $\mathrm{E}[Z]$, even though that is the only case we discuss. A typical formulation for f might be as a Chernoff-Hoeffding bound to estimate a tail event.

A (possibly promising) procedure for estimating a tail bound of the form $\mathrm{E}[F_{k_0}]$ is: ignore the conditioning by estimating $\mathrm{E}[f(x_1, x_2, \ldots, x_n)]$, and multiply the resulting bound by some modest constant to account for the conditioning. The basic proof strategy is as follows.

Since

$$\mathrm{E}[[f(x_1, \ldots, x_n)|Z \ge k_0]] = \frac{\mathrm{E}[f(x_1, \ldots, x_n) \cdot \chi\{Z \ge k_0\}]}{\mathrm{Prob}\{Z \ge k_0\}},$$

and

$$\frac{\mathrm{E}[f(x_1, \ldots, x_n) \cdot \chi\{Z \ge k_0\}]}{\mathrm{Prob}\{Z \ge k_0\}} \le \frac{\mathrm{E}[f(x_1, \ldots, x_n)]}{\mathrm{Prob}\{Z \ge k_0\}},$$

the hypothesis

$$\mathrm{E}[[f(x_1, \ldots, x_n)|Z = k_0]] \le ?\, \mathrm{E}[[f(x_1, \ldots, x_n)|Z \ge k_0]] \tag{2}$$

will imply that

$$\mathrm{E}[[f(x_1, \ldots, x_n)|Z = k_0]] \le ?\, \frac{\mathrm{E}[f(x_1, \ldots, x_n)]}{\mathrm{Prob}\{Z \ge k_0\}}.$$

Thus, bounding $\mathrm{E}[F_{k_0}]$ (from above) can be reduced to estimating the unconditioned $\mathrm{E}[f]$ (from above), estimating $\mathrm{Prob}\{Z \ge k_0\}$ (from below), and showing that F_k is non-decreasing (in some sense). For many applications, the underlying events may represent large deviations, and it may, therefore, suffice to establish these inequalities up to constant factors.

In practice, we might begin with a dependent sequence $w_1, w_2, \ldots, w_n$ of random variables that is statistically equivalent to a sequence $x_1, x_2, \ldots, x_n$ of independent random variables conditioned on some event $Z = k_0$. For example, the w-s might be the multinomial distribution for distributing n balls among n bins, which is statistically equivalent to n independent identically distributed Poisson distributions that are conditioned to sum to n.

The inequality in Display (2) can often be established by a variety of approaches. Sometimes the x's can be modeled in two ways – as specific random variables, or as the outcome of an increasing random process that evolves over time. In these cases, the notion of growth over time can give a direct proof of the inequality. In the previous example, for instance, the Poisson distributions $x_1, x_2, \ldots, x_n$ can be modeled by Poisson processes $x_1(t), x_2(t), \ldots, x_n(t)$ with arrival rate λ, and the stopping time T defined to be the first time t where $x_1(t) + x_2(t) + \cdots + x_n(t) = n$. Let $\tau = \mathrm{E}[T]$. Since the resulting distribution is independent of the value of T, we can condition on the subset where $T \le \tau$, so that

$$\mathrm{E}[f(w_1, w_2, \ldots, w_n)] = \mathrm{E}[[f(x_1(T), \ldots, x_n(T))|T \le \tau]].$$

Of course,

$$\mathrm{E}[[f(x_1(T), \ldots, x_n(T))|T \le \tau]] \le \mathrm{E}[[f(x_1(\tau), \ldots, x_n(\tau))|T \le \tau]].$$

Finally,

$$\mathrm{E}[[f(x_1(\tau),\ldots,x_n(\tau))|T \le \tau]] = \frac{\mathrm{E}[f(x_1(\tau),\ldots,x_n(\tau)) \cdot \chi\{T \le \tau\}]}{\mathrm{Prob}\{T \le \mathrm{E}[T]\}},$$

which shows that the weakly dependent w_i's can be replaced by independent Poisson random variables with mean 1, provided the resulting bound is doubled.

More abstractly, Inequality (2) can sometimes be established by identifying a mapping σ from a subset of the probability space P, for the unconstrained (or suitably conditioned) x_i-s, onto the probability space $\hat{P}$ defined by the x_i-s conditioned on the event $Z = k_0$. Typically, the domain is the x_i-s where $Z \ge k_0$. The requirement for the mapping is that for $\bar{x} \in \hat{P}$, for $\bar{y} \in \sigma^{-1}(\bar{x})$: $f(\bar{x}) \le f(\bar{y})$, and $\mathrm{Prob}_{\hat{P}}\{\bar{x}\} \le \alpha \mathrm{Prob}_P\{\sigma^{-1}(\bar{x})\}$, for some fixed constant α. The time evolution view, when it applies, gives this mapping implicitly: σ just maps a time evolution process back to its value at the earlier stopping time $T = \min\{t : Z(t) = k_0\}$.

There are also many other methods where simpler random variables can sometimes be substituted in quantifiable ways to fulfill the function of a more complex family of random variables.

Perhaps the most noteworthy case where time evolution does not seem to apply is when the x_i's are heterogeneous Bernoulli Trials, and Z is the sum of the x_i's. See Section A.8. In any case, if Inequality (2) can be established, the resulting estimation problems are unconditioned, and therefore more amenable to analysis by a larger class of standard probabilistic tools. If $k_0 = \mathrm{E}[Z]$, then the mean-median theory might give a convenient factor of 2. If not, it may be possible to (non-linearly) rescale the underlying random variables into (statistically) equivalent ones where the new Z is conditioned to be at its mean. This is easily done, for example, for Bernoulli Trials, which turns out to yield new Bernoulli Trials. See Section A.9.

Lastly, it is worth noting that if f is the step function

$$f(x_1,\ldots,x_n) = \begin{cases} 0, & \text{if } x_1 + \cdots + x_n < k_0; \\ 1, & \text{otherwise,} \end{cases}$$

then the bound is tight. Consequently, this formulation requires the factor $\frac{1}{\mathrm{Prob}\{Z \ge k_0\}}$, although it may actually be unnecessary for many applications.

4. OTHER ALGORITHMIC APPLICATIONS

4.1. Concurrent sorting for PRAM emulation

Ranade showed how to emulate a Common PRAM via an $n \times 2^n$ butterfly network of processors [15]. The emulation scheme is roughly structured as follows.

1 Each of the $n2^n$ processors constructs a memory request.

2 Each row of n processors performs a systolic (parallel) bubblesort to sort its n requests.

3 The first processor in each row injects its row's n requests into the routing network, smallest address first. The messages are routed to the correct row, whence they pass along the row to the correct column.

4 Each return message backs up along its path to its home processor.

In this algorithm, an $n2^n$-processor machine has a running time, for emulating one parallel step of an $n2^n$-processor Common PRAM, that is $\Theta(n)$ with very high probability and on average.

In the fast hashing work [16], steps 1,3, and 4 were shown to work in $O(n)$ expected time for a machine comprising $n \times 2^n$ switches, but with only one column of 2^n processors, where each processor runs n virtual processes, so that the $n2^n$ parallel computations can still be executed in $O(n)$ time. However, step 2 was not done locally by each individual processor when combining was needed. Rather, each processor used its row of n smart switches to perform a bubblesort as in Ranade's original scheme. We now show that step 2 can also be done locally, with the slowest of the 2^n sortings completing in $O(n)$ expected time (and with high probability).

The sorting is accomplished in three steps, where all memory accesses and processing are local to each individual processor.

Step 1) Hashing is used to group together the memory references to identical locations. This is necessary because combining operations generate common addresses that are therefore not independent. The common references within each pipeline are then combined.

This step runs in $O(n)$ time, both in the expected case and with high probability. The requisite resources for this step also turn out to be modest.

Step 2) Each processor's set of memory references (with one representative for each of the $k \le n$ combined address references) is partitioned among n local buckets. Simply stated, a reference with the hashed address (t, x) with $t \in [0, n-1]$ and $x \in [0, 2^n m - 1]$ (where m is the number of words per memory module), is placed in $Bucket[t]$.

This step clearly runs in $O(n)$ time.

Step 3) The contents of each Bucket (with an expected count of just $O(1)$) are then sorted locally with an efficient sorting algorithm such as mergesort. Lastly, the n sorted sequences are concatenated to produce the desired result.

Step 3 requires a performance proof, since we must assert that each of the 2^n processors can complete its individual task of sorting its own n numbers in $O(n)$ time.

THEOREM 4.1. *Let $[0, m-1)^n$ comprise all sequences of n values selected uniformly from $[0, m-1)$. Then for any fixed $c > 0$, there is a fixed d such that Step 3 sorts a fraction exceeding $1 - 2e^{-cn}$ of all $D \in [0, m-1)^n$ in dn steps, where $c = d - 2 - \ln d$.*

Proof. Given an instance $D \in [0, m-1)^n$, let b_j, for $j = 1, 2, \ldots, n$, be the number of items in D that are placed in $Bucket[j]$ by step 2). Then the running time for the n calls to a fast sorting algorithm is proportional to $T(D) = n + \sum_j \ln(b_j!)$. Let g be the generating function for $\mathcal{T} \equiv T - n$:

$$g(z) = \mathrm{E}[z^{\mathcal{T}}] \equiv \frac{1}{n^n} \sum_{\alpha_1 + \ldots + \alpha_n = n} \binom{n}{\alpha_1, \alpha_2, \ldots, \alpha_n} z^{\sum_i \ln(\alpha_i!)}.$$

Standard formulations of Chernoff probability estimates give

$$\mathrm{Prob}\{\mathcal{T} > dn\} \le e^{-\lambda dn} g(e^\lambda),\ \lambda \ge 0,$$

and we seek a λ that (roughly) minimizes $e^{-\lambda dn} g(e^\lambda)$, and need an estimate for the resulting value.

Let $\zeta_1, \zeta_2, \ldots, \zeta_n$ be n independent identical Poisson distributions with mean 1, so that $\mathrm{Prob}\{\zeta_j = \ell\} = \frac{e^{-1}}{\ell!}$. Finally, define $V = \sum_{j=1}^n \ln(\zeta_j!)$, and let $G(z) = \mathrm{E}[z^V]$. We have replaced the random mapping of n items into n buckets (where $Bucket[j]$ receives b_j items with $\mathrm{E}[b_j] = 1$) by n independent assignments where $Bucket[j]$ has ζ_j items, with an average count that is still $\mathrm{E}[\zeta_j] = 1$ item per bucket, for $j = 1, 2, \ldots, n$. As is well known, the ζ_j's and the b_j's have the same joint statistics if we condition the ζ's on the event $\sum_{j=1}^n \zeta_j = n$. Since n is the expectation for the sum of the n ζ_i-s, and the mean is the median for the Poisson distribution, we can simply compute the tail bound for the ζ_i-s and double the resulting bound since the expected work is increasing in the total number of bucket items.

Since the unconditioned ζ's are i.i.d., we have:

$$G(e^\lambda) = \mathrm{E}[e^{\lambda \ln(\zeta_1!)}]^n = \Big(\frac{1}{e} \sum_{j=0}^{\infty} \frac{(j!)^\lambda}{j!}\Big)^n.$$

Hence G can be used in the overestimate

$$\frac{1}{2}\mathrm{Prob}\{T - n > dn\} < \Big(e^{-\lambda d} \frac{1}{e} \sum_{j=0}^{\infty} \frac{1}{(j!)^{1-\lambda}}\Big)^n.$$

We claim that for $0 \le \lambda \le 1$,

$$\sum_{j=0}^{\infty} \frac{1}{(j!)^{1-\lambda}} \le \frac{1}{1-\lambda} \sum_{j=0}^{\infty} \frac{1}{j!} = \frac{e}{1-\lambda}.$$

The proof is given in Lemmas 6.1 and 6.2 of Section A.6. Now let λ minimize the estimate $(\frac{e^{-\lambda d}}{1-\lambda})^n$, whence $d = \frac{1}{1-\lambda}$. Substituting gives

$$\frac{1}{2}\text{Prob}\{T - n > dn\} < (de^{1-d})^n,$$

so $c = d - 2 - \ln d$. ∎

Interestingly, quadratic sorting, for each of these $O(1)$ (expected) sized data sets is just too slow. Fairly large deviations will probably occur in too many local bins for some of the 2^n sorting experiments.

For completeness, we observe that a little more work can eliminate the factor of 2 in Theorem 4.1. In particular, Lemma 7.1 of Section A.7 shows that the function $f(k) = \text{E}[z^{\sum_{j=1}^{n}\ln(\zeta_j)} \mid \zeta_1 + \cdots + \zeta_n](k)$ where $\zeta_1 + \cdots + \zeta_n = k$, is convex in k, so that Jensen's Inequality can be used as a substitute for the mean-median step in the derivation to attain the same bound without the extra factor of 2.

4.2. Double Hashing

In an analysis of double hashing by Lueker and Molodowitch [7], a subproblem arose that can be abstracted as follows. Let S be a collection of some subsets of $U \equiv \{x_1, x_2, \ldots, x_n\}$, where each element x_i is a Boolean variable. Interpret each $s \in S$ as a Boolean conjunction over its elements. Compute a tail bound for the number of subsets that evaluate to true, subject to the weak dependency condition that exactly m of x_i are true, with all possibilities equally likely. Their specific function count was not symmetric in the x_i's but was trivially monotone. In this case, a very easy first step is to exploit mean-median bounds by allowing each x_i to be true with probability m/n, and then compute (via additional observations) a tail bound for the Bernoulli case, and finally double the resulting bound. Instead, Lueker and Molodowitch used a Chernoff-Hoeffding estimate to make each x_i a Bernoulli Trial with probability $\frac{m}{n}(1 + \sqrt{\frac{\log n}{n}})$, or so, and accounted for the error where fewer than m of the x's are set to 1. While this is not at all difficult, it was a bit more trouble and slightly (but immaterially) less sharp.

5. CONCLUSIONS

This paper may have a bit more than doubled the number of non-trivial cases where the mean of a probability distribution is known to be the median, and has given a framework for establishing strong median estimates. We have also shown how to use median estimates to get yet another approach for attaining large deviation bounds for problems with weak dependencies. The results have, en passant, also contributed to the structural analysis of such basic probability distributions as Bernoulli Trials.

In a sense, divide-and-conquer was used mostly as an analytic tool, as opposed to an algorithmic device. From this perspective, the target of mean-median results can be seen to be much too restrictive. For example, approximate medians that are $\frac{1}{4} - \frac{3}{4}$ splitters of the probability measure would be perfectly adequate for the performance bounds as presented. On the other hand, the methods in this paper can be used to attain such results – sometimes quite easily – and median bounds can sometimes be combined to attain multiplicatively weaker estimates for more complex processes.

Finally, these methods were applied in several performance analyses that were new, either in content or simplicity. We also note that median bounds might find application in statistics, since median computations can be less sensitive to outliers than means.

APPENDIX

A.1. ANALYZING INTERPOLATION SEARCH

This subsection presents a mildly enhanced version of the first part of an analysis originally given by Perl, Itai, and Avni [11]. In particular, it includes the issue of rounding, distinguishes between successful and unsuccessful search, and does not require the algorithm to abandon logical interpolation at any point.

It should be noted that the following derivation would be simplified if the issue of rounding were ignored. Probably the easiest way to accommodate rounding is as an after the fact modification that adds 1 to the right hand side of the recurrence relation A.2 below. Such an approach would give a bound that is about 1.5 probes more than the analysis we now present.

Let the search parameters be as defined in Section 2.1.2. A probing under the assumption of successful search would require that at round t, one of the n_t keys be conditioned to be x, and take the rest to be random. A probing that expects the search to be unsuccessful would allow all n_t to be random. For this latter case, we can imagine a (virtual) key x to be in location $j + \frac{1}{2}$, where j is the unique index satisfying $T[j] < x < T[j+1]$. According to the assumptions underlying the two search schemes, the next interpolated probe location, which is computed at round t and probed at round $t+1$, should be, provided x has not been found in the first t probes,

$$s_{t+1} = l_t + p_t n_t + \frac{1}{2}$$

for unsuccessful search, and

$$s_{t+1} = l_t + p_t(n_t - 1) + 1$$

in the case of successful search.

Actual probing will be probabilistic when s_t is not an integer, so that $\lfloor s_t \rfloor$ is probed with probability $1 + \lfloor s_t \rfloor - s_t$, and $\lfloor s_t \rfloor + 1$ is probed with probability $s_t - \lfloor s_t \rfloor$. Let s_t^r denote a probe that includes the probabilistic rounding of the exact interpolation calculation s_t, and let $\epsilon_t = s_t^r - s_t$, so that ϵ is the probabilistic rounding adjustment for s_t. By construction, $\mathrm{E}[\epsilon_t] = 0$.

In the case of unsuccessful search, the interpolation scheme can yield probe locations satisfying $s_{t+1} - l_t < 1$ or $u_t - s_{t+1} < 1$, where probabilistic rounding would be senseless. In these circumstance, s_{t+1} must be rounded to the nearest integer. This issue is revisited at the end of this subsection.

An advantage of probabilistic rounding is that it does not disturb our basic expectations:

$$\mathrm{E}[|s_{t+1} - s_t^r| \;\Big|\; s_t^r, p_{t-1}, n_{t-1}] = \mathrm{E}[|s_{t+1}^r - s_t^r| \;\Big|\; s_t^r, p_{t-1}, n_{t-1}],$$

because adding the rounding fraction ϵ_{t+1} cannot change the sign of $s_{t+1} - s_t^r$ from positive to negative nor vice versa. Thus, the average of the absolute value of $s_{t+1} + \epsilon_{t+1} - s_t^r$ is simply the average of the positive values minus the average of the negative values, and each average has a zero contribution from ϵ_{t+1}. Furthermore, Corollary 2.1 ensures that such a rounding yields a good probabilistic splitting of the search interval.

Let the "true" location of x be s_* as quantified by each scheme. Of course, s_* is a random variable, which equals one (or $\frac{1}{2}$) plus the number of data items in the table $T[1..n]$ that are less than x. Similarly, it is convenient to let $\hat{n}_t$ be n_t or $n_t - 1$, depending on whether the search is designed for failure or success, and let Δ be $\frac{1}{2}$ or 1 as dictated by the respective schemes. Let ϵ_t be the rounding increment as used by each algorithm. For t indices that are larger than the number of probes actually used, we adopt the convenience of setting $\hat{n}_t$ to zero, and freeze all such l_t, u_t and s_t to equal the last location actually probed. Finally, let $F(t)$ be 1 if x has not yet been located (or not yet determined to be absent from the table) in the first t probes, and 0 otherwise. Given these definitions, the interpolation scheme can be write written as:

$$s_{t+1} = l_t + p_t \hat{n}_t + \Delta_t F(t).$$

Let the state information $S(t)$ comprise the values l_τ, u_τ, $T[l_\tau]$, $T[u_\tau]$ and the derived interpolation fraction p_τ as well as the computed index s_τ, Δ_τ and rounding fraction ϵ_τ, for the rounds τ, where $0 \le \tau \le t$. At round $t+1$, of course, ϵ_{t+1} is evaluated to compute s_{t+1}^r. Then $T[s_{t+1}^r]$ is probed and either the algorithm terminates, or s_{t+1}^r is assigned to one of l_{t+1} or u_{t+1} and the other variables are updated according to the new state information.

Perl, Itai, and Avni use the fact that the error $(s_{t+1} - s_*)$ in the index selection is just the difference between number of items, among the $\hat{n}_t$, that are actually less than x, and the expected count. Furthermore, this difference is statistically the same as the difference between $\hat{n}_t p_t$ and the sum of $\hat{n}_t$ Bernoulli trials, each with probability of success p_t.

Thus, $(s_{t+1} - s_*)^2$ is, in expectation, the variance for the sum of $\hat{n}_t$ Bernoulli trials, each with mean p_t, so that

$$\mathrm{E}[(s_{t+1} - s_*)^2 | S(t)] = \hat{n}_t p_t (1 - p_t).$$

Similarly,

$$\mathrm{E}[(s^r_{t+1} - s_*)^2 | S(t)] = \hat{n}_t p_t (1 - p_t) + var[\epsilon_{t+1}] \leq \hat{n}_t p_t (1 - p_t) + \frac{1}{4} F(t),$$

since the variances of independent random variables are additive, and the variance of a Bernoulli Trial is bounded by $\frac{1}{4}$. The factor $F(t)$ is a 0–1 multiplier to account for the fact that ϵ_t will not exist if the search has already terminated.

On the other hand,

$$s_{t+1} = \mathrm{E}[s_* | S(t)],$$

since s_{t+1} is by construction and by definition the expected location for the search key x when the known information comprises the state data in $S(t)$.

We need to use a fundamental fact about conditional expectations, which is that $\mathrm{E}[\mathrm{E}[f|S(t)]|S(s)] = \mathrm{E}[f|S(s)]$, for $s < t$. Further, for any function H, $\mathrm{E}[H(q_t)|S(t)] = H(q_t)$ when q_t is known information in the state set $S(t)$.

We also exploit Jensen's Inequality, which states that for any convex function G, such as $G(x) = x^2$ or $G(x) = |x|$, $\mathrm{E}[G(f)|S] \geq G(\mathrm{E}[f|S])$.

Putting all this together gives:

$$\begin{aligned}
\mathrm{E}[|s^r_t - s_{t+1}| \Big| S(t-1)]^2 &= \mathrm{E}[|s^r_t - \mathrm{E}[s_*|S(t)]| \Big| S(t-1)]^2 \\
&= \mathrm{E}[|\mathrm{E}[s^r_t - s_*|S(t)]| \Big| S(t-1)]^2 \\
&\leq \mathrm{E}[\mathrm{E}[|s^r_t - s_*| \Big| S(t)] \Big| S(t-1)]^2 \\
&\leq \mathrm{E}[|s^r_t - s_*| \Big| S(t-1)]^2 \\
&\leq \mathrm{E}[(s^r_t - s_*)^2 \Big| S(t-1)] \\
&\leq \hat{n}_{t-1} p_{t-1} (1 - p_{t-1}) + \frac{1}{4} F(t-1).
\end{aligned}$$

Now, s^r_{t-1} is either l_{t-1} or u_{t-1}; hence $|s^r_{t-1} - s_t|$ is either $p_{t-1}\hat{n}_{t-1} + \Delta F(t-1)$ or $(1 - p_{t-1})\hat{n}_{t-1} + \Delta F(t-1)$. Consequently,

$$\hat{n}_{t-1} p_{t-1} (1 - p_{t-1}) + \frac{1}{4} F(t-1) \leq |s^r_{t-1} - s_t|.$$

Thus,

$$\mathrm{E}[|s^r_t - s_{t+1}| \Big| S(t-1)]^2 \leq |s^r_{t-1} - s_t|. \tag{A.1}$$

Taking the expectation of both sides, letting

$$d_t = \mathrm{E}[|s^r_t - s^r_{t+1}|],$$

recalling that

$$\mathrm{E}[|s^r_t - s^r_{t+1}|] = \mathrm{E}[|s^r_t - s_{t+1}|],$$

and applying Jensen's inequality once more gives:

$$\begin{aligned}
d_t^2 &= \mathrm{E}[|s^r_t - s_{t+1}|]^2 = \mathrm{E}[\mathrm{E}[|s_t - s_{t+1}| \Big| S(t-1)]]^2 \\
&\leq \mathrm{E}[\mathrm{E}[|s_t - s_{t+1}| \Big| S(t-1)]^2] \leq \mathrm{E}[|s^r_{t-1} - s_t|] = d_{t-1},
\end{aligned}$$

so that

$$d_t^2 \leq d_{t-1}. \tag{A.2}$$

A.1.1. Successful search

For successful search, we can take $d_0 = \frac{n}{4}$, since $d_1^2 \leq \hat{n}_0 p_0(1-p_0) + \frac{1}{4} \leq n_0 \frac{1}{2}(1-\frac{1}{2})$. Writing $n = 2^h$, we see that after $\log_2 h$ iterations, we attain $d_{\log_2 \log_2 n} \leq 2$, which shows that

$$\mathrm{E}[|s^r_{\log_2 \log_2 n} - s^r_{1+\log_2 \log_2 n}|] < 2,$$

which includes, in this averaging, ranges of size zero due to earlier termination.

As before, let $i_0 = \log_2 \log_2 n + 1$. The reasoning of Section 2.1.2 applies to bound the expected number of probes for successful search in this more detailed analysis, and gives, as we now show, the same bound of

$$\lceil \log_2 \log_2 n \rceil + 3.$$

The only issue that remains to be addressed is the effect of rounding at the $(\log_2 \log_2 n + 1)$-th probe, because a rounding that increases the size of $|s_{i_0} - s^r_{i_0 - 1}|$ is more likely to have x in the interval of indices delineated by these two probe values. While it suffices to increase the resulting probe estimate by 1, no increase is needed.

The reasoning is as follows. Let

$$F_\xi(t) = \chi\{x \text{ is not found in the first } t-1 \text{ probes}\},$$

which indicates if a t-th actually probe takes place. Let the random variable τ to be the smallest index $\geq i_0$ for which $s^r_{\tau+1}$ lies between s^r_τ and $s^r_{\tau-1}$, or $s^r_\tau = x$. By definition, τ is i_0 if the search terminates prior to the i_0-th probe.

For expositional simplicity, suppose that $s_{i_0-1} < s_{i_0}$, where $i_0 = 1 + \lceil \log_2 \log_2 n \rceil$. Given these formulations, and assuming that $s_{i_0-1} < s_{i_0}$, the number of probes is bounded by

$$i_0 - 1 + F_\xi(i_0) + (\tau - i_0) + 0 \cdot \chi\big\{(s^r_\tau = \lfloor s_\tau \rfloor + 1) \text{ and } (T[s^r_\tau] = x)\big\} + (|s^r_\tau - s^r_{\tau-1}| - F_\xi(\tau))\chi\{T[\lfloor s_\tau \rfloor] \geq x\}.$$

This formulation is correct because first, the count is at least $\tau - 1 + F_\xi(i_0)$ if x is found among the first τ probes. Second, the $T[s_\tau]$ must be greater than x if the next probe location is to be less than s_τ. Then the reasoning is exactly as before. The decomposition via χ serves to strip away the dependency of the rounding for s^r_τ from the definition of τ. The point is that $\mathrm{E}[(|s^r_\tau - s^r_{\tau-1}| - F_\xi(\tau))\chi\{T[\lfloor s_\tau \rfloor] \geq x\}] = \mathrm{E}[(|s_\tau - s^r_{\tau-1}| - F_\xi(\tau))\chi\{T[\lfloor s_\tau \rfloor] \geq x\}]$, because both candidate indices for the rounding of s_τ have values that are larger than x. Hence the definition of τ does not condition the rounding of s_τ. Taking expectations of relevant terms gives:

$$\begin{aligned}
\mathrm{E}[\tau - i_0] &< 1; \\
\mathrm{E}[(|s^r_\tau - s^r_{\tau-1}| - F_\xi(\tau))\chi\{T[\lfloor s_\tau \rfloor] \geq x\}] &= \mathrm{E}[(|s_\tau - s^r_{\tau-1}| - F_\xi(\tau))\chi\{T[\lfloor s_\tau \rfloor] \geq x\}] \\
&\leq \mathrm{E}[(|s_\tau - s^r_{\tau-1}| - F_\xi(\tau))] \\
&\leq \mathrm{E}[|s_{i_0} - s^r_{i_0-1}|] - F(i_0 - 1),
\end{aligned}$$

where the very last line follows from the inequality $|s_\tau - s^r_{\tau-1}| - F_\xi(\tau) \leq |s_{i_0} - s^r_{i_0-1}| - F(i_0 - 1)$. Thus, the expected number of probes is bounded by $i_0 - 1 + \mathrm{E}[\tau - i_0] + \mathrm{E}[|s_{i_0} - s^r_{i_0-1}|] < \lceil \log_2 \log_2 n \rceil + 3$ as before.

A.1.2. Unsuccessful search

For unsuccessful search, the reasoning is almost the same as for successful search. Only two adaptations must be made to the derivation, which was conceptually organized as

$$i_0 - 1 + F(i_0 - 1) + \rho_{i_0} + (d_{i_0-1} - F(i_0-1)).$$

The first problem is that, in the case where balanced rounding occurs, it is not true that the probe to s_t^r will find a value that has, say, at least a 50% chance of being greater than x. Rather, it has at least a 50% chance of being greater than or equal to the largest entry less than x. This issue can be resolved by adding 1 to our previous estimate for the expected number of additional probes necessary to determine, after probe i_0, that s_* is between the two most recent probe locations.

Second, the issue of unbalanced rounding must be addressed. Suppose, for example, that the search interval comprises the n unprobed locations $3, 4, \ldots, n+2$ with values bounded by 5.1 and 6. The next probe location is $s = 2 + n\frac{x-5.1}{.9} + \frac{1}{2}$. If $s < 3$, probabilistic rounding can give location 2, which has already been probed. The rounding must be to 3, in this case. In reality, this rounding is advantageous, since it moves the computed index in a direction that will hasten termination.

A brief formalization of this observation is as follows. According to the interpolation scheme for unsuccessful search, the probe s_{t+1}^r will have such an unbalanced rounding if and only if $n_t p_t < \frac{1}{2}$, or $n_t(1-p_t) < \frac{1}{2}$. It is easy to see that in such a case, s_{t+1}^r will be the terminating probe precisely when no value among the n_t is less than x if $n_t p_t < \frac{1}{2}$, or when no value among the n_t is greater than x if $n_t(1-p_t) < \frac{1}{2}$. Thus, the probability that s_{t+1} is the terminating probe (conditioned on termination not having occurred earlier and the rounding being unbalanced) is $(1 - \frac{n_t p}{n_t})^{n_t}$, where $p = \min(p_t, 1-p_t)$. Now, $(1 - \frac{n_t p}{n_t})^{n_t} \geq \frac{1}{2}$ because the expression is increasing in the variable n_t when $n_t p$ is held fixed, and because $n_t p$ is bounded by $\frac{1}{2}$. Consequently, each such probe will terminate with a probability that is at least $(1 - \frac{1}{2n_t})^{n_t} \geq (1 - \frac{1}{2\cdot 1})^1 = \frac{1}{2}$, which is good. Moreover, once unbalanced rounding occurs, it continues for each subsequent probe, and the distance between consecutive probes, in these cases, will be fixed at 1, up to termination. Evidentially, termination occurs, on average, in less than two such steps, since the probability of completion is so high.

We can apply this observation to the first $i_0 + 1$ probes. Let $U(t)$ be the event that an unbalanced rounding occurs among the computations for the first t probes. Let $u = \text{Prob}\{U(i_0+1)\}$. Let $\mathcal{P}$ be the probe count. Then $\mathrm{E}[\mathcal{P} \cdot \chi\{U(i_0+1)\}] \leq (i_0+2)u$.

As for $\mathrm{E}[\mathcal{P} \cdot \chi\{\text{not}\, U(i_0+1)\}]$, our previous counting applies to get the overestimate $(i_0+2)(1-u)+1$. The (i_0+2) represents i_0 probes followed by an expected count of 2 to get three consecutive probes where the last lies between the previous two. The last term is a bound for $\mathrm{E}[|s_{i_0} - s_{i_0-1}|\chi\{\text{not}\, U(i_0+1)\}]$, which comes from setting $d_t = \mathrm{E}[|s_t^r - s_{t+1}^r|\chi\{\text{not}\, U(t+1)\}]$. This d_t satisfies the same recurrence as its predecessor; nothing changes. Notice, however, that there is no factor of $1-u$ in the answer.

The resulting sums give $\mathcal{P} < \lceil \log_2 \log_2 n \rceil + 4$.

For completeness, we note that $\mathrm{E}[|s_{i+1} - s_*| \mid S(i)]$ can also be computed directly by straightforward elementary combinatorics to get: $\mathrm{E}[|s_{i+1} - s_*| \mid S(i)] = 2(\hat{n}_i)p_i(1-p_i)\binom{\hat{n}_i - 1}{\lfloor \hat{n}_i p_i \rfloor} p_i^{\lfloor \hat{n}_i p_i \rfloor}(1-p_i)^{\hat{n}_i - 1 - \lfloor \hat{n}_i p_i \rfloor}$, which can give an (insignificant) improvement to the recurrence for d_i. On the other hand, tighter bounds for the probe count would have to come from an analysis such as that given by Gonnet et al, where recurrences are derived that account for the possibility of termination at each step of the probe procedure.

A.1.3. Computing $\mathrm{E}[|s_i - s_*| \mid S(i-1)]$ exactly

For convenience, we change notation and compute $\mathrm{E}[|X_n - np|]$, where X_n is the sum of n independent Bernoulli Trials which each have probability p of success.

Since $\mathrm{E}[X_n - np] = 0$, the negative and positive contributions must have the same magnitude, whence $\mathrm{E}[|X_n - np|] = 2\mathrm{E}[X_n - np \text{ and } X_n > np]$. So

$$\mathrm{E}[|X_n - np|] \;=\; 2\sum_{j>np}^{n} \binom{n}{j}(j-np)p^j(1-p)^{n-j}$$

$$\begin{aligned}
&= 2\sum_{j>\lfloor np\rfloor}^{n}\binom{n}{j} j p^j(1-p)^{n-j} - 2np\sum_{j>\lfloor np\rfloor}^{n}\binom{n}{j}p^j(1-p)^{n-j}\\
&= 2n\sum_{j>\lfloor np\rfloor}^{n}\binom{n-1}{j-1}p^j(1-p)^{n-j} - 2np\sum_{j>\lfloor np\rfloor}^{n}\left(\binom{n-1}{j-1}+\binom{n-1}{j}\right)p^j(1-p)^{n-j}\\
&= 2(1-p)n\sum_{j>\lfloor np\rfloor}^{n}\binom{n-1}{j-1}p^j(1-p)^{n-j} - 2np\sum_{j>\lfloor np\rfloor}^{n}\binom{n-1}{j}p^j(1-p)^{n-j}\\
&= 2pn(1-p)\sum_{j>\lfloor np\rfloor}^{n}\binom{n-1}{j-1}p^{j-1}(1-p)^{n-j} - 2np(1-p)\sum_{j>\lfloor np\rfloor}^{n}\binom{n-1}{j}p^j(1-p)^{n-j-1}\\
&= 2pn(1-p)\binom{n-1}{\lfloor np\rfloor}p^{\lfloor np\rfloor}(1-p)^{n-1-\lfloor np\rfloor}.
\end{aligned}$$

While this derivation is simple, it is not clear how easily the resulting expression can be bounded by $\sqrt{np(1-p)}$. The following naive but slightly tedious derivation is probably far from the simplest. It gives, for $n > 1$, the rather meticulous estimate:

$$2pn(1-p)\binom{n-1}{\lfloor np\rfloor}p^{\lfloor np\rfloor}(1-p)^{n-1-\lfloor np\rfloor} \le \sqrt{\frac{2}{e}np(1-p)}(1+\frac{1}{7n}).$$

Define

$$Factor(n,p) = \sqrt{2enp(1-p)}\binom{n-1}{\lfloor np\rfloor}p^{\lfloor np\rfloor}(1-p)^{n-1-\lfloor np\rfloor},$$

so that

$$\mathrm{E}[|X_n - np|] = \sqrt{\frac{2np(1-p)}{e}}Factor(n,p).$$

We wish to show that $Factor(n,p) < 1 + \frac{1}{7n}$, and that for p near 1 (or zero), $Factor$ is asymptotic to 1.

The first issue is to let $np = k + \epsilon$, where $k = \lfloor np\rfloor$, and determine, for n and k fixed, the value of ϵ that maximizes $Factor(n, \frac{k+\epsilon}{n})$. Taking the logarithmic derivative of $Factor$ will show that $\epsilon = \frac{1}{2}$ independent of k and n.

Next, let $x = np$. For the x-s of interest, define

$$Ratio(n,x) = \frac{Factor(n, \frac{x+1}{n})}{Factor(n, \frac{x}{n})}.$$

Then

$$\begin{aligned}
Ratio(n,x) &= \frac{x+1}{x+\frac{1}{2}}\left(\frac{x+1}{x}\right)^x \frac{n-\frac{1}{2}-x}{n-x}\left(\frac{n-x-1}{n-x}\right)^{n-x-1}\\
&= \frac{g(x)}{g(n-x-1)},
\end{aligned}$$

where $g(x) = \frac{x+1}{x+\frac{1}{2}}\left(\frac{x+1}{x}\right)^x$.

We need to show that $g(x)$ is increasing in x, which will ensure that the ratios are initially less than 1, and are increasing, which will guarantee that $Factor(n,p)$ is maximum for the extreme cases $p = \frac{1}{2n}$, $1 - \frac{1}{2n}$.

Taking the logarithmic derivative of g gives $\frac{g'(x)}{g(x)} = \log(x+1) - \log(x) - \frac{1}{x+\frac{1}{2}}$, which is the difference between $\int_x^{x+1}\frac{1}{x}dx$ and the area under the trapezoid defined by the line tangent to the curve $y = \frac{1}{x}$ at location $x + \frac{1}{2}$, with left and right vertical sides at x and $x+1$, and base along the horizontal axis. Evidentially, this difference is positive since the function $y = \frac{1}{x}$ is convex. Hence g is increasing.

Setting $p = \frac{1}{2n}, 1-\frac{1}{2n}$, shows that $Factor(n,p) \le \sqrt{e}(1-\frac{1}{2n})^{n-\frac{1}{2}}$. Finally, $\sqrt{e}(1-\frac{1}{2n})^{n-\frac{1}{2}}$ can be shown to be less than $1+\frac{1}{7n}$, for $n \ge 2$, by taking the logarithm of both expressions and expanding each in a Taylor's Series.

Gonnet et al show that $\mathrm{E}[n_{i+1}p_{i+1}(1-p_{i+1})|S(i)] < \sqrt{\frac{2}{\pi}n_i p_i(1-p_i)}$. Evidentially, each step in their derivation involves complicated asymptotics. In this case too, straight forward computation can yield an exact somewhat semiclosed combinatorial form, but turning the resulting identity into the desired inequality appears to be rather unappealing. However, it is tempting to conjecture that some such approach ought to be competitive with the current alternative.

A.2. PROOF OF THEOREM 2.4

Proof. Let $\sigma(t)$ be a measure preserving transformation on $[0,2\mu]$ such that $\sigma(t)=t$ for $t<\mu$, and $g(\sigma(t))$ is nonincreasing for $\mu \le t \le 2\mu$. Set $h(t)=\frac{1}{2}(g(t)+g(\sigma(2\mu-t)))$. Then h is symmetric with respect to the midpoint μ, and has one local maximum, which is at μ. Intuitively, h represents a rearrangement of "mass" density that defines g to achieve the desired symmetry. It is easy to see that $a2$ and $a3$ ensure that h represents a left shifting of the mass of g.

Since $F(t)$ is monotone increasing, $\int_0^{2\mu} F(t)g(t)dt \ge \int_0^{2\mu} F(t)h(t)dt$. Now,

$$\int_0^{2\mu} F(t)h(t)dt = \int_0^{2\mu} \frac{F(t)+F(2\mu-t)}{2} h(t)dt > \int_0^{2\mu} \frac{h(t)}{2}dt,$$

where the last integral represents a lossy shifting of mass the mass density for $\frac{F(t)+F(2\mu-t)}{2}$ away from the midpoint μ, since we know from Lemma 2.1 that the mean of F, over $[0,2\mu]$, exceeds $\frac{1}{2}$, and the region where the symmetrized $\frac{F(t)+F(2\mu-t)}{2}$ exceeds $\frac{1}{2}$ comprises a connected interval, because of the b) criteria, just as in the proof of Theorem 2.1.

Finally, since h is defined by a measure preserving transformation and symmetric averaging, we have:

$$\int_0^{\infty} F(t)g(t)dt > \int_0^{2\mu} \frac{g(t)}{2}dt + \int_{2\mu}^{\infty} F(2\mu)g(t)dt \ge \int_0^{\infty} \frac{g(t)}{2}dt = \frac{1}{2},$$

since g is a probability density function.

Condition $a3'$) simply states that $\log(g(t))$ is growing at least as fast as $\log(g(2\rho-t))$, for $t \in (0,\rho)$, which is sufficient to ensure the skew condition $a3$).

The conditions $a2''$) $a3''$) and $a4''$) can be established as follows. First, note that

$$\begin{aligned}
&\int_0^{2\mu} g(t)F(t)dt + F(2\mu)\int_{2\mu}^{\infty} g(t)dt - \int_0^{2\mu} g(2\mu-t)F(t)dt \\
&\quad = (G(t)-1)F(t)\Big|_0^{2\mu} + \int_0^{2\mu}(1-G(t))f(t)dt + F(2\mu)(1-G(2\mu)) + G(2\mu-t)F(t)\Big|_0^{2\mu} \\
&\qquad - \int_0^{2\mu} G(2\mu-t)f(t)dt \\
&\quad = \int_0^{2\mu}(1-G(t)-G(2\mu-t))f(t)dt \\
&\quad \ge 0, \quad \text{according to } a2'',
\end{aligned}$$

whence dividing by 2 and rearranging terms shows that

$$\int_0^{2\mu} \frac{g(t)F(t)}{2}dt \ge \int_0^{2\mu} \frac{g(2\mu-t)F(t)}{2}dt - \frac{F(2\mu)}{2}\int_{2\mu}^{\infty} g(t)dt.$$

Adding $\int_0^{2\mu} \frac{g(t)F(t)}{2}dt + \int_{2\mu}^{\infty} g(t)F(t)dt$ to both sides gives:

$$\int_0^{\infty} g(t)F(t)dt \;\geq\; \int_0^{2\mu} \frac{g(t)+g(2\mu-t)}{2}F(t)dt + \int_{2\mu}^{\infty} g(t)(F(t) - \frac{F(2\mu)}{2})dt. \tag{A.3}$$

Now,

$$\int_0^{2\mu} \frac{g(t)+g(2\mu-t)}{2}F(t)dt = \int_0^{2\mu} \frac{g(t)+g(2\mu-t)}{2}\frac{F(t)+F(2\mu-t)}{2}dt,$$

and the unimodality and symmetry of $\frac{g(t)+g(2\mu-t)}{2}$ ensures, by the reasoning used for Theorem 2.1, that replacing $\frac{F(t)+F(2\mu-t)}{2}$ by its average value on $[0,2\mu]$ will be lossy for the integral, since mass will be shifted away from the center $t=\mu$. Thus, appealing to Lemma 2.1 to attain the exact mean $\frac{1}{2} + \frac{1}{2\mu}\int_{2\mu}^{\infty}(t-2\mu)f(t)dt$ gives the inequality:

$$\int_0^{2\mu} \frac{g(t)+g(2\mu-t)}{2}F(t)dt \geq \frac{G(2\mu)}{2} + \frac{G(2\mu)}{2\mu}\int_{2\mu}^{\infty}(t-2\mu)f(t)dt. \tag{A.4}$$

Substituting (A.4) in (A.3) gives:

$$\begin{aligned}
\int_0^{\infty} g(t)F(t)dt \;&\geq\; \frac{G(2\mu)}{2} + \frac{G(2\mu)}{2\mu}\int_{2\mu}^{\infty}(1-F(t))dt + \int_{2\mu}^{\infty} g(t)(F(t) - \frac{F(2\mu)}{2})dt \\
&\geq\; \frac{G(2\mu)}{2} + \int_{2\mu}^{\infty}\frac{g(t)}{2}(1-F(t))dt + \int_{2\mu}^{\infty} g(t)(F(t) - \frac{F(2\mu)}{2})dt \\
&\geq\; \frac{G(2\mu)}{2} + \int_{2\mu}^{\infty} g(t)\frac{1+F(t)-F(2\mu)}{2}dt \qquad \text{(by } a4'' \text{)} \\
&\geq\; \frac{G(2\mu)}{2} + \int_{2\mu}^{\infty}\frac{g(t)}{2}dt = \frac{1}{2}.
\end{aligned}$$

∎

A.3. PROOF OF COROLLARY 2.3

Proof. Let $X_\circ(t)$ be the number of red balls so selected by time t, and $X_\bullet(t)$ be the number of black. Let $T_\bullet^+$ be the random stopping time when $X_\bullet$ becomes $b+1$: $T_\bullet^+(t) = \min\{t : X_\bullet(t) = b+1\}$.

By symmetry, it suffices to prove that Prob$\{X_\circ(T_\bullet^+) \geq r\} > \frac{1}{2}$. This probability is formulated as

$$\int_0^{\infty} g_\bullet^+(t)F_\circ(t)dt,$$

where
$$F_\circ(t) = \sum_{r\leq j\leq R}\binom{R}{j}(1-e^{-w_\circ t})^j(e^{-w_\circ t})^{R-j},$$

and
$$g_\bullet^+(t) = \binom{B}{b}(1-e^{-w_\bullet t})^b(e^{-w_\bullet t})^{B-b}w_\bullet(B-b).$$

For simplicity, we can take $\rho = 1$ by rescaling (time or) the weights $w_\circ$ and $w_\bullet$ by the factor ρ. Theorem 2.4 is applicable with conditions $a1$, $b1'$ and $b2$ used for $f \equiv \frac{dF_\circ}{dt}$. In particular,

$$\mu = \int_0^{\infty} t dF_\circ(t) \;=\; \frac{1}{w_\circ}(\frac{1}{R} + \frac{1}{R-1} + \cdots + \frac{1}{R+1-r})$$

$$< \frac{1}{w_\circ}(\ln R - \ln(R-b)) = 1.$$

The remaining conditions $a2$ and $a3'$, are checked as follows. The function $g_\bullet^+(t)$ is maximized at the time satisfying $0 = \frac{bw_\bullet e^{-w_\bullet t}}{1-e^{-w_\bullet t}} - (B-b)w_\bullet$. Substituting w for $w_\bullet$, and using the fact that $b = B(1-e^{-w_\bullet})$ gives $\frac{e^{-wt}}{1-e^{-wt}} = \frac{e^{-w}}{1-e^{-w}}$, which has the unique solution $t = 1 > \mu$, and thereby fulfills requirement $a2$.

Requirement $a3'$ states that

$$\frac{g'(t)}{g(t)} \geq -\frac{g'(2-t)}{g(2-t)} \geq 0,$$

for $t \in (0,1)$. The chief target inequality reads:

$$\frac{bw_\bullet e^{-w_\bullet t}}{1-e^{-w_\bullet t}} - (B-b)w_\bullet \geq ? - \frac{bw_\bullet e^{-w_\bullet (2-t)}}{1-e^{-w_\bullet (2-t)}} + (B-b)w_\bullet,$$

which can be simplified to

$$\frac{e^{-wt}}{1-e^{-wt}} + \frac{e^{-w(2-t)}}{1-e^{-w(2-t)}} \geq ?\, 2\frac{e^{-w}}{1-e^{-w}},$$

where we again substituted $b = B(1-e^{-w_\bullet})$, and set $w = w_\bullet$.

The target inequality achieves equality when $t = 1$. Differentiating (and noting that we have equality on the right) gives the new objective:

$$\frac{-we^{-wt}}{(1-e^{-wt})^2} + \frac{we^{-w(2-t)}}{(1-e^{-w(2-t)})^2} \leq ?\, 0.$$

Equality again occurs at $t = 1$. Differentiating once more gives the clearly achieved objective:

$$w^2\frac{e^{-wt}+e^{-2wt}}{(1-e^{-wt})^3} + w^2\frac{e^{-w(2-t)}+e^{-2w(2-t)}}{(1-e^{-w(2-t)})^3} \geq 0,$$

which holds since each term is positive. Integrating this inequality down from 1 twice establishes the original inequality as valid.

Finally, we must show that $-\frac{g'(2-t)}{g(2-t)} \geq 0$, for $t \in (0,1)$, which translates to

$$-\frac{bw_\bullet e^{-w_\bullet (2-t)}}{1-e^{-w_\bullet (2-t)}} + (B-b)w_\bullet \geq ?\, 0.$$

Substituting for B and $w_\bullet$, and simplifying gives:

$$\frac{e^{-w(2-t)}}{1-e^{-w(2-t)}} \leq ?\, \frac{e^{-w}}{1-e^{-w}}.$$

This relation holds as an equality when $t = 1$. But since $\frac{x}{1-x}$ is increasing for $x \in (0,1)$, it follows that $\frac{e^{-w(2-t)}}{1-e^{-w(2-t)}}$ must be decreasing as t decreases from 1, since $e^{-w(2-t)}$ is decreasing as t decreases, which verifies the bound.

As a sufficient set of requirements for Theorem 2.4 have been satisfied, it follows that

$$\text{Prob}\{X_\circ \geq r\} > \frac{1}{2} \quad \text{and} \quad \text{Prob}\{X_\bullet \geq b\} > \frac{1}{2}$$

as claimed. ▮

A.4. PROOF OF THEOREM 2.5

Proof. By definition,

$$\text{Prob}\{\zeta_n = k\} = \frac{\sum_{\substack{c_1,c_2,\ldots,c_{n+l}\in\{0,1\} \\ c_1+c_2+\cdots+c_n=k \\ c_1+c_2+\cdots+c_{n+l}=x+y}} \prod_{i=1}^{n+l} p_i^{c_i}(1-p_i)^{1-c_i}}{\sum_{\substack{c_1,c_2,\ldots,c_{n+l}\in\{0,1\} \\ c_1+c_2+\cdots+c_{n+l}=x+y}} \prod_{i=1}^{n+l} p_i^{c_i}(1-p_i)^{1-c_i}}. \tag{A.5}$$

We now show that it suffices to establish the bound when p_1 through p_n are the same, and p_{n+1} through p_{n+l} are also all equal (to a possibly different value). Let

$$\phi_i = \text{Prob}\{\zeta_n = i\}.$$

LEMMA 4.1. *Let $S_{x,y}$ be the set of all vectors $v \in \Re^{n+l}$ where $0 \le v_i \le 1$, for $i = 1, 2, \ldots, l+n$, $v_1+v_2+\cdots+v_n = x$, and $v_{n+1} + v_{n+2} + \cdots + v_{n+l} = y$. Let $B_{x,y}$ be the set of all conditional random variables ζ defined by equation (A.5) with the underlying $(p_1, p_2, \ldots, p_{l+n})$ belonging to $S_{x,y}$. We can form the real linear functionals on ζ by evaluating, for each vector $\ell =< \ell_1, \ell_2, \ldots, \ell_n >\in \Re^n$, $\ell(\zeta) = \sum_{i=0}^n \ell_i\phi_i$.*

Then $\ell(\zeta)$, for $\zeta \in B_{x,y}$, achieves its extremal values on some random variables ζ defined by vectors in $S_{x,y}$ where p_1 through p_n are restricted to the three values 0, 1, and some fixed α in $(0,1)$, and similarly p_{n+1} through p_{n+l} are restricted to the three values 0, 1, and some fixed β in $(0,1)$.

Proof. Consider ζ for $i = 0, 1, \ldots, n$, defined over the set of free variables

$$< p_1, p_2, \ldots, p_{n+l} >\in S_{x,y}.$$

Since $S_{x,y}$ is closed and bounded, the extrema are achieved in $S_{x,y}$.

Equation A.5 shows that the ϕ_i are symmetric in p_1, …, p_n and in p_{n+1}, …, p_{n+l}. Furthermore, they are first order rational functions (i.e., of the form $(ap_h + b)/(cp_h + d)$) in any single p_h. Since the denominator does not depend on i, $\ell(\zeta)$ is also a rational function that is symmetric in $p_1, \ldots, p_n$ and in $p_{n+1}, \ldots, p_{n+l}$, and is first order in each p. Thus, the dependence on p_j and p_k for, say, $j, k \le n$ can be exposed as follows: for some constants a, b, c, d, e, f,

$$\ell(\zeta) = \frac{a(p_j + p_k) + bp_jp_k + c}{d(p_j + p_k) + ep_jp_k + f}.$$

The extreme points (in $S_{x,y}$) for this expression can be located by freezing all variables other than p_j and p_k and seeking the extrema subject to the constraint that $p_j + p_k$ is held fixed, so that the p vector remains in $S_{x,y}$. Now, first order rational functions of the form $\frac{\alpha w+\beta}{\delta w+\gamma}$ have just one local extremum, which occurs at $w = -\gamma/\delta$. But $\ell(\zeta)$ is bounded, so its the extreme values must occur when the one free variable $w \equiv p_jp_k$ is itself at an extreme value, which is to say that either $p_j = p_k$, or at least one of the p's is set to an extreme value (0 or 1). (If b and e are zero, then these assignments cause no harm). It follows that these conclusions about p_j and p_k can/must hold for all such pairs $1 \le j < k \le n$, and the same holds for $n < j < k \le n + l$. Consequently, lemma is established. ∎

Clearly, probability evaluations such as $\text{Prob}\{\zeta_n \le x\} = \sum_{i=0}^x \phi_i$ are linear functionals on ζ. Thus, Lemma 4.1 restricts the probabilities that must be considered to establish our basic median bound. For completeness, we note that Hoeffding introduced this approach to analyze unconditioned Bernoulli Trials, which gives rise to linear expressions in p_j and p_k, rather than first order rational fractions [4].

We can now drop all Bernoulli Trials that are not random and adjust the integers x, y, n and l accordingly, so that it suffices to set $p_i = p$, for $1 \le i \le n$, and $p_i = q$, for $n < i \le n + l$. In this case,

$$\text{Prob}\{\zeta_n = k\} = \frac{\binom{n}{k}p^k(1-p)^{n-k}\binom{l}{x+y-k}q^{x+y-k}(1-q)^{l+k-x-y}}{\sum_{0\le j\le n}\binom{n}{j}p^j(1-p)^{n-j}\binom{l}{x+y-j}q^{x+y-j}(1-q)^{l+j-x-y}},$$

where $x = np$ and $y = lq$. Let $a_k = \binom{n}{k} p^k (1-p)^{n-k}$, and $b_k = \binom{l}{k} q^k (1-q)^{l-k}$.

We need to show that

$$|\sum_{0 \le k} (a_{np-k} b_{lq+k} - a_{np+k} b_{lq-k})| < a_{np} b_{lq},$$

for when we divide by the correct normalizing denominator

$$\sum_{0 \le k \le n} a_k b_{np+lq-k},$$

the resulting inequality says that

$$|\text{Prob}\{\zeta_n \le np\} - \text{Prob}\{\zeta_n \ge np\}| < \text{Prob}\{\zeta_n = np\},$$

which ensures that $x = np$ is the median value of ζ_n.

Now,

$$\sum_{0<k} (a_{np-k} b_{lq+k} - a_{np+k} b_{lq-k}) \;=\; \sum_{0<k} a_{np-k}(b_{lq+k} - b_{lq-k}) + \sum_{0<k} b_{lq-k}(a_{np-k} - a_{np+k}).$$

Consequently, it suffices to prove that each of the summations on the right is bounded in absolute value by $\frac{a_{np} b_{lq}}{2}$. The proof is completed below with the proof of Theorem 2.6, which shows that for any K: $\left| \sum_{0<k<K} (a_{np-k} - a_{np+k}) \right| < \frac{a_{np}}{2}$. Indeed,

$$|\sum_{0<k} b_{lq-k}(a_{np-k} - a_{np+k})| < b_{lq} \max_K |\sum_{0<k<K} (a_{np-k} - a_{np+k})|,$$

since the b_{lq-k} are monotonic, and Theorem 2.6 shows that

$$b_{lq} \max_K |\sum_{0<k<K} (a_{np-k} - a_{np+k})| < \frac{b_{lq} a_{np}}{2}.$$

Comparable arguments apply to $\sum_{0<k} a_{np-k}(b_{lq+k} - b_{lq-k})$. ∎

A.5. PROOF OF THEOREM 2.6

Proof. In view of Hoeffding's extremal bound for Bernoulli Trials and Lemma 4.1, we can suppose that the x_i's are identically distributed with probability of success $p = \frac{\mu}{n}$. We can also assume that $p < \frac{1}{2}$: the case where $p > \frac{1}{2}$ follows from replacing p by $1 - p$.

LEMMA 5.1. *Let p, μ, n, and a_i be defined as in Theorem 2.6. Then the following are true.*

1) For $p < \frac{1}{2}$: $\displaystyle\sum_{0<k\le\mu} a_{\mu-k} > \sum_{0<k\le(1-p)n} a_{\mu+k}$.

2) For $p < \frac{1}{2}$, $j \ge 0$: *if* $\dfrac{a_{\mu-j-1}}{a_{\mu-j}} < \dfrac{a_{\mu+j+1}}{a_{\mu+j}}$ *then* $\dfrac{a_{\mu-j-2}}{a_{\mu-j-1}} < \dfrac{a_{\mu+j+2}}{a_{\mu+j+1}}$.

Proof. 1) This follows from the Jogdeo-Samuels bound for Bernoulli Trials.

2) Note that $\frac{a_{\mu-k-1}}{a_{\mu-k}} = (\frac{1-p}{p})(\frac{\mu-k}{n(1-p)+k+1}) = \frac{1-\frac{k}{\mu}}{1+\frac{k+1}{(1-p)n}}$, and by symmetry, $\frac{a_{\mu+k+1}}{a_{\mu+k}} = \frac{1-\frac{k}{(1-p)n}}{1+\frac{k+1}{\mu}}$. Thus $\frac{a_{\mu-k-1}}{a_{\mu-k}} - \frac{a_{\mu+k+1}}{a_{\mu+k}} = \frac{\frac{1}{\mu} - \frac{1}{n(1-p)} - k(k+1)(\frac{1}{\mu^2} - \frac{1}{n^2(1-p)^2})}{\text{denominator}}$, which has a numerator that is strictly decreasing in k. Hence, once the difference becomes negative, it stays negative for larger k. ∎

Consequently, once $a_{\mu-j} - a_{\mu+j}$ is negative, the expression will remain negative for larger j. Suppose $p < \frac{1}{2}$, and let $\kappa = \max\{k : a_{\mu-k} - a_{\mu+k} \geq 0\}$. It follows that $\sum_{0<k\leq K}(a_{\mu-k} - a_{\mu+k})$ is maximized at $K = \kappa$.

The difficulty with the current inequality we now seek to establish is that it depends on two parameters, n and μ. We now show that the worst case for the inequality occurs in the limit where we have the Poisson distributions with the single parameter μ. That is,

$$\sum_{k=1}^{\kappa} \frac{a_{\mu-k} - a_{\mu+k}}{a_\mu} \leq ? \sum_{k=1}^{\kappa} \frac{p_{\mu-k} - p_{\mu+k}}{p_\mu},$$

where $\kappa = \max\{k : a_{\mu-k} - a_{\mu+k} \geq 0\}$, and $p_k = e^{-\mu}\frac{(\mu)^k}{k!}$. Then the proof can be completed by establishing the bound for the p_k values.

LEMMA 5.2. *Suppose that* $p < .5$ *and let* $\kappa = \max\{k : a_{\mu-k} - a_{\mu+k} \geq 0\}$. *Let* $p_k = e^{-\mu}\frac{(\mu)^k}{k!}$.

Then for $0 \leq k \leq \kappa$*:*

$$\frac{a_{\mu-k} - a_{\mu+k}}{a_\mu} \leq \frac{p_{\mu-k} - p_{\mu+k}}{p_\mu}.$$

Proof. A little manipulation gives the following:

$$\begin{aligned} \frac{a_{\mu-k}}{a_\mu} &= \frac{\prod_{j=0}^{k-1}(1-\frac{j}{\mu})}{\prod_{j=1}^{k}(1+\frac{j}{(1-p)n})}; \\ \frac{a_{\mu+k}}{a_\mu} &= \frac{\prod_{j=0}^{k-1}(1-\frac{j}{(1-p)n})}{\prod_{j=1}^{k}(1+\frac{j}{\mu})}; \\ \frac{p_{\mu-k}}{p_\mu} &= \prod_{j=0}^{k-1}(1-\frac{j}{\mu}); \\ \frac{p_{\mu+k}}{p_\mu} &= \frac{1}{\prod_{j=1}^{k}(1+\frac{j}{\mu})}. \end{aligned}$$

Thus, since

$$\frac{p_{\mu-k}}{p_\mu} - \frac{p_{\mu+k}}{p_\mu} = (\frac{a_{\mu-k}}{a_\mu})\prod_{j=1}^{k}(1+\frac{j}{(1-p)n}) - (\frac{a_{\mu+k}}{a_\mu})\frac{1}{\prod_{j=0}^{k-1}(1-\frac{j}{\mu})},$$

the conclusion

$$\sum_{k=1}^{\kappa} \frac{a_{\mu-k} - a_{\mu+k}}{a_\mu} \leq ? \sum_{k=1}^{\kappa} \frac{p_{\mu-k} - p_{\mu+k}}{p_\mu}$$

will be established if the "magnifier" $\prod_{j=1}^{k}(1+\frac{j}{(1-p)n})$ can be shown to exceed the "magnifier" $1/\prod_{j=0}^{k-1}(1-\frac{j}{\mu})$, for $k = 1, 2, \ldots, \kappa$. Multiplying both magnifiers, for $k = 1, 2, \ldots, \kappa$, by $\prod_{j=0}^{k-1}(1-\frac{j}{\mu})$ gives the target inequalities:

$$(1+\frac{k}{(1-p)n})\prod_{j=1}^{k-1}(1-\frac{j^2}{(1-p)^2n^2}) >? \, 1,$$

for $k \leq \kappa$.

From the definition of κ, it follows that for $k \leq \kappa$,

$$\frac{\prod_{j=0}^{k-1}(1-\frac{j}{\mu})}{\prod_{j=1}^{k}(1+\frac{j}{(1-p)n})} \geq \frac{\prod_{j=0}^{k-1}(1-\frac{j}{(1-p)n})}{\prod_{j=1}^{k}(1+\frac{j}{\mu})},$$

whence

$$\frac{\prod_{j=0}^{k-1}(1-\frac{j^2}{\mu^2})}{\prod_{j=0}^{k-1}(1-\frac{j^2}{(1-p)^2n^2})} \ge \frac{1+\frac{k}{(1-p)n}}{1+\frac{k}{\mu}}.$$

Applying the expansion $1-x = e^{\log(1-x)} = e^{-\sum_{1\le j} x^j/j}$, for $|x|<1$ with x set to j/μ and $j/((1-p)n)$ gives:

$$\frac{e^{-\sum_{j=1}^{\infty} S_k(2j)/(p)^{2j}}}{e^{-\sum_{j=1}^{\infty} S_k(2j)/((1-p)n)^{2j}}} \ge \frac{1+\frac{\kappa}{(1-p)n}}{1+\frac{\kappa}{\mu}},$$

where

$$S_k(\ell) \equiv \sum_{h=1}^{k-1} \frac{h^\ell}{j}.$$

Combining the two exponentials gives

$$e^{-\sum_{j=1}^{\infty} \frac{S_k(2j)((1-p)^{2j}-p^{2j})}{n^{2j}p^{2j}(1-p)^{2j}}} \ge \frac{1+\frac{k}{(1-p)n}}{1+\frac{k}{\mu}}.$$

Observe that the factor $\frac{(1-p)^r-p^r}{p^{r-1}} = (1-p)(\frac{1-p}{p})^{r-1} - p$ is increasing in r, since $p<1-p$. When $r=1$, the expression equals $1-2p$. Therefore the factor $\frac{(1-p)^{2j}-p^{2j}}{p^{2j}}$ can be replaced by $\frac{(1-2p)}{p}$, which will increase the exponent (since it is negative) to get

$$e^{-\sum_{j=1}^{\infty} \frac{S_k(2j)(1-2p)}{n^{2j}p(1-p)^{2j}}} \ge \frac{1+\frac{k}{(1-p)n}}{1+\frac{k}{\mu}}.$$

Raising both sides to the power $p/(1-2p)$ gives

$$e^{-\sum_{j=1}^{\infty} \frac{S_k(2j)}{n^{2j}(1-p)^{2j}}} \ge \frac{(1+\frac{k}{(1-p)n})^{p/(1-2p)}}{(1+\frac{k}{\mu})^{p/(1-2p)}}.$$

Now, $(1+\frac{k}{\mu})^{p/(1-2p)} = (1+\frac{k}{np})^{\frac{p}{1-p}(\frac{1-p}{1-2p})} < (1+\frac{k}{(1-p)n})^{\frac{1-p}{1-2p}}$, where we used the inequality $(1+x)^\alpha < 1+\alpha x$, which is valid for for $0\le\alpha\le 1$ and $x \ge -1$.

Applying this last inequality gives

$$e^{-\sum_{j=1}^{\infty} \frac{S_k(2j)}{n^{2j}(1-p)^{2j}}} \ge \frac{(1+\frac{k}{(1-p)n})^{p/(1-2p)}}{(1+\frac{k}{(1-p)n})^{(1-p)/(1-2p)}} = (1+\frac{k}{(1-p)n})^{-1}.$$

Reversing the expansion to present the exponential as a finite product gives

$$\prod_{h=1}^{k-1}(1-\frac{h^2}{(1-p)^2n^2}) > \frac{1}{1+\frac{k}{(1-p)n}},$$

which is our target inequality for $k \le \kappa$. ∎

We could have set $r=2$ in this derivation, thereby attaining the slightly stronger

$$(1+\frac{k}{(1-p)n})^p \prod_{j=1}^{k-1}(1-\frac{j^2}{(1-p)^2n^2}) > 1.$$

LEMMA 5.3. *Let the p_i-s and μ be as in Lemma 5.2. Let $\kappa = \max\{k : p_{\mu-k} - p_{\mu+k} \geq 0\}$. Then*

$$\lim_{\mu\to\infty} \sum_{0\leq j\leq\kappa} \frac{p_{\mu-j} - p_{\mu+j}}{p_\mu} = \int_0^{\sqrt{3}} e^{-x^2/2}(x - \frac{x^3}{3})dx = \frac{1+e^{-3/2}}{3} \approx .48209.$$

Proof. We will see that that $\kappa \approx \sqrt{3\mu}$, which allows the dominant errors to be readily identified. Formally, we could restrict κ to satisfy, say, $\kappa < 2\sqrt{3\mu}$. There follows:

$$\begin{aligned}
\sum_{j=1}^{\kappa} \frac{p_{\mu-j} - p_{\mu+j}}{p_\mu} &= \sum_{j=1}^{\kappa}\Big(\prod_{k=0}^{j-1}(1-\frac{k}{\mu}) - \frac{1}{\prod_{k=1}^{j}(1+\frac{k}{\mu})}\Big) \\
&= \sum_{j=1}^{\kappa} \frac{(1+\frac{j}{\mu})\prod_{k=0}^{j-1}(1-\frac{k^2}{\mu^2}) - 1}{\prod_{k=1}^{j}(1+\frac{k}{\mu})} \qquad \text{(A.6)} \\
&= \sum_{j=1}^{\kappa}\bigg(\Big((1+\frac{j}{\mu})e^{-\frac{(j-1)j(2j-1)}{6\mu^2} - O(\frac{j^5}{10\mu^4})} - 1\Big)e^{-\frac{j(j+1)}{2\mu} + \frac{j(j+1)(2j+1)}{12\mu^2} - O(\frac{j^4}{12\mu^3})}\bigg) \\
&= \sum_{j=1}^{\kappa}\bigg((\frac{j}{\mu} - \frac{(j-1)j(2j-1)}{6\mu^2} - O(\frac{j^4}{6\mu^3}) - O(\frac{j^5}{10\mu^4}))e^{-\frac{j(j+1)}{2\mu}}\Big(1+O(\frac{j(j+1)(2j+1)}{12\mu^2})\Big)\bigg) \\
&= \sum_{j=1}^{\kappa}\bigg((\frac{j}{\mu} - \frac{(j-1)j(2j-1)}{6\mu^2} - O(\frac{j^4}{6\mu^3}) - O(\frac{j^5}{10\mu^4}))e^{-\frac{j(j+1)}{2\mu}}\Big(1+O(\frac{j(j+1)(2j+1)}{12\mu^2})\Big)\bigg) \\
&= \int_0^{\kappa} e^{-y^2/2\mu}(\frac{y}{\mu} - \frac{y^3}{3\mu^2} + O(y^2/n^2))\mathrm{d}y,
\end{aligned}$$

which confirms that $\kappa \approx \sqrt{3\mu}$.

Rescaling $x = y\sqrt{\mu}$ gives $(1+o(1))\int_0^{\sqrt{3}} e^{-x^2/2}(x - \frac{x^3}{3})dx$. Evaluating the integral gives $(2e^{-3/2}+1)/3 \approx .48208677$. ▮

LEMMA 5.4. *Let the p_i-s, μ and κ be as in Lemma 5.3. Then*

$$\Big|\sum_{0\leq j\leq\kappa} \frac{p_{\mu-j} - p_{\mu+j}}{p_\mu} - \int_0^{\sqrt{3}} e^{-x^2/2}(x - \frac{x^3}{3})dx\Big| < \frac{200}{\mu}.$$

Proof. From (A.6), the requirement for κ can be formulated as the largest integer satisfying $(1+\frac{\kappa}{\mu})\prod_{k=0}^{\kappa-1}(1-\frac{k^2}{\mu^2}) \geq 1$. Since $1-x < e^{-x-x^2}$, for $x>0$, $(1+\frac{\kappa}{\mu})\prod_{k=0}^{\kappa-1}(1-\frac{k^2}{\mu^2}) < e^{\log(1+\frac{k}{\mu})}e^{\frac{-S_k(2)}{\mu^2} - \frac{S_k(4)}{\mu^4}}$, and therefore $\kappa < k$, where the k is any value satisfying:

$$e^{\log(1+\frac{k}{\mu})}e^{\frac{-S_k(2)}{\mu^2} - \frac{S_k(4)}{\mu^4}} \leq 1.$$

Expanding the logarithm up to three terms and taking the log of this inequality gives a sufficiency condition of

$$\frac{k}{\mu} - \frac{k^2}{2\mu^2} + \frac{k^3}{3\mu^3} - \frac{(k-1)k(2k-1)}{6\mu^2} - \frac{k^5/5 - k^4/2 + k^3/3 - k/30}{2\mu^4} \leq 0.$$

Simplifying gives

$$6\mu - 3k + \frac{2k^2}{\mu} - (k-1)(2k-1) - \frac{6k^4/5 - 3k^3 + 2k^2 - 1/5}{2\mu^2} \leq 0,$$

and

$$6\mu + \frac{2k^2}{\mu} - \frac{3k^4}{5\mu^2} - 1 + \frac{3k^3 - 2k^2 + 1/5}{2\mu^2} = 2k^2.$$

It follows that $\kappa < \sqrt{3\mu}$ provided

$$\frac{2k^2}{\mu} - \frac{3k^4}{5\mu^2} - 1 + \frac{3k^3}{2\mu^2}\Big|_{k=\sqrt{3\mu}} \le 0,$$

which holds if $\mu \ge 450$. For smaller μ, direct calculation verifies that indeed, $\kappa \le \sqrt{3\mu}$ for small μ as well as large.

Now that κ is tightly bounded, asymptotic expansions can be used to bound

$$\Big| \int_0^{\sqrt{3}} e^{-x^2/2}(x - \frac{x^3}{3})dx - \sum_{j=1}^{\kappa} \frac{p_{\mu-j} - p_{\mu+j}}{p_\mu} \Big| .$$

We get:

$$\begin{aligned}\Big| \int_0^{\sqrt{3}} e^{-x^2/2}(x - \frac{x^3}{3})dx - \sum_{j=1}^{\kappa} \frac{p_{\mu-j} - p_{\mu+j}}{p_\mu} \Big| \ \le\ & \sum_{j=1}^{\kappa} \Big| \frac{p_{\mu-j} - p_{\mu+j}}{p_\mu} - e^{-j(j+1)/2\mu}(\frac{j}{\mu} - \frac{(j-1)j(2j-1)}{6\mu^2}) \Big| \\ & + \Big| \int_0^{\sqrt{3}} e^{-x^2/2}(x - \frac{x^3}{3})dx \\ & \quad - \sum_{j=1}^{\kappa} e^{-j(j+1)/2\mu}(\frac{j}{\mu} - \frac{(j-1)j(2j-1)}{6\mu^2}) \Big| .\end{aligned}$$

It is not difficult to show that

$$\sum_{j=1}^{\kappa} \Big| \frac{p_{\mu-j} - p_{\mu+j}}{p_\mu} - e^{-j(j+1)/2\mu}(\frac{j}{\mu} - \frac{(j-1)j(2j-1)}{6\mu^2}) \Big| = O(1/\mu),$$

and that

$$\Big| \int_0^{\sqrt{3}} e^{-x^2/2}(x - \frac{x^3}{3})dx - \sum_{j=1}^{\kappa} e^{-j(j+1)/2\mu}(\frac{j}{\mu} - \frac{(j-1)j(2j-1)}{6\mu^2}) \Big| = O(1/\mu).$$

Casual assessment of the coefficients shows that the coefficients each are less than 100. ▮

To complete the proof of Theorem 2.6, we have to show that

$$\sum_{0 \le j \le \kappa} \frac{p_{\mu-j} - p_{\mu+j}}{p_\mu} \le \frac{1}{2}$$

for $\mu < 20000$, say, since the asymptotics of Lemma 5.3 and error bound of Lemma 5.4 ensure that the summation is at most $.483 + \frac{200}{\mu} < \frac{1}{2}$, when $\mu \ge 20000$.

For $\mu = 1, 2$ direct calculation gives

$$\sum_{0 \le j \le \kappa} \frac{p_{\mu-j} - p_{\mu+j}}{p_\mu} = \frac{1}{2}.$$

For larger values of μ, a computer program was written to verify Theorem 2.6 for Poisson distributions with integer means up to 1,000,000. ▮

A.6. ESTIMATING $\sum_{\ell=0}^{\infty} \frac{1}{(\ell!)^\gamma}$

LEMMA 6.1. *Suppose that $\gamma = 1-\lambda$ is in the interval $[0,1]$, and let $\gamma j = k+\delta_j$, where $0 \le \delta_j < 1$, and $j > 1$ and k are integers. Then $\frac{1}{(j!)^\gamma} \le \frac{1-\delta_j}{k!} + \frac{\delta_j}{(k+1)!}$.*

Proof. For notational simplicity, let $\delta = \delta_j$. Solving for γ gives $\gamma = \frac{k+\delta}{j}$. Substituting gives the target inequality $\frac{1}{(j!)^{\frac{k+\delta}{j}}} \le ? \frac{1-\delta}{k!} + \frac{\delta}{(k+1)!}$, or $\frac{1}{exp(\frac{k+\delta}{j}\ln(j!))} \le ? \frac{1-\delta}{k!} + \frac{\delta}{(k+1)!}$. The left hand side of this expression is maximized when j is as small as possible, which is to say that $j = k+1$. Thus it suffices to prove that $\frac{1}{(k+1)!^{\frac{k+\delta}{k+1}}} \le ? \frac{1-\delta}{k!} + \frac{\delta}{(k+1)!}$. Multiplying by $(k+1)!$ gives $(k+1)!^{\frac{1-\delta}{k+1}} \le ? (1-\delta)(k+1)+\delta = (1-\delta)k+1$, which has the form $(k+1)!^{\frac{x}{k+1}} \le ? kx+1$. Now the function $(k+1)!^{\frac{x}{k+1}}$ satisfies $(k+1)!^{\frac{x}{k+1}} \le kx+1$ for $x=0$ and $x=1$. But since $kx+1$ is a straight line and the exponential $(k+1)!^{\frac{x}{k+1}}$ is convex, the inequality must hold for all x in $[0,1]$. ∎

LEMMA 6.2. *Let w_k be the sum of the weights derived for $\frac{1}{k!}$ according to Lemma 6.1, so that, in the notation of Lemma 6.1, $w_k = \sum_{\gamma j \in [k,k+1)} (1-\delta_j) + \sum_{\gamma j \in [k-1,k)} \delta_j$. Then for $\ell \ge 1$,*

$$\sum_{k=0}^{\ell} w_k \le \frac{\ell+1}{\gamma}.$$

Proof. Let $j = \lceil \frac{\ell}{\gamma} \rceil = \frac{\ell}{\gamma} + \delta$, where $0 \le \delta < 1$. From Lemma 6.1, $\frac{1}{(j!)^\gamma} \le \frac{1-\delta_j}{\ell!} + \frac{\delta_j}{(\ell+1)!}$, where $\delta_j = \delta\gamma$, since $\gamma j = \ell + \delta\gamma$. Then

$$\begin{aligned}\sum_{k=0}^{\ell} w_k \le j+1-\delta_j &= \frac{\ell+\gamma+\delta\gamma(1-\gamma)}{\gamma} \\ &\le \frac{\ell+\gamma+\gamma(1-\gamma)}{\gamma} = \frac{\ell+1-(1-\gamma)^2}{\gamma} \le \frac{\ell+1}{\gamma}.\end{aligned}$$

∎

We can now write $\sum_{j=0}^{\infty} \frac{1}{j!^\gamma} \le \sum_{k=0}^{\infty} \frac{w_k}{k!}$, where w_k is the sum of the (appropriate subset of) interpolation weights $1-\delta_j$ and δ_j as derived for the factor $\frac{1}{k!}$ in Lemmas 6.1 and 6.2.

Evidentially, $w_0 \le \frac{1}{\gamma}$. In view of Lemma 6.2 and the definition of w_k, it follows that for $0 \le \lambda \le 1$, $\sum_{0 \le j} \frac{1}{(j!)^{1-\lambda}} \le \sum_{0 \le k} \frac{w_k}{k!} \le \sum_{0 \le k} \frac{1}{\gamma k!} = \frac{e}{1-\lambda}$, since the w_k-s constitute a skew of the uniform coefficient weightings $\frac{1}{\gamma}$ toward larger denominators.

A.7. A CONVEXITY CLAIM

LEMMA 7.1 (Folklore). *Let $\zeta_1, \zeta_2 \ldots, \zeta_n$ be n independent identically distributed Poisson random variables with mean 1 and let $V = \sum_j \ln(\zeta_j!)$. Set $y = \sum_j \zeta_j$, and define $f(y) = \mathrm{E}[z^V|y]$, for $z > 0$ and y a non-negative integer. Then f is strictly convex: $f(y) - 2f(y+1) + f(y+2) > 0$.*

Proof. We model $< \zeta_1, \zeta_2 \ldots, \zeta_n >$ as the outcome of a Poisson Process that distributes balls among n bins over time. Let $b_j(y)$ be the number of balls in bin j after y balls have been randomly dispersed among the n bins, with each bin equally likely to receive any ball, and define the random variable (a function in z that depends on the outcome of the experiments with y balls)

$$W(y) = z^{\sum_{j=1}^{n} \ln(b_j(y)!)}.$$

As defined, the random variables $W(y)$ and $W(y+1)$ are quite dependent, since the actual ball distributions only differ by the placement of the $(y+1)$-st ball. By construction and the properties of Poisson processes, $f(y) = \mathrm{E}[W(y)]$. Direct computation gives:

$$f(y) - 2f(y+1) + f(y+2) = \mathrm{E}[\mathrm{E}[W(y) - 2W(y+1) + W(y+2) \mid b_1(y), b_2(y), \ldots, b_n(y)]]$$

$$= \; \mathrm{E}\left[z^{\sum_{j=1}^{n} \ln(b_j(y)!)} \left(1 - \sum_{j=1}^{n} \frac{2}{n} z^{\ln(b_j(y)+1)} \right.\right.$$
$$\left.\left. + \sum_{i=1}^{n} \sum_{j>i}^{n} \frac{2}{n^2} z^{\ln(b_j(y)+1)+\ln(b_i(y)+1)} + \sum_{j=1}^{n} \frac{1}{n^2} z^{\ln(b_j(y)+1)+\ln(b_j(y)+2)} \right) \right]$$
$$> \; \mathrm{E}\left[W(y) \left(1 - \sum_{j=1}^{n} \frac{1}{n} z^{\ln(b_j(y)+1)} \right)^2 \right],$$

since $W > 0$. The positivity of the discrete second derivative is thus established. ▮

Consequently, $\mathrm{E}[z^V] = \mathrm{E}_y[\mathrm{E}[z^V|y]] = \mathrm{E}_y[f(y)] \geq f(\mathrm{E}_y[y]) = f(n)$, where E_y denotes the expectation over the probability space defined by y. In plain English, $E[z^V]$ is formed when the $\zeta_1, \zeta_2 \ldots, \zeta_n$ are i.i.d. Poisson distributed random variables with mean 1. If, instead, we use the uniform multinomial distribution that corresponds to distributing n balls among n bins, we get $f(n)$, which is a smaller function in z.

A.8. CONDITIONING ON THE NUMBER OF BERNOULLI SUCCESSES

We first establish the following auxiliary fact.

LEMMA 8.1 (Folklore). *Let* $\xi_k = \mathrm{Prob}\{b_1 = 1 | b_1 + b_2 + \cdots + b_n = k\}$. *Then* ξ_k *is monotone increasing in* k.

Proof. Let $\zeta_k = \mathrm{Prob}\{b_2 + b_3 + \cdots + b_n = k\}$. Recall that $p_i = \mathrm{Prob}\{b_i = 1\}$. Evidentially, $\xi_k = \frac{p_1 \zeta_{k-1}}{p_1 \zeta_{k-1} + (1-p_1)\zeta_k}$, and we therefore have the following equivalence: $\xi_{k+1} \geq \xi_k$ iff $\frac{\zeta_k}{\zeta_{k+1}} \geq \frac{\zeta_{k-1}}{\zeta_k}$. Thus we must show that $\zeta_k^2 \geq \zeta_{k+1}\zeta_{k-1}$. Put $\gamma = \prod_1^n (1-p_i)$, and let $r_i = p_i/(1-p_i)$. Then

$$\zeta_k = \gamma \sum_{i_1 < i_2 < \cdots < i_k} r_{i_1} r_{i_2} \cdots r_{i_k},$$

whence

$$\zeta_k^2 = \gamma^2 \sum_{j=0}^{k} \sum_{\substack{i_1 < i_2 < \\ \cdots < i_{2k-j}}} \left(\binom{2k-2j}{k-j} r_{i_1} r_{i_2} \cdots r_{i_{2k-j}} \sum_{\substack{\{s_1, s_2, \ldots, s_j\} \subset \\ \{i_1, \ldots, i_{2k-j}\}}} r_{s_1} r_{s_2} \cdots r_{s_j} \right).$$

However, $\zeta_{k-1}\zeta_{k+1}$ is equal to

$$\gamma^2 \sum_{j=0}^{k-1} \sum_{\substack{i_1 < i_2 < \\ \cdots < i_{2k-j}}} \left(\binom{2k-2j}{k-j-1} r_{i_1} r_{i_2} \cdots r_{i_{2k-j}} \sum_{\substack{\{s_1, s_2, \ldots, s_j\} \subset \\ \{i_1, \ldots, i_{2k-j}\}}} r_{s_1} r_{s_2} \cdots r_{s_j} \right).$$

Thus ζ_k^2 contains a superset of the terms in $\zeta_{k-1}\zeta_{k+1}$, and with corresponding coefficients of $\gamma^2 \binom{2k-2j}{k-j}$ as opposed to $\gamma^2 \binom{2k-2j}{k-j-1}$, which shows that the inequality is in fact strict, unless $\zeta_k = 0$. ▮

LEMMA 8.2 (Folklore). *Let* $b_1, b_2, \ldots, b_n$ *be* n *independent Bernoulli trials with* $\mathrm{E}[b_i] = p_i$. *Let* f *be a real valued* n*-ary function that is increasing in each coordinate. Let* $k = b_1 + \cdots + b_n$ *Then* $\mathrm{E}[f(b_1, \ldots, b_n)|k]$ *is increasing in* k.

Proof. The proof will be by induction on (n, k). The base cases are $n = k - 1$, and $k = 0$. In either instance, it is trivially seen that $\mathrm{E}[[f(b_1, \ldots, b_n)|b_1 + b_2 + \cdots + b_n = k+1]] \geq \mathrm{E}[[f(b_1, \ldots, b_n)|b_1 + b_2 + \cdots + b_n = k]]$.

So we may suppose that the inequality holds for $(n-1, k+1)$ and $(n-1, k)$. Let f denote $f(b_1, b_2, \ldots, b_n)$ and let $B = b_1 + b_2 + \cdots + b_n$. Then

$$\begin{aligned}
\mathrm{E}[[f|B = k+1]] &= \mathrm{E}[[f|b_1 = 1, b_2 + \cdots + b_n = k]]\xi_{k+1} + \mathrm{E}[[f|b_1 = 0, b_2 + \cdots + b_n = k+1]](1 - \xi_{k+1}) \\
&= \mathrm{E}[[f|b_1 = 1, b_2 + \cdots + b_n = k]]\xi_k \\
&\quad + \mathrm{E}[[f|b_1 = 1, b_2 + \cdots + b_n = k]](\xi_{k+1} - \xi_k) \\
&\quad + \mathrm{E}[[f|b_1 = 0, b_2 + \cdots + b_n = k+1]](1 - \xi_{k+1}) \\
&\geq \mathrm{E}[[f|b_1 = 1, b_2 + \cdots + b_n = k-1]]\xi_k \qquad (\dagger) \\
&\quad + \mathrm{E}[[f|b_1 = 0, b_2 + \cdots + b_n = k]](\xi_{k+1} - \xi_k) \qquad (\ddagger) \\
&\quad + \mathrm{E}[[f|b_1 = 0, b_2 + \cdots + b_n = k]](1 - \xi_{k+1}) \qquad (\dagger) \\
&\geq \mathrm{E}[[f|b_1 = 1, b_2 + \cdots + b_n = k-1]]\xi_k + \mathrm{E}[[f|b_1 = 0, b_2 + \cdots + b_n = k]](1 - \xi_k) \\
&\geq \mathrm{E}[[f|b_1 + b_2 + \cdots + b_n = k]].
\end{aligned}$$

Notes:

($\dagger$): by the induction hypothesis for $n-1$ Bernoulli Trials.
($\ddagger$): since f is nondecreasing and $\xi_{k+1} - \xi_k \geq 0$ as established in Lemma 8.1 ▮

As a consequence, if f is a non-negative function that is increasing in each coordinate, and $\mathrm{E}[b_1 + b_2 + \cdots + b_n]$ is the integer k, then

$$\begin{aligned}
\mathrm{E}[[f(b_1, \ldots, b_n)|b_1 + \cdots + b_n = k]] &\leq \mathrm{E}[[f(b_1, \ldots, b_n)|b_1 + \cdots + b_n \geq k]] \\
&\leq \frac{\mathrm{E}[f]}{\mathrm{Prob}\{b_1 + \cdots + b_n \geq k\}} \leq 2\mathrm{E}[f].
\end{aligned}$$

A.9. RESCALING BERNOULLI TRIALS

Let $X_n = x_1 + x_2 + \cdots + x_n$ be the sum of n independent Bernoulli Trials with mean μ, and let and $Y_m = x_{n+1} + x_{n+2} + \cdots + x_{n+m}$ be an analogous sum with mean ν.

We demonstrate the standard fact that the probability distribution of X conditioned on $X + Y = k$ is statistically equivalent to the outcome of some $\widehat{X}_n$ conditioned on $\widehat{X}_n + \widehat{Y}_m = k$, where $\widehat{X}_n$ and $\widehat{Y}_m$, where are analogously defined Bernoulli Trials with $\mathrm{E}[\widehat{X}_n + \widehat{Y}_m] = k$.

Let $\mathrm{Prob}\{x_i = 1\} = p_i$ and let $\mathrm{Prob}\{\hat{x}_i = 1\} = q_i$, where $\widehat{X}$ and $\widehat{Y}$ are the corresponding sums of the $\hat{x}_i$-s. Let ρ be a constant. Define q_i so that $\frac{q_i}{1-q_i} = \rho\frac{p_i}{1-p_i}$.

Evidentially, assigning ρ values in the interval $(0, \infty)$ maps $\mathrm{E}[\widehat{X}_n + \widehat{Y}_m]$ onto $(0, n+m)$, so there is a suitable ρ where $\mathrm{E}[\widehat{X}_n + \widehat{Y}_m] = k$. To see that the conditional statistics are unchanged for all $\rho \in (0, \infty)$, let $Q(\alpha_1, \alpha_2, \ldots, \alpha_k)$ be the (unconditioned) probability that $x_{\alpha_i} = 1$ for $i = 1, 2, \ldots, k$ where $1 \leq \alpha_1 < \alpha_2 < \cdots < \alpha_k \leq n+m$, and all other x_i-s are set to zero. Similarly, let $\hat{Q}(\alpha_1, \alpha_2, \ldots, \alpha_k)$ be the (unconditioned) probability that $\hat{x}_{\alpha_i} = 1$ for $i = 1, 2, \ldots, k$ where $1 \leq \alpha_1 < \alpha_2 < \cdots < \alpha_k \leq n+m$, and all other $\hat{x}_i$-s are set to zero. Then

$$\begin{aligned}
\hat{Q}(\alpha_1, \alpha_2, \ldots, \alpha_k) &= \Big(\prod_{i=1}^{k} \frac{q_{\alpha_i}}{1 - q_{\alpha_i}}\Big)\Big(\prod_{i=1}^{n+m}(1 - q_i)\Big) \\
&= Q(\alpha_1, \alpha_2, \ldots, \alpha_k)\rho^k \prod_{i=1}^{n+m} \frac{1 - q_i}{1 - p_i},
\end{aligned}$$

which shows that k-way successes for $\widehat{X}_n+\widehat{Y}_m$ have an unconditioned probability that is the same as that for $X+Y$, apart from a uniform rescaling factor of $\rho^k \prod_{i=1}^{n+m} \frac{1-q_i}{1-p_i}$. Consequently, both $X+Y$ and $\widehat{X}+\widehat{Y}$ exhibit the same conditional probability statistics.

ACKNOWLEDGMENTS

It is a pleasure to express my appreciation for helpful and stimulating discussions to Richard Cole, Don Coppersmith, A. & S. Karlin, Steve Samuels, Joel Spencer, and S.R.S. Varadhan. In addition, the referees provided valuable suggestions, and pointed out many instances where the wording could be improved. I hope that the corrections and elaborations provide some justification for their effort. Last but far from least, I would like to express my thanks to Bernard Moret for making this publication possible.

REFERENCES

1. Anderson, A., personal communication.
2. Cole, R., personal communication.
3. Gonnet, G., L. Rogers and J. George, *An algorithmic and complexity analysis of interpolation search*, Acta Inf., 13, 1 (1980), pp. 39–52.
4. Hoeffding, W., *On the distribution of the number of successes in independent trials*, Ann. Math. Stat., 27 (1956), pp. 713–721.
5. Hoeffding, W., *Probability Inequalities for Sums of Bounded Random Variables*, J. Am. Stat. Ass., 58 (1963), pp. 13–30.
6. Jogdeo, K. and S. Samuels, *Monotone Convergence of Binomial Probabilities and a Generalization of Ramanujan's Equation, Ann. Math. Stat.*, 39 (1968), pp.1191–1195.
7. Lueker, G. and M. Molodowitch, *More Analysis of Double Hashing*, Combinatorica, 13 (1993), pp. 83–96.
8. Marshall, A. W. and I. Olkin, *Inequalities: Theory of Majorization and Its Applications*, Acad. Press, 1979.
9. Mehlhorn, K. and A. Tsakalidis, *Dynamic Interpolation Search*, JACM, 40, 3 (1993), pp. 621–634.
10. Panconesi. A. and A. Srinivasan, *Randomized distributed edge coloring via an extension of the Chernoff-Hoeffding bounds*, SIAM J. Comput., 26 (1997), pp. 350–368.
11. Perl, Y., A. Itai and H. Avni, *Interpolation search–A* $\log\log N$ *search*, CACM, 21, 7 (1978), pp. 550–554.
12. Perl. Y. and E. M. Reingold, *Understanding the complexity of interpolation search*, IPL, 6 (1977), pp. 219–222.
13. Peterson, W. W., *Addressing for random access storage*, IBM J. Res. Dev. 1 (1957), pp. 130–146.
14. Ramanujan, S., *Collected Papers*, Cambridge University Press, 1927.
15. Ranade, A. G., *How To Emulate Shared Memory*, JCSS Vol 42, 3 (1991), pp, 307–326.
16. Siegel, A., *On universal classes of fast high performance hash functions, their time-space tradeoff, and their applications*, 30^{th} FOCS, (1989), pp. 20–25.
17. Siegel, A., *Median Bounds and their Application*, 10^{th} SODA, (1999), pp. 776–785.
18. Willard, D. E., *Searching unindexed and nonuniformly generated files in* $\log\log N$ *time*, SIAM J. Comput., 14 (1985), pp. 1013–1029.
19. Yao, A. C. and F. F. Yao, *The complexity of searching an ordered random table*, 17^{th} FOCS, (1976), pp. 173–175.