1 Introduction

In the preceding discussion, we assumed the credit instruments subject to default in a particular rating class were independent of each other. Of course this is almost always an unrealistic simplifying assumption. Historically, defaults do not occur independently of each other in time but rather appear to be correlated, with periods of high default rates and low default rates occurring more than would be expected if they were independent. One reasonable way to model this apparent dependence is to assume there exists a common "market" factor that affects the default likelihood of each creditor in a class in the same way. For example, an unfavorable market factor would make it more likely for each creditor to default, while a favorable market factor would make each creditor less likely to default. In this way the model would produce the clustering of defaults and non-defaults in time that we observe in market data. Thus default probability among creditors is correlated through the market factor. In this section we first introduce copulas and factor models and then use these as a way to model the dependence among creditors when considering defaults. We also introduce the Gaussian one-factor copula as a specific example. Finally we show how to form an exact two-sided confidence interval for the number of defaults in a particular class using this model and consider analogues of some of the various binomial confidence intervals covered previously. These intervals are compared by coverage probability and expected interval length.

2 Copulas and Factor Models

We continue as in the last section to consider a portfolio of debt instruments subject to default and so with a putative default probability. For now we will consider a portfolio of similarly rated instruments and so assume each one has the same default probability. Previously we had assumed that defaults occured independently of one another in time. As mentioned in the introduction, however, this is a simplifying assumption for computational purposes but is unrealistic in practice. The advantage this gave us was the reduction of the joint default probability distribution to the product of the marginal distributions, which were all the same. From there we could use a binomial model to calculate probabilities. Clearly though, we must add some dependence structure to the model. If we then have an estimate for the dependence (e.g., a correlation coefficient), as well as the marginal distributions of default probability, how do we relate this to the joint default probability distribution?

The problem of relating a multivariate distribution function to its lower dimensional marginals has been written about since at least Fréchet(1951). A seminal paper with regard to this problem was written by Sklar(1959), in which he introduced a class of functions he called *copulas*. A simple definition of a copula C is as follows, (see McNeil et al.(2005)):

Definition. *An n-dimensional copula C is a distribution function on the n-cube with standard uniform marginal distributions.*

Thus a copula C is a mapping of the form $C : [0,1]^n \rightarrow [0,1]$.

Copulas were first used,implicitly, to model portfolios of debt instruments by Vasicek(1987) and Gupton et al. (1997). Li(2000) was the first to use them explicitly to model defaults. They have since found extensive use in the credit modelling literature due to the necessity of capturing dependence in a portfolio and several properties we will now briefly review.

An extremely useful result with respect to modeling with copulas is Sklar's Theorem:

Theorem (Sklar 1959). *Let F be a joint distribution function with margins $F_1, \ldots, F_n$. Then there exists a copula $C : [0,1]^n \rightarrow [0,1]$ such that, for all $x_1, \ldots, x_n \in \mathbb{R}$,*

$$F(x_1, \ldots, x_n) = C(F_1(x_1), \ldots, F_n(x_n)). \tag{1}$$

If the margins are continous, then C is unique; otherwise C is uniquely determined on $\mathrm{Ran}\ F_1 \times \cdots \times \mathrm{Ran}\ F_n$*, where* $\mathrm{Ran}\ F_i$ *denotes the range of F_i. Conversely, if C is a copula and $F_1, \ldots, F_n$ are univariate distribution functions then the function F defined in* (1) *is a joint distribution function with margins $F_1, \ldots, F_n$.*

Thus for any n-dimensional distribution function F, with marginals $F_1, ..., F_n$, we have:

$$\begin{aligned} F(x_1, ..., x_n) &= P(X_1 \leq x_1, ..., X_n \leq x_n) \\ &= P(F_1(X_1) \leq F_1(x_1), ..., F_n(X_n) \leq F_n(x_n)) \\ &= P(U_1 \leq u_1, ..., U_n \leq u_n) \end{aligned}$$

where the $U_i, i = 1...n$ are standard uniform random variables. But this, as defined, is a copula function, so we can write:

$$F(x_1, ..., x_n) = C(F_1(x_1), ..., F_n(x_n))$$

Sklar also showed the converse is true, so that:

$$C(u_1, ..., u_n) = F(F_1^{-1}(u_1), ..., F_n^{-1}(u_n))$$

The advantage copulas provide in multivariate modeling is that we can specify the marginal distributions and dependence structure separately. The copula then *couples* the marginals to the joint distribution function. A specific example that will be used in what follows is the *Gaussian* copula. Given n random variables, if we assume standard normal distributions(Φ) for the marginals and define a correlation matrix, Σ, that specifies the dependence among the random variables, we can write the joint distribution function as the Gaussian copula:

$$\Phi_n(\Phi^{-1}(u_1), ..., \Phi^{-1}(u_n))$$

where Φ_n is the n-dimensional normal cumulative distribution function with correlation matrix Σ.

As mentioned previously, defaults tend to occur in what appears to be a correlated manner. One way to model this dependence is through *factor*, or *latent-variable*, models. Vasicek(1987) appears to be the first to use a Gaussian factor model to describe porfolios of debt instruments. We will use a simple one-factor model in which a "market" factor, M, affects each creditor in a portfolio in the same way. The market factor can be construed as the market, or economic, environment that is common to all market participants included in our portfolio. For example, a recession would be a negative economic environment, causing more defaults than usual, whereas an economic boom or so-called bull market, would be a positive economic environment resulting in relatively less defaults. Since each creditor is affected by this market factor, they will be correlated through it. Thus the factor model allows us a way to model the dependence inherent in the market. In addition, a major advantage of the one-factor model is that it allows us to condition on the market factor and then use conditional independence. This greatly enhances calculation speed, which, in the marketplace where time is of the essence, is vital. Besides the market factor M, each creditor j has a unique idiosyncratic factor, Z_j, that is independent of the market factor as well as the other creditors' idiosyncratic factors. So, following Hull and White (2006), for each credit j we define the random variable:

$$x_j = a_j M + \sqrt{1 - a_j^2} Z_j \tag{2}$$

where M and the Z_j's are independent random variables with mean 0 and variance 1. It follows that each x_j has mean 0 and variance 1 and that x_i and x_j have correlation $a_i a_j$.

Now we would like to relate the time until default for credit j, t_j, to the random variable x_j. To do this, let F_j be the cumulative distribution of x_j and Q_j be the cumulative distribution of t_j. Then we link x_j and t_j in the following way:

$$x_j = F_j^{-1}\left[Q_j(t_j)\right] \Longleftrightarrow t_j = Q_j^{-1}\left[F_j(x_j)\right]$$

Thus by introducing the x_j random variables, we keep the essential information from the time to default marginals while adding a correlation structure among them.

In addition to introducing a correlation structure, using a one-factor model simplifies computation by allowing for conditional independence of default times. So now let us consider the distribution of x_j conditional on the factor M:

$$\begin{aligned} P(x_j \leq x|M) &= P(a_j M + \sqrt{1-a_j^2} Z_j \leq x|M) \\ &= P\left(Z_j \leq \frac{x - a_j M}{\sqrt{1-a_j^2}}\right) \\ &= H_j\left[\frac{x - a_j M}{\sqrt{1-a_j^2}}\right] \end{aligned}$$

where H_j is the cumulative distribution of Z_j. Then it follows from above that:

$$Q_j(t|M) = P(t_j \leq t|M) = H_j \left[\frac{F_j^{-1}\left[Q_j(t)\right] - a_j M}{\sqrt{1-a_j^2}} \right]$$

Since these conditional distributions are independent, their joint distribution simplifies to their product. In order to calculate any unconditional joint probabilities we then integrate this product over the density of the factor M.

The Gaussian One Factor Model - the standard

According to Hull and White(2006), the standard model used in the market assumes M and the Z_j's are standard normal random variables. In this case the model is referred to as the Gaussian One-Factor Copula (GOFC). The standard market model further assumes that all the a_j's are equal. Also, the default rate is assumed to be the same for all j. The recovery rate, defined as the percentage of principal recovered in the event of default, is assumed to be a constant 40%.

3 Confidence Intervals for GOFC model

We now turn to the problem of forming confidence intervals when using the GOFC model defined above. Again we assume we have a portfolio of n credit instruments, all of the same rating class, and all having the same given probability of default over a one year time frame. Call this probability p_{def}. We will model the number of defaults in this portfolio over a one year time frame using the GOFC. Thus, as in equation (2) above, for each credit j, we associate the random variable x_j:

$$x_j = aM + \sqrt{1-a^2} Z_j$$

where M and Z_j are independent standard normal random variables. It follows that x_j is a standard normal random variable and that x_i and x_j have correlation a^2. Also note that we have conditional independence given M.

Similarly to above, we will link p_{def} to x in the following way:

$$x = \Phi^{-1}(p_{def})$$

Where Φ is the cumulative distribution for a standard normal random variable. Thus the probability of default is linked to a threshold z-value: if x_j falls below this value, default occurs. These are the marginal distributions for the n credits. Next, we use the gaussian copula function to link these marginals to a joint distribution:

$$\Phi_\Sigma(\Phi_{(1)}^{-1}(p_{def}), ..., \Phi_{(n)}^{-1}(p_{def}))$$

where Φ_n is the n-dimensional normal cumulative distribution function with correlation matrix Σ, where Σ is the matrix with diagonal entries equal to 1 and all other entries equal to a^2. We do not need to evaluate this joint distribution however. Recall that in the one-factor model, the marginals are independent conditional on M. Thus we can integrate the product of the marginals over M to obtain any required probability. In our model, the probability that any individual creditor j defaults is the probability that x_j falls below $\Phi^{-1}(p_{def})$. Let $\Phi^{-1}(p_{def}) = \theta$ Then:

$$\begin{aligned} P_\theta^j(default) &= P(x_j \leq \theta) \\ &= P(aM + \sqrt{1-a^2}Z_j \leq \theta) \\ &= P\left(Z_j \leq \frac{\theta - aM}{\sqrt{1-a^2}}\right) \\ &= \Phi\left[\frac{\theta - aM}{\sqrt{1-a^2}}\right] \end{aligned}$$

Note this last probability is fixed once we are given M; in particular, it is the same for each j. Hence we will drop the superscript in what follows.

Now, given n creditors, there are $\binom{n}{d}$ different ways of getting d defaults. Taking all of this into account and letting D be the random variable for number of defaults after one year, we can write the probability that there are d defaults after one year as:

$$P_\theta(D = d) = \binom{n}{d} \int \left(\Phi\left[\frac{\theta - aM}{\sqrt{1-a^2}}\right]\right)^d \left(1 - \Phi\left[\frac{\theta - aM}{\sqrt{1-a^2}}\right]\right)^{n-d} \phi(M)dM \tag{3}$$

where ϕ is the standard normal density function.

Next, we first attempt to form an exact 95% confidence interval before using the intervals discussed in the last section to get approximate intervals.

3.1 Exact Interval

Suppose again we have n creditors and we observe d defaults. We can form an exact 95% confidence interval for p_{def} analogously to the the exact Clopper-Pearson interval for binomial proportions. Define L as:

$$L \equiv \{\theta | P_\theta(D \geq d) = \sum_{x=d}^{n} P_\theta(D = x) = .025\} \tag{4}$$

and U as:

$$U \equiv \{\theta | P_\theta(D \leq d) = \sum_{x=0}^{d} P_\theta(D = x) = .025\} \tag{5}$$

Then transforming these threshold values back to probabilities via the normal density, we get the exact 95% confidence interval is $[\Phi(L), \Phi(U)]$. Of course exact here is in the sense of Clopper-Pearson and is in fact highly conservative, with coverage probabilities approaching 1 for many probabilities. To find L and U, we can use a bisection-type alogorithm as follows.

To get a starting value, note that in equation (3) above, we have the binomial probability:

$$\binom{n}{d} \left(\Phi \left[\frac{\theta - aM}{\sqrt{1-a^2}} \right] \right)^d \left(1 - \Phi \left[\frac{\theta - aM}{\sqrt{1-a^2}} \right] \right)^{n-d} \tag{6}$$

which is maximized at:

$$\Phi \left[\frac{\theta - aM}{\sqrt{1-a^2}} \right] = \frac{d}{n}$$

and thus, given any value of M, (6) is maximized at:

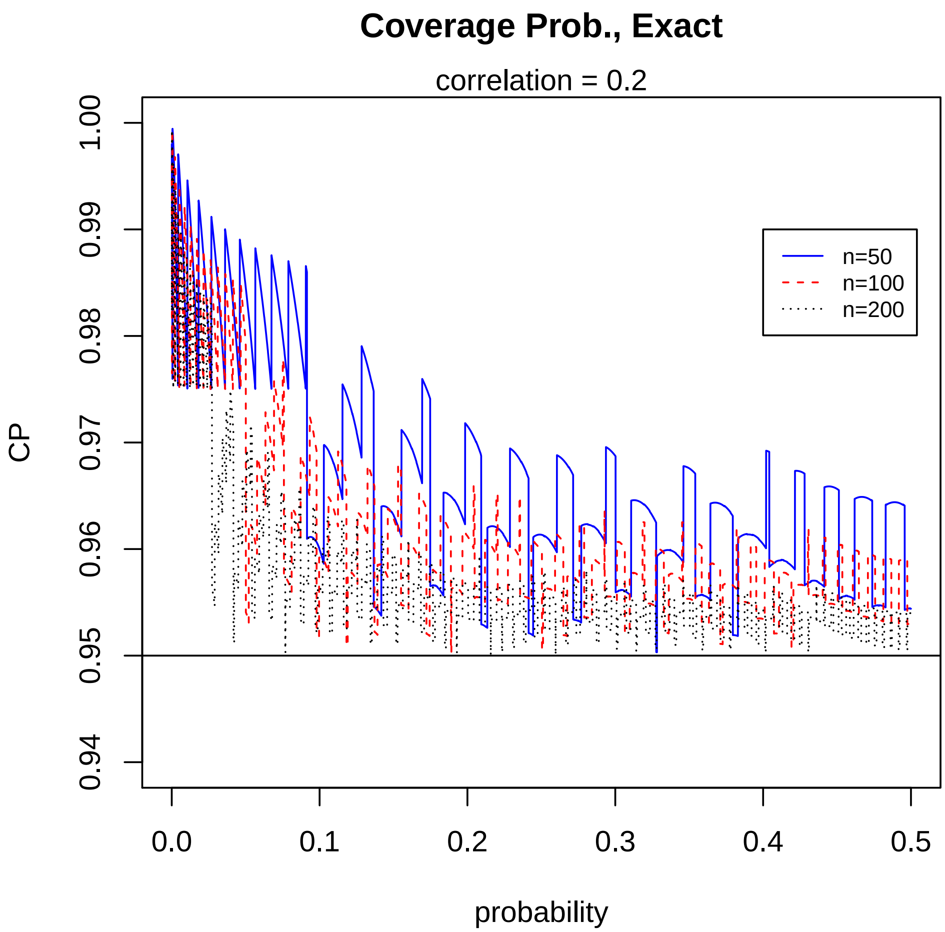
$$\hat{\theta} = \left(\sqrt{1-a^2} \right) \Phi^{-1} \left(\frac{d}{n} \right) + aM \tag{7}$$

Since the standard normal density, $\phi(x)$ reaches a maximum value of $\phi(0)$, we can set $M = 0$ above and use $\theta = \left(\sqrt{1-a^2}\right) \Phi^{-1} \left(\frac{d}{n}\right)$ as an initial value for the following algorithm to find L. Let ϵ be a small tolerance.

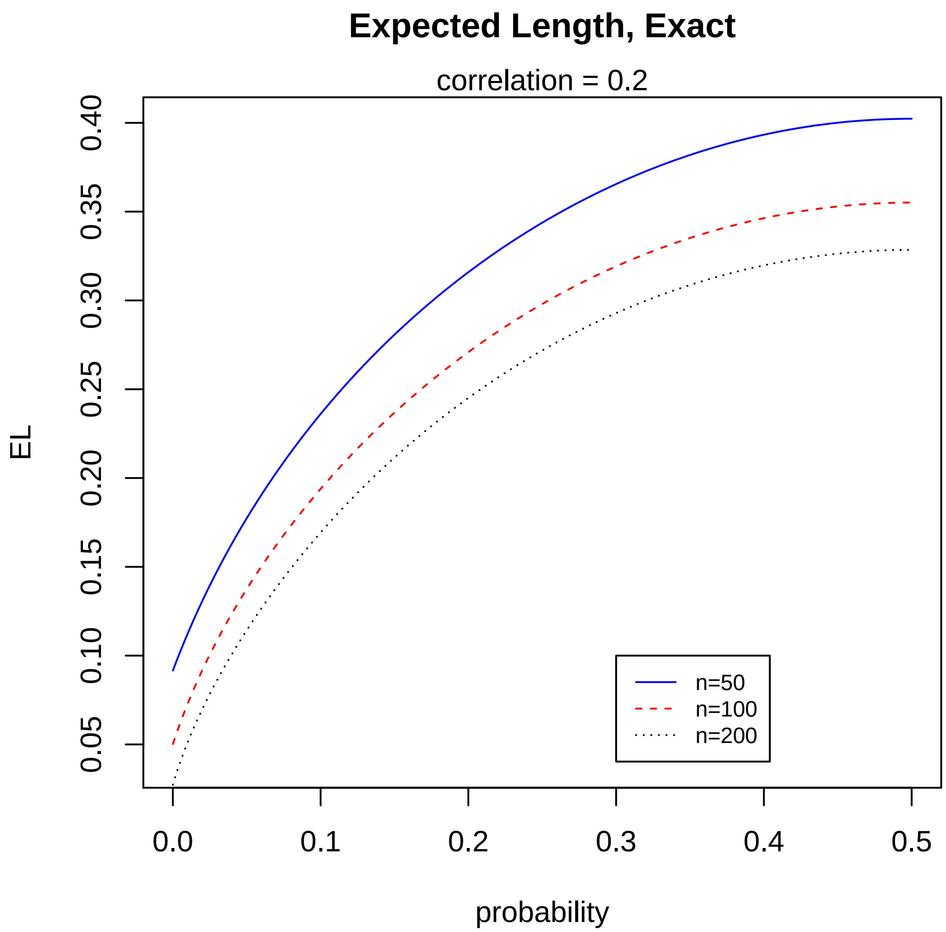
1. Set $\theta_l = \left(\sqrt{1-a^2}\right)\Phi^{-1}\left(\frac{d}{n}\right) + 3$ and $\theta_u = \left(\sqrt{1-a^2}\right)\Phi^{-1}\left(\frac{d}{n}\right) - 3$
2. Set $\theta_t = \frac{\theta_l+\theta_u}{2}$.
3. While $|P_{\theta_t}(D \geq d) - .025| > \epsilon$
   a) If $P_{\theta_t}(D \geq d) < .025$ then $\theta_u = \theta_t$. Else $\theta_l = \theta_t$.
   b) $\theta_t = \frac{\theta_l+\theta_u}{2}$
4. Return θ_t

.

An analogous algorithm can be used to find U. In practice, an initial value of 0 works well also.
Below, we see the examine the coverage probability and expected length for this exact interval for n values of 50,100, and 200, with a correlation value of $a = 0.2$. Recall that this is the correlation typically used in practice according the Hull and White(2006).

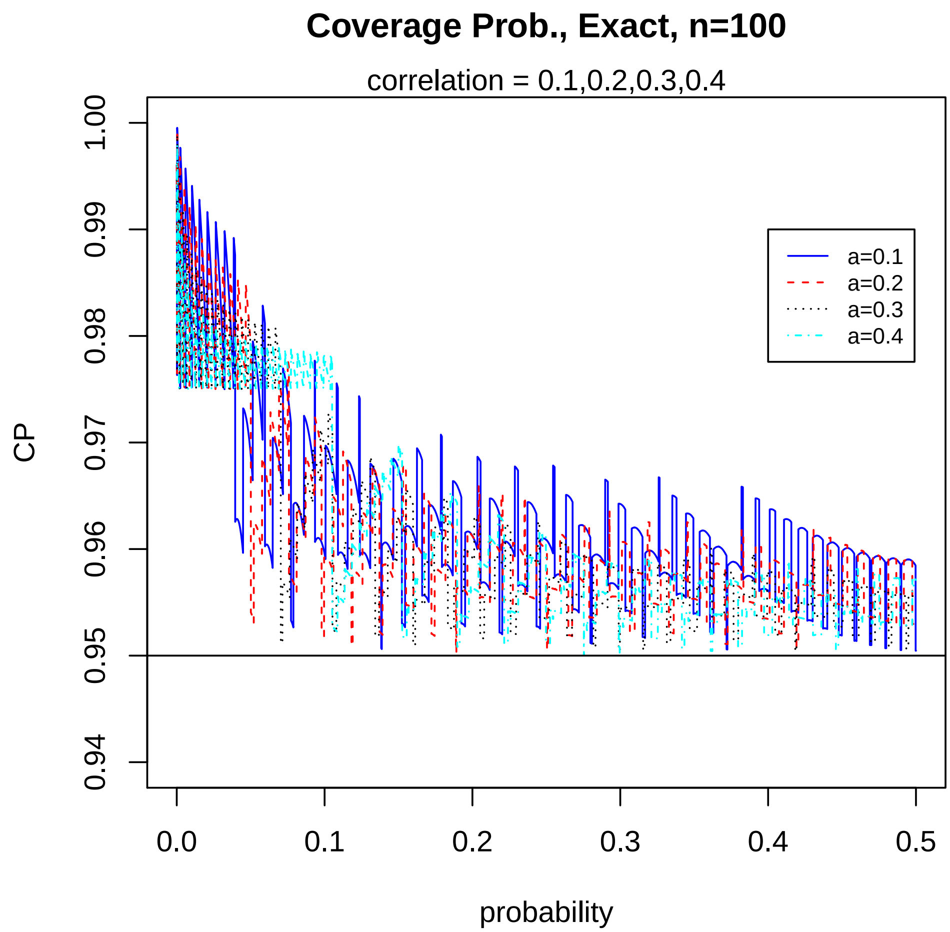


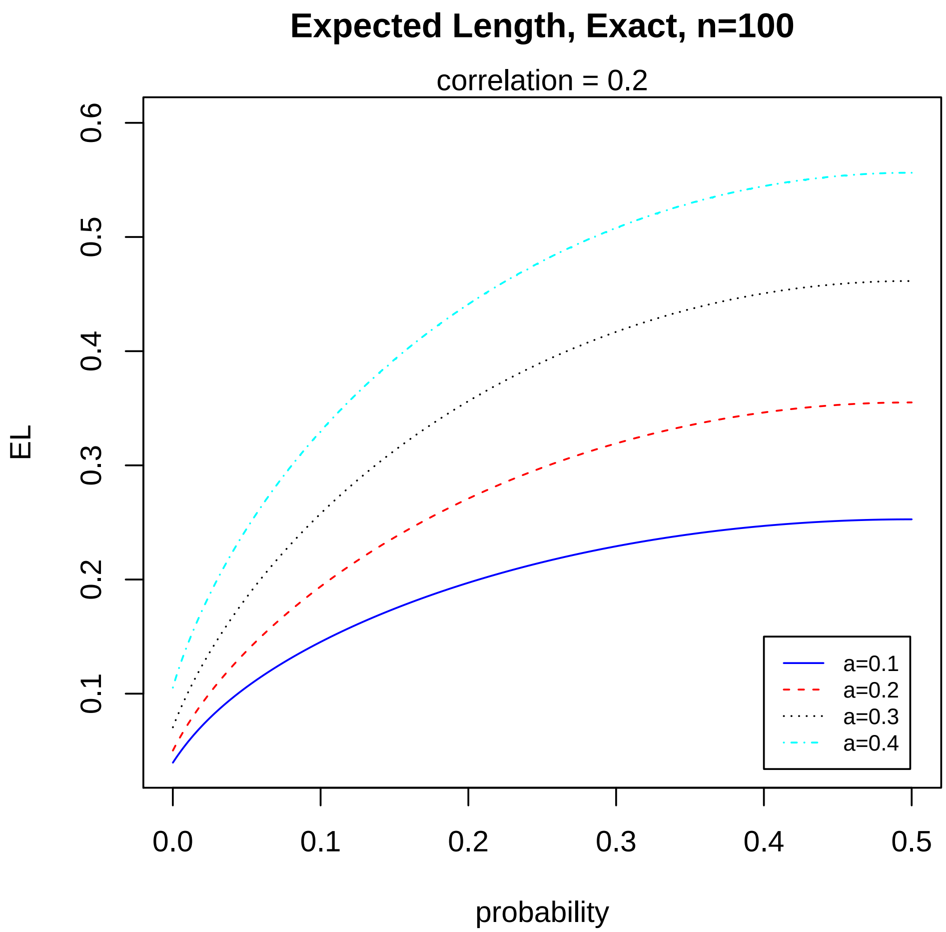
Coverage Prob., Exact
correlation = 0.2
CP
1.00
0.99
0.98
0.97
0.96
0.95
0.94
0.0
0.1
0.2
0.3
0.4
0.5
probability
n=50
n=100
n=200



The coverage probability of this interval is conservative for very small values of p, but becomes less so as p increases. As expected, it is at or above the nominal level for all values of p considered here. Note, however, the long expected length, especially for $n = 50$, where it becomes about 0.4. This does improve some as n increases, but still remains rather long.

Next, we look at the coverage probability and expected length for different values of the correlation coefficient a and $n = 100$. We choose $a = 0.1, 0.2, 0.3$ and 0.4. Higher values show a more pronounced effect of the relationship displayed. Similarly, increasing or decreasing n merely changes the degree of the effects shown for the values chosen.


Coverage Prob., Exact, n=100
correlation = 0.1,0.2,0.3,0.4
CP
probability
a=0.1
a=0.2
a=0.3
a=0.4
1.00
0.99
0.98
0.97
0.96
0.95
0.94
0.0
0.1
0.2
0.3
0.4
0.5



As a increases, we see that the coverage probability stays conservative and above the nominal level. The expected lengths become larger though for larger values of a, even going above 0.5 when $a = 0.4$ Thus maintaining this level of coverage comes at the expense of longer intervals.

3.2 Some Approximate Intervals

For approximate confidence intervals of default probability in the gaussian one factor model, we use independence conditional on M and the resulting binomial probabilities shown above in conjunction with the alternative binomial intervals considered in the previous section. Recall that $\hat{p} = d/n$, the observed default probability, is the maximum likelihood estimate of the default probability in the independent case and consider again equation (7). Then we have:

$$\hat{\theta} = \left(\sqrt{1-a^2}\right)\Phi^{-1}\left(\hat{p}\right) + aM$$

where we have stipulated the relationship $\theta = \Phi^{-1}(p_{def})$. So,

$$\begin{aligned} \hat{p}_{def} &= \Phi(\hat{\theta}) && (8) \\ &= \Phi\left[\left(\sqrt{1-a^2}\right)\Phi^{-1}\left(\hat{p}\right) + aM\right] && (9) \end{aligned}$$

Thus we have a maximum-likelihood type estimate after inverting. It is clear that for each value of M we have a different estimate of the probability of default and hence different confidence intervals. In what follows, we use the methods of interval formation discussed in the previous section along with the inversion above together with integration over M to form our approximate intervals. We first consider the Clopper-Pearson interval.

In the following, let z be the normal quantile appropriate for the confidence level selected, e.g. $z \approx 1.96$ for a 95% confidence level. Also, define equation (9) as:

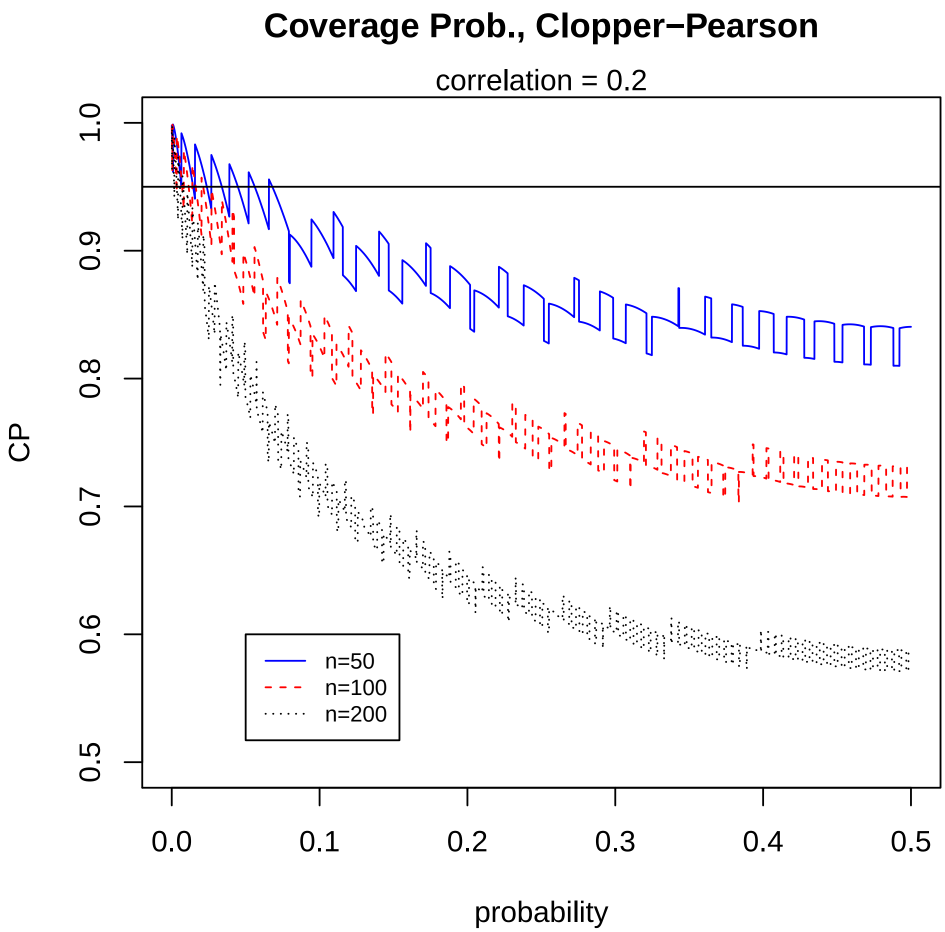
$$\mathcal{F}(x,m) \equiv \Phi\left[\left(\sqrt{1-a^2}\right)\Phi^{-1}\left(x\right) + am\right] \qquad (10)$$

3.2.1 Clopper-Pearson type Interval

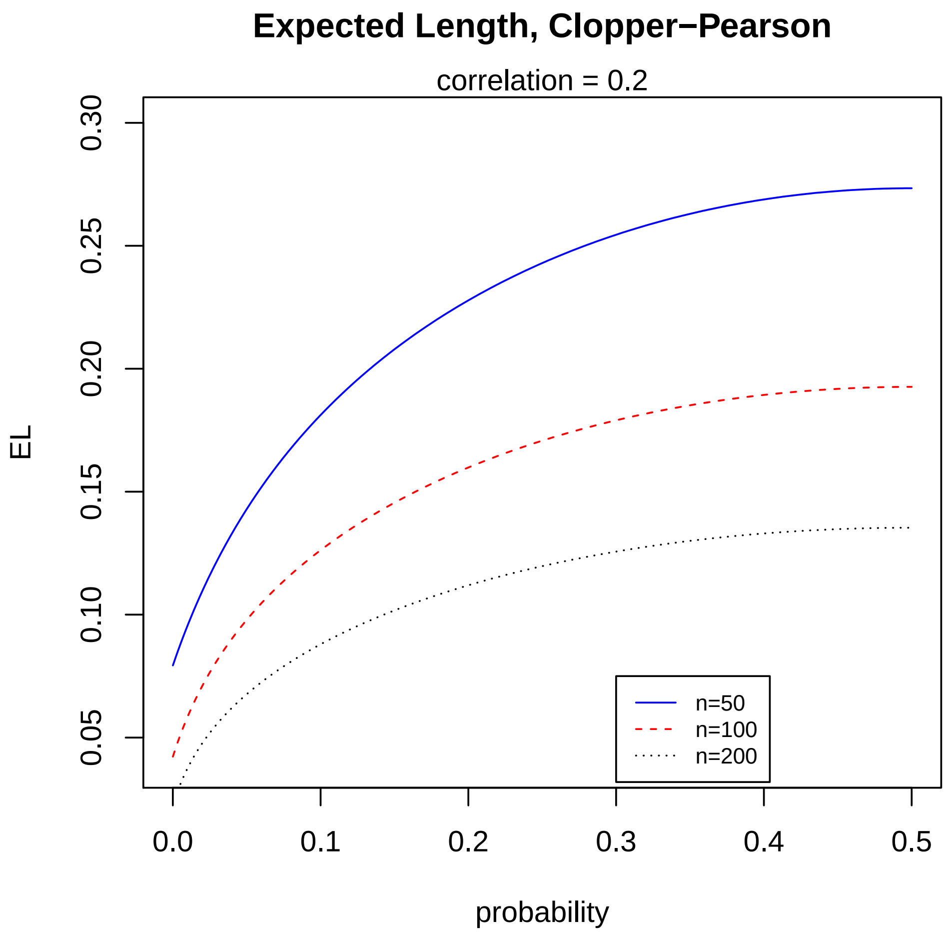
To form the lower bound of the Clopper-Pearson type intervals, we take the observed number of defaults, d, and form the normal Clopper-Pearson lower bound, L_{CP}, described previously. We then take the exptected value of the function $\mathcal{F}(L,M)$ with respect to M to get our new lower bound. Thus:

$$L_{CP}^{GOFC} = \int \mathcal{F}(L_{CP},m)\phi(m)dM$$

We define U_{CP}^{GOFC} in a similar way. Now we consider the performance of this interval in terms of coverage probability and expected length for various values of n and correlation $a = 2$.



Coverage Prob., Clopper−Pearson
correlation = 0.2
CP
1.0
0.9
0.8
0.7
0.6
0.5
0.0
0.1
0.2
0.3
0.4
0.5
probability
n=50
n=100
n=200



The expected length is better than the exact method, but the coverage probability is well below the nominal level for all values of n and most values of p. We thus do not consider the Clopper-Pearson method any further.

3.2.2 Wald-type Interval

We consider two types of Wald intervals. For the first, simply called Wald in the graph below, we again take the observed number of defaults and let L_W be the corresponding Wald lower bound. Then:

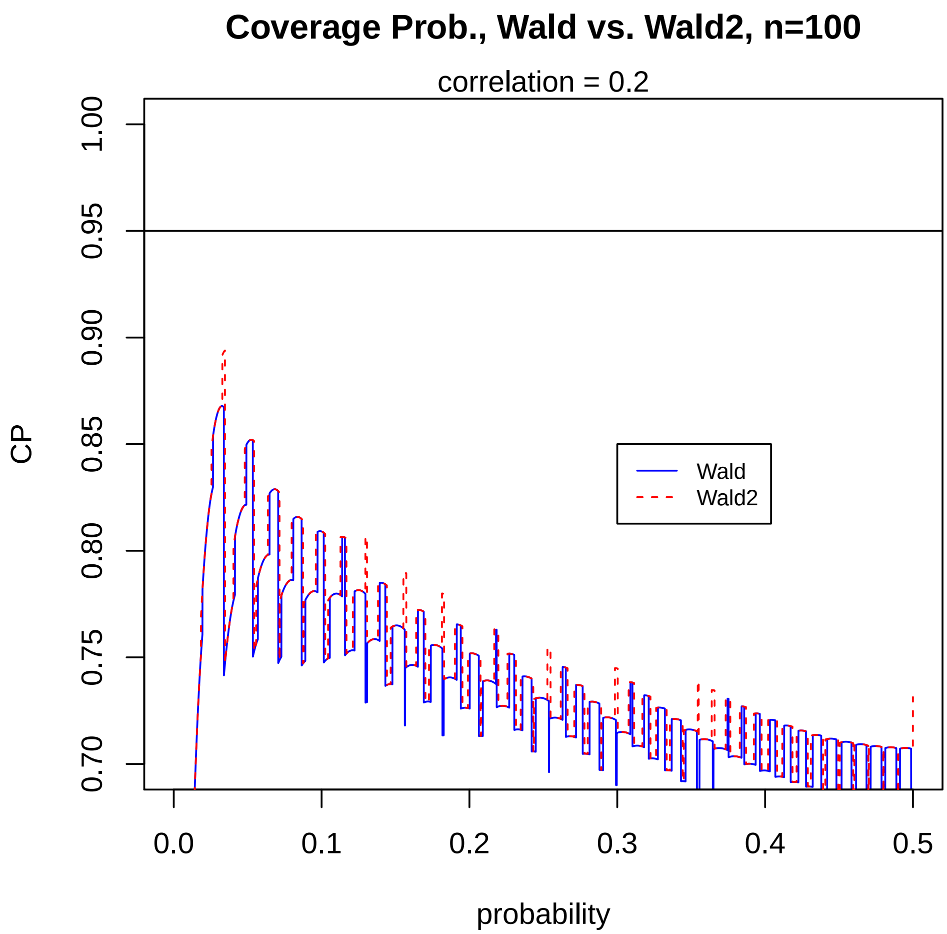
$$L_W^{GOFC} = \int \mathcal{F}(L_W, m)\phi(m)dM$$

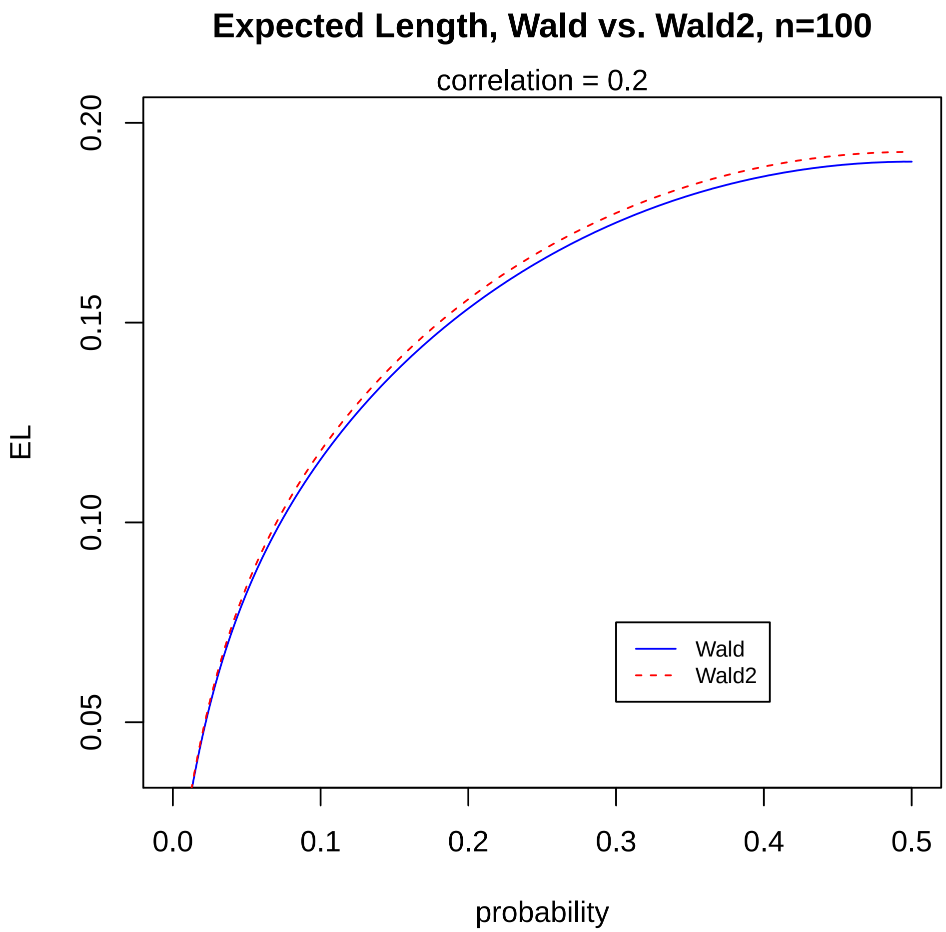
and similary for U_W^{GOFC}.

For the second, Wald2, we take $\hat{p} = d/n$ and define the center, C_{W2}^{GOFC}, of our interval by:

$$C_{W2}^{GOFC} = \int \mathcal{F}(\hat{p}, m)\phi(m)dM$$

We then form an interval by using this as our binomial estimate in a normal Wald interval. A comparison is shown below.





We see again very poor coverage probability for both these wald-type intervals and thus do not consider them further.

3.2.3 Agresti-Coull-type Interval

Recall that for the Agresti-Coull interval, we let $\tilde{X} = X + z^2/2$ and $\tilde{n} = n + z^2$, where X is the observed number of "successes" and n the number of trials. Then $\tilde{p} = \tilde{X}/\tilde{n}$ and $\tilde{q} = 1 - \tilde{p}$. The Agresti-Coull interval is then given by:

$$CI_{AC} = \tilde{p} \pm z(\tilde{p}\tilde{q})^{1/2}\tilde{n}^{-1/2}$$

Similar to the Wald intervals above, we again consider two AC- type intervals. For the first, AC, we make the usual modification, namely adding $z^2/2$ to the observed number of defaults and z^2 to the number of observations. Once this is done, we can get the lower AC bound, L_{AC}, and then form:

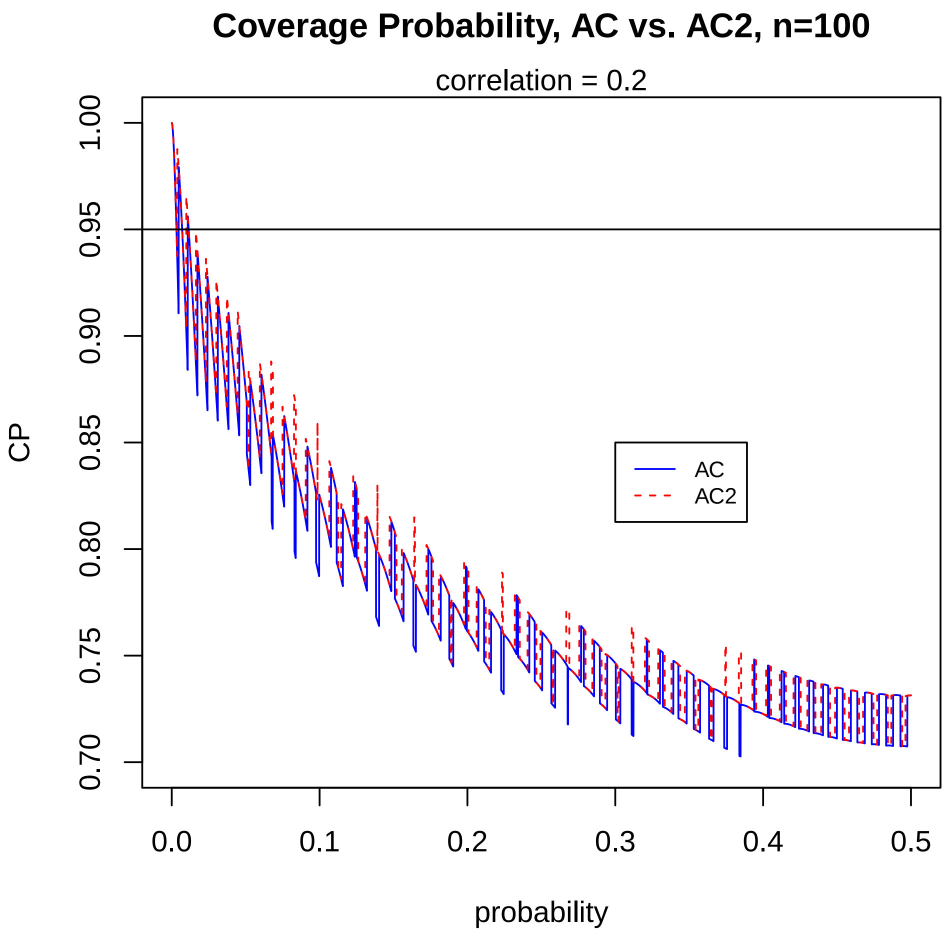
$$L_{AC}^{GOFC} = \int \mathcal{F}(L_{AC}, m)\phi(m)dM$$

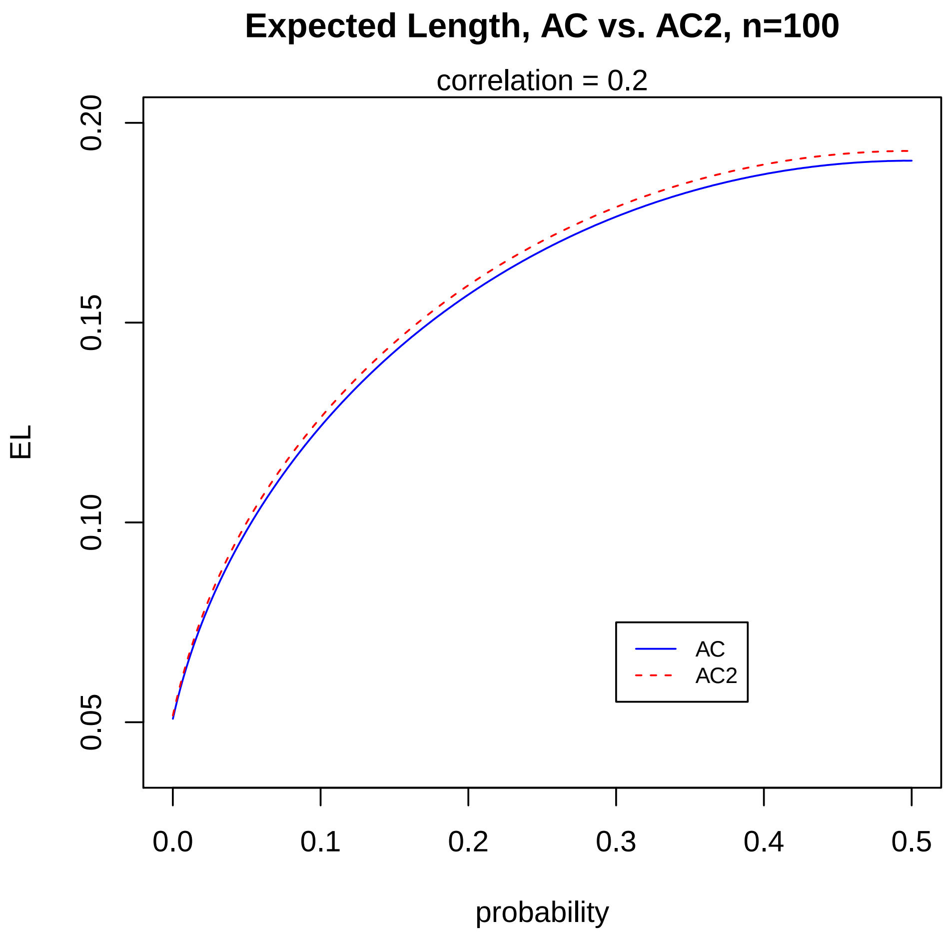
and similary for U_W^{GOFC}.

For AC2, we find the center, C_{AC2}^{GOFC} by:

$$C_{AC2}^{GOFC} = \int \mathcal{F}(\tilde{p}, m)\phi(m)dM$$

and then use this as our binomial estimate in a normal Agresti-Coull interval. Similar to the Wald-type intervals, these AC-type intervals have the same center and vary only in width.


Coverage Probability, AC vs. AC2, n=100
correlation = 0.2
CP
1.00
0.95
0.90
0.85
0.80
0.75
0.70
AC
AC2
0.0
0.1
0.2
0.3
0.4
0.5
probability



Although slightly better than Wald, the AC intervals still perform poorly in coverage probability. We next look at Score-type intervals.

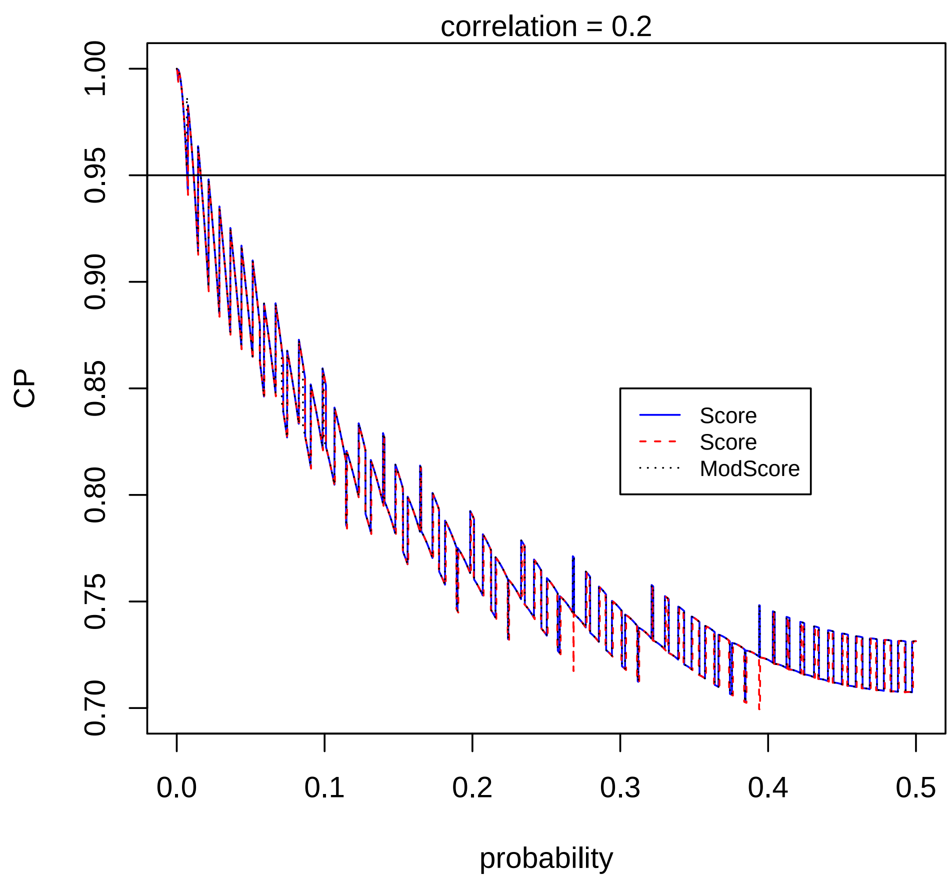
3.2.4 Score-type Interval

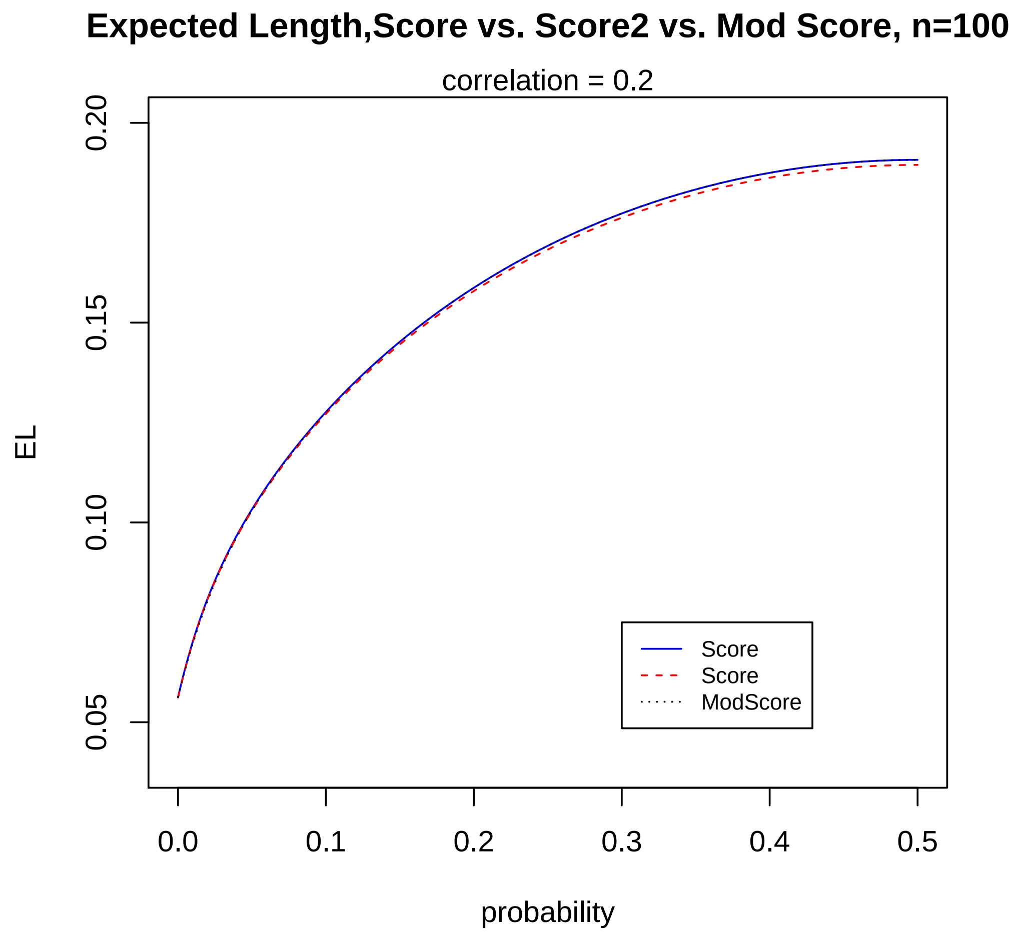
The Score interval is given by:

$$CI_W = \frac{X + z^2/2}{n + z^2} \pm \frac{zn^{\frac{1}{2}}}{n + z^2}(\hat{p}\hat{q} + z^2/(4n))^{\frac{1}{2}}$$

Similar to the Agresti-Coull-type interval, we consider two methods for the Score-type interval. We call Score the method of taking the expected value of the lower and upper bounds with respect to M and Score2 the method of estimating the probability first by taking the expected value with respect to M and then forming the interval. The details are omitted since the are exactly the same as the Wald and AC type intervals considered above. We also inlclude the a version of the modified Score interval discusses earlier for comparative purposes. This is formed in exactly the same way as Score, but with the bounds of the modifed Score in place of the regular Score:

Coverage Prob., Score vs. Score2 vs. Mod Score, n=100





The preformance of the Score-type intervals is nearly identical to the AC intervals. Interestingly, the Modified Score is no better here for small values of p than the regular Score.

3.2.5 Zhou-Li-type Interval

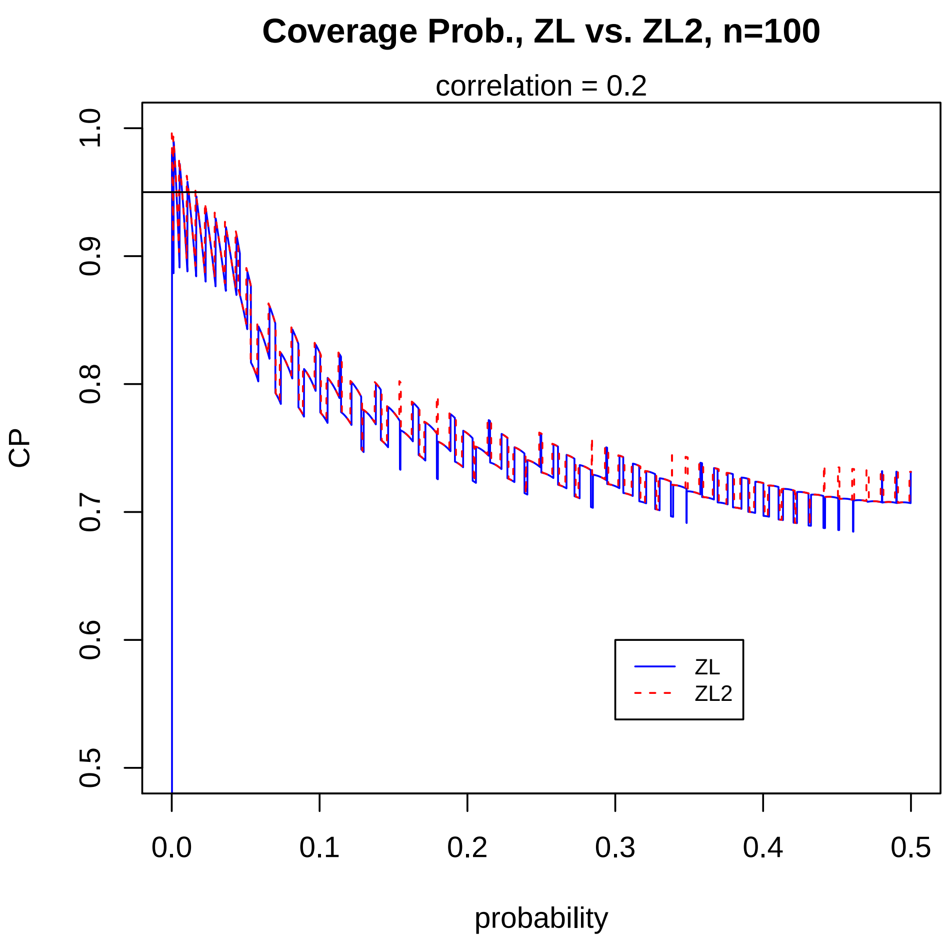
The ZL interval was introduced in the previous section as:

$CI_{ZL} =$

$$\left[\frac{\exp\left(\log\left(\frac{\hat{p}}{\hat{q}} - n^{-1/2}(\hat{p}\hat{q})^{-1/2}g^{-1}(z_{1-\alpha/2})\right)\right)}{1+\exp\left(\log\left(\frac{\hat{p}}{\hat{q}} - n^{-1/2}(\hat{p}\hat{q})^{-1/2}g^{-1}(z_{1-\alpha/2})\right)\right)}, \frac{\exp\left(\log\left(\frac{\hat{p}}{\hat{q}} - n^{-1/2}(\hat{p}\hat{q})^{-1/2}g^{-1}(z_{\alpha/2})\right)\right)}{1+\exp\left(\log\left(\frac{\hat{p}}{\hat{q}} - n^{-1/2}(\hat{p}\hat{q})^{-1/2}g^{-1}(z_{\alpha/2})\right)\right)}\right]$$

where the monotone transformation $g(x)$ was defined in the previous section.

We again form two intervals as in the case of AC and Score. ZL is formed by taking the expectation of the lower and upper bounds with respect to M and ZL2 is formed by taking the expected value of the probability estimate with respect to M and then forming the interval.



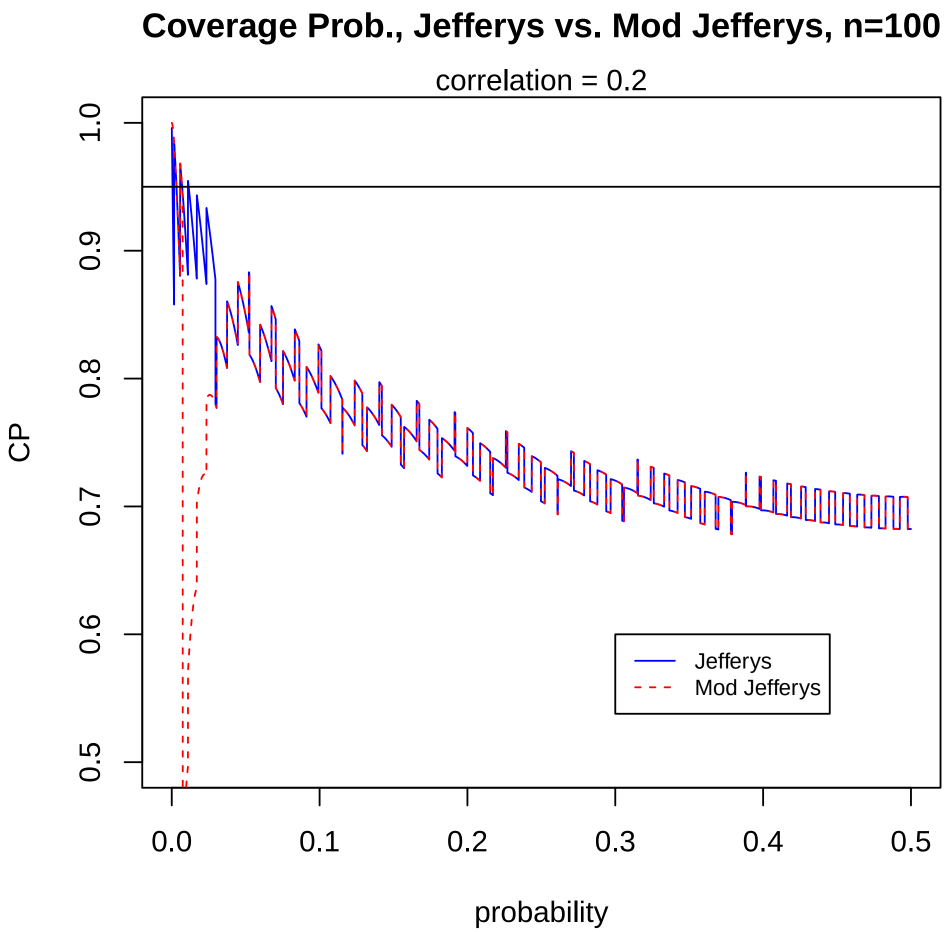
Once again, we see the same performance issues as with the other approximate intervals above.

3.2.6 Jefferys-type Interval

Let L_J and U_J be the lower and upper bounds, respectively, of an interval formed using Jefferys method discussed in the previous section. Recall that,

$$L_J(x) = B(\alpha/2; X + 1/2, n - X + 1/2),$$
$$U_J(x) = B(1 - \alpha/2; X + 1/2, n - X + 1/2)$$

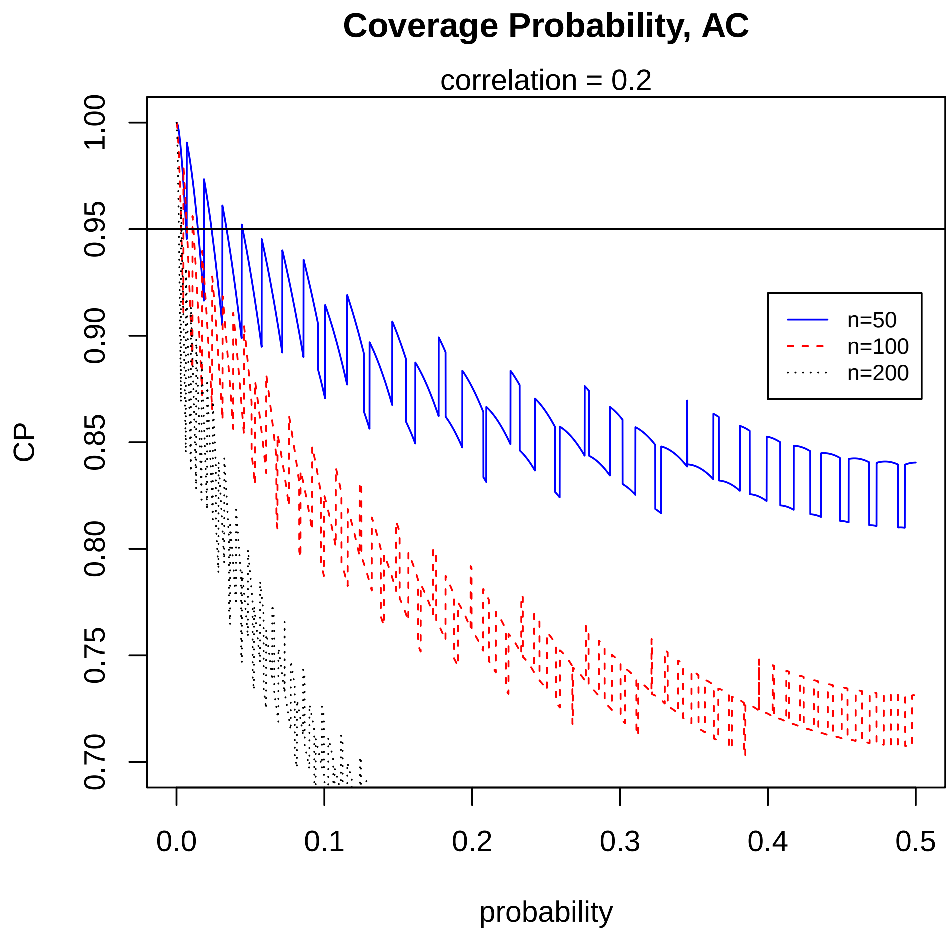
Where B is the quantile function of the Beta distribution. Then to form the lower interval we take the expected value with respect to M of $\Phi(\Phi^{-1}(L_J) \cdot \sqrt{1 - a^2} + aM)$. Similarly, for the upper bound we take the expected value with respect to M of $\Phi(\Phi^{-1}(U_J) \cdot \sqrt{1 - a^2} + aM)$. We also consider the modified Jefferys interval in an analogous way.

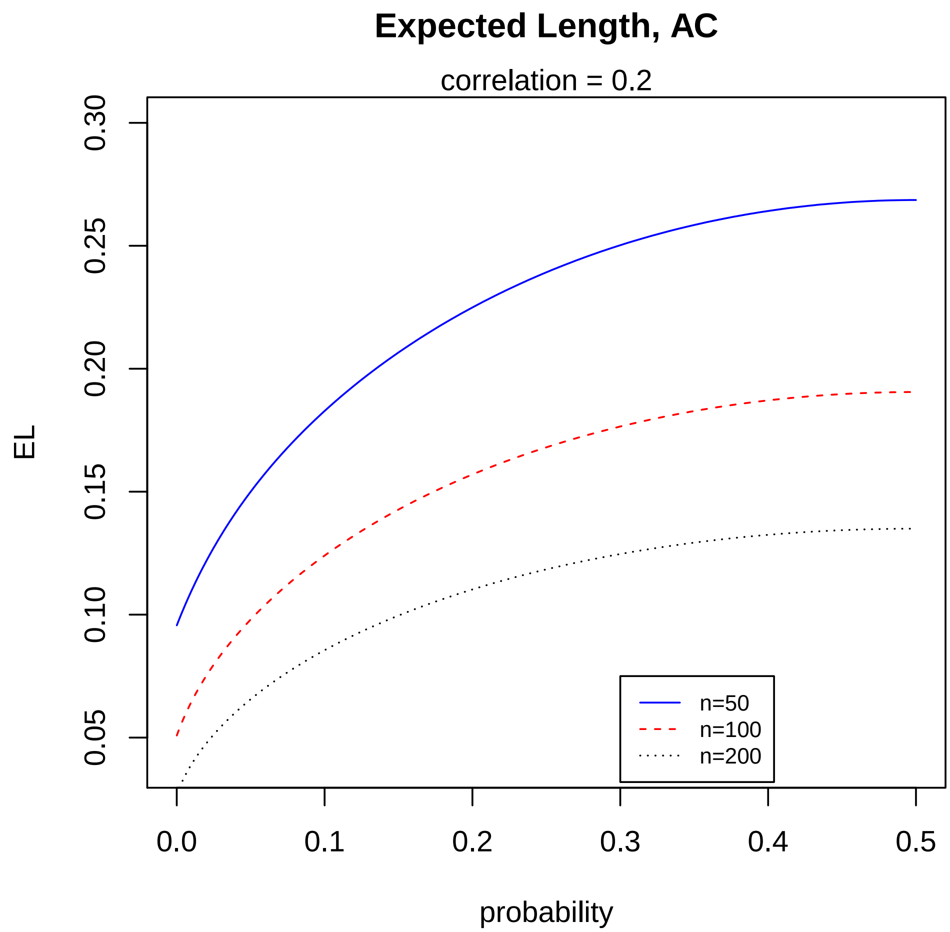


Interestingly, the modified Jefferys actually performs much worse for very low probability values which is contrary to the independent binomial case. Again we see effects similar to the other intervals above. We will not consider the Jeffreys intervals further.

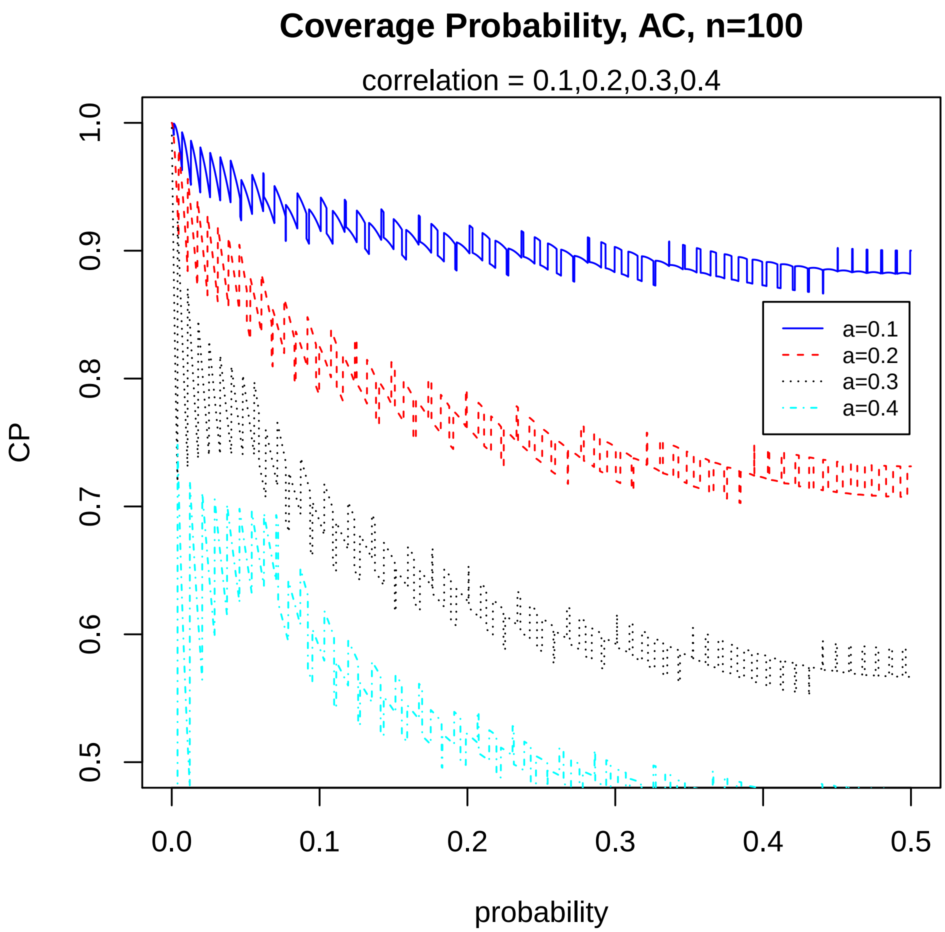
3.2.7 Approximate Interval Conclusions

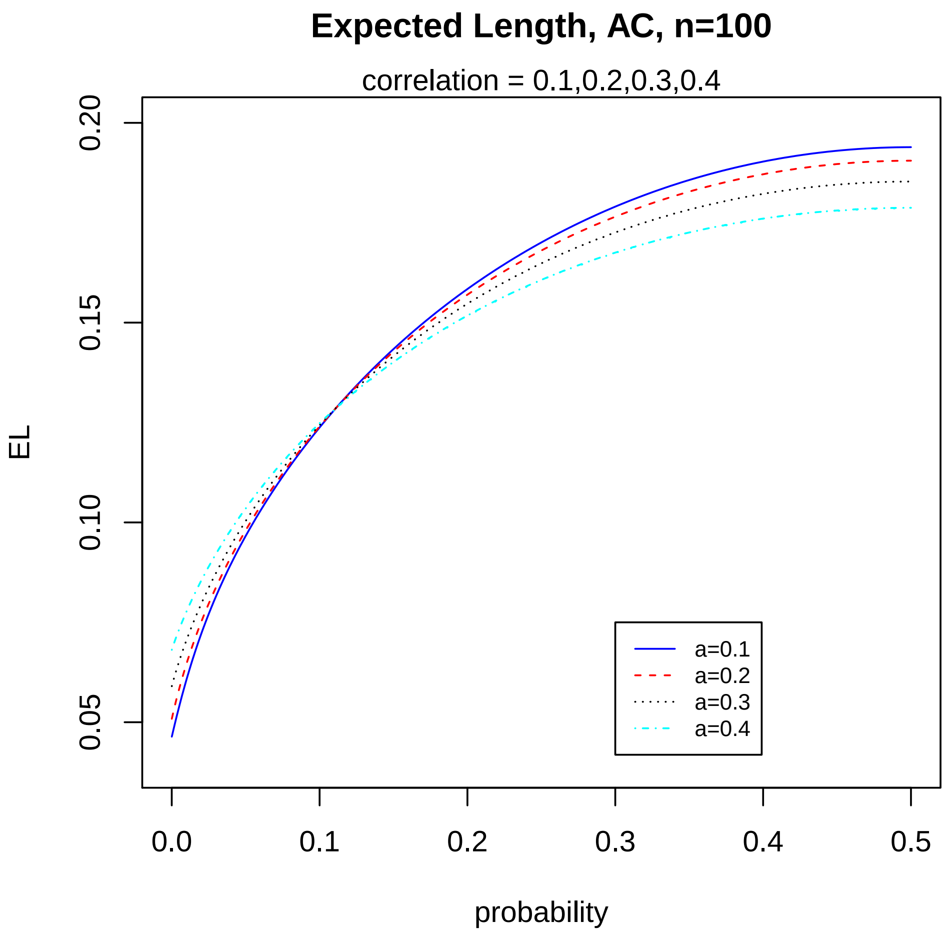
Again consider the Agrest-Coull approximate interval for the GOFC. We first look at the AC interval for different values of n, with a=0.2:


Coverage Probability, AC
correlation = 0.2
CP
probability
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.0
0.1
0.2
0.3
0.4
0.5
n=50
n=100
n=200



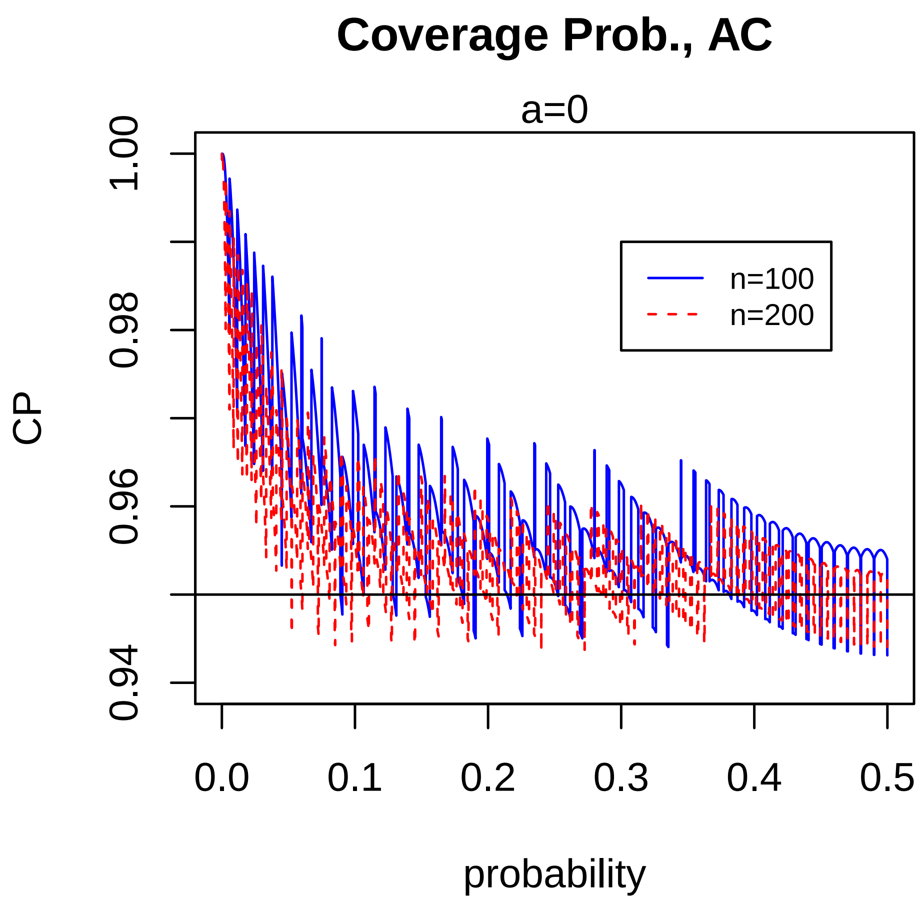
Note that the coverage probability becomes much worse as n increases, and the expected length decreases. This is the case for all the approximate intervals.
Next we look at the AC approximate interval for the GOFC with n fixed at 100 and varying values of a:


Coverage Probability, AC, n=100
correlation = 0.1,0.2,0.3,0.4
CP
probability
a=0.1
a=0.2
a=0.3
a=0.4
1.0
0.9
0.8
0.7
0.6
0.5
0.0
0.1
0.2
0.3
0.4
0.5



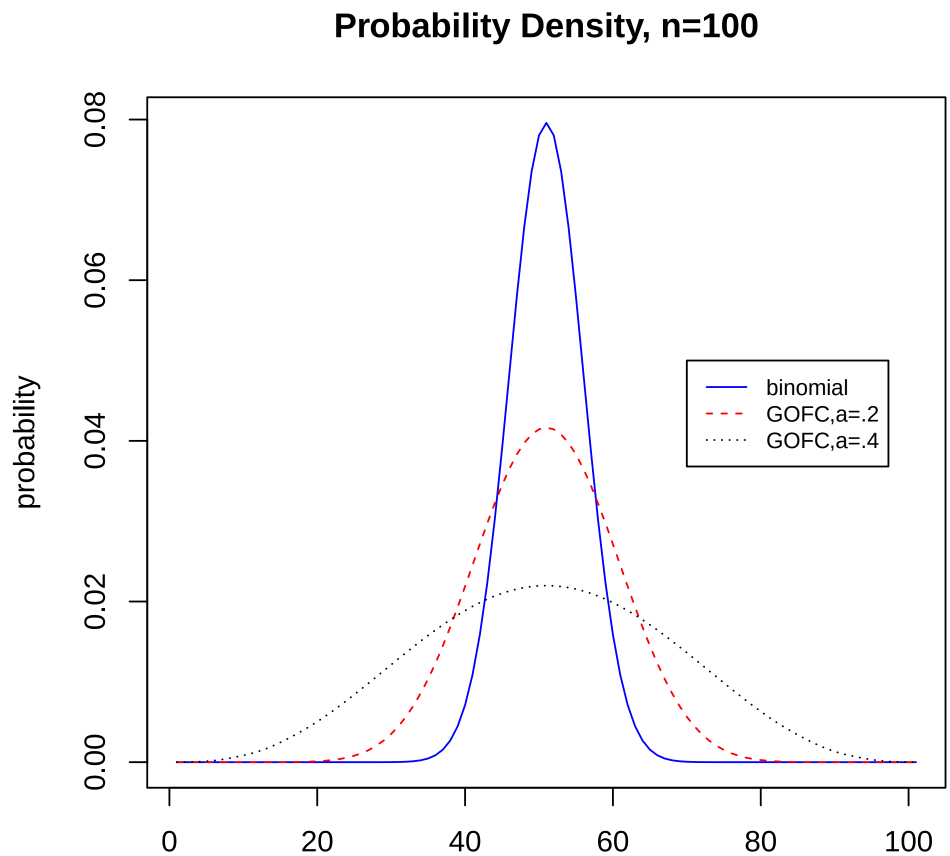
The coverage probability again becomes much worse as a increases from 0.1 to 0.4. The expected length does not change a great deal as a increases. Again, this is representative of all the approximate interval performances when used in the GOFC model.

Finally, we look at the AC approximate interval for the GOFC when $a = 0$ and $n = 100$ and $n = 200$:



We see that setting $a = 0$ in the model and using the AC approximate interval is basically identical to using the AC interval for the binomial proportion for $n = 100, 200$.

So why do the approximate intervals perform so poorly? We first look at the graph of the binomial distribution vs. the GOFC with $a = .2$ vs. the GOFC with $a = 0.4$ and $p = 0.5$:



We see that increasing the correlation flattens the density and puts more probability in the tails of the distribution. Thus the likelihood that we will observe a small or large $\hat{p}$ increases as we increase the correlation. In fact, the 95% symmetric density interval centered at $p = 0.5$ for the binomial distribution with $n = 100$ and $p = 0.5$ is (.39,.59), while the GOFC with $a = 0.2$ is (.32,.68), and the GOFC with $a = 0.4$ is (.18,.82). For the approximate confidence intervals considered above, the variance that is used to form the width of the interval is based on the binomial distribution, and thus is considerably shorter than what would be used if the correlation was taken into account. It follows that as we increase a we need to increase the width of our intervals, but the approximate intervals remain the same width. Thus the coverage probability must necessarily decrease. Similarly, as n increases, the approximate intervals decrease in width while the variance of the GOFC model does not change, resulting in decreased coverage probability.

On a final note, consider a related problem of n dependent Bernoulli random variables with probability of success p and correlation a^2. Letting $S_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, then $E(S_n) = p$ and,

$$\begin{aligned} Var(S_n) &= \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) + 2\sum_{i<j} Cov(X_i, X_j) \\ &= \frac{1}{n^2}(np(1-p) + 2p(1-p)n(n-1)a^2/2) \\ &= \frac{p(1-p)}{n}(1 + (n-1)a^2) \end{aligned}$$

So in addition to the variance we see in the independent case, $p(1-p)/n$, we have the additional term $\frac{n-1}{n}p(1-p)a^2$. When $n = 100$ and $p = 0.5$, the variance of S_n when $a = 0$ is .0025. When $a = .2$, the variance becomes 0.0124 and when $a = 0.4$ the variance is 0.0421. Miao and Gastwirth(2004) note that in the case of dependence, S_n converges to p, albeit more slowly due to the higher variance, as long as the X_i satisfy certain conditions such as m-dependence or mixing conditions. They suggest modifications to the intervals considered above by adding in the modified variance. Although their intervals work well in the cases they consider, they produce extremely conservative intervals when considered in the GOFC model and no modification proved successful in our attempts.

Finally we note that the expected length of the exact interval increases as a increases, but this is to be expected due to the increase in variance of the underlying distribution described above.

References

[1] Gupton G., Finger C., Bhatia, M., 1997. CreditMetrics - Technical Document, J.P. Morgan.

[2] Hull, J., White, A., 2006. "Valuing Credit Derivatives Using an Implied Copula Approach. *Journal of Derivatives.*, Vol. 14, No.2, (Winter 2006), pp. 8-28.

[3] Li, D., 2000. "On Default Correlation: A Copula Function Approach.", *Journal of Fixed Income*, 9(4):43-54.

[4] O'Kane, D., 2008. Modelling single-name and multi-name Credit Derivatives, John Wiley & Sons.