Paper CC-20

PROC FORMAT in DATA Step mathematics

Elizabeth Schreiber, DC SAS Users Group, Washington, DC

ABSTRACT

PROC FORMAT and data step mathematics can be used to bypass computational limitations to calculate probability estimates of exceedingly rare events. A client needed to assess the likelihood of finding a defect, given that one hadn't yet been found in thousands of tests. Standard binomial tables extend only to 500 trials. The formula cannot be calculated directly and even the numeric approximation was intractable given the available hardware. The numbers even exceeded the capacity of SAS®'s combination and factorial functions. A review of publications extending back through 1964, the application of mathematical methods to simplify calculations, and a custom-written PROC FORMAT and SAS data step led to an answer for the client … vanishingly small.

INTRODUCTION

The client was interested in examining the reliability of a process. The desired answer was the likelihood of finding a defect, given that one hadn't yet been found, and if one was ever found, the likelihood of finding another. The formulas and iterative solution algorithm are detailed in the later in this paper. Essentially, for each combination of r and n (successes and trials), successive "guesses" at p are made and the answer is the p for which the stopping criteria is "small enough." For small values of r and n, an iterative algorithm was accomplished in a DATA STEP using a DO loop and the COMB function. An initial guess at p was used to calculate Δp which is added to p to form the next guess.

THE CALCULATION CHALLENGE

To replicate the original table, the iterative procedure was accomplished in a DATA STEP using a DO loop and the COMB function to calculate nCr and PC SAS 9.1.3. Scaling the calculations from the n=500 maximum of the original paper to the n=5000 required by the client was not straightforward. The calculations require a computed value of nCr for n=4 to n=5,000. PC SAS 9.1.3 couldn't do the calculation. Using these n and r with the factorial function nCr=FACT (n) / (FACT(n-r)FACT(r)) also exceeded the capability.

Borrowing inspiration from the original paper, logarithms can be used and $nCr = 10^z$ where

$$z = \sum_{x=1}^{n} \log(x) - \left(\sum_{x=1}^{n-r} \log(x) + \sum_{x=1}^{r} \log(x) \right).$$

Using this arithmetic relationship, a DATA STEP was used to build a table of logarithms from 4 to 5000.

This data set was used as the input control data set for the FORMAT procedure to create the format SUMlog. The SUMlog format was applied to n, r, and n-r using the DATA STEP expressions below to calculate nCr for n=1 to 5,000 and from r=n/2 when n is even or r=(n-1)/2 for odd n to r.

```
lognF=input(put(n,SUMlog.),best32.);
lognrF=input(put(nminusr,SUMlog.),best32.);
logrF=input(put(r,SUMlog.),best32.);
lognCr=lognF-(lognrF+logrF);
nCr=10**lognCr;
```

```
data ctrl;
   format label best32.;
   retain fmtname 'SUMlog'
          type 'n'
          SUMlogn 0;
   do n=1 to 5000;
      logn=log10(n);
      SUMlogn=SUMlogn+logn;
         start=n;
         label=SUMlogn;
      output;
   end;
run;
```

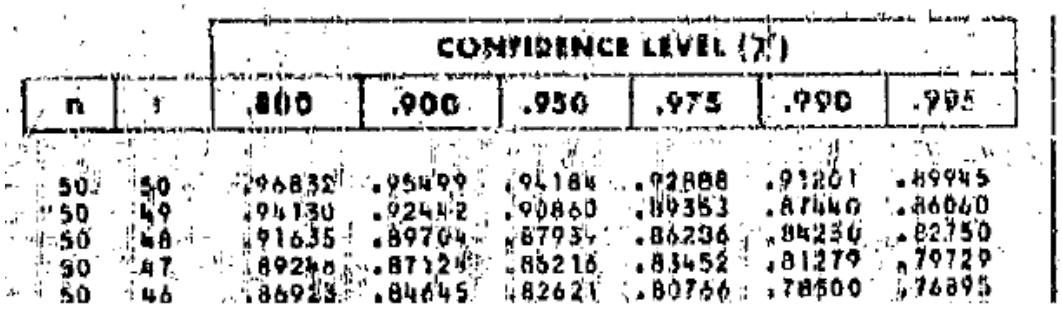
A data set of these n, r, and lognCr was the starting point for the iterative process used to produce each entry of the expanded binomial reliability table. Because that iterative process used the lognCr in additional calculations, this was sufficient. SAS 9.1.3 and Windows XP run into the same numeric limitations when doing the nCr = 10**lognCr calculation. To obtain human-readable nCr numbers in scientific notation for lognCr as high as 1503.2, additional data step lines are needed.

```
integer=int(lognCr);
remainder=lognCr-integer;
nCr_rem=10**remainder;
nCr_text=put(nCr_rem,best5.)||"E"||left(integer);
```

FORMULAS AND ITERATIVE ALGORITHM

Essentially, the client wanted a table similar to the binomial reliability table presented in Cooke, Lee, and Vanderbeck's 1964 publication (Cooke *et al.*, 1964, p.vi) which states:

> **Example**: A sample of size 50 is randomly selected from a population whose reliability we wish to predict. Forty-eight of the items tested are successful. Using the table, find a lower confidence limit so that the true population *p* will be equal to or greater than this value 90% of the time, i.e. if many samples are drawn and a lower confidence limit is computed from each sample, 90% of the time we would be correct in stating that the true population *p* is equal to or greater than this lower limit.
>
> Looking in the table for *n* = 50, *r* = 48; γ =.90 … we find that the lower limit is .89704.
>
> Hence, we are 90% confident that the true population reliability is at least .89



		CONFIDENCE LEVEL (γ)					
n	r	.800	.900	.950	.975	.990	.995
50	50	.96832	.95499	.94184	.92888	.91201	.89945
50	49	.94130	.92442	.90860	.89353	.87440	.86060
50	48	.91635	.89704	.87934	.86236	.84230	.82750
50	47	.89296	.87124	.85216	.83452	.81279	.79729
50	46	.86923	.84645	.82621	.80766	.78500	.76895

Figure 1. Binomial Reliability Table (Cooke 1964, p.7) ("Best Available Copy" enlarged)

MATHEMATICAL ALGORITHM

The binomial distribution is useful when analyzing attribute data (e.g., favorable or unfavorable, reliable or unreliable, etc.). The binomial distribution is defined as

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad \begin{array}{l} n = \text{number of items tested} \\ p = \text{proportion of favorable items in population} \\ x = \text{number of favorable items obtained in a sample of size } n \end{array} \tag{1}$$

Equation (1) gives the probability of obtaining exactly *x* favorable items from a sample of size *n* when the true population proportion of reliable (favorable) items is *p*. This assumes the probability of a favorable event, *p*, remains constant from sample to sample, and every item in the population is equally likely to be chosen.

We are interested in determining the worst that can be expected from an as-yet-to-be-taken sample. From our current sample we obtained *r* favorable events. An estimate $\hat{p}$ of the true population *p* is given by *r/n*. We can construct a limit which will be lower than *p* most of the time by finding the *lower confidence limit*. In calculating this confidence limit, the cumulative form of the binomial distribution, *F(x)*, is used (Mood *et al.* 1974, p. 220).

$$F(x) = \sum_{r}^{x} \binom{n}{x} p^x (1-p)^{n-x} \tag{2}$$

Equation (2) gives the probability of obtaining *r* or more favorable items from a sample of size *n* when the true percent of favorable items in the population is *p*. The lower one-sided confidence limit is obtained by solving the following equation for p_l:

$$\sum_{x=r}^{n} \binom{n}{x} p_l^x (1-p_l)^{n-x} = 1-\gamma \qquad p_l = \text{lower one-sided confidence limit} \tag{3}$$

For known *r*, *n*, and γ, p_l can be determined and we can state with confidence γ that the true population *p* will not be less than p_l. Thus, when we assign gamma to be .95, if we test many samples of size *n* and compute the lower confidence limit each time, *p* will be less than or equal to p_l about 95 times out of 100. We are 100γ% confident that the true population proportion of reliable items *p* is equal to or greater than our lower one-sided confidence limit p_l.

Given n, r, and γ we solve the following equation for p_l:

$$\sum_{x=r}^{n}\binom{n}{x}p_l^{x}(1-p_l)^{n-x} = 1\text{-}\gamma = \alpha \qquad p_l = \text{lower one-sided confidence limit} \tag{4}$$

An iterative procedure was used to determine p_l for γ=1-α. For each value of r, the p_l corresponding to each γ level was computed. When $r = n$, the expression p_hat=10**log10(alpha/n)) was used to solve p_l directly from the equation

$$\log p_l = \frac{\log \alpha}{n} \tag{5}$$

When $r < n$, the first estimate $\hat{p}_l$ of p_l was obtained from previously computed values.

The iterative procedure (modeled after Cooke *et al.*) consisted of finding the value of $\hat{p}_l$, such that the following would hold:

$$\sum_{x=r}^{n}\binom{n}{x}p^{x}(1-p)^{n-x} = \alpha \pm \varepsilon(\alpha) \quad \text{where } \varepsilon(\alpha) < 10^{-6} \tag{6}$$

Having the first estimate of $\hat{p}_l$ and using a second order Taylor series expansion (Wikipedia, 2010), the following equation was solved for Δp:

$$\alpha = B(\hat{p}_l \mid n,r) + B'(\hat{p}_l \mid n,r)\Delta p + \frac{B''(\hat{p}_l \mid n,r)\Delta p^2}{2!} \tag{7}$$

$$\text{where } B(\hat{p} \mid n,r) \text{ denotes} \sum_{x=r}^{n}\binom{n}{x}\hat{p}^{x}(1-\hat{p})^{n-x} \ .$$

Rearranging, grouping like terms, applying the quadratic equation, and solving for Δp, yields two roots.

$$\Delta p = \frac{-B'(\hat{p}_l \mid n,r) \pm \sqrt{B'(\hat{p}_l \mid n,r)B'(\hat{p}_l \mid n,r) - 4\dfrac{B''(\hat{p}_l \mid n,r)}{2!}\left(B(\hat{p}_l \mid n,r) - \alpha\right)}}{2\dfrac{B''(\hat{p}_l \mid n,r)}{2!}} \tag{8}$$

$$B'(\hat{p} \mid n,r) \text{ denotes} \frac{\partial}{\partial \hat{p}} \sum_{x=r}^{n}\binom{n}{x}\hat{p}^{x}(1-\hat{p})^{n-x} \text{ which can be written}$$

$$B'(\hat{p}_l \mid n,r) = \sum_{x=r}^{n}\binom{n}{x}\left[x\hat{p}_l^{\,x-1}(1-\hat{p}_l)^{n-x} - (n-x)\hat{p}_l^{\,x}(1-\hat{p}_l)^{n-x-1}\right],$$

$$\text{and } B''(\hat{p} \mid n,r) \text{ denotes} \frac{\partial^2}{\partial \hat{p}^2} \sum_{x=r}^{n}\binom{n}{x}\hat{p}^{x}(1-\hat{p})^{n-x} \text{ which can be written}$$

$$B''(\hat{p}_l \mid n,r) = \sum_{x=r}^{n}\binom{n}{x}\left[x(x-1)\hat{p}_l^{\,x-2}(1-\hat{p}_l)^{n-x} - 2x(n-x)\hat{p}_l^{\,x-1}(1-\hat{p}_l)^{n-x-1} + (n-x)(n-x-1)\hat{p}_l^{\,x}(1-\hat{p}_l)^{n-x-2}\right]$$

The value of Δp, which minimized $\left|\alpha - B(\hat{p}_l \mid n,r)\right|$, was chosen to correct the estimate. The next estimate $\hat{p}_l$ was $\hat{p}_l + \Delta p$. The process was repeated until the difference $\left|\alpha - B(\hat{p}_l \mid n,r)\right|$ was less than 0.000001. The estimate was rounded to 5 decimal places and printed in the body of the table.

Since the values of $\hat{p}_l$ were rounded, there is error in the final tabular value. There are also unexamined errors due limits of precision and rounding in the computations. However, this degree of accuracy was acceptable to the client.

The calculations require a computed value of nCr for n=4 to n=5,000 and from r=n/2 when n is even or r=(n-1)/2 for odd n to r. This area is shaded in Figure 1. The black shaded area indicates that area where the combination function computes a result using PC SAS 9.1.3. The largest result returned was COMB(1658,1402)=1.794987E308. Using the factorial function nCr=FACT(n)/(FACT(n-r)FACT(r)) doesn't work either; the largest value returned by the

factorial function is FACT(170)=7.2574E306.

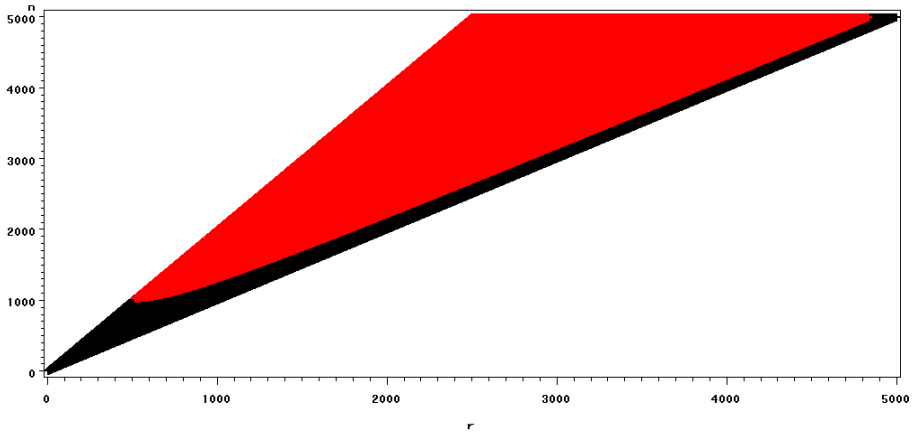


Figure 2. Region for which nCr (the number of combinations of n items chosen r at a time) is needed. The COMB(n,r) function returns values only for the black region. Another calculation method is used for the red region.

As stated above, logarithms can be used. Recall, the logarithm of the product of several numbers is equal to the sum of the logarithms of each. Also, recall that for all integer n >0, n factorial is the product of all positive integers between 1 and n inclusive (Connelly *et al.,* 1980).

Thus, $\binom{n}{r} = nCr = \frac{n!}{(n-r)!r!}$ can be rewritten as $nCr = \frac{\prod_{x=1}^{n} x}{\prod_{x=1}^{n-r} x \prod_{x=1}^{r} x}$ and taking the logarithm gives

$$\log(nCr) = \sum_{x=1}^{n} \log(x) - \left(\sum_{x=1}^{n-r} \log(x) + \sum_{x=1}^{r} \log(x) \right) \text{ and } nCr = 10^{z} \text{ where } z = \sum_{x=1}^{n} \log(x) - \left(\sum_{x=1}^{n-r} \log(x) + \sum_{x=1}^{r} \log(x) \right).$$

CONCLUSIONS

PROC FORMAT was successfully used in DATA STEP mathematics to answer the client's needs. Although the level of numeric inaccuracy is uncharacterized, the inaccuracy is larger when nCr is very large and relatively small for r near n. The client's primary interest was assessing in the likelihood of success in new trials, given that a failure hadn't yet been found in more than four thousand trials. In short, "What is the process reliability?" The table calculated with these methods answers these questions. For example: for n=4000, r=4000, and alpha=.05, we are 95% confident that at least 99.92% of trials will be successes; and for n=4000, r=3999 and alpha=.05, we are 95% confident that at least 99.88% of trials will be successes.

REFERENCES:

Connelly, James F., Fratangelo, Robert A. 1980. *Precalculus Mathematics a Functional Approach, 2nd ed.* New York: MacMillan Publishing Co., Inc.

Cooke, James R, Lee, Mark T, Vanderbeck, John P. 1964. *Binomial Reliability Table (Lower Confidence Limits for the Binomial Distribution).* China Lake, CA: Bureau of Naval Weapons.

Mood, Alexander M, Franklin A Graybill, Duane M Boes. 1974. *Introduction to the Theory of Statistics,3rd ed.* New York: McGraw Hill

Wikipedia. "Taylor Series" http://en.wikipedia.org/wiki/Taylor_series (August 7, 2010).

ACKNOWLEDGMENTS

Thanks are due to Rich Hanlen for bringing Cooke *et al.* to my attention and to ENSCO, Inc., whose client's needs in 2006 inspired the work. I'm also indebted to my parents who lent a 1955 text with the Taylor series expansion in delta notation and my son who re-validated the derivatives in this paper.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Elizabeth Schreiber
DC SAS Users Group www.dc-sug.org
Washington, DC