Balls in Boxes:
Variations on a Theme of Warren Ewens and Herbert Wilf

Shalosh B. EKHAD and Doron ZEILBERGER[1]

To Herbert Saul Wilf (b. June 13, 1931), on his 80-th birthday

Preface

There are r boys and n girls. Each boy must pick *one* girl to invite to be his date in the prom. Although each girl expects to get $R := r/n$ invitations, most likely, many of them would receive less, and many of them would receive more. Suppose that Nilini, the most "popular" girl, got as many as $m+1$ prom-invitations, is she indeed so popular, or did she just "luck-out"?

Each one of r students has to choose from n different parallel Calculus sections, taught by different professors. Although each professor expects to get $R := r/n$ students signing-up, most likely, many of them would receive less, and many of them would receive more. Suppose that Prof. Niles, the most "popular" professor got as many as $m+1$ students, is Prof. Niles justified in assuming that she is more popular than her peers, or did she just "luck-out"?

It is Saturday night, and there are r people who have to decide where to dine, and they have n restaurants to choose from. Although each restaurant expects to get $R := r/n$ diners, most likely, many of them would receive less, and many of them would receive more. Suppose that the Nevada Diner, the most "popular" restaurant, got as many as $m+1$ diners, can they congratulate themselves for the quality of their food, or ambiance, or location, or can they only congratulate themselves for being lucky?

Each one of r cases of acute lymphocitic leukemia has to choose one of n towns (artificially made all with equal-populations) where to happen. Although each town expects to get $R := r/n$ cases, most likely, many of them would receive less, and many of them would receive more. Suppose that the Illinois town Niles had $m+1$ cases of that disease, do its people have to be concerned about their environment, or is it only Lady Luck's fault?

Of course all these questions have the same answer, and typically one talks about r balls being placed, uniformly at random, in n boxes, where the largest number of balls that landed at the same box was $m+1$. Yet another way: A monkey is typing an r-letter word using a keyboard of an alphabet with n letters, and the most frequent letter showed-up $m+1$ times. Does the typing

[1] Department of Mathematics, Rutgers University (New Brunswick), Hill Center-Busch Campus, 110 Frelinghuysen Rd., Piscataway, NJ 08854-8019, USA. `zeilberg at math dot rutgers dot edu` ,
`http://www.math.rutgers.edu/~zeilberg/` . First Version: June 13, 2011. This version: Dec. 14, 2011. We wish to thank Eugene Zima for helpful corrections. Accompanied by Maple package
`http://www.math.rutgers.edu/~zeilberg/tokhniot/BallsInBoxes` . Sample input and output can be gotten from: `http://www.math.rutgers.edu/~zeilberg/mamarim/mamarimhtml/bib.html` .
Supported in part by the National Science Foundation of the United States of America.

monkey have a particular fondness for that letter, or is he a truly uniformly-at-random monkey who does not play favorites with the letters?

Asking the Right Question

As Herb Wilf pointed out so eloquently in his wonderful talk at the conference W80 (celebrating his 80th birthday) (based, in part, on [EW]), using the depressing disease formulation, the right questions are **not**:

"What is the probability that Nilini would get so many ($m+1$ of them) prom-invitations?"

"What is the probability that Prof. Niles would get so many ($m+1$ of them) students?"

"What is the probability that the Nevada Diner would get so many ($m+1$ of them) diners?"

"What is the probability that Niles, IL would get so many ($m+1$ of them) cases of acute lymphocitic leukemia?"

Even though this is the wrong question (whose answer would make Nilini, Prof. Niles and the Nevada Diner's successes go to their heads, and would make the real-estate prices in Niles, IL, plummet), because it is so tiny, and seemingly extremely unlikely to be "due to chance", let's answer this question anyway.

The *a priori* probability of Nilini getting $m+1$ or more prom-invitations, using the *Poisson Approximation* is:

$$e^{-R}(\sum_{i=m+1}^{\infty} \frac{R^i}{i!}) = e^{-R}(e^R - \sum_{i=0}^{m} \frac{R^i}{i!}) = 1 - e^{-R}\sum_{i=0}^{m} \frac{R^i}{i!} \quad ,$$

indeed very small if m is considerably larger than R.

But *a priori* we don't know who would be the "lucky champion" (or the unlucky town), the **right** question to ask is:

The Right Question: Given r, n, and m, compute (if possible exactly, but at least approximately):

$P(r, n, m) :=$ the probability that *every* box got $\leq m$ balls.

Getting the Right Answer to the Right Question, as Fast as Possible

In [EW], Ewens and Wilf present a beautiful, *fast* ($O(mn)$), algorithm for computing the *exact* value of $P(r, n, m)$, that employs a method that is described in the Nijenhuis-Wilf classic [NW] (but that has been around for a long time, and rediscovered several times, e.g. by one of us ([Z1]), and before that by J.C.P. Miller, and according to Don Knuth the method goes back to Euler. At any rate, [EW] does not claim novelty for the method, only for *applying* it to the present problem).

The *specific* real-life examples given in [EW] were:

1. (Niles, IL): $r = 14400, n = 9000$, (so $R = 8/5$), $m = 7$. Using their method, they got (in less than one second!) the value

$$P(14400, 9000, 7) = 0.095395913167130399971555481626\ldots \quad ,$$

meaning that the probability that *every* town in the US, of the size of Niles, IL, would get no more than 7 cases is less than ten percent. So with probability 0.904604086832869600002844451837, *some* town (of the same size, assuming, artificially that the US has been divided into towns of that size) somewhere, in the US, would get *at least* eight cases. There is (most probably) nothing wrong with their water, or their air-quality, the only one that they may blame is Lady Luck!

For comparison, the *a priori* probability that Niles,IL would get 8 or more cases is roughly:

$$1 - e^{-1.6} \sum_{i=0}^{7} \frac{1.6^i}{i!} = 0.00026044\ldots \quad ,$$

a real reason for (unjustified!) concern.

2. (Churchill County, NV): $r = 8000, n = 12000$, (so $R = 2/3$), $m = 11$. Using their method, they got (in less than one second!) the value

$$P(8000, 12000, 11) = 0.999999895529647647310726013392\ldots \quad ,$$

so it is extremely likely that *every* district got at most 11 cases, and the probability that *some* district got 12 or more cases is indeed small, namely

$$1 - P(8000, 12000, 11) = 0.104470 \cdot 10^{-6} \quad ,$$

so these people should indeed panic.

For comparison, the *a priori* probability that Churchill County, NV, would get 12 or more cases is roughly:

$$1 - e^{-2/3} \sum_{i=0}^{11} \frac{(2/3)^i}{i!} = .870586315 \cdot 10^{-11} \quad ,$$

in that case people would have been right to be concerned, but for the wrong reason!

The Maple package `BallsInBoxes`

This article is accompanied by the Maple package `BallsInBoxes` available from:
`http://www.math.rutgers.edu/~zeilberg/tokhniot/BallsInBoxes` .

Lots of sample input and output files can be gotten from:
`http://www.math.rutgers.edu/~zeilberg/mamarim/mamarimhtml/bib.html` .

How to Compute $P(r,n,m)$ Exactly?

Easy! As Ewens and Wilf point out in [EW], and Herb Wilf mentioned in his talk, there is an obvious, explicit, "answer"

$$P(r,n,m) = \frac{1}{n^r} \sum \frac{r!}{r_1! r_2! \dots r_n!} \quad ,$$

where the sum ranges over the set of n-tuples of integers

$$A(r,n,m) := \{(r_1, r_2, \dots, r_n) \mid 0 \le r_1, \dots, r_n \le m \quad , \quad r_1 + r_2 + \dots + r_n = r\} \quad .$$

So "all" we need, in order to get the *exact* answer, is to construct the set $A(r,n,m)$ and add-up all the multinomial coefficients.

Of course, there is a better way. As it is well-known (see [EW]), and easy to see, writing

$$P(r,n,m) = \frac{r!}{n^r} \sum_{(r_1,\dots,r_n) \in A(r,n,m)} \frac{1}{r_1! r_2! \dots r_n!} \quad ,$$

the $\sum$ is the coefficient of x^r in the expansion of

$$\left(\sum_{i=0}^{m} \frac{x^i}{i!} \right)^n \quad ,$$

so all we need is to go to Maple, and type (once r, n, and m have been assigned numerical values)

```
r!/n**r*coeff(add(x**i/i!,i=0..m)**n,x,r);
```

.

This works well for small n and r, but, please, **don't even try** to apply it to the first case of [EW], $(r = 14400, n = 9000, m = 7)$, Maple would crash!

Ewens and Wilf's brilliant idea was to use the Euler-Miller-(Nijenhuis-Wilf)-Zeilberger-... "quick" method for expanding a power of a polynomial, and get an *answer* in less than a second!

[We implemented this method in Procedure `Prnm(r,n,m)` of `BallsInBoxes`].

While their method indeed takes less than a second (in Maple) for $r = 14400, n = 9000$ (and $7 \le m \le 12$), it takes quite a bit longer for $r = 144000, n = 90000$, and we are willing to bet that for $r = 10^8, n = 10^8$ it would be hopeless to get an *exact answer*, *even* with this fast algorithm.

But why this obsession with *exact* answers? Hello, this is *applied* mathematics, and the epidemiological data is, of course, *approximate* to begin with, and we make lots of unrealistic assumptions (e.g. that the US is divided into 9000 towns, each exactly the size of Niles, IL.) . All we need to know is, "are that many diseases likely to be due to pure chance, or is it a cause for concern?", *Yes or No?*, *Ja oder Nein?*, *Oui ou Non?*, *Ken o Lo?*.

Enumeration Digression

It would be nice to get a more compact (than the huge multisum above) (symbolic) "answer", or "formula", in terms of the *symbols* r, n and m. This seems to be hopeless. But fixing, positive integers a, b and m, one can ask for a "formula" (or whatever), in n, for the quantity $P(an, bn, m)$ that can be written as $B(a, b, m; n)/(an)^{bn}$ where

$$B(a, b, m; n) := (an)! \sum_{(r_1,\dots,r_n) \in A(an,bn;m)} \frac{1}{r_1! r_2! \dots r_n!} \quad ,$$

the cardinality of the *natural* combinatorial set consisting of placing an balls in bn boxes in such a way that no box receives more than m balls. Equivalently, all *words* in a bn-letter alphabet, of length an, where no letter occurs more than m times. For example, when $a = b = m = 1$, we have the deep theorem:

$$B(1, 1, 1; n) = n! \quad .$$

Equivalently, $e(n) = B(1, 1, 1; n)$ is a solution of the *linear recurrence equation with polynomial coefficients*

$$e(n + 1) - (n + 1)e(n) = 0 \quad , \quad (n \geq 0) \quad ,$$

subject to the *initial condition* $e(0) = 1$.

It turns out that, thanks to the not-as-famous-as-it-should-be *Almkvist-Zeilberger* algorithm [AZ] (an important component of the deservedly famous *Wilf-Zeilberger Algorithmic Proof Theory*), one can find similar recurrences (albeit of higher order, so it is no longer "closed-form", in n) for the sequences $B(a, b, m; n)$ for any *fixed* triple of positive integers, a, b, m.

(See Procedures `Recabm` and `RacabmV` in the Maple package `BallsInBoxes`).

Indeed, since $B(a, b, m; n)$ is $(an)!$ times the coefficient of x^{an} in

$$\left(\sum_{i=0}^{m} \frac{x^i}{i!} \right)^{bn} \quad ,$$

it can be expressed, (thanks to *Cauchy*), as

$$\frac{(an)!}{2\pi i} \oint_{|z|=1} \frac{\left(\sum_{i=0}^{m} \frac{z^i}{i!} \right)^{bn}}{z^{an+1}} \, dz, \qquad (Cauchy)$$

and this is game for the Almkvist-Zeilberger algorithm, that has been incorporated into `BallsInBoxes`. See the web-book

`http://www.math.rutgers.edu/~zeilberg/tokhniot/oBallsInBoxes2`

for these recurrences for $1 \leq a, b \leq 3$ and $1 \leq m \leq 6$.

Asymptotics

Once the first-named author of the present article computed a recurrence, it can go on, thanks to the *Birkhoff-Trzcinski method* ([WimZ][Z2]), to get very good asymptotics! So now we can get a very precise asymptotic formula (in n) (to any desired order!) for $P(an, bn, m)$, that turns out to be very good for large, and even not-so-large n, and for *any* desired a, b, m. Procedure `Asyabm` in our Maple package `BallsInBoxes` finds such asymptotic formulas. See

`http://www.math.rutgers.edu/~zeilberg/tokhniot/oBallsInBoxes1`

for asymptotic formulas, derived by combining Almkvist-Zeilberger with `AsyRec` (also included in `BallsInBoxes` in order to make the latter self-contained.)

This works for *every* m, and *every* a and b, in principle! In practice, as m gets larger than 10, the recurrences become very high order, and take a very long time to derive.

But as long as $m \leq 8$ and even (in fact, especially) when n is very large, this method is much faster than the method of [EW] ($O(mn)$ with large n is not that small!). Granted, it does not give you an *exact* answer, but neither do they (in spite of their claim, see below!) .

But let's be pragmatic and forget about our purity and obsession with "exact" answers. Since we know from "general nonsense" that the desired probability

$$C(a, b, m; n) := P(an, bn, m) \quad (= B(a, b, m; n)/(an)^{bn})$$

behaves asymptotically as

$$C(a, b, m; n) \asymp \mu^n (c_0 + O(1/n)) \quad ,$$

for *some* numbers μ and c_0, all we have to do is crank out (e.g.) the 200-th and 201-th term and estimate μ to be $C(a, b, m; 201)/C(a, b, m; 200)$, and then estimate c_0 to be $C(a, b, m; 200)/\mu^{200}$. Using Least Squares one can do even better, and also estimate higher order asymptotics (but we don't bother, enough is enough!).

Procedure `AsyabmEmpir` in our Maple package `BallsInBoxes` uses this method, and gets very good results!

For example, for the Niles, IL, example, in order to get estimates for $P(14400, 9000, m)$, typing

```
evalf(subs(n=1800,AsyabmEmpir(8,5,m,200,n)));
```

for $m = 7, 8, 9, 10, 11, 12$ yields (almost instantaneously)

$m = 7$: $0.09540287131\ldots$ (the exact value being: $0.095395913167\ldots$) ,

$m = 8$: $0.664971462304\ldots$ (the exact value being: $0.66495441\ldots$) ,

$m = 9$: $0.9378712268719\ldots$ (the exact value being: $0.93786433\ldots$) ,

$m = 10$: $0.990845139\ldots$ (the exact value being: $0.9908433\ldots$) ,

$m = 11$: $0.998789295\ldots$ (the exact value being: $0.99878892861\ldots$) .

The advantage of the present approach is that we can handle very large n, for example, with the same effort we can compute

```
evalf(subs(n=180000,AsyabmEmpir(8,5,m,200,n)))
```

getting, for example, that $P(1440000, 900000, 11)$ is very close to 0.88554890636027. The method used in [EW] (i.e. typing `Prnm(1440000,900000,11);` in `BallsInBoxes`) would take forever!

Caveat Emptor

There is another problem with the $O(mn)$ method described in [EW]. Sure enough, it works well for the examples given there, namely $P(14400, 9000, m)$ for $6 \leq m \leq 12$ and $P(8000, 12000, m)$ for $4 \leq m \leq 8$.

This is corroborated by our implementation of that method, (Procedure `Prnm(r,n,m)` in `BallsInBoxes`).

Typing (once `BallsInBoxes` has been read onto a Maple session):

```
t0:=time():  Prnm(14400,9000,9) , time()-t0;
```

returns

$0.937864339305858219725360911354, 0.884$

that tells you the desired value (we set `Digits` to be 30), and that it took 0.884 seconds to compute that value.

But now try:

```
t0:=time():  Prnm(1000,100,15), time()-t0;
```

and get in 0.108 seconds (real fast!)

$-0.728465229161818857989128673465 \cdot 10^{50}$.

"Something is rotten in the State of Denmark!" We learned in kindergarten that a *probability* has to be between 0 and 1, so a negative probability, especially one with 50 decimal digits, is a bit fishy. Of course, the problem is that [EW]'s "exact" result is not really *exact*, as it uses floating-point arithmetic.

Big deal, since we work in Maple, let's increase the system variable `Digits` (the number of digits used in floating-point calculations), and type the following line:

```
evalf(Prnm(1000,100,15),80);
```

getting 5.71860506564981..., a little bit better! (the probability is now less than six, and at least it is positive!), but still nonsense.

`Digits:=83` still gives you nonsense, and it only starts to "behave" at `Digits:=90`.

Now let's multiply the inputs, r and n by 10, and take $m = 22$ and try to evaluate $P(10000, 1000, 22)$. Even `Digits:=250` still gives nonsense! Only Digits:=310 gives you something reasonable and (hopefully) correct.

The way to overcome this problem is to keep upping `Digits` until you get close answers with both `Digits` and, say, `Digits+100`. This is implemented in Procedure
`PrnmReliable(r,n,m,k)` in `BallsInBoxes`, if one desires an accuracy of k decimal digits. This is *reliable* indeed, but **not** exact, and *not* rigorous, since it uses numerical heuristics. The exact answer is a *rational number*, that is implemented in Procedure `PrnmExact(r,n,m)` of `BallsInBoxes`.

The Cost of Exactness

If you type

```
t0:=time():PrnmExact(14400,9000,7):  time()-t0;
```

you would get in 42 seconds (no longer that fast!) a *rational number* whose numerator and denominator are *exact* integers with 54207 digits.

See

`http://www.math.rutgers.edu/~zeilberg/tokhniot/oBallsInBoxes7a`

for the outputs (and timings) of `PrnmExact(14400,9000,m);` for m between 6 and 12 and see

`http://www.math.rutgers.edu/~zeilberg/tokhniot/oBallsInBoxes7b`

for the outputs (and timings) of `PrnmExact(8000,12000,m);` for m between 4 and 8. No longer fast at all! (2535 and 248 seconds respectively).

Let's Keep It Simple: An Ode to the Poisson Approximation

At the end of [EW], the authors state:

" A Poisson Approximation is also possible but it may be inaccurate, particularly around the tails of the distribution. Our exact method is fast and does not suffer from any of those problems."

Being curious, we tried it out, to see if it is indeed so bad. Surprise, it is terrific! But let's first review the Poisson approximation as we understand it.

The probability of any particular box (of the n boxes) getting $\leq m$ ball is, roughly, using the Poisson approximation ($R := r/n$):

$$e^{-R} \sum_{i=0}^{m} \frac{R^i}{i!} \quad .$$

Of course the n events are **not** independent, but let's pretend that they are. The probability that *every* box got $\leq m$ balls is approximated by

$$Q(r,n,m) := \left(e^{-R} \sum_{i=0}^{m} \frac{R^i}{i!} \right)^n \quad .$$

[$Q(r,n,m)$ is implemented by procedure `PrnmPA(r,n,m)` in `BallsInBoxes`. It is as fast as lightning!]

Ewens and Wilf are very right when they claim that $P(r,n,m)$ and $Q(r,n,m)$ are very far apart around the "tail" of the distribution, but who cares about the tail? Definitely not a scientist and even not an applied mathematician. It turns out, empirically (and we did extensive numerical testing, see Procedure `HowGoodPA1(R0,N0,Incr,M0,m,eps)` in `BallsInBoxes`), that whenever $P(r,n,m)$ is not extremely small, it is very well approximated by $Q(r,n,m)$, and using the latter (it is so much faster!) gives very good approximations, and enables one to construct the "center" of the probability distribution (i.e. ignoring the tails) very accurately. See

`http://www.math.rutgers.edu/~zeilberg/tokhniot/oBallsInBoxes4` ,

and

`http://www.math.rutgers.edu/~zeilberg/tokhniot/oBallsInBoxes5` ,

for comparisons (and timings!, the Poisson Approximation wins!) .

In particular, the estimates for the *expectation*, *standard deviation*, and even the higher moments match extremely well!

Another (empirical!) proof of the fitness of the Poisson Approximation can be seen in:

`http://www.math.rutgers.edu/~zeilberg/tokhniot/oBallsInBoxes1`

where the (rigorous!) asymptotic formulas derived, via `AsyRec`, from the recurrences obtained via the Almkvist-Zeilberger algorithm are very close to those predicted by the Poisson Approximation (except for very small m, corresponding to the "tail").

The Full Probability Distribution of the Random Variable "Maximum Number of Balls in the Same Box"

It would be useful, for given positive integers a and b, to know how the probability distribution "maximum number of balls in the same box when throwing an balls into bn boxes" behaves. One can "empirically" construct (without arbitrarily improbable tail) the distribution of the random variable "maximum number of balls in the same box" when an balls are uniformly-at-random placed in bn boxes (Let's call it $X_n(a,b)$, and X_n for short) using

$$Pr(X_n = m) = P(an, bn, m) - P(an, bn, m-1) \quad .$$

First, and foremost, what is the expectation, μ_n, of this random variable? Second, what is the standard deviation, σ_n?, skewness?, kurtosis?, and it would be even nice to know higher α-coefficients (alias moments of $Z_n := (X_n - \mu_n)/\sigma_n$), as asymptotic formulas in n.

For the expectation, μ_n, Procedure `AveFormula(a,b,n,d,L,k)` uses the more accurate "empirical approach" and Maple's built-in Least-Squares command, to obtain the following empirical (symbolic!) estimates for the expectation.

$a = 1, b = 1$: `evalf(AveFormula(1,1,n,1,300,1000,10),10);` yields that

μ_n is roughly $2.293850526 + (0.4735983525) \cdot \log n$

$a = 2, b = 1$: `evalf(AveFormula(2,1,n,1,300,1000,10),10);` yields that

μ_n is roughly $3.963420618 + (0.5834252496) \cdot \log n$

$a = 1, b = 2$: `evalf(AveFormula(1,2,n,1,300,1000,10),10);` yields that

μ_n is roughly $1.640094145 + (0.3873602232) \cdot \log n$.

Note that for $a = 1, b = 1$, the approximation to μ_n can be written
$2.293850526 + (1.090500507) \cdot \log_{10} n$, so a "rule-of-thumb" estimate for the expectation when n balls are thrown into n boxes is a bit more than 2 plus the number of (decimal) digits.

Procedure `NuskhaPA1(R,n,K,d)` uses the Poisson Approximation to guess polynomials in $\log n$ of degree d fitting the average, standard deviation, and higher moments, as asymptotic expressions in n, for nR balls thrown into n boxes, where R is now any (numeric) *rational* number. Even $d = 1$ seems to give a fairly good fit, so they all seem to be (roughly) linear in $\log n$.

Procedure `SmallestmPA`

Procedure `SmallestmPA(r,n,conf)` gives you the smallest m for which, with confidence `conf`, you can deduce that the high value of m is **not** due to chance (using the Poisson Approximation). For example

```
SmallestmPA(14400,9000,.99);
```

yields 10, meaning that if a town the size of Niles, IL got 10 or more cases, then with probability > 0.99 it is not just bad luck. If you want to be %99.99-sure of being a victim of the environment rather than of Lady Luck, type:

```
SmallestmPA(14400,9000,.9999);
```

and get 13, meaning that if you had 13 cases, then with probability larger than 0.9999 it is not due to chance.

The Minimum Number of Balls that Landed in the Same Box, Procedure `LargestmPA`

An equally interesting, and harder to compute, random variable is the *minimum number of balls that landed in the same box*, but the Poisson Approximation handles it equally well. Analogous to `SmallestmPA`, we have, in `BallsInBoxes`, Procedure `LargestmPA(r,n,conf)` that tells you the largest m for which you can't blame luck for getting m or less balls.

For example, if there are 10000 students that have to decide between 100 different calculus sections,

`LargestmPA(10000,100,.99);`

that happens to be 66, tells you that any section that only has 66 students or less, with probability > 0.99, it is because that professor (or time slot, e.g. if it is an 8:00am class) is not popular, and you can't blame bad luck.

`LargestmPA(10000,100,.9999);`

that outputs 57, tells you that anyone who only had ≤ 57 students enrolled is unpopular with probability $>$ %99.99, and can't blame bad luck.

On the other end, going back to the original problem,
`SmallestmPA(10000,100,.99);` yields 139, telling you that any section for which 139 or more students signed up is *probably* (with prob. > 0.99) due to the popularity of that section, while `SmallestmPA(10000,100,.9999);` yields 151.

Final Comments

1. One can possibly (using the *saddle-point method*) get asymptotic formulas from the contour integral (*Cauchy*), but this is not *our* cup-of-tea, so we leave it to other people.

2. Another "back-of-the-envelope" "Poisson Approximation" is to argue that since the probability of any individual box getting strictly more than m balls is roughly (recall that $R = r/n$)

$$e^{-R} \sum_{i=m+1}^{\infty} \frac{R^i}{i!} = e^{-R}(e^R - \sum_{i=0}^{m} \frac{R^i}{i!}) = 1 - e^{-R} \sum_{i=0}^{m} \frac{R^i}{i!},$$

by the *linearity of expectation*, the expected number of *lucky* (or *unlucky* if the balls are diseases) boxes exceeding m balls is roughly

$$n \left(1 - e^{-R} \sum_{i=0}^{m} \frac{R^i}{i!} \right) \quad .$$

In the case of Niles, IL, the expected number of towns that would get 8 or more cases is:

$$9000 \left(1 - e^{-1.6} \sum_{i=0}^{7} \frac{(1.6)^i}{i!} \right) = 2.343961376410372 \quad ,$$

so it is not at all surprising that at least one town got as many as 8 cases. On the other hand, in the other example $r = 8000, n = 12000, m = 12$, the expected number of unfortunate counties is:

$$12000 \left(1 - e^{-(2/3)} \sum_{i=0}^{12} \frac{(2/3)^i}{i!} \right) = 0.533706802 \cdot 10^{-8} \quad ,$$

so it is indeed a reason for concern.

Conclusion

We completely agree with Ewens and Wilf that simulation takes way too long, and is not that accurate, and that *their* method is far superior to it. But we strongly disagree with their dismissal of the Poisson Approximation. In fact, we used their ingenious method to conduct extensive empirical (numerical) testing that established that the Poisson Approximation, that they dismissed as "inaccurate", is, as a matter of fact, sufficiently accurate, and far more reliable, in addition to being yet-much-faster! It is much safer to use the Poisson Approximation than to use their "exact" method (in floating-point arithmetic), and when one uses *truly* exact calculations, in rational arithmetic, their "fast" method becomes *anything but.*

Even when the floating-point problem is addressed by using multiple precision (`PrnmReliable` discussed above), their fast algorithm becomes slow for very large r and n, while the Poisson Approximation is almost instantaneous even for very large r and n, and *any* m.

So while we believe that the algorithm in [EW] is not as *useful* as the Poisson Approximation, it sure was *meta-useful*, since it enabled us to conduct extensive numerical testing that showed, *once and for all*, that it is far less useful then the latter.

Additional evidence comes from our own symbolic approach (fully rigorous for $m \leq 9$ and semi-rigorous for higher values of m), that establishes the adequacy of the Poisson Approximation for *symbolic* n.

Finally, as we have already pointed out, since the data that one gets in applications is always *approximate* to begin with, insisting on an "exact" answer, even when it is easy to compute, is unnecessary.

Coda: But We, Enumerators, Do Care About Exact Results!

Our point, in this article, was that for *applications* to statistics, the Poisson Approximation suffices. But *we* are *not* statisticians. We are *enumerators*, and we do like exact results! The approach of [EW] enables us to know, for example, in less than one second the **exact** number of ways that 1001 balls can be placed in 1001 boxes such that no box received more than 7 balls. Just type (in `BallsInBoxes`)

```
(1001**1001)*PrnmExact(1001,1001,7);
```

and get a beautiful **exact** integer with 3004 digits!

Typing

```
(1001**1001)*PrnmPA(1001,1001,7);
```

will give you something fairly close (the ratio being $0.9997852\ldots$)

but for a **pure** enumerator, this is very unsatisfactory. So long live exact answers!, but *not* in statistics.

References

[AZ] Gert Almkvist and Doron Zeilberger, *The Method of Differentiating Under The Integral Sign*, J. Symbolic Computation **10** (1990), 571-591. [Available on-line from:
`http://www.math.rutgers.edu/~zeilberg/mamarim/mamarimPDF/duis.pdf`]

[EW] Warren J. Ewens and Herbert S. Wilf, *Computing the distribution of the maximum in balls-and-boxes problems with application to clusters of disease cases*, Proc. National Academy of Science (USA) **104(7)** (July 3, 2007), 11189-1191 . [Available on-line from:
`http://www.pnas.org/content/104/27/11189.full.pdf`]

[WimZ] Jet Wimp and Doron Zeilberger, *Resurrecting the asymptotics of linear recurrences*, J. Math. Anal. Appl. **111** (1985), 162-177. [Available on-line from:
`http://www.math.rutgers.edu/~zeilberg/mamarimY/WimpZeilberger1985.pdf`]

[Z1] Doron Zeilberger, *The J.C.P. Miller Recurrence for exponentiating a polynomial, and its q-Analog*, J. Difference Eqs. and Appl. **1** (1995), 57-60. [Available on-line from:
`http://www.math.rutgers.edu/~zeilberg/mamarim/mamarimhtml/power.html`]

[Z2] Doron Zeilberger *AsyRec: A Maple package for Computing the Asymptotics of Solutions of Linear Recurrence Equations with Polynomial Coefficients*, The Personal Journal of Shalosh B. Ekhad and Doron Zeilberger `http://www.math.rutgers.edu/~zeilberg/pj.html`, April 6, 2008. [Article and package available on-line from:
`http://www.math.rutgers.edu/~zeilberg/mamarim/mamarimhtml/asy.html`]