Seminar Report

The accuracy of Normal approximation by the Central Limit Theorem

Martin Westberg

Gustav Sundkvist Olsson

Nicolas Kuiper

Steven Lennartsson

Group 3

Probability, VT 2022

Course code: MMA306

Instructor: Ying Ni

Abstract

In this paper we will be looking at the *Central Limit Theorem* and investigate how big of a sample size is needed in order to get a decent Normal approximation. We start by providing background and definitions of necessary probability concepts, probability distributions and the Central Limit Theorem itself. We will then examine how big the absolute and relative errors are for three different probability distributions when varying the sample size. For this paper we have chosen Poisson-,Gamma- and Beta distributions and we have used the program *MATLAB* to help with varying the sample size and get visual representations.

Our primary source of reference is Wackerly *et alia* (2008) [1]. Appendix 1 presents the *MATLAB* code used.

Keywords: Central Limit Theorem, Normal distribution, Probability distributions: Poisson, Gamma, Beta

INDEX

1. Introduction

The Central Limit Theorem is a fundamental theorem of probability theory that we will be inspecting in this paper by implementing MATLAB code to conduct experiments. After presenting the necessary theory behind the concepts and parameters used, we will investigate what sample size gives a good approximation when using the Central Limit Theorem with different distributions. Combined with selected parameter values that make up visually distinct graphs, the idea is observing if this can affect what sample size is appropriate.

If we expect a certain average value from some probability distribution, and we make enough separate measurements of the average value by doing several experiments, we will observe that the measured average values will have a normal distribution around the center of the expected value. This is meaningful because if the distribution of the samples are unknown then the Central Limit Theorem can be used to do statistical analysis anyway. Different sample sizes are accurate or less accurate by our intentions, so it's important to know what sample size should be used to get a good enough approximation.

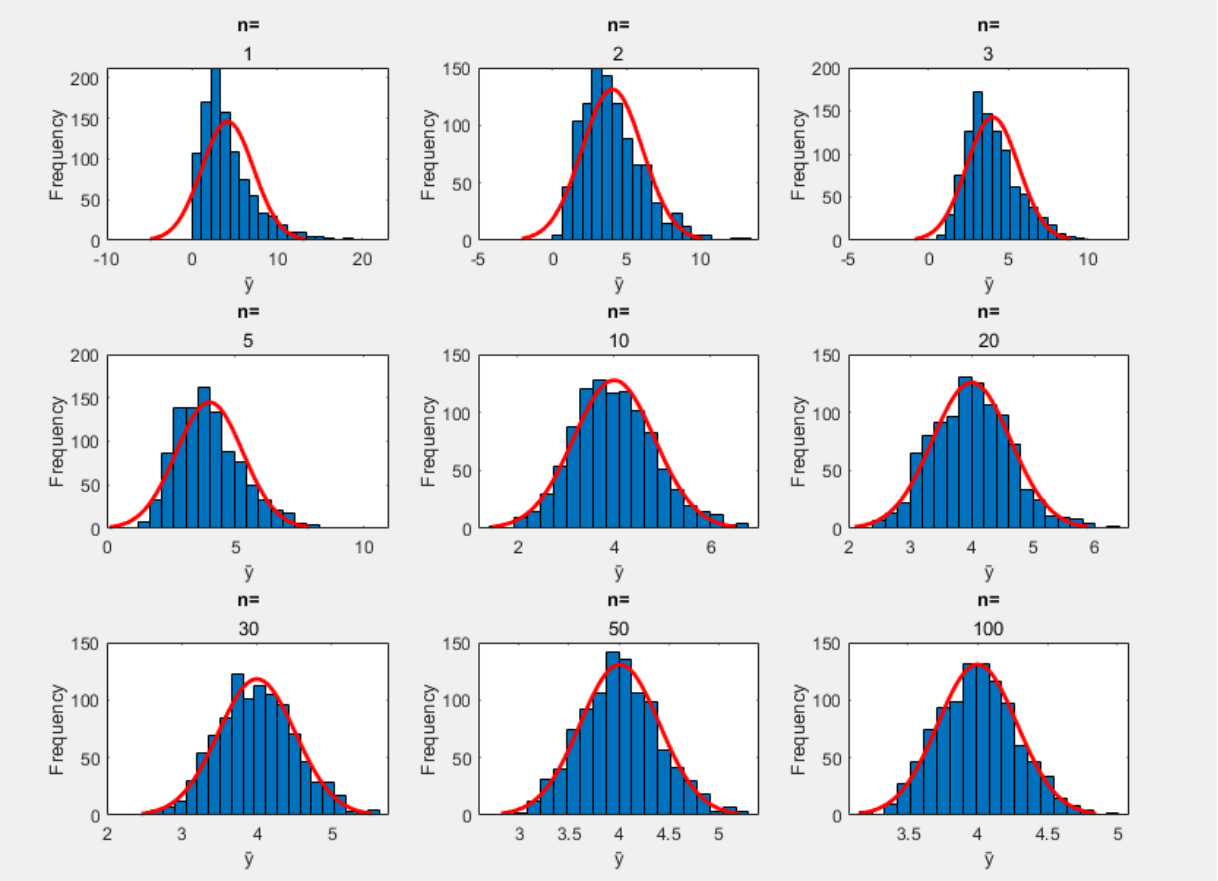


Figure 1: A visual example of how a Gamma distribution with $\alpha = 2, \beta = 2$ *can be approximated to a normal distribution when increasing sample size* n

2. Definitions and theorems

In this section we provide theorems and definitions of underlying probability concepts needed to understand the two central components of this paper: The Normal distribution and the Central Limit Theorem. We also give definitions of the three probability distributions used. The theorems and definitions are provided from [1] D. D. Wackerly, W. Mendenhall, and R. L. Scheaffer. *Mathematical Statistics with Applications, 7th Edition*. Thomson Learning, 2008.

Definition 1: Probability distribution function

Let Y denote any random variable. The (cumulative) *distribution function* of Y, denoted by $F(y)$, is defined as

$$F(y) = P(Y \leq y) \text{ for } -\infty < y < \infty$$

Definition 2: Normal distribution

A random variable Y has a *normal probability distribution* with parameters μ and σ iff (if and only if), for $\sigma > 0$ and $-\infty < \mu < \infty$, the density function of Y is:

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(y-\mu)^2/(2\sigma^2)} \text{ for } -\infty < y < \infty$$

Definition 3: Mean, Standard deviation, Variance

If Y is a normally distributed random variable with parameters μ and σ, then:

- *Expected value* or *mean* $E[Y] = \mu$
- *Variance* $V(Y) = E[Y^2] - E[Y] = \sigma^2$
- *Standard deviation* $\sigma = \sqrt{\sigma^2}$

Definition 4: Standardization

A normal distribution Y can be transformed to a standard normal random variable Z.
When Y is standardized, the mean becomes 0 and the standard deviation becomes 1. This is done by subtracting the mean from the original value and dividing the difference by the standard deviation. When standardization is used together with probability, the other variables also need to be standardized in the same way.

$$Z = \frac{x - \mu}{\sigma}$$

Definition 5: Poisson distribution

A random variable Y has a *Poisson distribution* with parameter $\lambda > 0$ iff the density function of Y is:

$$f(y) = \frac{\lambda^y}{y!} \cdot e^{-\lambda} \text{ for } y = 0, 1, 2, \ldots, \lambda > 0$$

And for any Poisson distribution:

$$E[Y] = \mu = \lambda \text{ and } V[Y] = \sigma^2 = \lambda$$

Definition 6: Gamma distribution

A random variable Y has a *Gamma distribution* with parameters $\alpha > 0$ and $\beta > 0$ iff the density function of Y is:

$$f(y) = \begin{cases} \frac{y^{\alpha-1} e^{-y/\beta}}{\beta^{\alpha}\Gamma(\alpha)}, 0 \le y < \infty \\ 0, \quad elsewhere \end{cases}$$

where the *Gamma function* $\Gamma(\alpha)$ is defined as:

$$\Gamma(\alpha) = \int_0^{\infty} y^{\alpha-1} e^{-y} dy$$

And for any Gamma distribution:

$$E[Y] = \mu = \alpha\beta \text{ and } V[Y] = \sigma^2 = \alpha\beta^2$$

Definition 7: Exponential distribution

A random variable Y has an *Exponential distribution* with parameter $\beta > 0$ iff the density function of Y is:

$$f(y) = \begin{cases} \frac{1}{\beta} \cdot e^{\frac{-y}{\beta}}, & 0 \le y < \infty \\ 0, & elsewhere \end{cases}$$

And for any exponential distribution:

$$\mu = E[Y] = \beta \text{ and } V[Y] = \sigma^2 = \beta^2$$

Definition 8: Beta distribution

A random variable Y has a *Beta probability distribution* with parameters $\alpha > 0$ and $\beta > 0$ iff the density function of Y is:

$$f(y) = \begin{cases} \frac{(1-y)^{\beta-1}}{B(\alpha,\beta)}, & 0 \le y < \infty \\ 0, & elsewhere \end{cases}$$

where the *Beta function* $\beta(\alpha, \beta)$ is defined as:

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

And for any Beta distribution:

$$E[Y] = \frac{\alpha}{\alpha + \beta} \text{ and } V[Y] = \sigma^2 = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

Theorem 1: Sample mean

Let $Y_1, Y_2, \ldots, Y_n$ be a random sample of size n from a normal distribution with mean μ and variance σ^2. The sample mean, $\overline{Y}$, is given by:

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

which is normally distributed with mean $\mu_{\overline{Y}} = \mu$ and variance $\sigma^2_{\overline{Y}} = \frac{\sigma^2}{n}$

Theorem 2: Central Limit Theorem

Let $Y_1, Y_2, \ldots, Y_n$ be independent and identically distributed random variables with

$E[Y_i] = \mu$ and $V[Y_i] = \sigma^2 < \infty$

We define:

$$U_n = \frac{\sum_{i=1}^{n} Y_i - n\mu}{\sigma\sqrt{n}} = \frac{\overline{Y} - \mu}{\frac{\sigma}{\sqrt{n}}} \text{ where } \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

Then the distribution function of U_n converges to the standard normal distribution function as $n \to \infty$. That is:

$$\lim_{n\to\infty} P(U_n \le u) = \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} e^{\frac{-t^2}{2}} dt \quad \text{for all } u.$$

3. Method

We have chosen to examine two different distributions for each type (Poisson, Gamma and Beta). The parameters we have chosen for each distribution are:

- $\lambda = 3$ for the first Poisson distribution and $\lambda = 10$ for the second one
- $\alpha = 2$ and $\beta = 2$ for the first Gamma distribution and $\alpha = 10$ and $\beta = 2$ for the second one
- $\alpha = 0.5$ and $\beta = 0.5$ for the first Beta distribution and $\alpha = 10$ and $\beta = 1$ for the second one

These parameters are chosen to have probability distributions that look distinctly different from each other as well as having some that don't resemble a normal distribution at all before the approximation. They look like the following:

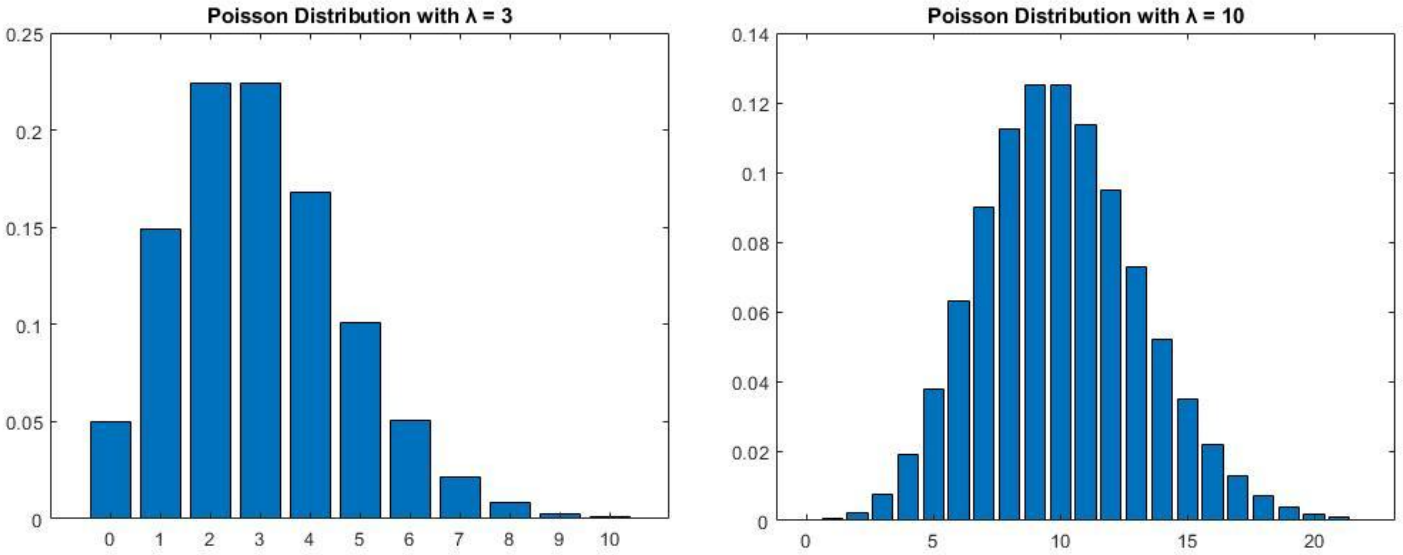
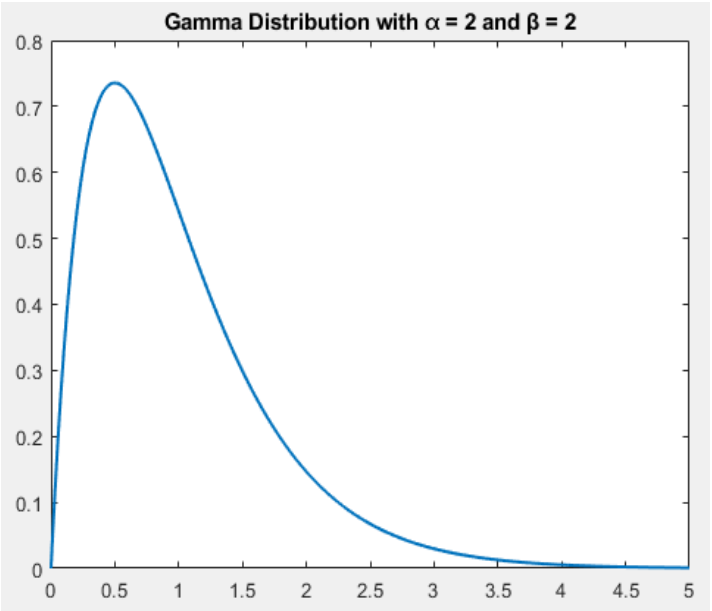


Figure 2: The two Poisson distributions before normal approximation



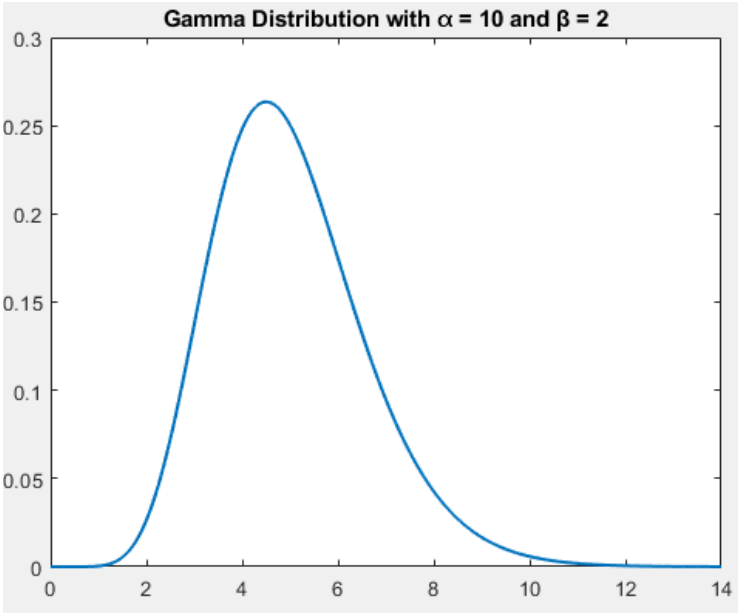
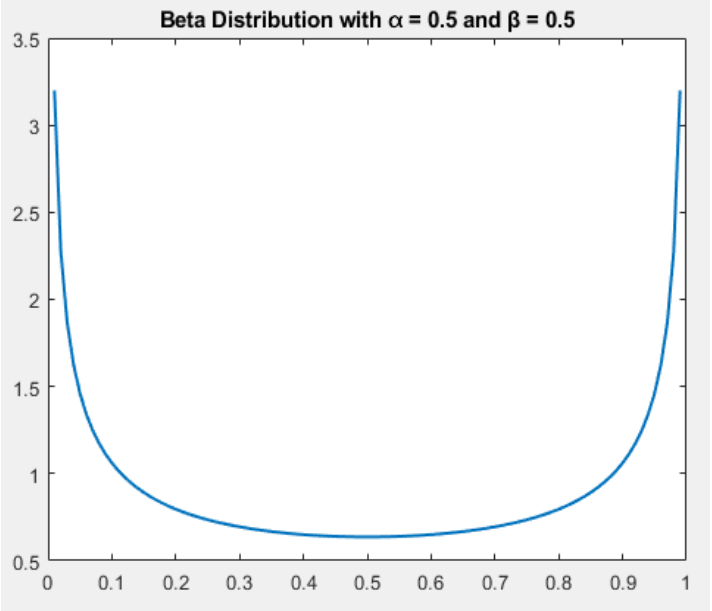


Figure3: The two Gamma distributions before normal approximation



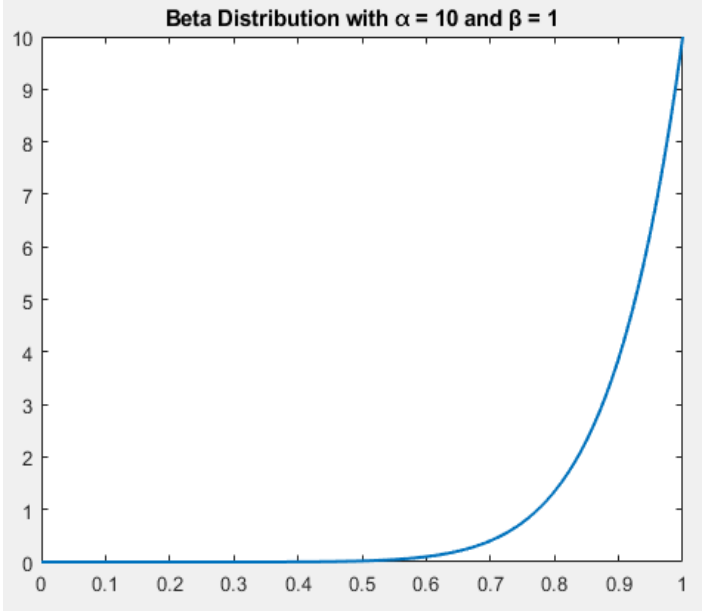


Figure 4: The two Beta distributions before normal distribution approximation

To measure the accuracy of the Central Limit Theorem normal approximation, we start by generating n samples of a distribution of choice Y_1 and take the mean of these samples. We store that value and redo this process many times, we have chosen to do it 1000 times. The sample size is now referred to as n and the set of means is referred to as the samples.

The new distribution Y_2, which is the normal approximation of these 1000 samples, will have mean μ and variance $\frac{\sigma^2}{n}$. By the Central Limit Theorem, it can be approximated by

$$Y_2 \sim N\left(\mu, \sqrt{\frac{\sigma^2}{n}}\right)$$

Now consider an interval $I = [a, b]$. The percentage of samples in this interval will be the number of samples in the interval divided by the number of samples i.e. 1000. If we know how many percentage points the approximation gives, we can then define the absolute error as

$$|approximation - Percentage\ of\ samples|$$

and the relative error as

$$\frac{|approximation - Percentage\ of\ samples|}{approximation}$$

We choose to divide by the approximation rather than the percentage of samples because sometimes the latter can be zero so this is more convenient.

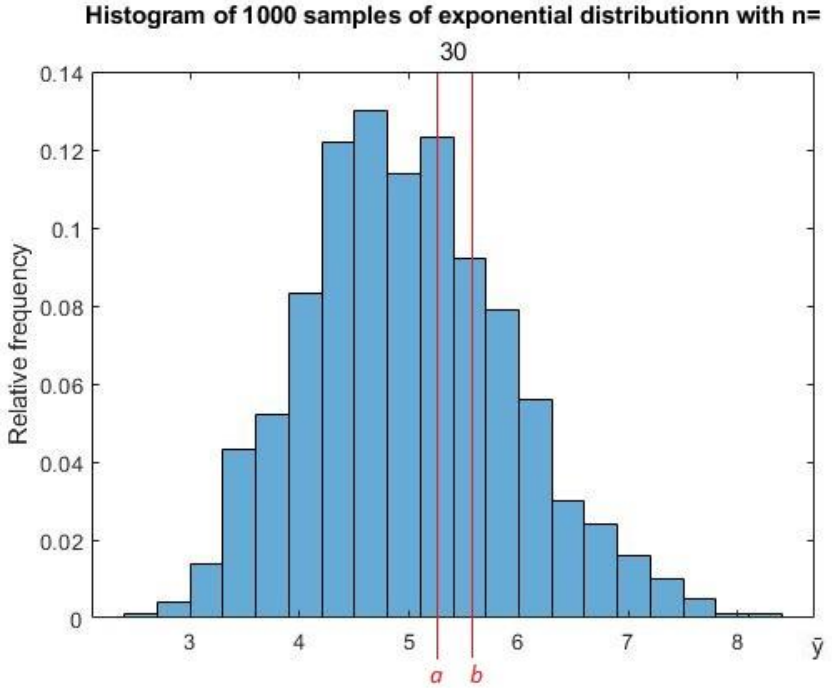


Figure 5: Histogram of samples of exponential distribution with interval [a,b] marked

To measure these errors, we will standardize the normal approximation. This is because we know which probabilities correspond to which intervals by looking at probability tables with these values. We can store these values beforehand and use them for any set of samples we want to look at. So the interval in the set of samples will be compared to an equivalent interval in the standard normal distribution. However, knowing what interval to use in the set of samples can be hard. This is because for any set of samples, the mean and the variance will be different since it depends on the mean and the variance of the distribution that the samples originally come from. It is easier to start with an interval on the standard normal distribution and calculate what is an equivalent interval in the set of samples.

For the following example we want to look at the top 33rd percentile in the normal approximation and compare that with the set of samples. The normal approximation follows

$$Y_2 \sim N\left(\mu, \sqrt{\frac{\sigma^2}{n}}\right)$$

For $P\left(Y_2 \geq y\right) = 0.33 \Rightarrow /standardize/$

$$\Rightarrow P\left(Z \geq \frac{y-\mu}{\sqrt{\frac{\sigma^2}{n}}}\right) = P(Z \geq 0.44) = 0.33 \Leftrightarrow \frac{y-\mu}{\sqrt{\frac{\sigma^2}{n}}} = 0.44$$

μ, σ^2, n are all known constants which means that we solve for y and get where on the $\bar{y}$ axis we should measure from. That is the percentage of samples with mean greater than y compared to the approximation 0.33.

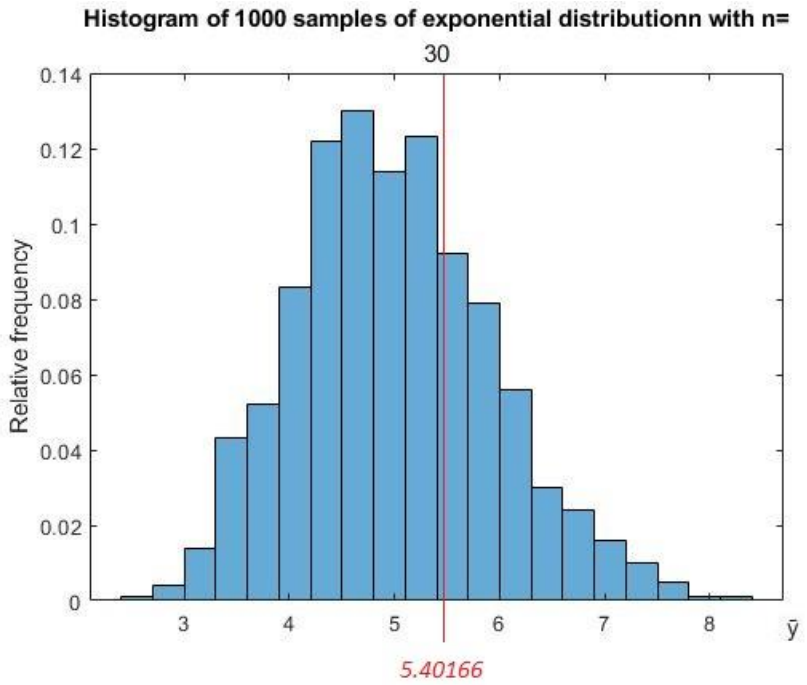


Figure 6: For an exponential distribution with mean 5, the 33rd percentile in the approximation is compared to the percentage of samples with mean greater than 5.40166

This process can be done for any interval and any sample size. We have chosen to look at 11 intervals:

I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8	I_9	I_{10}	I_{11}
[0,0.2]	[0.2,0.4]	[0.4,0.6]	[0.6,0.8]	[0.8,1]	[1,1.2]	[1.2,1.4]	[1.4,1.6]	[1.6,1.8]	[1.8,2]	[2,∞)
0.0793	0.0761	0.0703	0.0624	0.0532	0.0436	0.0343	0.0260	0.0189	0.0131	0.0228

Table 1: The first row is the notation for each interval, second row is each interval and third row is the probabilities for each interval for a standard normal distribution

We have made functions in MATLAB for each distribution we are looking at that takes the distributions parameters, number of samples that should be used, the sample size n, the intervals as arguments. It creates sample means for this distribution and calculates the absolute error when approximating the resulting distribution of those sample means with a normal approximation in the way we have described. This code can be found in appendix 1.

These errors will be different each time since the samples are taken from a random distribution. Therefore, if we only do this once then the results can be misleading. To get

more reliable results we instead do this 100 times and take the average of the errors for each interval. This is implemented in a MATLAB script that can also be found in appendix 1. Furthermore, this script calculates the relative error for each interval and in the end returns the average of those 100 relative errors.

4. Results

The following tables show the resulting average absolute and relative errors when approximating 100 Gamma distributions with $\alpha = 2, \beta = 2$. Each column is for different intervals, shown in table 1, and each row is for different sample sizes n. The last column shows the average error over all intervals for the corresponding sample size.

	I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8	I_9	I_{10}	I_{11}	AVERAGE
1	0.0110	0.0147	0.0192	0.0203	0.0175	0.0147	0.0111	0.0074	0.0046	0.0028	0.0235	0.0133
2	0.0076	0.0127	0.0141	0.0157	0.0141	0.0107	0.0088	0.0058	0.0040	0.0028	0.0184	0.0104
3	0.0074	0.0098	0.0122	0.0130	0.0119	0.0089	0.0083	0.0042	0.0038	0.0038	0.0179	0.0092
5	0.0068	0.0093	0.0098	0.0111	0.0096	0.0084	0.0064	0.0049	0.0034	0.0031	0.0139	0.0079
10	0.0067	0.0071	0.0086	0.0083	0.0085	0.0069	0.0063	0.0044	0.0037	0.0029	0.0102	0.0067
20	0.0073	0.0063	0.0075	0.0075	0.0068	0.0059	0.0050	0.0035	0.0039	0.0027	0.0083	0.0059
30	0.0080	0.0066	0.0066	0.0077	0.0067	0.0064	0.0052	0.0041	0.0031	0.0031	0.0069	0.0058
40	0.0074	0.0073	0.0073	0.0061	0.0059	0.0055	0.0051	0.0040	0.0032	0.0028	0.0067	0.0056
50	0.0067	0.0064	0.0067	0.0069	0.0059	0.0057	0.0052	0.0044	0.0032	0.0026	0.0056	0.0054
60	0.0074	0.0063	0.0080	0.0058	0.0061	0.0057	0.0042	0.0041	0.0035	0.0028	0.0052	0.0054
80	0.0064	0.0073	0.0078	0.0064	0.0057	0.0054	0.0045	0.0040	0.0039	0.0026	0.0053	0.0054
100	0.0059	0.0073	0.0065	0.0062	0.0055	0.0055	0.0039	0.0040	0.0033	0.0030	0.0045	0.0051

Table 2: Average absolute errors for each interval when approximating 100 Gamma distributions with $\alpha = 2, \beta = 2$

	I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8	I_9	I_{10}	I_{11}	AVERAGE
1	0.1390	0.1930	0.2727	0.3253	0.3281	0.3376	0.3241	0.2858	0.2418	0.2173	1.0289	0.3358
2	0.0953	0.1665	0.2001	0.2510	0.2655	0.2446	0.2574	0.2223	0.2092	0.2150	0.8057	0.2666
3	0.0938	0.1293	0.1739	0.2086	0.2228	0.2034	0.2426	0.1623	0.2014	0.2531	0.7846	0.2433
5	0.0862	0.1223	0.1394	0.1781	0.1800	0.1936	0.1876	0.1865	0.1823	0.2347	0.6101	0.2092
10	0.0848	0.0938	0.1220	0.1323	0.1602	0.1575	0.1850	0.1700	0.1976	0.2191	0.4457	0.1789
20	0.0926	0.0822	0.1071	0.1207	0.1282	0.1355	0.1444	0.1362	0.2057	0.2095	0.3632	0.1568
30	0.1004	0.0866	0.0938	0.1239	0.1250	0.1470	0.1514	0.1569	0.1617	0.2340	0.3009	0.1529
40	0.0932	0.0955	0.1037	0.0975	0.1109	0.1268	0.1475	0.1527	0.1703	0.2099	0.2929	0.1455
50	0.0840	0.0847	0.0958	0.1103	0.1115	0.1313	0.1507	0.1677	0.1618	0.1985	0.2461	0.1402
60	0.0937	0.0821	0.1145	0.0936	0.1144	0.1316	0.1220	0.1581	0.1857	0.2108	0.2296	0.1396
80	0.0813	0.0955	0.1104	0.1026	0.1069	0.1229	0.1324	0.1538	0.2070	0.2014	0.2319	0.1406
100	0.0749	0.0955	0.0926	0.0988	0.1043	0.1254	0.1134	0.1535	0.1745	0.2322	0.1979	0.1330

Table 3: Average relative errors for each interval when approximating 100 Gamma distributions with $\alpha = 2, \beta = 2$.

Here are similar tables when approximating 100 beta distributions with $\alpha = 0.5,\ \beta = 0.5$.

	I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8	I_9	I_{10}	I_{11}	AVERAGE
1	0.0344	0.0303	0.0225	0.0112	0.0081	0.0282	0.0987	0.0185	0.0189	0.0131	0.0228	0.0279
2	0.0177	0.0102	0.0141	0.0121	0.0082	0.0054	0.0058	0.0105	0.0146	0.0198	0.0228	0.0128
3	0.0110	0.0089	0.0068	0.0091	0.0064	0.0049	0.0047	0.0043	0.0044	0.0045	0.0034	0.0062
5	0.0067	0.0073	0.0063	0.0062	0.0057	0.0052	0.0049	0.0044	0.0040	0.0028	0.0036	0.0052
10	0.0069	0.0064	0.0058	0.0060	0.0055	0.0056	0.0048	0.0046	0.0033	0.0030	0.0039	0.0051
20	0.0061	0.0060	0.0056	0.0065	0.0055	0.0056	0.0046	0.0040	0.0035	0.0032	0.0038	0.0050
30	0.0068	0.0069	0.0065	0.0061	0.0060	0.0055	0.0046	0.0041	0.0034	0.0032	0.0040	0.0052
40	0.0061	0.0070	0.0063	0.0067	0.0051	0.0053	0.0048	0.0033	0.0035	0.0028	0.0042	0.0050
50	0.0067	0.0072	0.0067	0.0063	0.0057	0.0052	0.0041	0.0039	0.0034	0.0030	0.0038	0.0051
60	0.0064	0.0057	0.0057	0.0067	0.0051	0.0057	0.0044	0.0038	0.0032	0.0027	0.0041	0.0049
80	0.0067	0.0068	0.0062	0.0058	0.0061	0.0053	0.0038	0.0041	0.0035	0.0028	0.0042	0.0050
100	0.0073	0.0068	0.0066	0.0053	0.0057	0.0052	0.0047	0.0039	0.0034	0.0031	0.0036	0.0051

Table 4: Average absolute errors for each interval when approximating 100 Beta distributions with $\alpha = 0.5,\ \beta = 0.5$

	I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8	I_9	I_{10}	I_{11}	AVERAGE
1	0.4337	0.3982	0.3205	0.1803	0.1524	0.6472	2.8767	0.7112	1.0000	1.0000	1.0000	0.7927
2	0.2229	0.1335	0.2002	0.1946	0.1550	0.1244	0.1689	0.4019	0.7709	1.5076	1.0000	0.4436
3	0.1392	0.1167	0.0965	0.1459	0.1210	0.1120	0.1381	0.1654	0.2331	0.3447	0.1496	0.1602
5	0.0839	0.0966	0.0891	0.0992	0.1072	0.1197	0.1425	0.1708	0.2097	0.2157	0.1584	0.1357
10	0.0873	0.0844	0.0827	0.0964	0.1036	0.1291	0.1411	0.1777	0.1770	0.2261	0.1701	0.1341
20	0.0773	0.0782	0.0797	0.1046	0.1039	0.1282	0.1342	0.1527	0.1863	0.2479	0.1682	0.1328
30	0.0860	0.0904	0.0923	0.0970	0.1134	0.1257	0.1349	0.1577	0.1822	0.2423	0.1738	0.1360
40	0.0769	0.0924	0.0893	0.1079	0.0963	0.1206	0.1396	0.1258	0.1829	0.2124	0.1822	0.1296
50	0.0841	0.0949	0.0949	0.1016	0.1070	0.1182	0.1206	0.1500	0.1792	0.2256	0.1649	0.1310
60	0.0808	0.0745	0.0816	0.1068	0.0949	0.1311	0.1292	0.1442	0.1687	0.2067	0.1814	0.1273
80	0.0848	0.0891	0.0888	0.0936	0.1153	0.1216	0.1094	0.1588	0.1870	0.2111	0.1833	0.1312
100	0.0926	0.0898	0.0936	0.0850	0.1079	0.1183	0.1365	0.1504	0.1772	0.2400	0.1584	0.1318

Table 5: Average relative errors for each interval when approximating 100 Beta distributions with $\alpha = 0.5,\ \beta = 0.5$

The following tables show the average absolute and relative errors for three different distributions each with different parameters over all intervals. The intervals used are the same as in table 1. The rows are for different samples sizes n, specified in the first column.

	$Y \sim gam(2,2)$	$Y \sim gam(10,2)$	$Y \sim beta(0.5,0.5)$	$Y \sim beta(10,1)$	$Y \sim poi(3)$	$Y \sim poi(20)$
1	0.0133	0.0081	0.0279	0.0324	0.0551	0.0140
2	0.0104	0.0067	0.0128	0.0199	0.0451	0.0160
3	0.0092	0.0062	0.0062	0.0147	0.0338	0.0142
5	0.0079	0.0056	0.0052	0.0109	0.0167	0.0106
10	0.0067	0.0055	0.0051	0.0080	0.0076	0.0064
20	0.0059	0.0055	0.0050	0.0065	0.0141	0.0068
30	0.0058	0.0054	0.0052	0.0059	0.0084	0.0053
40	0.0056	0.0051	0.0050	0.0059	0.0072	0.0062
50	0.0054	0.0048	0.0051	0.0056	0.0098	0.0060
60	0.0054	0.0051	0.0049	0.0057	0.0076	0.0055
80	0.0054	0.0051	0.0050	0.0054	0.0068	0.0050
100	0.0051	0.0049	0.0051	0.0053	0.0075	0.0053

Table 6: Average absolute errors for different distributions.

	$Y \sim gam(2,2)$	$Y \sim gam(10,2)$	$Y \sim beta(0.5,0.5)$	$Y \sim beta(10,1)$	$Y \sim poi(3)$	$Y \sim poi(20)$
1	0.3358	0.2133	0.7927	0.8376	1.1495	0.3266
2	0.2666	0.1762	0.4436	0.5776	1.0018	0.3333
3	0.2433	0.1630	0.1602	0.4333	0.7698	0.3180
5	0.2092	0.1509	0.1357	0.3033	0.4236	0.2191
10	0.1789	0.1463	0.1341	0.2177	0.1870	0.1590
20	0.1568	0.1425	0.1328	0.1732	0.3181	0.1626
30	0.1529	0.1422	0.1360	0.1598	0.1984	0.1385
40	0.1455	0.1338	0.1296	0.1547	0.1796	0.1552
50	0.1402	0.1248	0.1310	0.1454	0.2201	0.1482
60	0.1396	0.1329	0.1273	0.1496	0.1871	0.1365
80	0.1406	0.1314	0.1312	0.1413	0.1598	0.1322
100	0.1330	0.1290	0.1318	0.1386	0.1726	0.1347

Table 7: Average relative errors for different distributions.

5. Discussion

The aim of this paper was to investigate what sample size is appropriate when approximating a distribution using the Central Limit Theorem. By looking at tables 6 and 7, we can see that this sample size is different depending on what distribution the samples come from.

A Beta distribution with $\alpha = 0.5, \beta = 0.5$ does not get much better already after $n = 5$, even if the sample size is increased to 100. But a beta distribution with $\alpha = 10, \beta = 1$ is much worse than the previous example at $n = 5$ and does not become better until large n's such as $n = 30$ or even higher.

For a Poisson distribution with $\lambda = 3$,we see that the average relative error is a lot higher than the others at n = 1, 2, 3 and 5 and plateaus at ca 17% rather than 13% within our chosen range of sample size n. For this specific distribution maybe we would need a n of 200, 500 or even 1000 to get a decent approximation with an average relative error of under 14% like all the other distributions. This is most likely the case because the Poisson distribution is discrete while Beta and Gamma are continuous.

When using the Central Limit Theorem in practice, this matters less because the distribution that the samples come from are often unknown. We observe that when $n = 20$ or when $n = 30$, the errors are quite small. A sample size of 20 is sufficient for many distributions that we have tested but overall, a sample size of 30 is a more safe option. According to some theory sample size has to be at least 30 to get a good approximation and this is very much in line with our findings. More importantly, the approximations do not get much better with higher sample sizes than these. There are diminishing marginal returns when $n \geq 30$ for all these distributions. Taking higher sample sizes will likely be inefficient since the benefit is so small.

By looking at table 3 we can see that the relative errors vary for different intervals such that the errors closer to the tail of the distribution are larger. This holds true for every distribution we have tested when $n \geq 5$. This is an indication that the Normal approximation is better when approximating intervals closer to the mean, at least if a lower relative error is wanted. When $n < 5$, this is not as consistent. For distributions that look very different from a Normal distribution, it seems like the absolute and relative error can vary in other ways when close to the tails. An example of this is shown in table 5.

6. Conclusion

As the theory suggests, we can have any distribution - discrete or continuous - and by picking samples with sample size n and taking the averages of the samples, letting n go towards infinity; if we then make a histogram with the frequencies of the averages, the means conform to a normal distribution. We don't have to get close to infinity to see this.

In the case of all distributions we have chosen, all of them except the Beta distribution with parameters (10,1) and the two Poisson distributions start to resemble a normal distribution and the average relative error starts to even out after a sample size of 10. The average relative error starts to even out for all distributions with a sample size of 30, and going above this number is inefficient due to diminishing returns.

Therefore, with our results we can conclude that the optimal and most safe choice of sample size for all distributions examined is 30.

The concept of *Central Limit Theorem* is embodied in Figure 1 where the Gamma distribution approaches the looks of a normal distribution. The presented conclusions, in conjunction with *Theorem 2: Central Limit Theorem*, has a useful quality - We can accurately predict the characteristics of populations. In finance, *CLT* is implemented in analysis, such as from a large industry sector taking samples of companies to estimate performance, when looking into every company may be too far fetched with a limited amount of time. The analyst can select 30 companies and confidently evaluate.

7. References

[1] D. D. Wackerly, W. Mendenhall, and R. L. Scheaffer. *Mathematical Statistics with Applications, 7th Edition*. Thomson Learning, 2008.

8. Appendix 1

The following are the MATLAB functions used to calculate the absolute errors when approximating sample means for specific distributions.

Gamma distribution:

```
function errors = cltgam2(input1,input2,numberofsamples,n,p1,p2)

options = optimset('Display','off');
[size1 size2]=size(p1);

for i=1:numberofsamples
    for k=1:n
        samplesofvariable(k)=random('gam',input1,input2);
    end
    averageofvariable(i)=mean(samplesofvariable);
end

for j=1:size2
    if j==size2
        f1=@(x) (((x-(input1*input2))/sqrt((input1*(input2^2))/n))-p1(j));
        x1=fsolve(f1,0,options);
        errors(j)=abs(p2(j)-(nnz(averageofvariable>=x1)/numberofsamples));
    else
     f1=@(x) (((x-(input1*input2))/sqrt((input1*(input2^2))/n))-p1(j));
        x1=fsolve(f1,0,options);
        f2=@(x) (((x-(input1*input2))/sqrt((input1*(input2^2))/n))-p1(j+1));
        x2=fsolve(f2,0,options);

errors(j)=abs((p2(j)-p2(j+1))-((nnz(averageofvariable>=x1)-nnz(averageofvaria
ble>=x2))/numberofsamples));
    end
end
```

Beta distribution:

```
function errors = cltbeta2(input1,input2,numberofsamples,n,p1,p2)

options = optimset('Display','off');
[size1 size2]=size(p1);

for i=1:numberofsamples
    for k=1:n
        samplesofvariable(k)=random('beta',input1,input2);
    end
    averageofvariable(i)=mean(samplesofvariable);
end

for j=1:size2
    if j==size2
        f1=@(x) (((x-(input1/(input1+input2)))/sqrt((input1*input2)/
          (((input1+input2)^2)*(input1+input2+1)*n)))-p1(j));
        x1=fsolve(f1,0,options);
        errors(j)=abs(p2(j)-(nnz(averageofvariable>=x1)/numberofsamples));
    else
        f1=@(x) (((x-(input1/(input1+input2)))/sqrt((input1*input2)/
          (((input1+input2)^2)*(input1+input2+1)*n)))-p1(j));
        x1=fsolve(f1,0,options);
        f2=@(x) (((x-(input1/(input1+input2)))/sqrt((input1*input2)/
          (((input1+input2)^2)*(input1+input2+1)*n)))-p1(j+1));
        x2=fsolve(f2,0,options);

errors(j)=abs((p2(j)-p2(j+1))-((nnz(averageofvariable>=x1)-nnz(averageofvaria
ble>=x2))/numberofsamples));
    end
end
```

Poisson distribution:

```
function errors = cltpoiss2(input1,numberofsamples,n,p1,p2)

options = optimset('Display','off');
[size1 size2]=size(p1);

for i=1:numberofsamples
    for k=1:n
        samplesofvariable(k)=random('poiss',input1);
    end
    averageofvariable(i)=mean(samplesofvariable);
end

for j=1:size2
    if j==size2
        f1=@(x) (((x-input1)/sqrt(input1/n))-p1(j));
        x1=fsolve(f1,0,options);
        errors(j)=abs(p2(j)-(nnz(averageofvariable>=x1)/numberofsamples));
    else
        f1=@(x) (((x-input1)/sqrt(input1/n))-p1(j));
        x1=fsolve(f1,0,options);
        f2=@(x) (((x-input1)/sqrt(input1/n))-p1(j+1));
        x2=fsolve(f2,0,options);

errors(j)=abs((p2(j)-p2(j+1))-((nnz(averageofvariable>=x1)-nnz(averageofvaria
ble>=x2))/numberofsamples));
    end
end
```

Next is the MATLAB script that runs these functions N number of times and then calculates the mean error for each interval as well as the mean over all the intervals:

SCRIPT

```
clear abserrors relerrors addedabserrors addedrelerrors totalabserror
absmeans relmeans
n=[1 2 3 5 10 20 30 40 50 60 80 100]; N=100;
[size1 size2]=size(n);
p1=[0 0.2 0.4 0.6 0.8 1 1.2 1.4 1.6 1.8 2];
p2=[0.5  0.4207  0.3446  0.2743  0.2119  0.1587  0.1151  0.0808  0.0548  0.0359
0.0228];
p3=[0.0793  0.0761  0.0703  0.0624  0.0532  0.0436  0.0343  0.0260  0.0189  0.0131
0.0228];
for j=1:N
    for k=1:size2
        abserrors(k,:)=cltgam2(2,2,1000,n(k),p1,p2);
      %NOTE: This line is changed if a different distribution should be used.
        relerrors(k,:)=abserrors(k,:)./p3;
    end
    if j==1
        addedabserrors=abserrors;
        addedrelerrors=relerrors;
    elseif j==N
        addedabserrors=addedabserrors+abserrors;
        addedrelerrors=addedrelerrors+relerrors;
        abserrors=addedabserrors./N;
        relerrors=addedrelerrors./N;
    else
        addedabserrors=addedabserrors+abserrors;
        addedrelerrors=addedrelerrors+relerrors;
    end
end
for k=1:size2
    totalabserror(k)=sum(abserrors(k,:));
    absmeans(k)=mean(abserrors(k,:));
    relmeans(k)=mean(relerrors(k,:));
end
abserrors=[n' abserrors totalabserror' absmeans']
relerrors=[n' relerrors relmeans']
```