Derivation of Nonparametric Confidence Bounds for Proportion

By Allan Mense, Ph.D., PE, CRE, Principal Engineering Fellow

References:

1) Statistical Theory with Engineering Applications by A. Hald, Chap 21, Wiley (1952),
2) Statistical Theory and Methodology for Science & Engineering by K. Brownlee, Chap 3, Sec. 3.5, Wiley (1965)
3) Reliability Engineering Handbook by B. Dodson & D. Nolan, Chap. 3, Quality Publishing (1999),
4) Statistical Methods for Reliability Data by Meeker & Escobar, Chap 3, Wiley (1998)
5) Nonparametric Statistical Methods by M. Hollander & D. Wolfe, Chap 1, pg 32, Wiley (1999)

Introduction:

When one measures the outcomes of a series of binary events and calculates say a proportion of failures <p> = r/n where r = number of failure in n trials, there is always a question of what is the confidence interval around this proportion.

The classical method relies on calculating the cumulative binomial distribution.

These bounds, labeled $[p_L,p_U]$, depend on the confidence level $(1-\alpha)$ desired for each bound, the number of failures and of course the number of trials.

In most textbooks the procedure is assume the sample size is large enough to call upon the central limit theorem and use a normal approximation that in its least accurate form looks like

$$[p_L, p_U]= \left[< p > -Z_{1-\alpha^*}\sqrt{\frac{< p > (1- < p >)}{n}}, < p > +Z_{1-\alpha}\sqrt{\frac{< p > (1- < p >)}{n}} \right]$$

Let α be the significance level for the upper bound and α^* the significance level for the lower bound. The confidence level is therefore $1 - \alpha - \alpha^*$. A more general statement would take the form of the probability statement below.

$$\Pr\left\{ p_L(\alpha^*, r, n) < p < p_U(\alpha, r, n) \right\} = 1 - \alpha^* - \alpha$$

Given a hypothetical proportion in a *population* we wish to know what is the chance of that population value, p, being within the above interval. The sample proportion <p> = r/n and we use this information in finding p_L and p_U.

To find the upper $(1-\alpha)100\%$ confidence bound, $p_U(\alpha)$, we need to pose the following question: If the population proportion equals p_U, *what is the probability that a sample of size, n, has r or fewer failures.* Call this probability α? The equation that needs to be solved is setting the CDF of the binomial distribution equal to α and finding p_U. That is

given $\sum_{i=0}^{r}\binom{n}{i}p_U^i(1-p_U)^{n-i}=\alpha$, solve iteratively to find the p_U value. Example: r=2, n=20, <p>=2/20=0.1, Find the 95% upper confidence bound (α=0.05). Using SOLVER in Excel gives p_U = .2826.

This is a way to get the exact (nonparametric)answer.

To find the lowest (1-α*)100% confidence level value we pose the opposite question: *what value of proportion, p_L, will produce r failures or more in n trials with a probability* α^*. Again the equation to solve is given by $\sum_{i=r}^{n}\binom{n}{i}p_L^i(1-p_L)^{n-i}=\alpha^*$ (or stated in another way $\sum_{i=0}^{r-1}\binom{n}{i}p_L^i(1-p_L)^{n-i}=1-\alpha^*$. Again it can be solved iteratively for p_L. The lower 95% confidence bound is found for the above example to be p_L=0.0181.

To summarize the equations to be solved;

$$\Pr\{p\le r/n\,|p_U\}=\sum_{i=0}^{r}\binom{n}{i}p_U^i(1-p_U)^{n-i}=\alpha$$

$$\Pr\{p\ge r/n\,|p_L\}=\sum_{i=r}^{n}\binom{n}{i}p_L^i(1-p_L)^{n-i}=\alpha^*, \text{ or}$$

$$=1-\sum_{i=0}^{r-1}\binom{n}{i}p_L^i(1-p_L)^{n-i}=\alpha^*$$

The purpose of this note is to calculate the nonparametric expressions for p_L and p_U, i.e. show

$$p_L=\frac{r}{r+(n-r+1)F^{-1}_{1-\alpha^*,2(n-r+1),2r}},\ p_U=\frac{(r+1)F^{-1}_{1-\alpha,2(r+1),2(n-r)}}{(n-r)+(r+1)F^{-1}_{1-\alpha,2(r+1),2(n-r)}}$$

or

$$p_L=1-B_{1-\alpha^*}(n-r+1,r),\ p_U=1-B_{\alpha}(n-r,r+1)$$

where the F-distribution (actually the inverse of the CDF of the F-distribution) and B is the inverse beta distribution. These distributions are used instead of summing the terms in the binomial distribution. This makes calculations ever so much easier as excel has the FINV and BETAINV functions.

A short numerical example helps here.

Example 1: Calculate a 80% confidence (α=α*=.1) interval for n=40 tests and r=2 (# failures). Note <p>=r/n=2/40 =.05 (The Excel function FINV calculates F from highest to lowest values so its argument in the FINV function uses α not 1−α for the probability argument). Plugging in numbers to the formulas shown above and

noting that $F^{-1}_{.1,78,4}$=3.783, and $F^{-1}_{.1,6,76}$=1.853, one obtains for this 80% confidence interval $\frac{2}{2+(40-2+1)*3.783} \le p \le \frac{(2+1)1.853}{(40-2)+(2+1)1.853}$ or Pr{0.013<p<0.128}=.8.
Instead of using the FINV function one can use the BETAINV function in excel. Using the latter function gives the same answer of course!

n =	40	alpha =	0.1
r =	2	alpha* =	0.1
Using BETAINV in excel			
P_L =	0.013375	P_U =	0.127628

In statistics courses one learns to use the normal approximation to the binomial distribution. This is not the correct way to solve for the confidence intervals for binary data. It is taught because it is easy to teach not because it is accurate. To demonstrate this consider the above numerical problem. Using the normal approximation one obtains p_L=<p>-1.282*(.1*0.9/40)$^{1/2}$ = 0.006 and p_U=0.094. This gives a confidence interval 0.006<p<0.094 which is smaller than the exact interval by (0.094-0.006)/(0.128-0.013)=0.77. That is, the approximation gives a tighter interval than the exact calculation. What a mistake to make! Underestimating the interval by almost 25%. The correct answer in this counting problem is given by the nonparametric calculation of the interval. The sample size n=40 is considered by most "statistics users" to be "large enough" to use the normal approximation (Appendix C). It is not!

Derivation of the expression p_U and p_L:

Consider the integral

$$\int_p^1 y^{r+1-1}(1-y)^{n-r-1}dy \equiv I_{1-p}(r+1,n-r)$$

This is proportional to the "upper tail" of the cumulative beta distribution (see appendix A) with parameters r+1 and n-r.

The pdf of the beta distribution is of course $f_\beta(y|r+1,n-r) = y^{r+1-1}(1-y)^{n-r-1}/B(r+1,n-r)$.

Integrate I_{1-p} by parts. After the first integration,

$$I_{1-p}(r+1,n-r) = \frac{p^r(1-p)^{n-r}}{n-r} + \frac{r}{n-r}\int_p^1 y^{r-1}(1-y)^{n-r}dy .$$

Performing integration by parts again gives

$$I_{1-p}(r+1,n-r) = \frac{p^r(1-p)^{n-r}}{n-r} + \frac{r}{n-r}I_{1-p}(r,n-r+1)$$

$$= \frac{p^r(1-p)^{n-r}}{n-r} + \frac{p^{r-1}(1-p)^{n-r+1}x}{(n-r)(n-r+1)} + \frac{r(r-1)}{(n-r)(n-r+1)}\int_p^1 y^{r-2}(1-y)^{n-r+1}dy$$

Multiplying the above expression by $n\binom{n-1}{r}$ gives a more familiar expression

$$n\binom{n-1}{r}I_{1-p}(r+1,n-r)=\frac{n!}{(n-r-1)!r!}\left\{\frac{p^r(1-p)^{n-r}}{n-r}+\frac{p^{r-1}(1-p)^{n-r+1}r}{(n-r)(n-r+1)}+\frac{x(x-1)}{(n-x)(n-x+1)}\int_p^1 y^{r-2}(1-y)^{n-r+1}dy\right\}$$

$$=\frac{n!}{(n-r)!r!}p^r(1-p)^{n-r}+\frac{n!}{(n-r+1)!(r-1)!}p^{r-1}(1-p)^{n-r+1}+\mathrm{L}$$

These terms begin to look familiar.

Performing this integration r-2 more times produces the result

$$n\binom{n-1}{r}I_{1-p}(r+1,n-r)=n\binom{n-1}{r}\int_p^1 y^{r+1-1}(1-y)^{n-r-1}dy\equiv\sum_{\nu=0}^{r}\binom{n}{\nu}p^\nu(1-p)^{n-\nu}\equiv P\{r\}$$

P{r} is the cumulative mass probability for the binomial distribution for r failures in n trials given we know p. It is also called the "lower tail" of the binomial distribution and it will be shown below to equal the "upper tail" of the beta distribution. ***This is a key insight.*** Rewriting the above expression by differencing two integrals with limits (0,1) and (0,p) respectively one obtains,

$$P\{r\}=\frac{1}{B(r+1,n-r)}\left[\int_0^1 y^{r+1-1}(1-y)^{n-r-1}dy-\int_0^p y^{r+1-1}(1-y)^{n-r-1}dy\right]$$

and defining B and B_P one obtains,

$$B(r+1,n-r)=\int_0^1 y^{r+1-1}(1-y)^{n-r-1}dy=\frac{\Gamma(r+1)\Gamma(n-r)}{\Gamma(n+1)}=\frac{r!(n-r-1)!}{n!}$$

$$B_p(r+1,n-r)=\int_0^P y^{r+1-1}(1-y)^{n-r-1}dy$$

the cumulative probability of the binomial distribution from 0 to r is also given by

$$P\{r\}=\left[B(r+1,n-r)-B_p(r+1,n-r)\right]\frac{n!}{(n-r-1)!r!}=$$

$$\left[B(r+1,n-r)-B_p(r+1,n-r)\right]\frac{\Gamma(n+1)}{\Gamma(n-r)\Gamma(r+1)}=\frac{B(r+1,n-r)-B_p(r+1,n-r)}{B(r+1,n-r)}=P\{r\}$$

$$P(r)=1-\beta_p(r+1,n-r),$$

where $\beta_P(a,b)$ is the incomplete beta distribution with parameters (a,b).

[*Note how the factorials are just the complete Beta function. Remember this form for P{r}.*].

If you have software or tables that compute the incomplete beta distribution, $\beta_P(a,b)$ then use those values to compute the needed P(r). Many times however on has only the tables for the F-distribution and not the tables for the beta distribution.

Now explore the form of the F-distribution. The pdf is generally written as shown below.

$$f_F(z\,|\,\nu_1,\nu_2)=\frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)}\nu_1^{\frac{\nu_1}{2}}\nu_2^{\frac{\nu_2}{2}}\frac{z^{\frac{\nu_1}{2}-1}}{\left(\nu_2+\nu_1 z\right)^{\frac{\nu_1+\nu_2}{2}}},0\le z\le\infty$$

Using the variable $y=\nu_1 z/(\nu_2+\nu_1 z)$ transforms the F-distribution into the form (see Appendix A)

$$f_F(y\,|\,\nu_1,\nu_2)=\frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)}y^{\frac{\nu_1}{2}-1}(1-y)^{\frac{\nu_2}{2}-1},0\le y\le 1$$

The cumulative F-distribution is given by integrating over y from 0 to y_p

$$F_F(y_p\,|\,\nu_1,\nu_2)=\frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)}\int_0^{y_p}z^{\frac{\nu_1}{2}-1}(1-z)^{\frac{\nu_2}{2}-1}dz=\frac{B_{y_p}\left(\frac{\nu_1}{2},\frac{\nu_2}{2}\right)}{B\left(\frac{\nu_1}{2},\frac{\nu_2}{2}\right)}$$

Which looks a lot like the CDF of the Beta distribution if one lets $\nu_1 / 2 = (r+1)$ and $\nu_2 / 2 = (n-r)$. Thus the function $B_P(r+1,n-r)$ is related to the CDF of the F-distribution by the relation

$F((n-r)/(r+1)*p/(1-p)|2(r+1),2(n-r)) = B_p(r+1,n-r)/\ B(r+1,n-r) = \beta_p(r+1,n-r)$

Note the argument of the F distribution is given by
$((n-r)/(r+1))*(p/(1-p))$

Now

$$\frac{B(r+1,n-r)-B_P(r+1,n-r)}{B(r+1,n-r)}\equiv P\{r\}=1-\ B_p(r+1,n-r)/\ B(r+1,n-r)=1-\beta_P(r+1,n-r)=1-\alpha$$

From this one immediately sees that

$$P\{r\}=1-\Pr\left\{F(2(r+1),2(n-r))<\left(\frac{n-r}{r+1}\right)\frac{p}{1-p}\right\}=\alpha$$

$$\Rightarrow\Pr\left\{F(2(r+1),2(n-r))<\left(\frac{n-r}{r+1}\right)\frac{p_U}{1-p_U}\right\}=1-\alpha$$

Since $\Pr\{F<F_{1-\alpha}\}=1-\alpha$, this implies $\left(\frac{n-r}{r+1}\right)\frac{p_U}{1-p_U}=F_{1-\alpha}^{-1}(2(r+1),2(n-r))$

More practically one uses the FINV function in Excel. Solving for p gives

$$\frac{n-r}{r+1}\frac{p_U}{1-p_U} = F^{-1}_{1-\alpha,2(r+1),2(n-r)}$$

$$p_U = \frac{(r+1)F^{-1}_{1-\alpha,2(r+1),2(n-r)}}{(n-r)+(r+1)F^{-1}_{1-\alpha,2(r+1),2(n-r)}}$$ QED

Just to make the situation more complicated on finds in practice that finding the inverse using excel requires noting that $F^{-1}_{1-\alpha,2(r+1),2(n-r)}$ =FINV(α,2(r+1),2(n-r)). Excel calculates the CDF from largest to smallest values of α. See Appendix B

Now it is a little more complex calculating p_L. $\sum_{i=r}^{n}\binom{n}{i}p_L^i(1-p_L)^{n-i} = \alpha^*$ and to use P{r} we need to rewrite this expression as
$1-\sum_{i=0}^{r-1}\binom{n}{i}p_L^i(1-p_L)^{n-i}$ = 1-P{r-1} = α*. Using

$$P\{r-1\} = 1-P\left\{Y < \frac{n-r+1}{r}\frac{p}{1-p}\right\}$$

Setting P{r-1} = 1-α*, p=p_L, and noting the identity for the CDF of the F distribution which is F(α∗,2r,2(n-r+1)) = 1/F(1-α∗,2(n-r+1),2r). Solve for p_L using r →r-1 one finds,

$$\frac{n-r+1}{r}\frac{p_L}{1-p_L} = F^{-1}_{\alpha^*}(2(r),2(n-r+1)) = 1\Big/F^{-1}_{1-\alpha^*}(2(n-r+1),2r)$$

$$p_L = \frac{(r)F^{-1}_{\alpha^*}(2(r),2(n-r+1))}{(n-r+1)+(r)F^{-1}_{\alpha^*}(2(r),2(n-r+1))} = \frac{r}{r+(n-r+1)F^{-1}_{1-\alpha^*}(2(n-r+1),2r)}$$ QED

Thus I have shown that the lower and upper bounds for the nonparametric distribution for binary data can be found without recourse to summing over the binomial distribution. It requires instead the use of the FINV function in excel and also somewhat unfortunately learning the idiosyncrasies of the FINV function in excel, e.g.
$F_{1-\alpha^*,2(n-r+1),2r}$ =FINV(α*,2(n-r+1),2r).

Summary
The bounds to the (1−α−α∗)100% confidence interval are given by the expressions shown below.

$$p_L = \frac{r}{r+(n-r+1)F_{1-\alpha^*,2(n-r+1),2r}},\ p_U = \frac{(r+1)F_{1-\alpha,2(r+1),2(n-r)}}{(n-r)+(r+1)F_{1-\alpha,2(r+1),2(n-r)}}$$

Appendix A

Fisher's F-Distribution

$$f(x \mid \nu_1, \nu_2) = \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)} \nu_1^{\frac{\nu_1}{2}} \nu_2^{\frac{\nu_2}{2}} \frac{x^{\frac{\nu_1}{2}-1}}{(\nu_2+\nu_1 x)^{\frac{\nu_1+\nu_2}{2}}}, 0 \le x \le \infty$$

Let

$$y \equiv \frac{\nu_1 x}{(\nu_2+\nu_1 x)}, \frac{dy}{dx} = \frac{(\nu_2+\nu_1 x)\nu_1 - \nu_1 x \nu_1}{(\nu_2+\nu_1 x)^2} = \frac{\nu_1 \nu_2}{(\nu_2+\nu_1 x)^2}$$

$$f(y \mid \nu_1, \nu_2) = \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)} y^{\frac{\nu_1}{2}-1}(1-y)^{\frac{\nu_2}{2}-1}, 0 \le y \le 1$$

$$P\{y\} = \int_0^y \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)} z^{\frac{\nu_1}{2}-1}(1-z)^{\frac{\nu_2}{2}-1} dz = \frac{B_y\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)}{B\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)}$$

Gamma Function (Complete):

$$\Gamma(x) \equiv \int_0^\infty e^{-t} t^{x-1} dt, \text{ x=real}$$

$$\Gamma(n+1) = n! = n\Gamma(n), \quad n = \text{integer} \ge 1$$

Beta Function:

Complete Beta Function

$$B(\alpha, \beta) \equiv \int_0^1 y^{\alpha-1}(1-y)^{\beta-1} dy = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}, \text{ where } \alpha, \beta \ge 1$$

$$\frac{1}{B(n,m)} = m\binom{n+m-1}{n-1} = n\binom{n+m-1}{m-1}, \text{where n,m integers} \ge 1$$

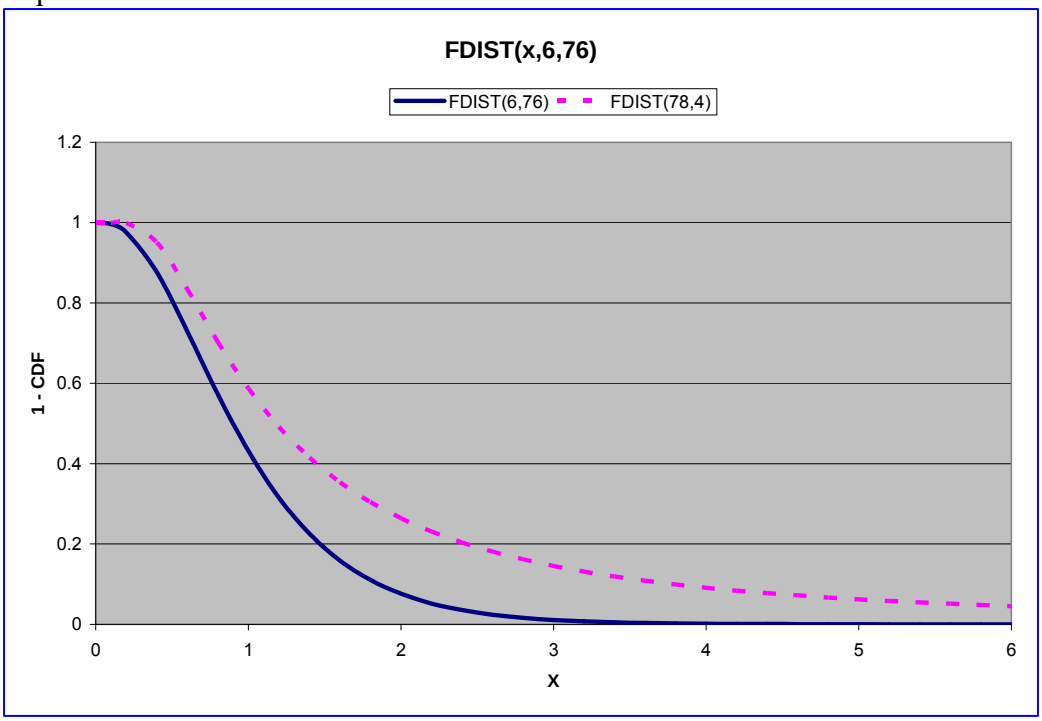
$$\binom{\text{a}}{\text{b}} \equiv \frac{a!}{(a-b)!b!}, \text{where a,b are integers } \ge 0.$$

Incomplete Beta Function, $B_p(\alpha, \beta)$

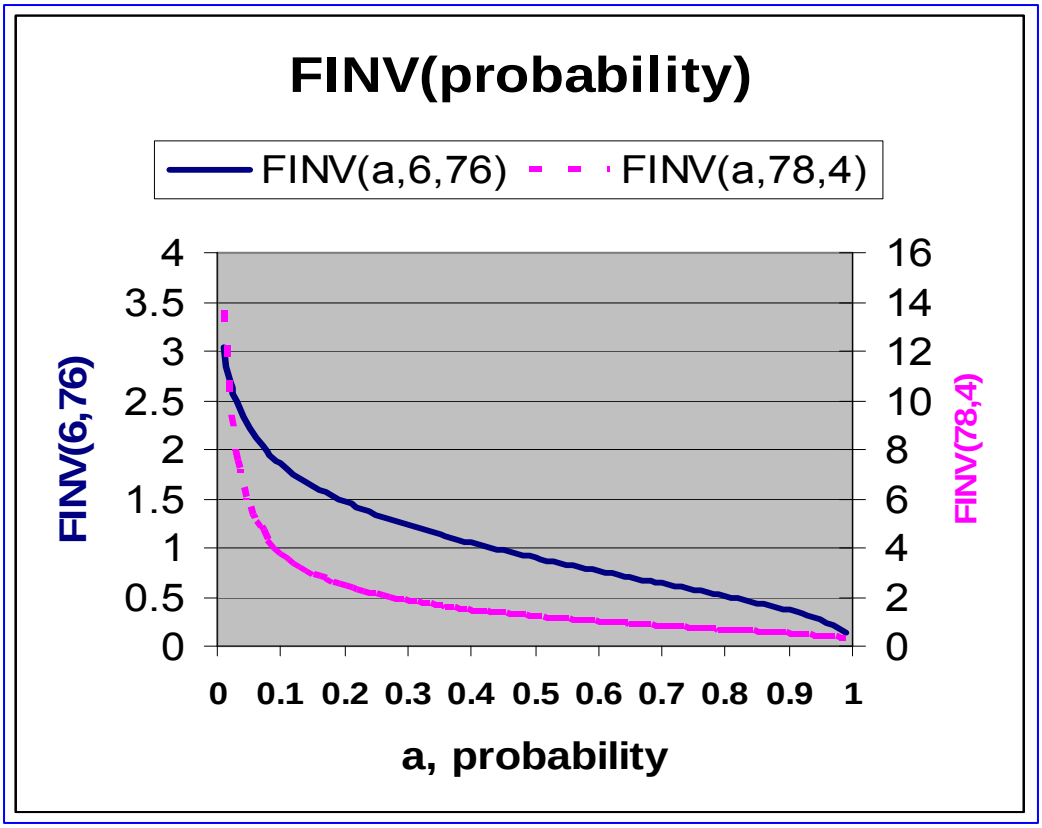
$$B_p(\alpha, \beta) \equiv \int_0^p y^{\alpha-1}(1-y)^{\beta-1} dy = B(\alpha, \beta) - \int_p^1 y^{\alpha-1}(1-y)^{\beta-1} dy$$

Appendix B

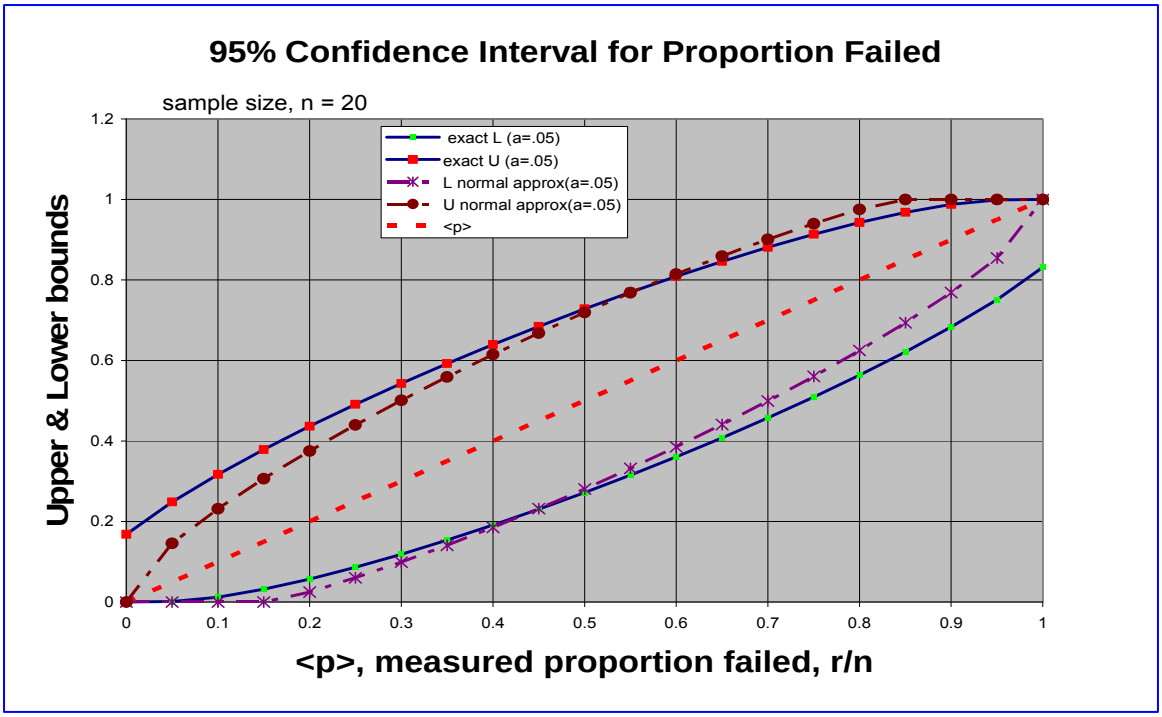
A plot of the CDF for two F-distributions is shown below. Note that Excel actually



calculates 1- CDF for this distribution. The inverse of the (1-CDF) is shown below.



Appendix C



Shown above is plot of exact and normal approximation for the 95% confidence interval for a sample size of 20 with the abscissa = <p>=r/n=# measured failures in n trials. Note the normal approximation, when it is valid, always gives smaller intervals than the exact nonparametric calculation. The chart below shows the same calculation for a sample size n=100. The two calculations are much closer, except for small <p>~0 or large <p>~1 where the differences are non negligible.