Paper 181-2011

Exact Confidence Intervals for Risk-Adjusted Rates Versus Trouble in River City

Ted Clay, Clay Software & Statistics, Ashland, Oregon

ABSTRACT

When confidence intervals for risk-adjusted rates are based on large-sample approximations, with truncation in case the lower or upper limit falls outside the range of 0-100%, you know there is Trouble in River City. This paper presents an exact calculation of confidence intervals for risk-adjusted rates, applied to the evaluation of hospital performance. From subjects with known probabilities of an event, the method takes the inverse of the binomial distribution generalized to the case of unequal probabilities. Like the logistic model, the confidence interval is calculated assuming that a group effect is additive in the logit domain, causing a shift of the probabilities as a group. When the exact upper tail probability for the observed events is α/2, the mean of the shifted probabilities is the lower (1-α)% confidence limit on the underlying rate. This generalization of the Clopper-Pearson confidence interval method compares favorably with intervals based on the Poisson and normal distributions. With mid-P adjustment, the resulting intervals have coverage probability close to the nominal probability.

INTRODUCTION

The desire for quality improvement in medical care has led to public reporting of the performance of hospitals and other providers. Confidence intervals, consisting of a lower and upper limit, are used to communicate the extent of knowledge about a given hospital's performance. By convention, these are 95% confidence intervals, indicating that the probability is 0.95 that the true performance is between the lower and upper limits. This can be used to classify a hospital as "better", "average" or "worse", based on where the hospital confidence interval lies relative to a benchmark rate. When the outcome being measured is sensitive to differences in patient-level risk, the confidence interval should be adjusted for the patient-level risk factors, to create a level playing field. The "risk-adjusted rate" is an estimate of what the hospital's performance would have been if the patients at that hospital had a risk profile like the "average" hospital. If that were the case, the hospital's patient outcomes would in theory only reflect the quality of care at that hospital, plus random variation, rather than a difference due to having more or less risky patients.

When the complete data set of all hospitals is available, hierarchical statistical models can be used, in which patient-level risk and hospital rates are estimated in the same model. In some circumstances, this may not be practical or possible. For example, a very large database (either geographically or over time) may be used to accurately estimate the parameters for patient-level risk factors, which then may be made available either in a publication or built into a software module to compute patient-level risk. Also, even when hierarchical methods are possible, some approaches favor the statistical simplicity of first developing a model on the patient level, then summarizing to the hospital level as a subsequent step. Assuming the outcome is an event which either does or does not occur, two examples of a patient-level model are ordinary logistic regression, or simply calculating rates of the event within strata defined by risk categories. In either case, the result is a set of patient-level estimated probabilities of the outcome. The usual next step is to sum the events and probabilities to obtain an observed (O) and expected (E) total on the hospital level, and multiply the O/E ratio times a global rate (G) to get a risk-adjusted rate (R). So,

$$R = G * (O/E) \qquad (1)$$

E is assumed to be fixed, and a confidence interval is calculated for O. More correctly stated, a confidence interval for the "true" underlying hospital rate is estimated based on the observed number of events. If the confidence interval for O is contains E, the hospital is considered "average". Otherwise it is classified as an outlier "above" or "below" average. This paper presents a new exact method of calculating the variability of the observed events, free of distributional assumptions. The resulting confidence intervals are valid with small samples and, after a standard adjustment, have estimated coverage probability close to the nominal value of 95%.

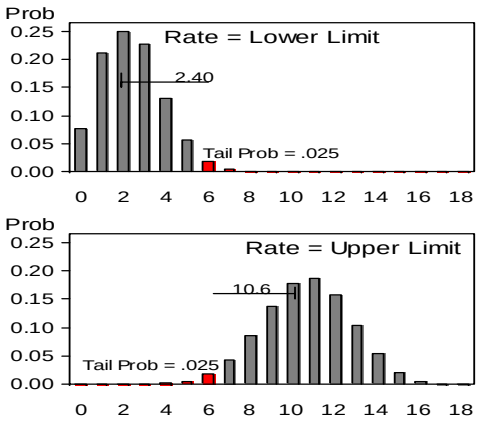
We will use "K" to represent the observed events instead of "O". The rest of this paper is organized as follows: Section 1 explains how the exact confidence interval is calculated. Section 2 covers the effect of diversity of risk. Section 3 discusses coverage probability and the well-known mid-P adjustment. Section 4 compares the exact method to other methods. This is followed by a discussion and conclusion.

SECTION 1 – EXACT CONFIDENCE INTERVAL CALCULATION

STATISTICAL CONTEXT

By definition there is an inverse relationship between a p-value and a confidence interval. Given the observed data and a model with a single parameter, the confidence interval contains all values of the parameter which would not be rejected because the p-value is significant. A two-sided statistical test corresponds to a two-sided confidence interval. In the case of a 95% two-sided confidence interval, the probability (α) of the underlying parameter being outside the interval is .05, with a balanced probability .025 of being in either the upper or lower tail.

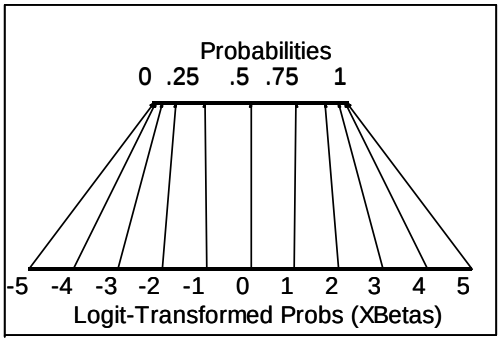
Figure 1: Binomial confidence interval, with K=6, N=18



First we will discuss this approach in a special case, in which all patients in a hospital have the same probability of the event. This reduces to the Clopper-Pearson method of calculating a confidence interval for a binomial proportion [Reference 1]. In this model, there are N subjects, each with an identical probability P of an event, and an observed total of K events. The probability of observing K events or less, the lower tail probability (LTP), is given by the binomial distribution, which in SAS® is the function PROBBNML(P,N,K). The upper tail probability (UTP), of observing K events or more, is given by 1-PROBBNML(P,N,K-1). With given values of K and N, the method searches for the value of P which causes the tail probability to be α/2. This is the confidence limit. Like a rubber-band with increasing tension, it is clear that any more extreme value of P, as a null hypothesis, would cause a reduction in the tail probability to below .025 and would be rejected by a two-tailed statistical test. The Clopper-Pearson method is known as "exact" because of the exact computation based on the binomial distribution. Figure 1 shows the binomial distribution when the value of P is at the two confidence limits, where the tail probabilities equal .025.

Now we will generalize this to the case where each subject can have a different probability P_i of the event. This "generalized binomial" probability calculation is done by a SAS® macro %GENBINOM (see "Access to Macros" section). A call to this macro takes the form %GENBINOM(N,K,P_) in which P_ is now an array of probabilities of length N. While a function can only produce a single value, this macro stores LTP, UTP, and the probabilities of observing each specific numbers of events.

Figure 2: Logistic transformation of probabilities



Just as the Clopper-Pearson method shifts the value of the single parameter P, in the more general case we need to shift the entire collection of individual probabilities. How we chose to shift the probabilities is open for discussion --- any monotone increasing transformation would be possible. One way to shift the set of probabilities P_ is by adding a constant in the logit-transformed domain. The shifted probabilities S_i are calculated by transforming from a probability to an XBeta, back into a probability:

$$XBeta = \log(P_i/(1-P_i)) + constant,$$
$$S_i = \exp(XBeta) / (1 + \exp(XBeta)). \quad (2)$$

One reason to use this particular transformation is that this is the one used in the logistic regression model, in which effects are assumed to be additive in the logit domain. So this method makes no assumptions other than those already used in logistic regression. The shift can be interpreted as a "group effect" parameter in a logistic model. If S_ is the array of shifted probabilities, the desired group effect value has been found when the upper or lower tail probability from %GENBINOM(N,K,S_) equals .025.

To summarize the statistical model, given a set of estimated probabilities P_i, assumed to be known and invariant, the statistical model is that the observed events Yi ~ Binary(S_i), where logit(S_i) = logit(P_i) + group effect, and Logit(P)=log(P/(1-P). The group effect is assumed to be the only unknown parameter. When the exact upper tail

probability for the observed events is α/2, the mean of the shifted probabilities is the lower (1-α)% confidence limit on the underlying rate, and conversely for the upper confidence limit.

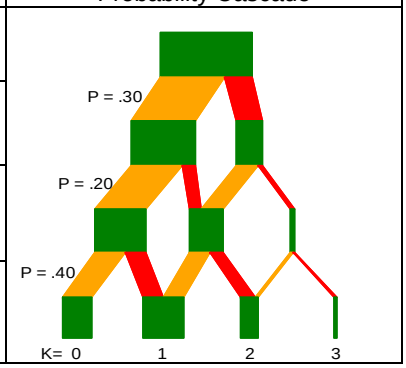
In the special case of K=0, a lower confidence limit cannot be calculated, and is given the value of zero. The upper confidence limit is calculated as with higher values of K. A comparable approach handles K=N.

CALCULATING THE GENERALIZED BINOMIAL PROBABILITIES

The Key Concept: When subject i is added, the total number of events will either increase by 1 if the event occurs with subject i, with probability P_i, or it will stay the same if the event does not occur with subject i, with probability $1-P_i$. So the probability of observing K events after subject i is added can easily be calculated from the probabilities before subject i is added.

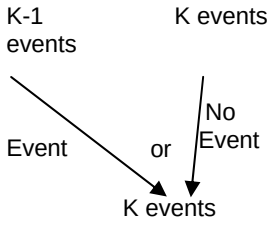
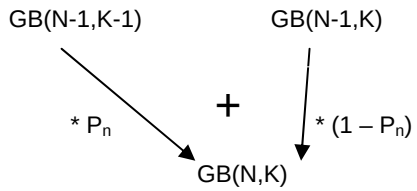
In this discussion, let P_ represent the array of individual probabilities, shifted or not, and let GB(N,K,P_) be the generalized binomial probability of observing exactly K events in N subjects with probability array P_. (These are not cumulative probabilities). For brevity, GB(N,K,P_) is written as GB(N,K). The following shows the relationship between GB(N,K) for various values of K and N.

Table 1: Changing probabilities as number of subjects N increases

		Probability of Observing K Events With N Subjects				
N	Change	K=0 Events	K=1 Event	K=2 Events	K=3 Events	Probability Cascade
0	No Subjects	1.00				
1	Add Subject 1	GB(1,0)	GB(1,1)			
2	Add Subject 2	GB(2,0)	GB(2,1)	GB(2,2)		
3	Add Subject 3	GB(3,0)	GB(3,1)	GB(3,2)	GB(3,3)	

The width of the rectangles in the Probability Cascade figure indicates the size of the probability in the corresponding (left-aligned) cells to the left, for a specific example where the subjects have the probabilities as shown. Rotating the rectangles 90 degrees creates a histogram comparable to Figure 1.

Figure 3: Changes of state, retrospective view from state with K events in N subjects.

Changes of State	Probabilities
	

The probability cascade algorithm works by induction, calculating the solution on each row from the probabilities already calculated on the row above. Formally,

$$GB(N,K) = GB(N-1,K-1) * Pn + GB(N-1,K)*(1-Pn). \quad (3)$$

On the edges of the triangle the above expression reduces to a single term. Starting at the top with the trivial case of N=0, the probability GB(0,0) = 1.0. (If there are no subjects, there are no events with certainty.) From there, all GB(N,K) probabilities are calculated, proceeding down through the triangle one row at a time. The goal is to obtain the set of probabilities on the bottom line. Table 1 with its internal relationships is a weighted version of Pascal's Triangle. This elegant algorithm for calculating the exact probability was published by Luft and Brown in 1993 [Reference 2], who attributed the idea to a conversation with renowned statistician John Tukey.

For some applications, the probabilities GB(N,K) for each value of K are needed in their own right. An example is the calculation of coverage probability discussed later in this paper. But for p-values or for the confidence interval calculation, one is interested in the cumulative probabilities. The lower tail probability (LTP) is the sum of GB(N,x) over x=0 to K. The upper tail probability (UTP) is calculated by subtracting the LTP of K-1 events from 1. Using the full form of the GB function, where the array of probabilities P_ is explicitly shown as an argument, we have the following formula:

$$LTP(K) = \sum_{x=0}^{K} GB(N,x,P\_) \qquad (4)$$

UTP(K) = 1 - LTP(K-1)

Note that these quantities can be calculated from the probabilities for K or fewer events, so there is no need calculate the columns of Table 1 to the right of K. Also for efficiency, problems in which K is predominantly above N/2 can be redefined by negating the definition of the event.

SEARCHING FOR THE EXACT CONFIDENCE INTERVAL

In calculating the confidence interval, a search algorithm applies trial values for the shifting constant. Each trial value produces a shifted set of probabilities, which in turn produce a revised exact tail probability (upper or lower). The search terminates when the tail probability is arbitrarily close to the target value of α/2. Any search algorithm can be used, such as an interval-splitting binary search. We implemented a search algorithm which is a mixture of the secant method and the false position method, both described in Wikipedia. This algorithm sacrifices some efficiency in return for a high degree of robustness. This algorithm is implemented in the %SEARCH macro (See Access to Macros section).

Table 2: Calculation times

K / N	Time (sec)
5 / 1000	0.016
50 / 10,000	0.95
500 / 100,000	163.7

Execution times of exact 95% confidence intervals. Times are reasonable even with very large problems, as shown in Table 2 using an accuracy of .0000001. Computer used: Intel Core 2 Duo running at 2.1 GHz.

CODE OUTLINE

The follow is a pseudo-code version of the program, modified to be similar to the explanation in this article. Assume the input data set has two variables: Prob and Event. The following calculates the upper confidence limit only.

```
PROC TRANSPOSE DATA=in OUT=probs PREFIX=prob;
 VAR prob;
PROC SUMMARY DATA=in;
 VAR event;
 OUTPUT OUT=counts(rename=(_freq_=N)) SUM=K;
DATA _null_;
 SET sums;
 CALL SYMPUT('N',N);

DATA out;
 MERGE counts probs;
 ARRAY P_ (&N) prob1-prob&N;
 ARRAY S_ (&N) _temporary_;
 ARRAY GB_(0:&N,0:&N) _temporary_;
 GB_(0,0)=1; * Fill top left cell;

DO shift = SEARCHING UNTIL( Lower_Tail = .025 );
  DO i = 1 to N;
     * Get shifted probability S_;
     xbeta = log(P_(i)/(1-P_(i))) + SHIFT;
     s_(i) = exp(xbeta)/(1 + exp(xbeta));
     * Calculate row i of generalized binomial probs;
     DO j = 0 to i;
         GB_(i,j) = GB_(i-1,j-1)*S_(i) +
                    GB_(i-1,j)*(1-S_(i));
     END;
```

Prepares the data as a single observation with counts (K and N) and all individual probabilities as variables.

Used in the array statements below.

Searches for the value of SHIFT which makes Lower_Tail = .025 (+/- a very small error). This syntax does not currently exist in the data step language.

Calculates shifted probabilities using Equation 2.

Does the “Probability Cascade” using Equation 3. A single-term version applies at the edges (not shown).

```
  END;

  * Lower tail probability is cumulative sum 0 to K;
  Lower_Tail = 0;
  DO j = 0 to K;
     Lower_Tail = Lower_Tail + GB_(N,j);
  END;
  * Mid-P adjustment;
  if 0<K<N then
  Lower_Tail = Lower_Tail - .5 * GB_(N,K);
END;

* Confidence limit is sum of shifted probabilities;
Upper_Limit=0;
DO i = 1 to N;
  Upper_Limit = Upper_Limit + S_(i);
END;
DROP prob:;
run;
```

Sums the "bottom line" to get the tail probability.

Applies mid-P adjustment, explained in Section 3 below.

End of the DO … SEARCHING loop

After the search has found the correct shift to satisfy the UNTIL condition, sum the shifted individual probabilities to get the confidence limit.

The above is not working code because the "searching" feature on DO … UNTIL does not exist. (Hopefully it will.) In the actual program, processing is done within BY-groups, the roles of Proc Transpose and Proc Summary are incorporated into the data step, searching is done by a macro, S_ is not an array, the GB_ array is one-dimensional, and the lower confidence limit is also calculated. For efficiency, the values of GB_ are not calculated past column K.

SECTION 2 – THE EFFECT OF DIVERSITY OF RISK

The first question about the above method might be whether it matters that the method uses the individual probabilities of the event, or would the confidence interval from the simpler Clopper-Pearson confidence interval be basically the same. In addressing this question (and used again later in examining coverage probability) we used the following four risk distributions taken from real-world data:

1. Equal probabilities.
2. PPR: The probability of a potentially preventable readmission predicted by the rate within strata defined by APR-DRG and severity of illness, using 2008 California discharges.
3. IQI09: The probability of death in pancreatic resection patients, an AHRQ inpatient quality indicator, among 2008 California discharges.
4. MPM3: The probability of death in an ICU as predicted by the MPM3 model, using data from CHART [Reference 3] during 2009.

To modify the number of subjects while holding the distribution constant, The KDE Procedure estimated a probability density distribution for the X-Betas. Cumulative percents were calculated, and these values were interpolated at N equally-spaced cumulative percent points between the minimum and maximum.

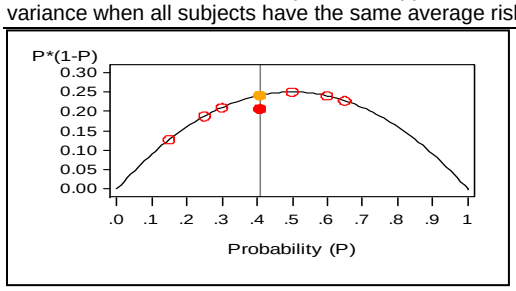
Table 3: Characteristics of risk distributions

	Probs	X-Betas = log(P/(1-P))			Exact 95% C.I.s for K=5 and N=100		
Distribution	Mean	Mean	Std Dev	Skew-ness	Lower Limit	Upper LImit	Width
EQUAL	n/a	n/a	0	n/a	1.64	11.28	9.64
PPR	0.081	-2.67	0.84	-0.95	1.65	11.13	9.47
IQI09	0.039	-3.58	0.84	1.01	1.66	11.04	9.37
MPM3	0.125	-2.53	1.34	0.52	2.04	10.35	8.32

The equal probabilities case provides a baseline of zero risk diversity. The PPR and IQI09 risk distributions have almost identical standard deviations, but IQI09 is skewed to the right and PPR is skewed to the left. MPM3 has the largest standard deviation. On the right, the pattern is clear that as the diversity of risk increases, the confidence intervals become narrower. The same pattern emerges in Figure 8 in Section 4.

A straightforward way to understand the above is to consider the following plot of the function P*(1-P), which is the formula for the variance of a single binomial trial (that is, subject) with probability P.

Figure 4: Variance reduction with risk diversity. Empty circles: individual subjects with diverse risk. Lower dot: mean variance when there is diversity of risk. Upper dot: mean variance when all subjects have the same average risk.



Note that the mean variance is reduced when there is diversity of risk. The conclusion is that methods which do not take diversity of risk into account will produce wider confidence intervals, and this effect increases with the amount of diversity of risk. Consider an extreme example: 10 patients, 4 of whom are almost certain to die, and 6 of whom will almost certainly live. There is very little doubt how many will die, so the confidence interval should be extremely narrow around 4, compared to a wider interval if all 10 patients had the same 40% risk of death.

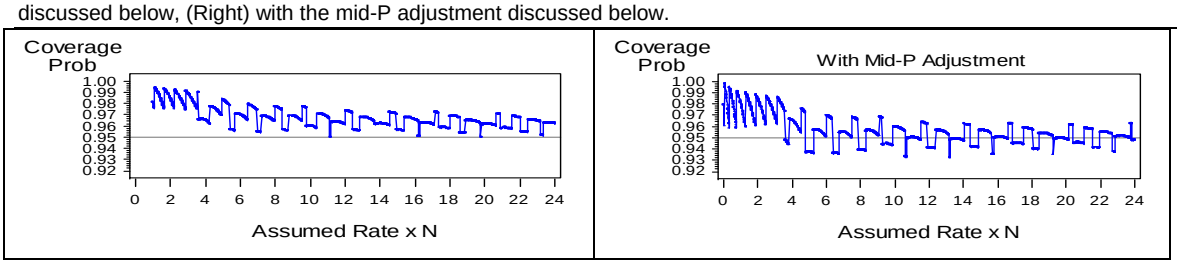
SECTION 3 – COVERAGE PROBABILITY AND THE MID-P ADJUSTMENT

Just as gas mileage is a way of evaluating automobiles, coverage probability is a way of evaluating a confidence intervals method. Coverage probability is defined as the probability that the confidence interval method will produce a confidence interval which actually contains the underlying parameter value. We want the coverage probability of a 95% confidence interval method to be about 95%.

While gas mileage depends on speed, coverage probability depends on the assumed value of the underlying rate parameter. A confidence interval method produces a confidence interval for each of the N+1 values of K from 0 to N. For any assumed value of the underlying rate, some confidence intervals contain this value and others do not. Also based on the assumed underlying rate, each value of K has a certain probability of occurring. Coverage probability is the sum of the probabilities of the values of K whose confidence intervals do in fact contain the underlying parameter value.

[Technical Point] about calculating coverage probability: To obtain the probability of observing K events with an assumed underlying rate, we used the generalized binomial probabilities, GB(N,K,P_) where the probabilities P_ were shifted to have a mean equal to the assumed underlying rate. In the statistical literature on coverage probability, [References 4,6,7], these studies use the binomial distribution for weighting coverage probability. In the current study we use the generalized binomial distribution, which uses the information available regarding diversity of risk and becomes more tightly distributed as risk diversity increases. This is reflected in the narrower confidence intervals seen in Section 2. This has the effect of increasing the coverage probability of any method which does not become narrower with increased risk diversity, such as the Poisson distribution discussed in Section 4.

Figure 5: Coverage probability for the exact method with N=100, for the PPR distribution, (Left) without the mid-P adjustment discussed below, (Right) with the mid-P adjustment discussed below.



On the left side Figure 5, there are two aspects to note: the shape and the location. Both are related to the fact that the outcomes are discrete integer counts. First, the function has a wild “saw-tooth” shape, which has been studied extensively [Reference 6]. Secondly, the coverage probability is too high, always above 0.95, indicating that the confidence intervals are too wide (“overly conservative”). This problem occurs with any confidence interval method based on integer outcomes, where there is a non-zero probability of observing exactly K events. The problem stems from the fact that the probability of K events is included in both the upper and lower tail areas (Refer to Figure 1 where K=6). Thus, the probability of observing K events is in-effect **double-counted**.

The standard solution, known as the "Mid-P adjustment", only includes 50% of the probability of observing exactly K events in the tail area. [Reference 5] With this adjustment, the tail probabilities defined in Equation (4) are redefined as follows:

$$LTP(K) = \sum_{x=0}^{K-1} GB(N,x,P\_) + .5 * GB(N,K,P\_) \qquad (5)$$

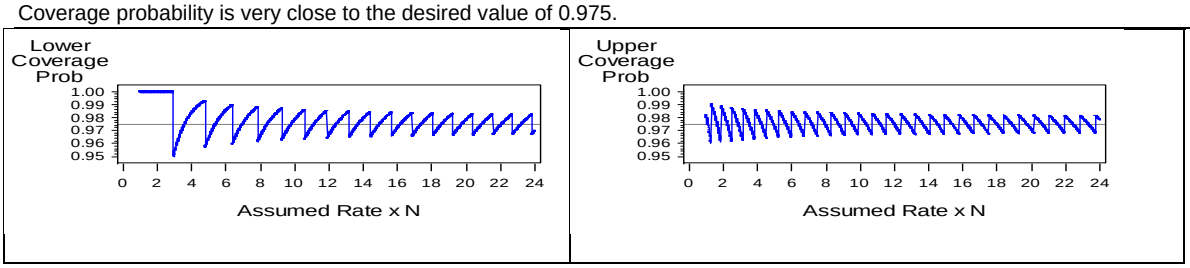
$$UTP(K) = 1 - LTP(K)$$

The result is shown on the right side of Figure 5. Note that the average coverage probability is close to the desired level of 0.95.

We apply the mid-P adjustment only when K > 0. The justification is that in the K=0 case, the probability of observing fewer than K events is zero, so no overlapping of tail areas is possible. Also, the inverted statistical test question becomes "What is the probability of observing zero events?" which seems more relevant than "What is half the probability of observing zero events?"

For deeper understanding, it is useful know whether the confidence interval is under- or over-estimating the location of the parameter, or our concern may be with one side of the confidence interval or the other for hypothesis testing. The lower coverage probability (LCP) is the probability that the true parameter is above the lower confidence limit, and conversely the upper coverage probability (UCP) is the probability that the true parameter value is below the upper confidence limit. For a 95% confidence interval the desired value of both LCP and UCP is 0.975. Figure 6 indicates that the upper and lower coverage probability functions both average around 0.975. A side-benefit is that the functions are amenable to statistical summarization, compared to the two-sided coverage probability functions in Figure 5 above.

Figure 6: Upper and lower coverage probability of exact method with Mid-P adjustment, (N=100, PPR distribution). Coverage probability is very close to the desired value of 0.975.



Because CP = LCP + UCP – 1, the two parts of Figure 6 sum to the right side of Figure 5 (subtracting 1). Interference patterns create the wild shape of the coverage probability functions.

SECTION 4 – COMPARISON OF METHODS

We evaluated how closely the exact confidence intervals are approximated by methods based on the Poisson distribution and the normal distribution.

FOUR METHODS DEFINED

Method 1 is the subject of this paper.
Method 2 "Shifted Normal" is a variant of Method 1 which uses the same shifting of probabilities and search algorithm, but the tail probabilities are estimated using the normal distribution.
Method 3 uses the Poisson distribution to estimate a confidence interval on the underlying mean number of events.
Method 4 "Fixed-Width Normal" uses the normal distribution to calculate a symmetrical confidence interval, using the variance calculated from the individual risk probabilities.

Method 2 is an unpublished method included to examine the effect of the assumed "normal approximation of the binomial". Method 3 is used by the NHSN branch within the CDC to evaluate rates of hospital-acquired infections. Method 4 is used by the AHRQ IQI/PSI software for evaluating hospital quality and safety.

To enable "apples-to-apples" comparisons, each method has a "wide" and "narrow" version. The narrow versions generally come closer to achieving the desired 95% coverage probability. The following table shows the definitions of the wide and narrow versions of each of the above methods. In further discussion we will append "W" or "N" to denote the wide or narrow version. All of these methods produce numbers between 0 and N, and θ represents the underlying rate between 0 and 1.

Table 4: Wide and narrow versions of methods being compared

Method	Wide Version	Narrow Version
1. Exact Using Generalized Binomial	1W. no Mid-P adjustment	1N. with Mid-P adjustment
2. Shifted Normal	2W. with continuity correction	2N no continuity correction
3. Poisson	3W. no Mid-P adjustment	3N with Mid-P adjustment
4. Fixed-Width Normal	4W. with continuity correction	4N no continuity correction

Note: The term "continuity correction" usually describes the adjustment of confidence intervals based on continuous distributions (e.g., normal) to account for discreteness. Some authors also use it to describe the mid-P adjustment, which is applied to a discrete distribution (e.g. binomial or Poisson).

Method 1W: The exact method based on the generalized binomial distribution, defined as follows:

CLupper(K,N) = θ*N | LTP(θ) = α/2
CLlower(K,N) = θ*N | UTP(θ) = α/2

Where LPT and UPT are defined by equations (4) above, or equivalently,

$$\text{LTP}(\theta) = \sum_{x=0}^{K} \text{GB}(N,x,S\_(\theta)), \text{ and } \text{UTP}(\theta) = 1 - \sum_{x=0}^{K-1} \text{GB}(N,x,S\_(\theta))$$

where S_(θ) is the set of probabilities P_ shifted in the logistic domain to have a mean of θ.

Method 1N: The exact method with mid-P adjustment. Like Method 1W except that GB(N,K,S_(θ)) / 2 is subtracted from both LTP(θ) and UTP(θ).

Method 2W: Shifted Normal with continuity correction. This method uses a simple formula in place of the exact generalized binomial probability algorithm in Method 1, and like the exact method this requires a search algorithm. The method is as follows:

CLupper(K,N) = θ*N | LTP(θ,K) = α/2
CLlower(K,N) = θ*N | UTP(θ,K) = α/2

Where LPT and UPT are defined by

LTP(θ,K) = ProbNorm((K+.5 - Exp(θ))/(sqrt(Var(θ)))
UTP(θ,K) = 1-ProbNorm((K -.5 - Exp(θ))/(sqrt(Var(θ)))

Where $\text{Exp}(\theta) = \sum_{i=0}^{N} S\_(\theta)_i$, $\text{Var}(\theta) = \sum_{i=0}^{N} S\_(\theta)_i(1-S\_(\theta)_i)$,

and S_(θ) are the probabilities P_ shifted in the logistic domain to have a mean of θ.

Method 2N: Shifted Normal without continuity correction. Like the above but without the use of the .5 continuity correction.

Method 3W: Poisson distribution.
Lower CL for the number of events is lambda such that the probability of observing K events or more with a Poisson distribution with mean lambda is α/2. Conversely for the upper CI. Because of the relationship between the Poisson and Inverse Gamma distributions, Poisson(Gaminv(Prob,K+1),K) = Prob. Therefore the confidence interval can be calculated from the formulas:

CLupper = Gaminv(1-α/2,K+1)
CLlower = Gaminv(α/2,K)

Method 3N: Poisson distribution with mid-P adjustment. Using a search algorithm to find

CLupper(K) = lamda | LTP(lamda,K) = α/2
CLlower(K) = lamda | UTP(lamda,K) = α/2

Where LTP(lamda,K) = Poisson(lamda,K-1) + .5 * (Poisson(lamda,K) – Poisson(lamda,K-1)),
and UTP(lamda,K) = 1 – LTP(lamda).
The Poisson(lamda,K) function returns the probability of observing K or fewer events. "Lamda" is the conventional name for the mean parameter for the Poisson.

Method 4W: Fixed-Width Normal with continuity correction.

CLupper = (K +.5 + 1.96*sqrt(Variance))
CLlower = (K - .5 - 1.96*sqrt(Variance)),

where Variance = $\sum_{i=0}^{N} P_i(1-P_i)$

Method 4N: Fixed-Width Normal. Like the above, without the .5 continuity correction.

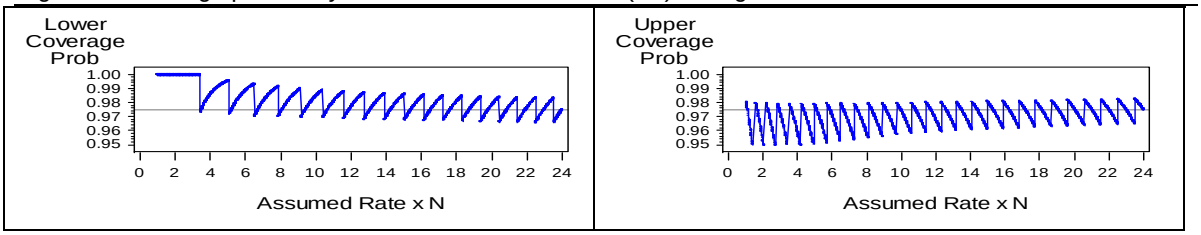
The methods described above differ in their dependency on K, N and the distributions of probabilities:

- The Exact method and the Shifted Normal method are dependent upon K, N and the shape of the distribution of the logit-transformed probabilities, but not location.
- The Poisson method depends only upon K.
- The Fixed-Width Normal method is dependent upon K, N and both the location and shape of the distribution of the probabilities.

EVALUATION OF METHOD 2: SHIFTED NORMAL

The shifted normal method has good overall mean coverage probability near .95 (not shown), but the problems are evident when you look at each end separately.

Figure 7: Coverage probability of Shifted Normal Method (2N). Using PPR distribution with N=50.

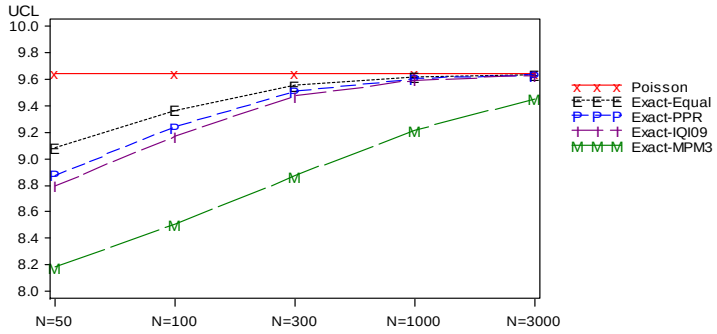


These results indicate that the shifted normal has good coverage probability when the number of events based on the assumed rate is higher than about 20. Below that, the interval is increasingly below the exact confidence interval, being too wide at the low end, and too narrow at the upper end.

EVALUATION OF METHOD 3: POISSON

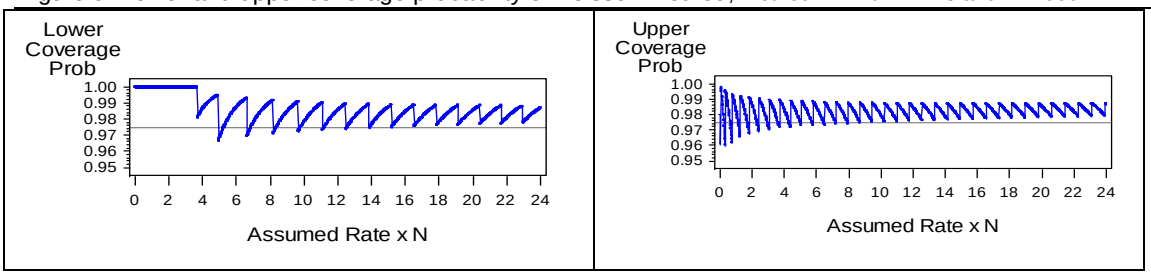
The Poisson approximation assumes that N is infinitely high and can be ignored. It is expected to work well when the number of events is low in comparison to the number of subjects. Therefore we chose a low number of events and increased N.

Figure 8: Upper 95% confidence limits, by Exact and Poisson methods (Methods 1N and 3N), for K=4, by N and by risk distribution.



Because the Poisson distribution only depends on K, the confidence interval based on it never varies as N increases or the risk distribution changes. As expected, with the MPM3 (M) distribution of probabilities, which has the largest risk diversity, there is the greatest difference between the Exact and Poisson methods. The exact confidence intervals with the PPR (P) and IQI09 (I) distributions differ slightly, possibly due to the difference in the skewness of these two distributions. If one wants to use the Poisson method and be within 1% of the exact confidence limit, Figure 8 shows that in the K=4 case and the MPM3 distribution, the sample size must be more than 3,000. When N=1000, which most researchers would consider high relative to K=4, one might expect the large-sample assumptions of the Poisson method to hold true, but with a high degree of risk diversity this expectation is not met, as seen in Figure 9.

Figure 9: Lower and upper coverage probability of Poisson method, Method 4N with MPM3 and N=1000.



As expected, when the risk diversity is high as with the MPM3 risk distribution, the Poisson method yields wider confidence intervals with upper and lower coverage probability above the target value of 0.975.

EVALUATION OF METHOD 4: FIXED-WIDTH NORMAL

This method has a "bad news / good news" story to it. First the bad news. With this method, a symmetrical confidence interval is calculated based on the expected probabilities, and it is rigidly shifted up or down to center on the observed number of events. It does not fit the standard theoretical definition of a confidence interval mentioned in Section 1. One problem is that the confidence intervals are symmetrical, which does not conform to the asymmetrical distribution of probabilities confined to the 0 to 1 interval. The most obvious problem is that it can sometimes produce impossible confidence limits outside the 0 to 1 interval. Truncation is required to provide cosmetic improvement, using statements like the following:

```
IF LIMIT < 0 THEN LIMIT = 0;
```

Now for the good news. We return to the topic of hospital classification, in which the goal is to classify hospitals as "average", "below average" or "above average. If the only consideration is the use of the confidence interval in classifying a hospital, all that matters is whether the confidence interval for the O/E ratio contains 1.0, or equivalently, whether the confidence interval around the observed contains the expected number of events. As it turns out, the one and only situation where this confidence interval method gives a valid confidence limit is when the true underlying rate is equal to the expected rate. In that situation, the shifted probabilities used in Methods 1 and 2 would be have a shift of zero, and would be exactly where they started, equal to the unshifted probabilities used in Method 4. So for the purposes of classification, this method is reasonable (exactly the same as Method 2, which has minor deficiencies). When a hospital is classified as significantly below average, the lower confidence limit might get truncated, but the upper limit, which determines the classification, is approximately correct.

Looking more closely at the "bad news", there are damages at the other end of the confidence intervals, away from the expected rate. Using Method 4N, we examined the specific case of K=7, with N=300 and the MPM3 risk distribution, in both the K<E and E<K situation. In the K<E case, by searching we found that when the expected number of events is precisely 13.29, the Method 4N normal confidence interval is [0.71 – 13.29], putting the hospital just on the edge between "average" and "below average. To assess the inaccuracy of the lower limit of 0.71, we adjusted the probability level of the exact confidence interval (Method 1N) to give the same lower limit. The probability level had to be increased to 99.9999%. In the E<K case, when the expected number of events is 3.68 the Method 4N confidence interval is 3.68 – 10.32, which is on the edge between "average" and "above average". The probability level of the exact confidence interval had to be lowered to 75% to give the same upper limit. Without extensive testing, it appears that this method is intended only for the purpose of comparing observed to expected rates.

DISCUSSION

The lack of good statistical methods for confidence intervals for risk-adjusted rates has resulted in difficult trade-offs and compromises, which may be having a widespread impact on the world of hospital performance evaluation. Selecting the normal distribution method using the expected probabilities (Method 4N) achieves the need for good classification of hospital performance, but at the expense of any other uses of the confidence intervals, such as to compare hospitals to a benchmark performance goal rather than to an expected rate. The Poisson method avoids the problems of symmetry and truncation, but makes no use of the information contained in the individual probabilities, producing wider confidence intervals which may misclassify the performance of some outlier hospitals as inliers, especially when there is high diversity of risk. In addition to the situations already discussed, the author has found that invalid confidence limits greater than 100% are sometimes produced (then truncated, of course) by the publicly available macro which summarizes the CAHPS hospital surveys. These are only the situations of which this author is aware. Clearly there is Trouble in River City.

The first step in the right direction would be to explore the SAS® macros included with this paper, which are easy to use and written to behave like a SAS® procedure. See "Access to Macros" section. These macros are currently in use by the California Office of Statewide Health Planning and Development (OSHPD).

Looking down the road, in the real world of the management of SAS® programs, convenience and institutional support from SAS Institute count for a lot. A huge advance in spreading the use of these methods would happen if they were made available in a SAS® procedure. SAS® already has a very close match in Proc Freq, which produces Clopper-Pearson confidence intervals, and has a WEIGHT statement. A single option could be added to cause the weights to be interpreted as probabilities, implementing the "generalized binomial" distribution. More widely applicable, the data step language needs a searching feature, such as a "SEARCHING" keyword on the DO statement with UNTIL.

CONCLUSION

There is clearly a need for improved statistical methods for calculating confidence intervals for risk-adjusted rates. Commonly used methods produce problematic confidence intervals. Users of statistical methodology no longer have to choose which type of problem to accept, or worry about whether sample sizes are adequate or distributional assumptions are valid. The exact calculation based on the generalized binomial distribution avoids these potential pitfalls. The evaluation of hospital performance using confidence intervals for risk-adjusted rates would be significantly improved through the use of the exact confidence interval method.

ACCESS TO MACROS

The macros are packaged in EXACTCI_Macros.zip, downloadable from the SAS Community web site. Go to http://www.sascommunity.org/wiki/Special:ListFiles. Search for media name “exactci”, click on EXACTCI_Macros.zip in the list which appears, then click on EXACTCI_Macros.zip at the top of the page to download. The documentation section of each macro contains instructions for use. The search capability on the main page of sascommunity.org may also work. If these sources do not work, feel free to contact the author for the macros.

REFERENCES

(1) CJ Clopper and ES Pearson, "The use of confidence or fiducial limits illustrated in the case of the binomial." Biometrika 26:404-413, 1934.

(2) Harold S Luft, Byron Wm. Brown, "Calculating the Probability of Rare Events: Why Settle for an Approximation?" Health Services Research, 1993 Oct; 28(4): 419-39. Pub Med ID: 8407336
Explains the algorithm for calculating the exact probability of the generalized binomial distribution.

(3) CHART: California Hospital Assessment and Reporting Task Force, a voluntary public reporting of acute hospital process and outcome measures, reported at www.calhospitalcompare.org. The supporting team is at the Phillip Lee Institute for Health Policy Studies at UC San Francisco.

(4) Robert G. Newcombe, "Two-Sided Confidence Intervals for the Single Proportion: Comparison of Seven Methods" Statistics in Medicine 17, 857-872 (1998)
Discusses coverage probability and the benefit of mid-P adjustment in the Clopper-Pearson confidence interval.

(5) G. Berry, P. Armitage, “Mid-P confidence intervals: a brief review” The Statistician, 44, No 4, pp. 417-423, (1995)
Discusses mid-P tests and confidence intervals, including handling of the K=0 case.

(6) Stein Emil Vollset, “Confidence Intervals for a Binomial Proportion”, Statistics in Medicine, Vol 12, 809-824 (1993)

(7) Lawrence D Brown, T. Tony Cai, Anirban DasGupta, “Interval Estimation for a Binomial Proportion”, Statistical Science, Vol. 16, No. 2, 101-133 (2001)

AUTHOR INFORMATION

Ted Clay
Clay Software & Statistics
168 Meade St.
Ashland, OR 97520
Email: tclay@ashlandhome.net

The author is a statistical consultant for the CHART Project (Reference 3).