

XIV. STATISTICAL DISTRIBUTIONS AND RELATED FUNCTIONS.....	1
EXCEL PROVIDED STATISTICAL DISTRIBUTIONS:.....	1
EQUIVALENCES TO OTHER CUMULATIVE STATISTICAL DISTRIBUTIONS:	2
INFORMATION ON THE BASIC EXCEL STATISTICAL DISTRIBUTIONS:.....	3
DENSITY FUNCTIONS AND EQUIVALENCES:.....	4
CUMULATIVE DISTRIBUTION FUNCTIONS AND P VALUES:.....	5
CUMULATIVE DISTRIBUTION TAIL AREAS:.....	5
EXCEL FUNCTION TAIL AREAS	5
INVERSE DISTRIBUTION FUNCTIONS:	6
EXCEL 2000 AND EARLIER VERSIONS:	7
EXCEL 2002, 2003 AND 2007	7
REPORTED COMPLAINTS AND PROBLEMS:	8
PRIOR TEST RESULTS:.....	8

XIV. STATISTICAL DISTRIBUTIONS AND RELATED FUNCTIONS

EXCEL PROVIDED STATISTICAL DISTRIBUTIONS:

Excel along with other statistical computer programs, supplies probability values from the common statistical distributions. There are generally three applications, one for a value of the density function, two, a probability from the cumulative density function given an x value and three, an x value from the inverse of the cumulative function, given a probability value. Excel provides the following statistical distribution functions.

Table 14-1: Provided Statistical Distributions

DISTRIBUTION	REFERENCE	NUMBER OF PARAMETERS	DENSITY	CUMMULATIVE	INVERSE
Beta	1	5		BETADIST	BETAINV*
Binomial	2	3	BINOMDIST*	BINOMDIST*	CRITBINOM*
Chi-Square	3	2		CHIDIST	CHINV*
Exponential	4	2	EXPONDIST	EXPONDIST	
F	5	3		FDIST	FINV*
Gamma	6	3	GAMMADIST	GAMMADIST	GAMMAINV*
Hyper geometric	7	4	HYPGEOMDIST*		
Log Normal	8	3		LOGNORMDIST*	LOGINV^
Negative Binomial	9	3	NEGBINOMDIST*		
Normal (with parameters)	10	3	NORMDIST*	NORMDIST*	NORMINV^

DISTRIBUTION	REFERENCE	NUMBER OF PARAMETERS	DENSITY	CUMMULATIVE	INVERSE
Normal (z values)	11	1		NORMSDIST*	NORMSINV^
Poisson	12	2	POISSON*	POISSON*	
T	13	2		TDIST	TINV*
Weibull	14	3	WEIBULL	WEIBULL	

Most of these functions remain unchanged from the Excel 97 version to the Excel 2007 version. Some were changed. A * after the function, indicates that the function was internally changed for the Excel 2003 version. A ^ indicates that the function was changed for the Excel 2002 version. There were no changes to NORMDIST and NORMINV, but these functions call NORMSDIST and/or NORMSINV, which were changed. The changes are all described in Microsoft's Knowledge Base Articles, freely accessible from the Internet. There was no KBA or any indication that there was a change in any of the above functions for Excel 2007. Note C is an index to relevant Knowledge Base Articles, by topic.

EQUIVALENCES TO OTHER CUMULATIVE STATISTICAL DISTRIBUTIONS:

There are theoretical equivalences between probability distribution functions. These equivalences can be used to obtain values of functions not explicitly given in Excel. For example the cumulative negative binomial can be calculated from BINOMDIST. The beta density can be obtained from BINOMDIST if the alpha and beta values are integers. The following is a list of some of cumulative distribution relationships.

Provided by Jerry Lewis (2004).

$$\text{BetaDist}(x,a,b) = 1 - \text{BetaDist}(1-x,b,a)$$

$$\text{BinomDist}(x,n,p,\text{TRUE}) = 1 - \text{BetaDist}(p,x+1,n-x)$$

$$\text{ChiDist}(x,n) = 1 - \text{GammaDist}(x,n/2,2,\text{TRUE})$$

$$\text{ExponDist}(x,u,c) = \text{Weibull}(x,1,1/u,c) \text{ \{c can be either TRUE or FALSE\}}$$

$$\text{ExponDist}(x,u,c) = \text{GammaDist}(x,1,1/u,c) \text{ \{c can be either TRUE or FALSE\}}$$

$$\text{ErfC}(x) = \text{ChiDist}(2*x^2,1) \text{ \{ChiDist is much more accurate than ErfC for large x\}}$$

$$\text{Erf}(x) = \text{GammaDist}(x^2,1/2,1,\text{TRUE})$$

$$\text{FDist}(x,n1,n2) = \text{BetaDist}(n2/(n2+n1*x),n2/2,n1/2)$$

$$\text{LognormDist}(x,m,s) = \text{NormDist}(\ln(x),m,s,\text{TRUE})$$

$$\text{NegbinomDist}(f,s,p) = \text{BinomDist}(s,s+f,p,\text{FALSE})*s/(s+f)$$

$$\text{NegbinomDist}(f,s,p) = \text{BetaDist}(p,s,f+1) - \text{BetaDist}(p,s,f)$$

$$\text{Negbinom cdf}(f,s,p) = \text{BetaDist}(p,s,f+1)$$

$$\text{NormSDist}(-|x|) = \text{ChiDist}(x^2,1)/2 \text{ \{ChiDist is much more accurate than NormSDist for large x prior to 2003\}}$$

$$\text{Poisson}(x,u,\text{TRUE}) = 1 - \text{GammaDist}(u,x+1,1,\text{TRUE})$$

$$\text{Poisson}(x,u,\text{TRUE}) = \text{ChiDist}(2*u,2*(x+1))$$

$$\text{TDist}(x,n,2) = \text{FDist}(x^2,1,n)$$

$TDist(x,1,2) = 2 \cdot \text{atan}(1/x)/\pi$
 $TDist(x,1,2) = 1 - 2 \cdot \text{atan}(x)/\pi$
 $TDist(x,2,2) = 1 - x/\text{Sqrt}(2+x^2)$
 $TDist(x,3,2) = 1 - 2 \cdot \text{atan}(x/\text{sqrt}(3))/\pi - 2 \cdot x \cdot \text{sqrt}(3)/(\pi \cdot (3+x^2))$
 $TDist(x,4,2) = 1 - x(6+x^2)/(4+x^2)^{(3/2)}$
 $Weibull(x,a,b,c) = \text{ExponDist}(x^a, b^{-a}, c)$ {c can be either TRUE or FALSE}
 $Weibull(x,a,b,c) = \text{GammaDist}(x^a, 1, b^a, c)$ {c can be either TRUE or FALSE}

Abramowitz (1964) provides some equivalences:

Hermite Polynomial, eqs 26.2.31 and 26.2.32 (uses normal densities (z))
 Hh Function, eqs 26.2.33 and 26.2.34
 Tetrachoric function, eq 26.2.35 (powers of normal density)
 Confluent Hypergeometric function, 26.2.36 to 26.2.39 (normal cumulative and density)
 Parabolic Cylinder Function, eq 26.2.40 (powers of normal density)
 Pearson's Incomplete gamma, eq 26.4.20 (Chi-square P value)
 Pearson's Type III, eq 26.4.22 (Chi-square P value)
 Generalized Laguerre Polynomials, eq 26.4.24 (Series of Chi-square terms)

From these you can derive opposite tail areas, inverses, binomial and Poisson confidence limits, etc. Note that the Beta and Gamma distributions are a basic core of many others.

INFORMATION ON THE BASIC EXCEL STATISTICAL DISTRIBUTIONS:

Help provides some information on these functions. It describes when the function is used, what the input parameters are, the allowable ranges of the input parameters, and under what conditions an error response occurs.

The Help subpart of Excel 2003 was improved over the Excel 2000 and earlier versions. Help for the 2003 and 2007 versions now give examples that are actual Excel worksheets. The search method is also easier to use, and easier to find out information. However the help articles remain essentially the same as in earlier versions. The supporting long index list, characteristic of Excel 2000 is not visible after or before a search for Excel 2003 and 2007.

The following table is a brief summary of what additional information help will provide on how the function develops a numerical result.

Table 14-2: Help Information Provided in Excel 2000, 2003 and 2007

Function	Distribution	Density Equation	Cumulative Equation	Information on how the cumulative is computed
BETADIST	1	No	No	No
BINOMDIST	2	Yes	Yes	Summing density terms
CHIDIST	3	No	No	No
EXPONDIST	4	Yes	Yes	By equation
FDIST	5	No	No	No
GAMMADIST	6	Yes	No	No

Function	Distribution	Density Equation	Cumulative Equation	Information on how the cumulative is computed
HYPGEOMDIST	7	Yes	No cumulative function	
LOGNORMDIST	8	No	Yes	From NORMSDIST
NEGBINOMDIST	9	Yes	No cumulative function	
NORMDIST	10	Yes	No	From NORMSDIST
NORMSDIST	11	Yes	No	No ¹
POISSON	12	Yes	Yes	Summing density terms
TDIST	13	No	No	No ²
WEIBULL	14	Yes	Yes	By equation

Microsoft issued a number of Knowledge Base Articles in the last quarter of 2003, describing the changes made to these distributions. Some of these KBAs describe the way the distribution is computed and what the limitations are with reference to accuracy. This useful information was not put into the help part of Excel.

Note C, is an index of KBA numbers related to these distributions. This can be helpful in finding further information on the provided distributions.

DENSITY FUNCTIONS AND EQUIVALENCES:

The discrete functions 2, 7, 9 and 12 have individual event probabilities that represent a density value for that event.

The five continuous functions 4, 6, 10, 11 & 14 have direct density equations that represent the differential of the cumulative function. The concept of an individual event probability for a continuous variable (real number) does not exist.

Where a density is needed, and the function listed in table 14-1 does not provide it, the density equation given in Help can be put in as a cell equation.

The six other distributions 1, 3, 5, 8, 11 & 13, do have density functions, but do not have explicit Excel functions for it. The densities of 3, 8 and 11 can be obtained as follows:

- 3 $\text{Chisq}(x,v) = \text{GAMMADIST}(x, v/2, 2, \text{FALSE})$
- 11 $\text{Normal}(z) = \text{NORMDIST}(z, 1, 0, \text{FALSE})$
- 8 $\text{Lognormal}(z) = \text{NORMDIST}(\text{LOG}(z), 1, 0, \text{FALSE})$

The remaining distributions 1, 5 and 13 have closely related density functions. Distributions 5 and 13 can be obtained from the beta (1) density. The binomial distribution (2) can under limited circumstances provide beta (1) densities. The limitation is that both a and b parameters in the beta distribution (1) have to be integers. T

¹ Tests indicate that Microsoft uses approximating polynomials.

² Tests on TDIST outputs indicate that it uses the BETADIST function internally, and truncates input fractional df values to integers.

distribution (13) and F distribution (5) densities can be calculated, from the beta (1) density but they only can be done for even degrees of freedom when the beta (1) density comes from (2). In theory the binomial distribution (2) exists for non-integer inputs, but Excel converts all the discrete distribution (2, 7 & 9) input parameters to integers (except the single event probability).

CUMULATIVE DISTRIBUTION FUNCTIONS AND P VALUES:

The cumulative distributions are either sums (discrete) or integrals (continuous) of the density (discrete event probability) functions. For distributions 4 & 14, exact equations representing the integrals exist. For distributions 2, 7, 9 & 12 sums of discrete event probability terms can be constructed to obtain cumulative values. The accuracy for these four distributions depends on the ability to get accurate discrete event probability values.

For distributions 1, 3, 5, 8, 10, 11 & 13 infinite series have to be used to obtain true cumulative values. There also exist sets of approximating functions that were constructed in the 1950s and 1960s as approximations, since there are problems with the series summations of terms of these distributions.

By convention, $P(x)$ represents the integral of the density function from minus infinity to some value. $P(x)$ is called the lower left tail area. $Q(x)$ represents the integral of the density function from x to plus infinity. $Q(x)$ is called the upper right tail area. By convention the normal and t distributions exist for both negative and positive x values. For all the other distributions, x can only be from zero to some positive value. With this convention then:

$$P(x) + Q(x) = 1$$

$$\text{and } P(x) = p$$

$$\text{and } Q(x) = q$$

This is sometimes confusing, since textbooks refer to the probability as a p value (alpha is a q value) when referring to the area of the upper tails. I will use the textbook conventions here and refer to any probability value as a p value, regardless if it is a lower or upper tail value. Specific references to p and q values imply that the p value is the area under the density function to the left of a reference statistic, and q is the area to the right of the statistic.

The chi-square test for variance is unusual, in that the largest variance is always put in the numerator. As a result the distribution is folded about the calculated test Chi-square value and only the area above the calculated Chi-square value (q) exists. Rarely is the area of the left-hand tail needed. The meaning of a left tail (when the chi value does not come from a ratio of variances) is somewhat in dispute, since it implies a “too perfect” occurrence.

CUMULATIVE DISTRIBUTION TAIL AREAS: **EXCEL FUNCTION TAIL AREAS**

Given the tail area conventions stated above, the following table describes what the actual outputs are of the statistical functions that return a probability value. The term

complement is the computation of one minus the probability value. The complement is always the least accurate value, since it is limited to no more than 15 decimal digits, including preceding nine's.

Table 14-3: Obtaining Distribution Tail Area Values

Function	P value	q value	Comments
BETADIST	Output	Complement output. Also can be done by interchanging a and b, and putting in x input = 1 - x.	
BINOMDIST	Output	Set #successes as #failure	Exchanging successes and failures
CHIDIST	Complement the output	Output	
EXPONDIST	Calculates q, returns 1 minus q.	Complement the output or directly calculate the value, using the equation shown in help.	
FDIST	Complement the output	Output	
GAMMADIST	Complement the output	Output	
HYPGEOMDIST	Output. Represents probability of event occurring.	Represents probability of event not occurring. Invert successes and failures.	Exchanging successes and failures
NEGBINOMDIST	Output. Represents probability of event occurring.	Represents probability of event not occurring. Invert successes and failures.	Exchanging successes and failures
NORMSDIST	Output	Change sign of z.	Function is symmetrical. Can input minus z values for small p values.
NORMDIST	Output	Complement output. Function will not take a negative standard deviation value.	
LOGNORMDIST	Output	Complement output. Function will not take a negative standard deviation value.	
TDIST	Complement output	Output	Function is symmetrical. Small p values can be obtained by symmetry.

INVERSE DISTRIBUTION FUNCTIONS:

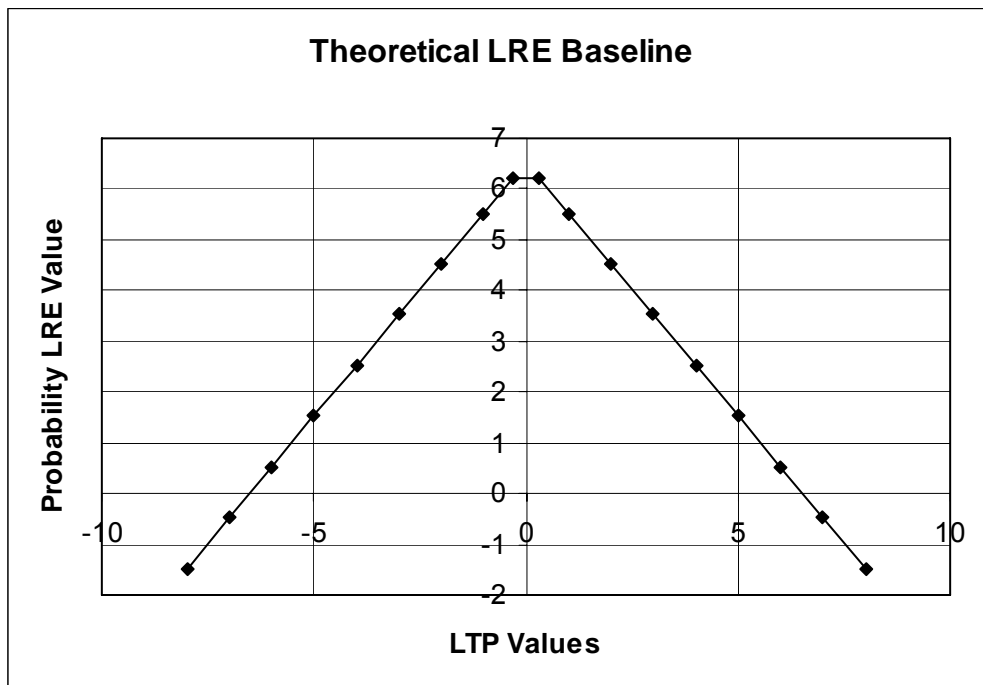
By convention, the inverse function is defined as a calculated value of x such that P(x) equals a preset or known value of p. In this case it is either p or q, whichever is the output of the DIST function. The sequence, {parameters > DIST function > p or q value > INV function > parameters} results in the same parameter set. The gap between the two parameter-sets is a measure of consistency.

The inverse functions for distributions 1, 3, 5, 6, 9, 10, 11 & 13 are obtained by iteration on the cumulative distribution until the computed p value differs from the input p value by a small value.

EXCEL 2000 AND EARLIER VERSIONS:

For distribution 1, Microsoft states that the iteration loop closure is 3E-07. This is an allowable error that limits the accuracy of these inverse functions. Assuming the error in the cumulative distributions is smaller than this (would have to be for computing stability), the actual LRE values of the computed parameter (an x value) would be close to the theoretical relationship shown in figure 14-1 (see section 11 for definitions of LRE and LTP). This figure actually is of the theoretical reflected p value, having values from 0 to 0.5 for p values from 0 to 0.5 and values from 0.5 to 0 for p values from 0.5 to 1. As each of the inverse functions are discussed, a similar chart showing actual LRE values will be shown. (See figure 14-16 in section 14-1 to see how this holds up with actual data.) In general there is considerable scatter, but the actual points for the Excel 2000 distributions appear to lie above or about the theoretical lines.

Figure 14-1: Theoretical Inverse Function Accuracy



If the differences were based on a relative difference (relative to the actual p value), then the theoretical baseline would be a horizontal line at the 6.1 LRE value. This simple change would have prevented a lot of criticism on Excel’s accuracies.

EXCEL 2002, 2003 AND 2007

Microsoft made a change to the inversion looping routine in NORMSDIST for Excel 2002. The change is described in KBA 826772. For Excel 2003 the change was made to all the other inverse distributions. The change was only to the allowable difference parameter value, from 3E-07 to an unknown smaller value. There was no change for the 2007 version.

REPORTED COMPLAINTS AND PROBLEMS:

The following is a summary table of complaints or problems with the above distributions for the pre Excel 2003 versions, giving the application, the problem and a fix or workaround. Statistical distributions are hard to evaluate for accuracy and it is very time consuming to do so. The easiest approach is to just pick a few points where the distributions have large errors and then totally condemn Excel's distributions as being unusable (Knüsel 1998c).

Table 14-4: Reported Faults

Application or Function	Problem	Source	Comments
Distribution Consistency Tests (Forward-Inverse) on Statistical Distributions	Systematic deviations, discontinuities and deviations near p=0 and p=1	CISE 27/99	Systemic deviations and discontinuities are really small in terms of LRE values. Some of the CISE report claims are not valid.
.....Accuracy of Excel Probability Functions.	Errors in computed results.	CISE 27/99, Knüsel 1998c	Valid.
Inverse distributions	Errors	RSS 1996	Valid.
Probability Distribution Tails	Failure to compute extreme tails of distribution functions	Cox	Valid.
Discrete Probability Distributions	Unable to handle very large number of cases.	Cox 2000	Valid. Fixed in Excel 2003.

These need to be viewed under the results of accuracy tests for each function as discussed below. In general one has to separate the performance of the direct cumulative distribution functions from the performance of the inverse cumulative distribution functions. Some of the direct cumulative distribution functions have very good tail performance and accuracy well above the accuracy of the corresponding inverse function.

PRIOR TEST RESULTS:

The CISE group of tests (CISE 27/99) were based on two methods:

“Forward-Inverse technique for evaluating the consistency: The method is shown for the Chi Square distribution:

$$\text{CHIINV}(\text{CHIDIST}(x1, df), df) = x2$$

If both functions are precise, then for a given df value, $x1 = x2$. However they are not precise, and there is a difference between $x1$ and $x2$. The difference provides a metric

which can be used to assess the accuracy of the pair. When further information is needed, then the two separate functions have to be separately tested.

Direct Comparison to a True Value for evaluating accuracy: For the individual distributions

P1 = CHIDIST(x1, df), the Excel function

P2 = function calculated by the IMSL function (IMSL Fortran 90 Math/Library (version 3.0) provided with Digital Visual Fortran (version 5.0 D Professional Edition) from Digital Equipment Corporation)".

The CISE report (CISE 27/99) only provided graphs of accuracies for some selected distributions, and only then for some selected parameter values. These graphs had a high “noise” content, because of the erratic termination of the forward-inverse method for increments in p values. There is also a lot of erratic behavior in a sequence of output values from a single function, making a plot of the values look like a lot of discontinuities in the computed values.

Good LRE values could not be developed from the CISE graphs.

The forward-inverse method could only test the pair of functions, and could not tell which one was bad. Since the distribution functions are used separately and not as pairs, the results reported by CISE were not useful.

The reported tests on statistical distribution accuracies by Knüsel and also by McCullough and Wilson are given in 14-5. Only the 20 functions and parameter values are listed that resulted in low LRE values out of 69 reported.. Criticisms on the functions were primarily situations where the function would not provide a value. There were no reported accuracies on the density functions when numerical values occurred.

Table 14-5 can be confusing since reported prior tests were all based on Knüsel’s DOS ELV program, which only gives p values to 6 decimal figures. As discussed in section 13, the ELV values are in some cases inaccurate.

Table 14-5: Summary of Reported Prior Tests on Excel 2003 Distributions

Row	Source	Distribution			
1		Poisson	K	Lambda	
2	Knüsel, (1998)	Poisson	100	200	
7	McCullough and Wilson(2002)	Poisson	0	200	
8	McCullough and Wilson(2002)	Poisson	10	200	
9	McCullough and Wilson(2002)	Poisson	50	200	
10	McCullough and Wilson(2002)	Poisson	100	200	
11	McCullough and Wilson(2002)	Poisson	103	200	
12	McCullough and Wilson(2002)	Poisson	104	200	
13	McCullough and Wilson(2002)	Poisson	110	200	
18		Binomial	K	n	p
19	Knüsel, (1998)	Binomial	400	1030	0.5
25	McCullough and Wilson(2002)	Binomial	10	1030	0.5
26	McCullough and Wilson(2002)	Binomial	50	1030	0.5

Row	Source	Distribution			
27	McCullough and Wilson(2002)	Binomial	100	1030	0.5
28	McCullough and Wilson(2002)	Binomial	200	1030	0.5
29	McCullough and Wilson(2002)	Binomial	390	1030	0.5
30	McCullough and Wilson(2002)	Binomial	391	1030	0.5
31	McCullough and Wilson(2002)	Binomial	400	1030	0.5
74		Inverse Beta	p	a	b
75	McCullough and Wilson(2002)	Inverse Beta	0.001	5	2
76	McCullough and Wilson(2002)	Inverse Beta	1E-06	5	2
77	McCullough and Wilson(2002)	Inverse Beta	0.001	10	100
78	McCullough and Wilson(2002)	Inverse Beta	1E-06	10	100

Table 14-5 Continued (matched by row number)

Row	Excel 2003	Reported Exact	Error, LRE	Smith	LRE on Reported Exact
2	0.00000E+00	3.72364E-15	0.00	3.72364E-15	6.49
7	0.00000E+00	1.38390E-87	0.00	1.38390E-87	5.60
8	0.00000E+00	4.10960E-71	0.00	4.10958E-71	5.44
9	0.00000E+00	6.81580E-37	0.00	6.81585E-37	5.16
10	0.00000E+00	3.72364E-15	0.00	3.72364E-15	6.49
11	0.00000E+00	2.89160E-14	0.00	2.89165E-14	4.79
12	2.72538E-14	5.61700E-14	0.29	5.61703E-14	5.28
13	2.45237E-12	2.48130E-12	1.93	2.48129E-12	5.25
19	3.86553E-13	3.89735E-13	2.09	3.89735E-13	6.20
25	0.00000E+00	3.11100E-287	0.00	3.11137E-287	3.93
26	0.00000E+00	3.94100E-225	0.00	3.94085E-225	4.41
27	0.00000E+00	1.39400E-169	0.00	1.39413E-169	4.02
28	0.00000E+00	5.45780E-92	0.00	5.45781E-92	5.79
29	0.00000E+00	3.18200E-15	0.00	3.18196E-15	4.94
30	2.05902E-15	5.24100E-15	0.22	5.24099E-15	5.59
31	3.86553E-13	3.89735E-13	2.09	3.89735E-13	6.20
75	1.81396E-01	1.81386E-01	4.24	1.81386E-01	6.13
76	4.29688E-02	4.44270E-02	1.48	4.44270E-02	5.97
77	2.79465E-02	2.79460E-02	4.77	2.79460E-02	6.35
78	1.17188E-02	1.21490E-02	1.45	1.21491E-02	5.03

As you can see, there were 18 LRE values less than 4 out of 69 reported test values. The values in the “LRE on Reported Exact” are consistent with the ELV reported 6 digits and the fact that ELV does not always give exact 6 digits.