An Exploration of the Exact Distribution and Probabilities for Sample Means

1. INTRODUCTION

One of the authors regularly teaches an upper division undergraduate two semester sequence in probability and statistics theory using the textbook Probability and Statistical Inference written by Hogg and Tanis (2010). The students are secondary mathematics majors and applied statistics undergraduate majors. When introducing the Central Limit Theorem (CLT) concept, the textbook uses the distribution of the sum of independent uniform random variables for samples sizes $n=2$ and $n=4$. These sampling distributions can be approximated using the normal distribution. Several graphs are displayed in an attempt to show the 'closeness' of the exact distributions to their normal-based approximations. It is explained that finding exact distributions can be very tedious (p. 258), hence the relatively simple-to-compute normal-based approximations.

After providing information on the CLT, probability and statistics textbooks commonly ask students to approximate answers to probability questions involving the sum or mean of independent samples drawn from probability density functions (pdf's). The approximations are based on parameter estimates, transformations, and probability tables.

The process generally stops here, depriving students of an additional learning opportunity. Students can still benefit from the process of evaluating integrals of exact pdf's to calculate probabilities. They can then determine the accuracy of the normal approximations. As normal-based methods are commonly used in statistical inference, students would greatly benefit from a better understanding of the limitations of the normal-based approximations.

The purpose of this article is to demonstrate how the inverse LaPlace transformation (ILT) can be used in Mathematica or similar computer algebra systems to calculate exact sampling distributions for means as well as the errors associated with using the analogous normal approximations.

It should be noted that students who take calculus at the University of Northern Colorado have lab assignments that use Mathematica commands. The students are given the commands and are asked to analyze the output. Similarly, students in the probability and statistics sequence are given the commands and asked to analyze the output of the approximation methods described in this article.

This methodology can be expanded to include exact sampling distributions of sums. Because these methods are demonstrated using computer software, they are less tedious and less prone to error than manual computations and allow extensive calculations that are often not manually tractable. This methodology can benefit classes of advanced undergraduate or master's level mathematical statistics courses at the level of Hogg, McKean, and Craig (2005) or Bain and Engelhardt (2000).

2. INVERSE LAPLACE TRANSFORM METHODOLOGY

Computing exact sampling distributions is largely absent from modern textbooks because of the complexity and tedium of doing so, as well as the widely-accepted use of normal approximations. Computer algebra systems (CAS) like Mathematica can be used to automate the hand calculations to relieve some of the tedium.

The following three definitions will be used throughout the paper:

Definition 1. Moment Generating Function (MGF), $M_X(t)$, for a continuous random variable X with pdf $f(x)$ with support space, S, is given by $M_X(t) = \int_S e^{tx} f(x) dx$ provided the integral exists for some interval $-h < t < h$ where $h > 0$.

Definition 2. The LaPlace transform for a function $F(t)$ is defined to be $L\{F(t)\} = \int_0^\infty e^{-\theta t} F(t) dt = L(\theta)$ provided the integral exists.

Definition 3. The inverse LaPlace transform (ILT) for a function $L(\theta)$, sometimes called the Bromwich integral, is given by the line integral $h(x) = \frac{1}{2\pi i} \lim_{x\to\infty} \int_{\alpha-ix}^{\alpha+ix} e^{x\theta} L(\theta) d\theta$. The following are sufficient conditions for the existence of the ILT: $h(x)$ is piecewise continuous and an exponentially-restricted function such that $L(h(t)) = L(\theta)$.

Rooney (1955) presents the mathematical background for the ILT process when working with real-valued functions. Existence conditions and properties of the inversion operator are given. Many common pdf's satisfy the existence conditions if (1) the MGF can be found and (2) the inversion process returns the initial pdf.

To use the ILT process in Mathematica to determine an exact sampling distribution of the mean of a random sample of size n taken from the given pdf $f(x)$, the steps below can be used. (For the reader's convenience, Mathematica code is in bold typeface.)

1. Assign the pdf to a function, say $f[x\_]$ with support space $S=(a,b)$.

2. Compute the MGF and assign it to **M[t_].**

   M[t_]:=Integrate[Exp[x t] f[x],{x,a,b}]

   The **M[t]** command can be used to view the MGF.

3. To compute the ILT, t is replaced with $-s/n$ in the function M[t_], thus converting the MGF into a Laplace transform. Taking this function to the power of n before invoking the ILT routine results in the exact distribution of the mean of a random sample of size n.

   ILT=InverseLaplaceTransform[M[-s/*n*]^*n*,s,x]

The MGF is taken to the *n*th power consistent with finding the MGF for the mean of a sample of independent observations of size *n*.

4. Because the distribution of the mean is a piecewise function, the ILT will have to be parsed accordingly. The following two lines of code operate on the ILT and parse out each polynomial piece for each domain segment of length 1/*n* using the HeavisideTheta function. The HeavisideTheta function in Mathematica is an indicator function for pdf's that have piecewise intervals. This allows the piecewise function to define the pdf for all interval domain segments in the output.

   ILT /. HeavisideTheta[e_] :> Piecewise[{{1,e > 0}}]

   PiecewiseExpand[%]

   The resulting output contains the exact sampling distribution, parsed over domain segments.

5. Construct a functional definition for each of the *n* polynomial pieces that appear in the output.

   gm0[x_], gm1[x_], ..., gm(n-1)[x_]

   For notation purposes, the exact distribution of the mean may be defined as the following:

$$gm[x] = \begin{cases} gmi[x], & i/n < x \le (i+1)/n, i = 0, 1, 2.., n-1 \\ 0, & otherwise \end{cases}$$

Appendix A.1 contains the equivalent code in the CAS, Maple for the ILT process described above.

Example: Consider the pdf $f(x) = 3x^2, 0 < x < 1, 0$ elsewhere. The parameters are $\mu = 0.75, \sigma^2 = 0.0375,$ and skewness $\gamma_1 \sim -0.86066.$

Consider the case where n=3 and the pdf of the mean of the random sample is to be found. The following list gives the steps in sequence to find that pdf:

1. Assign the pdf, $3x^2$ to *f*[*x*_].

   f[x_]:=3 x^2

2. Compute the MGF and assign it to M[t_].

   M[t_]:=Integrate[Exp[x t] f[x],{x,0,1}]

   Using the command **M[t]** the following output results.

$$\frac{3\left(-2 + e^t(2 - 2t + t^2)\right)}{t^3}$$

3. Invoke the ILT routine using the third power of the MGF

   ILT3=InverseLaplaceTransform[M[-s/3]^3,s,x]

4. The following two lines of code operate on ILT3 and parse out each polynomial piece.

ILT3 /. HeavisideTheta[e_] :> Piecewise[{{1,e > 0}}]

PiecewiseExpand[%]

The output is shown below as 8th degree polynomials defined over specific regions.

$$\begin{cases} \frac{59049x^8}{560} & x \le \frac{1}{3} \\ -\frac{81}{560}(7 - 120x + 840x^2 - 3024x^3 + 5670x^4 - 4536x^5 + 1458x^8) & \frac{1}{3} < x \le \frac{2}{3} \\ \frac{729}{560}(49 - 200x + 280x^2 - 336x^3 + 630x^4 - 504x^5 + 81x^8) & \frac{2}{3} < x \le 1 \\ 0 & \text{True} \end{cases}$$

5. Construct a functional definition for each of the *n* polynomial pieces that appear in the output.

   gm30[x_]:=(59049/560) x^8

   gm31[x_]:=(-81/560) (7 -120 x +840 x^2 -3024x^3 +5670 x^4 -4536 x^5 +1458 x^8)

   gm32[x_]:=(729/560) (49 -200 x +280 x^2 -336 x^3 +630 x^4 -504 x^5 +81 x^8)

   Once the functional definitions have been assigned, they can be used for finding mean, variance, probabilities, etc.

There are several validation checks of the ILT process that can be examined to demonstrate the accuracy of that process. Two such processes have been included in Appendices A.2 and A.3.

The exact distribution for the sample sum can be obtained directly by using the ILT process described above where t is replaced by $-s$ rather than $-s/n$ (see Appendix A.4 for more information).

The following graph, Figure 1, shows the exact distributions for the mean compared with the original pdf for three values of *n*: 3, 5, and 10. This, and all other plots in this article, was created using a series of basic **PLOT** commands in Mathematica. The pdfs were assigned to functions and integrated over the desired range to yield the probabilities. As can be seen in the graph, the curve associated with the largest sample size (n=10) is more symmetric and thereby more closely resembles a bell shape than the curve associated with the smallest sample size (n=3) as one might expect according to the CLT.

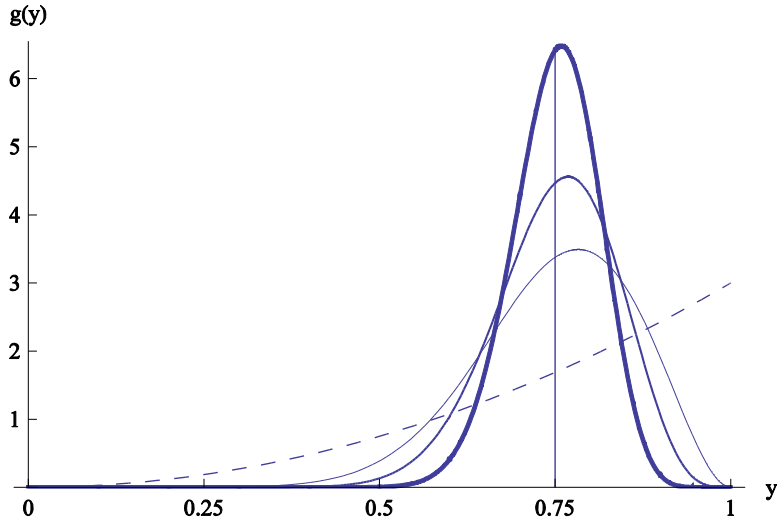

Figure 1: Sampling distribution of the mean, for samples of size *n*= 3, 5, 10 (in increasing order of thickness). The $3x^2$ pdf is dashed.

3. AN ACADEMIC EXERCISE

Given the capability of obtaining the exact distributions of the mean of random samples from the given pdf, consider the following academic question adapted from the CLT section of Hogg, McKean, and Craig (2005).

Compute an approximate probability that the mean of a random sample of size *n*=5 from a distribution having pdf $f(x)=3x^2, 0<x<1, 0$ *elsewhere*, is between 3/5 and 4/5.

The above question could be expanded as follows:

(1) Using the normal distribution (CLT), approximate the probability $P[3/5 < \bar{X} < 4/5]$.

(2) Determine the exact sampling distribution for the sample mean and calculate the exact probability $P[3/5 < \bar{X} < 4/5]$.

(3) Using the results of parts (1) and (2), compare the two probabilities.

The additional parts (1), (2), and (3) above could further be expanded to include additional sample sizes. Figure 2 below graphically depicts the academic question.

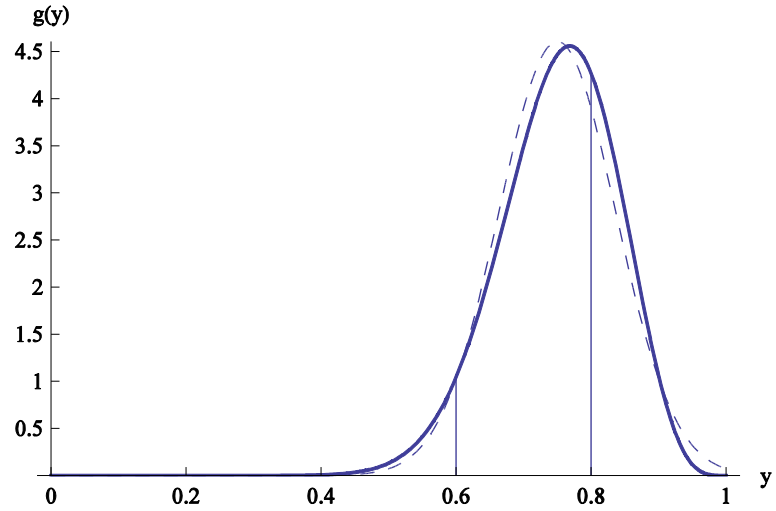

Figure 2: Sampling distribution of the mean for samples of size *n*=5 (*gm5* bold), with the corresponding normal *(nm5* dashed). The academic exercise limits of integration are identified.

The portion of the 14th degree polynomial for this distribution that is defined over the interval $3/5 < \bar{X} < 4/5$ includes *gm53[x]* and is coded as

gm53[x]:=(1/448448)(-482306682 + 5137065780 x - 23219070875 x^2 + 60276352500 x^3

- 110607997500 x^4 + 177101925000 x^5 - 247325203125 x^6 + 237290625000 x^7

- 129035156250 x^8 + 70382812500 x^9 - 87978515625 x^10 + 53320312500 x^11

- 4882812500 x^14)

The other portions of the distribution, *gm50[x]*, *gm51[x]*, *gm52[x]*, and *gm54[x]*, are omitted as they are not needed for calculations within the interval $3/5 < \bar{X} < 4/5$. Had the desired interval for calculations covered multiple portions of the distribution, those additional portions would need to be included in the calculations. See the calculation of the area in Appendix A.3 for an example of a calculation that covers multiple portions of the sampling distribution.

For part (1) of the academic exercise, assume the approximating normal has been defined as follows:

$\mu = m = 3/4$

$\frac{\sigma^2}{n} = v = (3/80)/5$

m:=3/4

v:=(3/80)/5

***nm5*[x_]:=(1/Sqrt[2 Pi v]) Exp[-0.5 (x-m)^2/v]**

The normal-based approximation is found by the following code:

approx_answer=NIntegrate[*nm5*[x],{x,3/5,4/5}]

The answer is 0.67652.

For part (2) of the academic exercise, use the *gm53[x]* portion of the sampling distribution of the mean, to compute the exact probability with the following code:

exact_answer=NIntegrate[*gm53*[x],{x,3/5,4/5}]

The answer is 0.64592.

To answer part (3) of the academic exercise, compare the two probabilities found above.

Let the normal approximation error (NAE) be defined as the normal approximation minus the exact answer found by the ILT process. For the distribution of the mean, the error for $P[a < \bar{X} < b],\ 0 < a < b < 1$ is found by $NAE = \int_a^b \big(nm(t) - gm(t)\big)dt$. Note that $\int_{-\infty}^{\infty} \big(nm(t) - gm(t)\big)dt = 0$.

For the $3x^2$ pdf, the NAE for *n*=5 is computed as follows:

$$NAE_5 = \int_{3/5}^{4/5} \big(nm5(t) - gm5(t)\big)dt = 0.67652 - 0.64592 = 0.03060$$

which accounts for the difference between the exact and approximate probabilities. That concludes the answers to parts (1), (2), and (3) of the academic exercise for *n*=5.

Next, consider the NAE's for the $P[3/5 < \bar{X} < 4/5]$ academic question calculated for samples of size 10, 15, 20, 25 and 30.

Table 1: Exact and normal approximation for $P[3/5 < \bar{X} < 4/5]$ and NAE's by sample size.

	***n*=5**	***n*=10**	***n*=15**	***n*=20**	***n*=25**	***n*=30**
Normal approximation	0.67652	0.78574	0.84000	0.87563	0.90159	0.92134
Exact probability	0.64592	0.77499	0.83770	0.87709	0.90479	0.92531
Normal approximation error (NAE)	0.03060	0.01075	0.00230	-.00146	-.00320	-.00397

Notice that for larger sample sizes, the NAE's become small quickly. The limit of the magnitude of the NAE's goes to zero, as required by the Central Limit Theorem. However, the magnitude of the NAE's, as sample sizes increase, may not go to zero monotonically, as is the case in line three of Table 1 above.

It is important to note that when using various CAS systems to perform integrations of the pdf for the sample mean, there can be convergence problems with the integral calculations due to the small interval domains of length 1/*n*. Such convergence problems can be avoided by performing the desired

calculations using the pdf of the sample sum, which always has unit interval domains, and then converting the resulting distribution function using the sample size. Appendix A.4 provides an example of computer coding for computing the distribution of the sample sum as well as demonstrating how to convert that distribution into the distribution of the sample mean, avoiding the potential convergence problems described above.

Given the results in Table 1, a natural extension of this exercise is to consider the maximum error incurred by using the Central Limit Theorem when calculating approximations for probabilities for the sample mean.

4. MAXIMUM ABSOLUTE NORMAL APPROXIMATION ERRORS

To examine the closeness of the normal approximation for probability statements for the sample mean, consider the difference between the normal approximation and the actual probability for the exact distribution of the sample mean. This process involves the following three steps:

1. For the given pdf and a sample size *n*, the cumulative error (CE) is defined for the sampling distributions of the mean as $CE(x) = \int_{-\infty}^{x} \big(nm(t) - gm(t)\big)dt,\ -\infty < x < \infty.$

   The approximating normal is defined over the interval $-\infty < x < \infty$ and the pdf is defined over the interval $0 < x < 1$. Consequently, the CE function is defined over the interval $-\infty < x < \infty$ and is applied over the appropriate domain region. Also note that $\lim_{x\to\infty} CE(x) = 0$.

   The ILT process results in points of intersection between the *gm5* and *nm5* functions at the approximate domain values *x*=0.07, 0.605, 0.764, and 0.904 as shown in Figures 3. These domain values will be used in subsequent calculations involving the CE function.

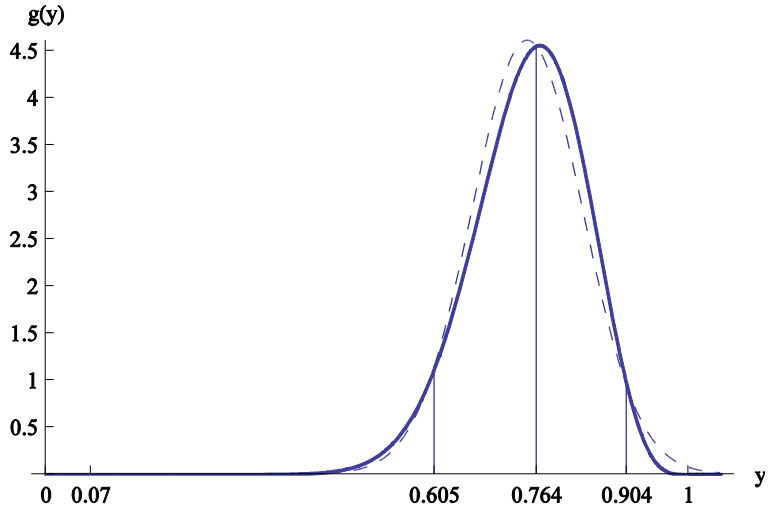

Figure 3: Sampling distribution of the mean (*gm5* bold) intersections with normal (*nm5* dashed). Domain values are approximate.

Consider the following CE graph constructed by calculating CE values across domain values in .005 increments with smaller increments in the vicinity of relative extrema. To produce the CE function, a list of $\{x,CE(x)\}$ is created and graphed with **ListLinePlot** for sample sizes n=5, 15, and 30. The CE relative extrema occur at the domain values where intersections occur (as shown in Figure 3).

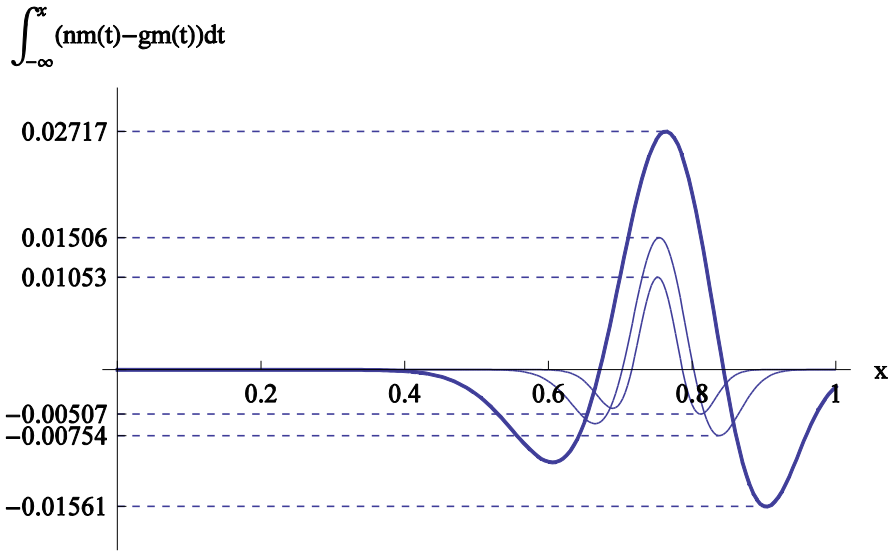

Figure 4: Cumulative error functions, n=5 (bold), 15, 30, with absolute extrema identified.

2. Next, consider the interval probability statement ($P[a < \bar{X} < b], 0 < a < b < 1$) encountered when working with the sample mean similar to the probability statement in the academic exercise in Section 3. Using the CE definition, a method for calculating the NAE for interval probabilities is developed. The NAE for sample size n is as follows:

$$NAE_n = \int_a^b \big(nm(t) - gm(t)\big)dt$$

$$= 0 + \int_a^b \big(nm(t) - gm(t)\big)dt$$

$$= \int_{-\infty}^0 nm(t)dt - \int_{-\infty}^0 nm(t)dt + \int_0^b \big(nm(t) - gm(t)\big)dt - \int_0^a \big(nm(t) - gm(t)\big)dt$$

$$= \int_{-\infty}^0 nm(t)dt + \int_0^b \big(nm(t) - gm(t)\big)dt - \int_{-\infty}^0 nm(t)dt - \int_0^a \big(nm(t) - gm(t)\big)dt$$

$= CE(b) - CE(a)$.

Thus, for any sample size $NAE\{P[a < \bar{X} < b]\} = CE(b) - CE(a)$.

3. It is possible to determine the worst possible case for accuracy when using the CLT and the normal distribution to approximate probabilities involving the sample mean. This case will be referred to as the Maximum Absolute Normal Approximation Error. The absolute value is used since the CLT may underestimate or overestimate the desired probability. Of interest is the magnitude of that error and that this maximum absolute NAE depends on sample size and skewness of the original distribution.

   For interval probabilities referring to Figure 4, $CE(0.764)$ is the absolute maximum of the CE function and $CE(0.904)$ is the absolute minimum. For the interval probability $P[a < \bar{X} < b] = P[0.764 < \bar{X} < 0.904] = CE(0.904) - CE(0.764) \cong -0.01561 - 0.02717 = -0.04278$ is calculated as the numerically largest NAE possible for $0 < a < b < 1$. The absolute value of this result (0.04278) is termed the maximum absolute NAE for interval probabilities, for the given pdf and the *n*=5 sampling situation.

   For larger samples of size *n*=10, 15, 20, 25, and 30, the maximum absolute NAE for interval probability statements are calculated in a similar manner as for *n*=5. The results are displayed in Table 2.

Table 2: Computed maximum absolute NAE, by sample size for $3x^2$ pdf.

	***n*=5**	***n*=10**	***n*=15**	***n*=20**	***n*=25**	***n*=30**
Maximum |NAE| for interval probabilities	0.04279	0.02830	0.02260	0.01934	0.01718	0.01560

As expected, maximum absolute NAE are smaller for larger sample sizes.

To further demonstrate this method, a brief comparison of the maximum absolute NAE for five other pdf's is presented. These additional pdf's have the absolute value of skewness, $|E[(X - \mu)^3]/\sigma^3|$, that range from 0 to 2. See Figure 5 below.

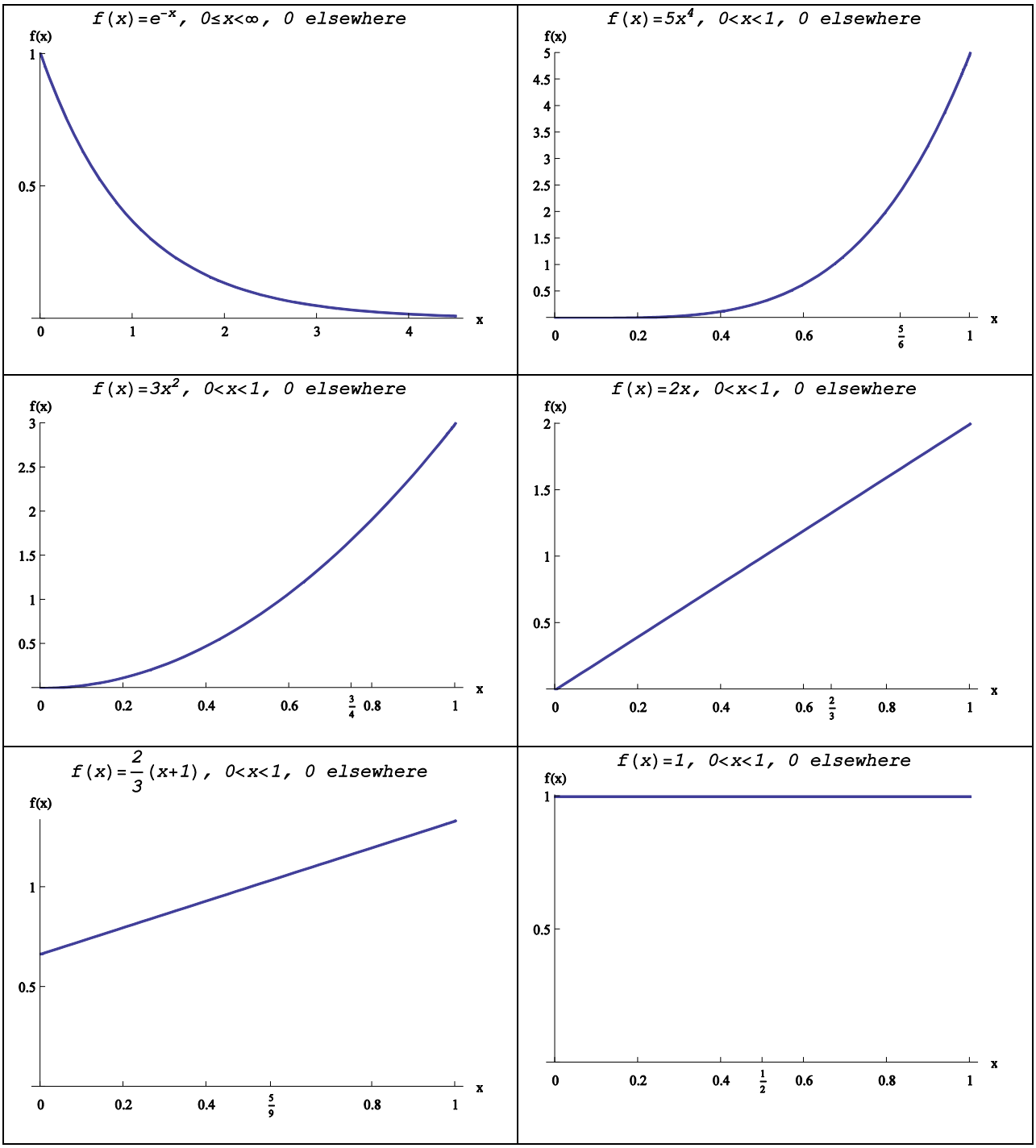

Figure 5: All pdf's, pictured by skewness (ordered left to right): 2, -1.183, -0.861, -0.566, -0.229, 0 (decimals approximate).

Table 3 below presents the maximum absolute NAE's for each pdf. The pdf used throughout this paper, $3x^2$, is also included in the table for comparison purposes. Also note that $5x^4$ values terminate at n=15 due to the complexity of the polynomials produced by the ILT process for larger sample sizes. The

computational complexity of the ILT process for a combination of high degree polynomials (>3) at larger sample sizes (>15) should be seen as a limitation to exclusive use of this method for all situations. It is rather suggested that this be used as a very effective teaching tool for better conceptual understanding of the exact sampling distributions vs. normal approximations using basic examples with relatively small sample sizes ($n \leq 15$).

Table 3. Computed maximum absolute NAE, by pdf and sample size.

	***n*=5**	***n*=10**	***n*=15**	***n*=20**	***n*=25**	***n*=30**
Exp	0.10013	0.06721	0.05374	0.04599	0.04081	0.03705
$\mathbf{5x^4}$	0.05956	0.03928	0.03134			
$\mathbf{3x^2}$	0.04279	0.02830	0.02260	0.01934	0.01718	0.01560
$\mathbf{2x}$	0.02839	0.01867	0.01488	0.01269	0.01130	0.01026
$\mathbf{\frac{2}{3}(x+1)}$	0.01491	0.00883	0.00671	0.00558	0.00488	0.00438
Uniform	0.01142	0.00562	0.00372	0.00278	0.00222	0.00185

The tabled values are graphed in Figure 6.

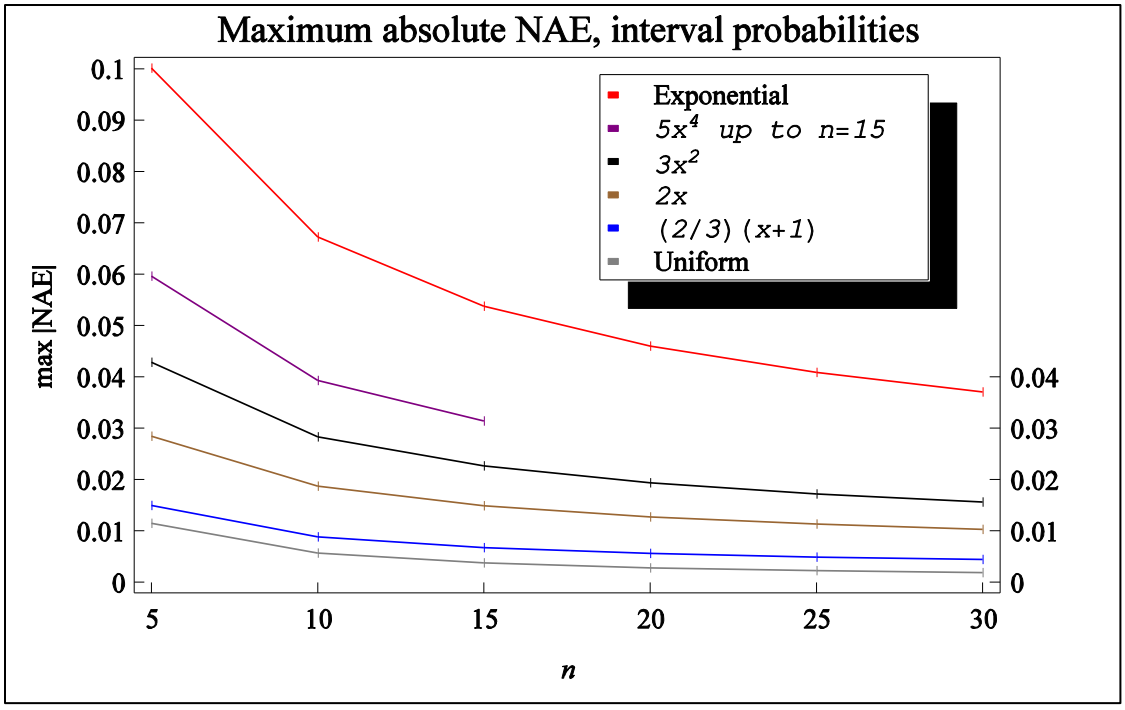

Figure 6: Maximum absolute NAE's for interval probabilities (by pdf).

Notice that the size of the maximum absolute NAE decreases with the magnitude of skewness for any sample size. Notice also that for all pdf's, the maximum absolute NAE decreases as sample size increases.

5. SUMMARY

In summary, when studying the Central Limit Theorem and normal approximations, students can benefit from the process of evaluating exact sampling distributions to find probabilities. This process can help students better understand the level of accuracy attained by using the normal approximation. This process can be facilitated using a CAS such as Mathematica. The academic exercise from Section 3 is one example as to how this process can be explored.

This paper demonstrates that students can determine exact answers to probability questions and, coupled with CLT approximations, find normal approximation errors for interval probability statements. More importantly, this paper presents the results of extensive computations (as seen in Table 2) to show the trend of maximum absolute NAE, by sample size, for the study case pdf, $f(x) = 3x^2$.

This paper also demonstrates that the techniques presented can be extended to other pdf's, selected for their differences in skewness, to show that absolute maximum NAE's vary by the absolute value of skewness, for various sample sizes (as shown in Table 3 and Figure 6).

6. REFERENCES

Bain, Lee. J. and Engelhardt, Max, (2000), Introduction to Probability and Mathematical Statistics (2nd edition), Boston, MA, PWS-Kent Publishing.

Belinfante, J. G. F. (2007), *"Random Samples from a Uniform Distribution,"* http://www..math.gatech.edu/~belinfan/3770su08/pdf/u-sample.nb.pdf.

Hogg, R. V., McKean, J. W., and Craig, A. T. (2005), Introduction to Mathematical Statistics (6th edition), Upper Saddle River, NJ: Prentice Hall.

Hogg, R.V. and Tanis, Elliot (2010), Probability and Statistical Inference (8th edition), Upper Saddle River, NJ: Prentice Hall.

Rooney, P. G. (1955), *"On an inversion formula for the Laplace transform,"* Canadian Journal of Mathematics, VII (1), pp. 101-115.

APPENDICES

A.1: Maple Code for the ILT Process

1. **f:= x→3 x^2**
2. **M:=t→integrate($e^{x\,t}$f(x), x=0..1)**
3. **_EnvUseHeavisideAsUnitStep:= true**
4. **inttrans[invlaplace] (M(-s/n)^n, s,x)**
5. **simplify (convert (inttrans[invlaplace] (M(-s/n)^n, s, x), piecewise))**

A.2: ILT Methodology Validation #1

The first validation check compares the exact sampling distributions for the mean of random samples of size *n*=2 or 3 derived by the ILT process, with the sampling distribution obtained by the mathematical statistics "pdf to pdf" (Jacobian) transformation process. Theoretically, this validation method extends to sample sizes larger than *n*=3. However, it quickly becomes intractable.

Steps:

(1) Define a transformation.

(2) Calculate the Jacobian.

(3) Form the joint distribution.

(4) Define the transformed space.

(5) Integrate out excess variables to yield the distribution of the mean.

(6) Compare the results found using the Jacobian process to the results found using the ILT process.

Example: For the sampling distribution of the mean of a sample of size *n*=3 from the $3x^2$ pdf, the ILT results derived in Section 2 are *gm30, gm31, and gm32*. These functions are defined over domain regions (0,1/3), (1/3,2/3), and (2/3,1), respectively.

Steps:

(1) Define a transformation: $Y = \frac{(X_1+X_2+X_3)}{3}, Y_2 = X_2, Y_3 = X_3$

(2) Calculate the Jacobian: answer is 3

(3) Form the joint distribution: $gnew(Y, Y_2, Y_3) = |J|(3(3Y - Y_2 - Y_3)^2\ 3(Y_2)^2\ 3(Y_3)^2)$

(4) Transformed space: bounded by $0 < Y_2 < 1,\ 0 < Y_3 < 1$, and two planes with equations below (see Figure A1)

Ytop[Y2_,Y3_]:=Y2/3 +Y3/3 + 1/3

Ybot[Y2_,Y3_]:=Y2/3 +Y3/3

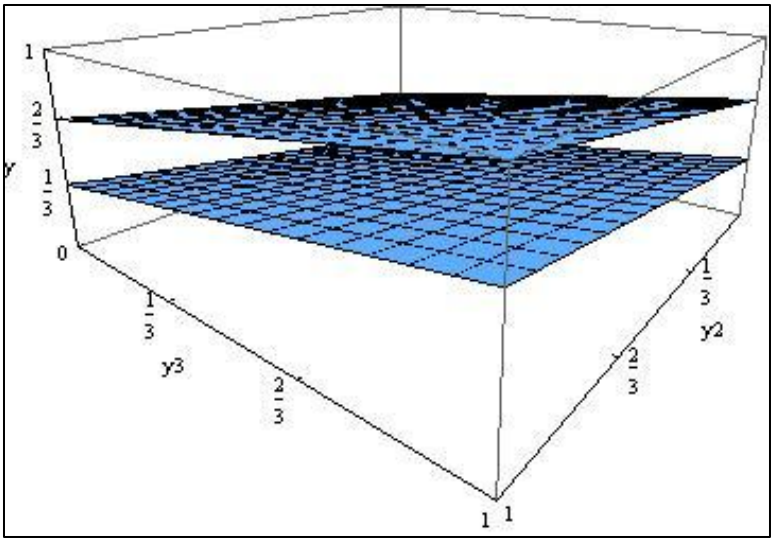

Figure A.1: Integration region for *n*=3 sampling. Transformation space is between the planes.

(5) Integrate out Y_2 and Y_3 resulting in the exact distribution of the mean.

Let $gnew(Y, Y_2, Y_3) = 3^4(3Y - Y_2 - Y_3)^2\ (Y_2)^2\ (Y_3)^2$.

$i0 = \int_0^{3Y} \int_0^{3Y-Y_3} gnew(Y, Y_2, Y_3) dY_2 dY_3, 0 < Y < \frac{1}{3}.$

$$i1 = \int_0^1 \int_0^1 gnew(Y, Y_2, Y_3) dY_2 dY_3 - \int_{3Y-1}^1 \int_{3Y-Y_3}^1 gnew(Y, Y_2, Y_3) dY_2 dY_3$$

$$- \int_0^{3Y-1} \int_0^{3Y-Y_3-1} gnew(Y, Y_2, Y_3) dY_2 dY_3, \frac{1}{3} < Y < \frac{2}{3}.$$

$i2 = \int_{3Y-2}^1 \int_{3Y-Y_3-1}^1 gnew(Y, Y_2, Y_3) dY_2 dY_3, \frac{2}{3} < Y < 1.$

The following code performs the integrations, with the pieces defined on the appropriate domain segments.

```
I0[Y_]=Integrate[gnew[Y,Y2,Y3],{Y3,0,3 Y},{Y2,0,3 Y-Y3}]
I1[Y_]=Integrate[gnew[Y,Y2,Y3],{Y3,0,1},{Y2,0,1}]
        -Integrate[gnew[Y,Y2,Y3],{Y3,3 Y-1,1},{Y2,3 Y-Y3,1}]
        -Integrate[gnew[Y,Y2,Y3],{Y3,0,3 Y-1},{Y2,0,3 Y-Y3-1}]
I2[Y_]=Integrate[gnew[Y,Y2,Y3],{Y3,3 Y-2,1},{Y2,3 Y-Y3-1,1}]
```

(6) Compare the results: The *gm30*, *gm31*, and *gm32* functions found by the ILT process are compared with *I0*, *I1*, and *I2* found by the Jacobian process. The two results are identical demonstrating the equality between exact sampling distributions derived by two very different processes. Similar calculations are used for the *n*=2 sampling situation also yielding equal results.

A.3: ILT Methodology Validation #2

The second ILT validation involves computing the area, mean, and variance of the sampling distributions produced by the ILT process.

Consider the ILT process as described in Section 2. Part 5 of that process constructs a functional definition for each of the *n* polynomial pieces.

gm0[x_], gm1[x_], …, gm(n-1)[x_]

For notation purposes, the exact distribution of the mean may be defined as the following:

$$gm[x] = \begin{cases} gmi[x], & i/n < x \leq (i+1)/n, i = 0, 1, 2.., n-1 \\ 0, & otherwise \end{cases}$$

Area: The area is computed for the distribution obtained using the ILT process by integrating each polynomial piece over its entire domain and summing the results. The result of the sum should equal one. The following code gives an example of what the code might look like for the first three polynomial pieces.

```
area=NIntegrate[gm0[x],{x,0,1/n}]
     +NIntegrate[gm1[x],{x,1/n,2/n}]
     +NIntegrate[gm2[x],{x,2/n,3/n}]
```

Mean: The mean is computed for the distribution obtained using the ILT process to ensure that it equals its theoretical mean by integrating the product of *x* and the polynomial piece $gmi[x\_]$ over the entire domain of that piece and then summing the results. The following code gives an example of what the code might look like for the first three polynomial pieces.

```
mean=NIntegrate[x gm0[x],{x,0,1/n}]
     +NIntegrate[x gm1[x],{x,1/n,2/n}]
     +NIntegrate[x gm2[x],{x,2/n,3/n}]
```

Variance: The variance is computed for the distribution obtained using the ILT process to ensure that it equals its theoretical variance by integrating the function $(x-\mu)^2 gmi[x\_]$ over the entire domain of that piece and then summing the results. The following code gives an example of what the code might look like for the first three polynomial pieces.

```
variance=NIntegrate[(x-mean)^2 gm0[x],{x,0,1/n}]
         +NIntegrate[(x-mean)^2 gm1[x],{x,1/n,2/n}]
         +NIntegrate[(x-mean)^2 gm2[x],{x,2/n,3/n}]
```

Example: For the sampling distribution of the mean of a sample of size n=3 from the $3x^2$ pdf, the ILT results derived in Section 2 are *gm30, gm31, and gm32*. These functions are defined over domain regions (0,1/3), (1/3,2/3), and (2/3,1), respectively.

The computational checks are then performed. First, the area is computed to ensure that it equals one.

```
area=NIntegrate[gm30[x],{x,0,1/3}]
     +NIntegrate[gm31[x],{x,1/3,2/3}]
     +NIntegrate[gm32[x],{x,2/3,1}]
```

The resulting area is one, exactly. Next, the mean is computed to ensure that it equals 3/4 =0 .75.

```
mean=NIntegrate[x gm30[x],{x,0,1/3}]
      +NIntegrate[x gm31[x],{x,1/3,2/3}]
      +NIntegrate[x gm32[x],{x,2/3,1}]
```

The resulting mean is 0.75, exactly. Next, the variance is computed to ensure that it equals (3/80)/3=0.0125.

```
variance=NIntegrate[(x-(3/4))^2 gm30[x],{x,0,1/3}]
         +NIntegrate[(x-(3/4))^2 gm31[x],{x,1/3,2/3}]
         +NIntegrate[(x-(3/4))^2 gm32[x],{x,2/3,1}]
```

The resulting variance is 0.0125, exactly.

A.4: Large Sample Means Methodology

As mentioned in Section 3, convergence problems can occur when using the ILT process to find the distribution for a sample mean with large sample sizes due to the small interval domains of length 1/*n*. To avoid these convergence problems, find the distribution of the sample sum and then convert it to the distribution of the sample mean. The steps are very similar to those described in Section 2. Replace Steps 3-5 from Section 2 with the following Steps 3-5.

3. To compute the ILT, *t* is replaced with –*s* in the function M[t_], thus converting the MGF into a Laplace transform. Taking this function to the power of *n* before invoking the ILT routine results in the exact distribution of the sum of a random sample of size *n*.

   ILT=InverseLaplaceTransform[M[-s]^*n*,s,x]

   where the MGF is taken to the *n*th power consistent with finding the MGF for the sum of a sample of size *n*.

4. Because the distribution of the sum is a piecewise function, the ILT will have to be parsed accordingly.

ILT /. HeavisideTheta[e_] :> Piecewise[{{1,e > 0}}]

PiecewiseExpand[%]

The resulting output contains the exact sampling distribution, parsed over domain segments.

5. Construct a functional definition for each of the *n* polynomial pieces that appear in the output.

   gs0[x_], gs1[x_], …, gs(n-1)[x_]

   For notation purposes, the exact distribution of the sum may be defined as the following:

$$gs[x] = \begin{cases} gsi[x], & i < x \le i+1, i = 0, 1, 2.., n-1 \\ 0, & otherwise \end{cases}$$

Next, use the transformation $x' = (1/n)x$. After the transformation, x' is replaced by *x*. The following computer code will accomplish the transformation from the distribution of sums to the distribution of means:

gm0[x_]:= n gs0[n x]

gm1[x_]:= n gs1[n x]

…

This process is continued for **gmi[x_]:= n gsi[n x]** for *n*=sample size and *i*=0, 1, …, *n*-1

The resulting distribution of the mean is the following.

$$gm[x] = \begin{cases} n\ gsi[n\ x], \frac{i}{n} < x \le \frac{(i+1)}{n}, & i = 0,1,2, \dots, n-1 \\ 0, otherwise \end{cases}$$

By performing the integration on the distribution of the sample sum first, and then converting the resulting distribution to that of a sample mean, integration using domains of length 1/*n* is avoided in favor of unit-length domains, eliminating the convergence problems that small domain-length integration might cause in CAS algorithms.