LESSON: SAMPLING DISTRIBUTION OF $\bar{X}$
This lesson includes an overview of the subject, instructor notes, and example exercises using Minitab.

Sampling Distribution of $\bar{X}$

Lesson Overview

The sampling distribution of the sample mean, denoted $\bar{X}$, is a concept that is required to understand and relate to introductory statistical inference, which includes hypothesis testing and confidence intervals. If repeated random samples are chosen from the same population, the values of the sample mean, denoted $\bar{x}$, will vary from sample to sample. Note that these samples are typically drawn without replacement where each observation can be sampled from the population only once. The sampling distribution $\bar{X}$ is the distribution of these sample mean $\bar{x}$ values, for a large number of samples.

If the original population has a normal distribution with mean μ and standard deviation σ, then plotting the sample means for the simple random samples (SRS), each containing n observations, will produce a sampling distribution that also follows a normal distribution. If the original population does not have a normal distribution, but each SRS has a large n (where most texts suggest $n > 30$), then plotting the sample means will produce a sampling distribution that has an *approximate* normal distribution. This result is called the **Central Limit Theorem**.

Prerequisites

This lesson helps to connect descriptive statistics, which is introduced in the ***Describing Data Numerically*** and ***Describing Data Graphically*** lessons, to inferential statistics. The ***Normal Distribution*** lesson is required to understand this lesson. Also, the **Sampling** lesson explains the concept of obtaining a simple random sample.

Learning Targets

This lesson teaches students:

- To recognize that there is variability due to sampling—repeated random samples from the same population will produce variable results, such as variable $\bar{x}$'s.
- About the concept of a sampling distribution of a sample mean.
- That the sampling distribution of $\bar{X}$ is normally distributed if the original population being sampled from is normally distributed.
- The implication of the Central Limit Theorem. Namely, students will learn that the sampling distribution of $\bar{X}$ is *approximately* normally distributed if the original population being sampled from is non-normal, given the sample size *n* is large.
- That if the original population being sampled from has mean μ and standard deviation σ, then the sampling distribution of $\bar{X}$ has mean $\mu_{\bar{X}} = \mu$ and $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$.
- To determine probabilities associated with $\bar{X}$ using a standard normal probability table and Minitab.

Time Required

It will take the instructor 30 minutes in class to introduce the ***Sampling Distribution of*** $\bar{X}$ lesson and go through exercises by hand and in Minitab to help students grasp the concept of the normal shape of $\bar{X}$ and the Central Limit Theorem. We recommend starting the activity sheet in class so that students can ask the instructor questions while working on it. The exercises on the activity sheet will take 60 minutes, and can be used as homework or quiz problems.

Materials Required

- Minitab 17 or Minitab Express
- Internet access for simulation and data-generating activities.

Assessment

The activity sheet contains exercises for students to assess their understanding of the learning targets for this lesson.

Possible Extensions

The instructor will be ready to proceed with the ***Population Mean Hypothesis Testing for Large Samples*** and ***Population Mean Confidence Intervals for Large Samples*** lessons after finishing this lesson.

References

Bunnies, Dragons and the 'Normal' World: https://www.youtube.com/watch?v=jvoxEYmQHNM

Instructor Notes with Examples

The Sampling Distribution of a Sample Mean

- The sample mean $\bar{x}$ is a ***statistic*** whose value is the average of sample data drawn from a population.
- For random samples of size *n* taken from a given population, the random variable $\bar{X}$ is the collection of these sample means, the $\bar{x}$'s. Like any random variable, $\bar{X}$ has a probability distribution associated with it; i.e., shape, mean, standard deviation.
- The probability distribution created by plotting sample means, the $\bar{x}$'s, is the **sampling distribution of the mean $\bar{X}$**.
- The sampling distribution of $\bar{X}$ depends on the:
  - distribution of the original population (e.g., normal, skewed, uniform, symmetric)
  - sample size *n*
  - method of sample selection

Suppose we take numerous simple random samples of a given size *n* from a normal distribution with population mean μ and standard deviation σ. Then we compute the mean for each of those samples. Some of these sample means will be less than μ and some will be greater than it, thus giving us the sampling distribution.

If we plot the sample means using a histogram, we will see that they are normally distributed, where the mean and standard deviation of the sampling distribution $\bar{X}$ are:

$$\mu_{\bar{X}} = \mu \text{ and } \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Let $X_1, X_2, X_3, \ldots, X_n$ be independent normally distributed random variables with mean μ and standard deviation σ.

The distribution of the sample mean $\bar{X}$ is *exactly* normally distributed with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.

The larger the sample size *n*, the closer the distribution will converge about the true population mean μ.

Example 1

Consider a pack of *n* = 4 batteries. The battery lifetime is normally distributed with a mean of 10 hours and a standard deviation of 2 hours.

(a) If we randomly select one battery from the pack, what is the probability that the lifetime of this battery is less than 9 hours?

Let X represent the lifetime of one battery selected from the pack. Since X is normally distributed, we can determine its z-score and determine the desired probability with the standard normal table.

$$P(X < 9) = P\left(\frac{X - 10}{2} < \frac{9 - 10}{2}\right) = P(Z < -0.5) \cong \mathbf{0.30854}$$

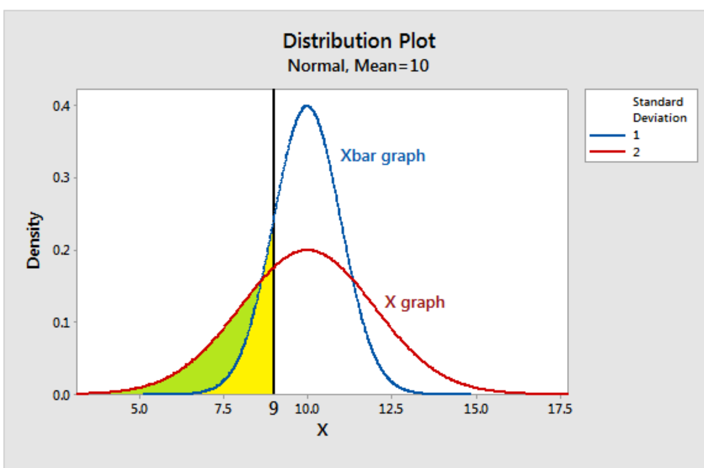
(b) What is the probability that the *average* lifetime $\bar{X}$ of a pack of four batteries is less than 9 hours? For example, one battery could last 9.5 hours, another 8.2 hours, another 11.6 hours, etc.

Let $\bar{X}$ represent the average lifetime of the pack of 4 batteries. Since X is normally distributed, then $\bar{X}$ is also normally distributed with mean μ = 10 and standard deviation $2/\sqrt{4}$ = 1.

$$P(\bar{X} < 9) = P\left(Z < \frac{9 - 10}{1}\right) = P(Z < -1) \cong \mathbf{0.15866}$$

The graph below shows the distributions of X and $\bar{X}$ for n = 4, where X represents the battery lifetimes and $\bar{X}$ represents the average lifetimes.

You can visibly see the difference in the probabilities by looking at the area under the curve to the left of the value 9. The individual observations are more variable than the means, and therefore the probability for the individual battery lifetime problem above (0.30854) is greater than the probability for the average lifetime (0.15866).



The Central Limit Theorem

Not all variables are normally distributed. In part, the normal distribution's popularity is due to its role in the **Central Limit Theorem (CLT).** Before giving a theoretical definition of this theorem, the following cartoon clip does a good job of displaying the CLT and what happens to averages of random samples as the sample size increases.

The video, *Bunnies, Dragons and the 'Normal' World*, is an animation that was displayed in the *New York Times* to illustrate the Central Limit Theorem:
https://www.youtube.com/watch?v=jvoxEYmQHNM

- The **Central Limit Theorem** is the heart of probability theory**.**
- The theorem states that the sampling distribution of $\bar{X}$ can be approximated by a **normal distribution** when the sample size *n* is "sufficiently large," irrespective of the shape of the original population distribution.
- As the sample size *n* increases, the corresponding distribution of the sample means will "converge" around the true population mean μ.

The symbolic explanation of the Central Limit Theorem is:

Let $X_1, X_2, X_3, \ldots, X_n$ be independent non-normal random variables with identical distributions with mean μ and standard deviation σ. If *n* is "large" enough (*n* > 30 suggested in most texts), then: The distribution of the sample mean $\bar{X}$ is *approximately* normally distributed with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$. The larger the sample size *n*, the more normally distributed the sampling distribution will be and the closer the distribution will converge about the true population mean μ.

The activity sheet will use simulations to help students derive these facts on their own so that they better understand the Central Limit Theorem. Although easily stated, the CLT is difficult to understand without graphics of $\bar{X}$ for increasing sample sizes *n*.
The following example considers averaging non-normal random variables for increasing sample sizes. Histograms will be used to display the shape of the distribution of $\bar{X}$.

Example 2

Let X_1 be a random variable representing the outcome of rolling a fair 6-sided dice. The random variable can take on the values *x* = 1, 2, 3, 4, 5, 6 with probability 1/6 for each.

To graph the distribution of X_1, we will use Minitab to sample 500 data points from the integer distribution *x* = 1, 2, 3, 4, 5, 6 with equally likely probabilities.

Minitab 17

1. Using a blank Minitab worksheet, choose **Calc > Random Data > Integer**.
2. In **Number of rows of data to generate,** enter *500***.**
3. In **Store in column(s)**, enter *C1*.
4. In **Minimum value**, enter *1*.
5. In **Maximum value**, enter *6*.
6. Click **OK.**

Minitab Express

1. Using a blank worksheet, open the generate random data dialog box.
    - Mac: **Data > Generate Random Data**
    - PC: **DATA > Random Data**
2. In **Number of columns to generate**, enter *1*.
3. In **Number of rows in each column**, enter *500*.
4. From **Distribution**, select **Integer**.
5. In **Minimum value**, enter *1*.
6. In **Maximum value**, enter *6*.
7. Click **OK**.

Column C1 now contains 500 randomly sampled values with equally likely probabilities chosen from *x* = 1, 2, 3, 4, 5, 6. Take a look at the column to make sure the distribution of values makes sense – just imagine these are the outcomes from someone rolling a die 500 times.

We can create a histogram for column C1 to see the shape of the distribution. We expect to see approximately the same heights for values *x* = 1, 2, 3, 4, 5, 6 around the value $\frac{1}{6} * 500 \cong 83.3$.

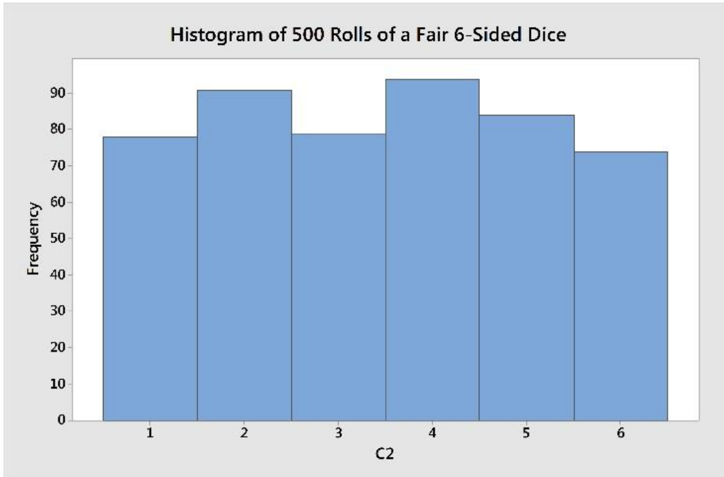
Minitab 17

1. Choose **Graph > Histogram**.
2. Choose **Simple**, then click **OK**.
3. In **Graph variables**, enter *C1*.
4. Click **OK.**

Minitab Express

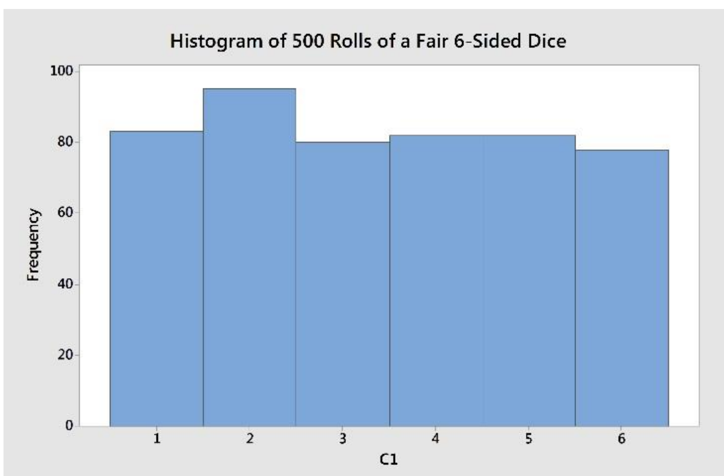
1. Open the histogram dialog box.
    - Mac: **Graphs > Histogram > Simple**
    - PC: **GRAPHS > Histogram > Simple**

2 In **Variables**, enter *C1*.
3 Click **OK.**



This graph is fairly flat with the value of 2 having a higher frequency than expected. This is not unusual for a random process, such as rolling a die. We can definitely see, though, that the distribution does not exhibit the shape of a normal curve.

In order to examine the shape of the average $\bar{X}$ of two die rolls, we'll create another 500 die rolls the column C2 in Minitab. Its histogram is:



The **Sampling Distribution for $\bar{X}$ when *n* = 2** and the initial distribution is discrete uniform for *x* = 1, 2, 3, 4, 5, 6.

Before creating a histogram for the average of these two rolls, we need to use Minitab to calculate the average of columns C1 and C2.

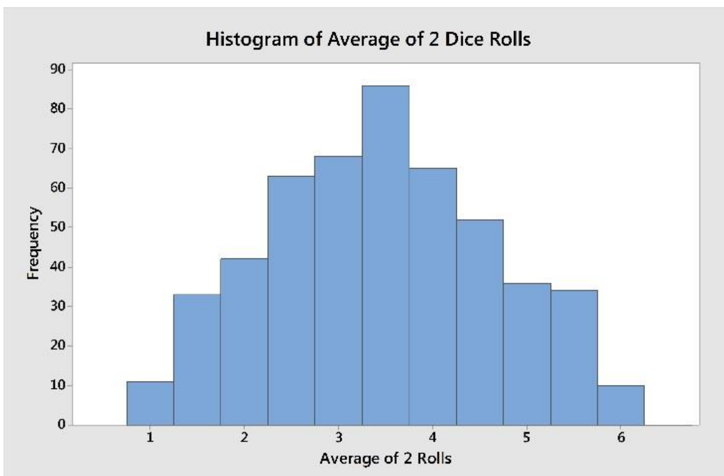
1 Name column C3 as "Average of 2 Dice Rolls."

2 Open the dialog.
  - Minitab 17: Choose **Calc > Calculator**. For **Store result in variable**, enter *C3*.
  - Minitab Express Mac and PC: Choose **Data > Formula**.

3 For **Expression**, enter (C1 + C2) / 2.
4 Click **OK.**

Column "Average of 2 Dice Rolls" contains 500 averages of two dice. Take a look down the column to make sure the distribution of values makes sense. These are the outcomes from someone rolling two dice and averaging their values. Observing a 1 or a 6 as an average is the most unlikely value, though you may see about 13 or 14 of each.

We will make a histogram for the column "Average of 2 Dice Rolls" to see the shape of the distribution. We expect the **highest bin to be around the average 3.5** since 6 dice roll averages ((1, 6); (2, 5); (3, 4); (4, 3); (5, 2); (6, 1)) yield 3.5. Minitab histogram plot instructions are given in **Example 2** of this lesson.

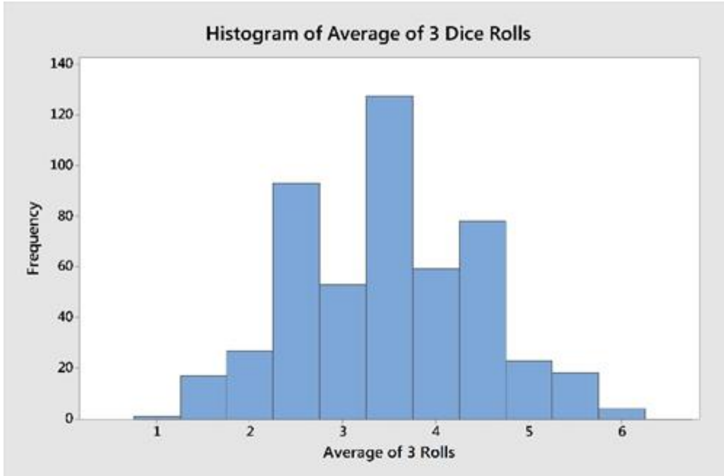


Facts about $\bar{X}$ when n = 2

- The mean of the distribution $\bar{X}$ is the same as the mean of X_1 or X_2.
- The standard deviation of $\bar{X}$ is less than the standard deviation of X_1 or X_2.
- The shape of $\bar{X}$ is converging toward the mean 3.5 and is taking on more of a normal distribution shape.

The **Sampling Distribution for $\bar{X}$ when n = 3** and the initial distribution is discrete uniform for x = 1, 2, 3, 4, 5, 6.

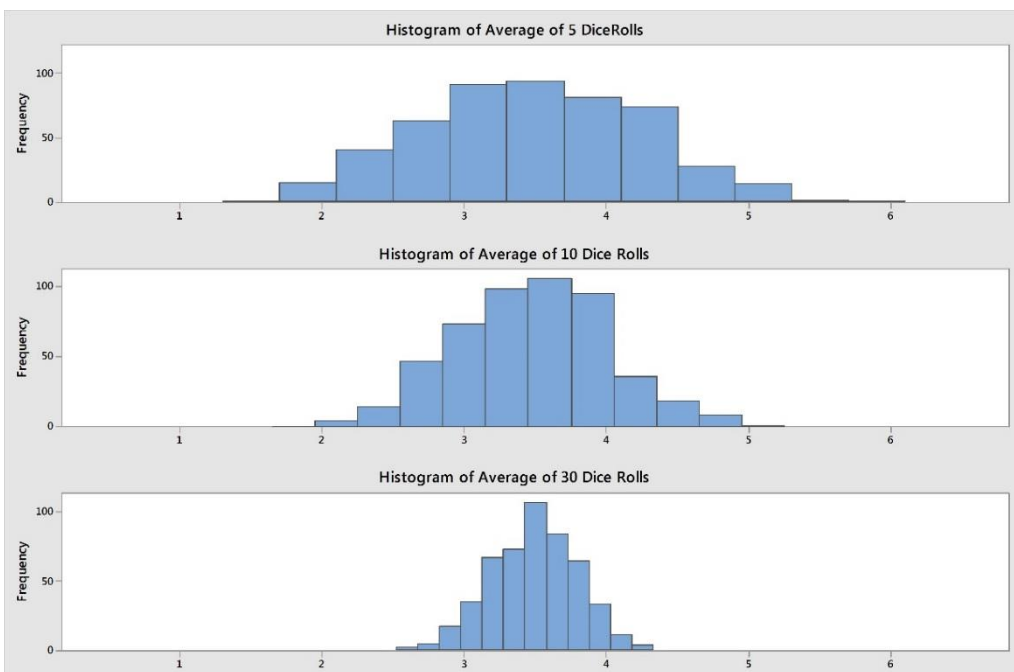
Let's average three random variables – X_1, X_2, and X_3 – where each is the outcome from rolling a fair 6-sided die. In Minitab, make a new column of 500 die tosses, then average the 3 random variables in a new column as shown previously. Last, make a histogram of $\bar{X}$ when n = 3.



Facts about $\bar{X}$ when n = 3

- The mean of the distribution $\bar{X}$ is the same as the mean of X_1, X_2, X_3, and $\bar{X}_2$.
- The standard deviation of $\bar{X}$ is less than the standard deviation of X_1, X_2, X_3, and $\bar{X}_2$.
- The shape of $\bar{X}$ is converging toward the mean 3.5 and is taking on more of a normal distribution shape.

Here are histograms for $\bar{X}$ for n = 5, 10, and 30.



The mean and standard deviation values for the histograms shown above are as follows:

Variable	N	Mean	StDev
C1	500	3.438	1.701
C2	500	3.474	1.667
Average of 2 Rolls	500	3.456	1.203
Average of 3 Rolls	500	3.487	0.977
Average of 5 Rolls	500	3.474	0.759
Average of 10 Rolls	500	3.472	0.532
Average of 30 Rolls	500	3.489	0.301

To recap the Central Limit Theorem from earlier in this lesson:

Let X_1, X_2, X_3, …, X_n be independent non-normal random variables with identical distributions with mean μ and standard deviation σ. If n is "large" enough (n > 30 suggested in most texts), then:

The distribution of the sample mean $\bar{X}$ is *approximately* normally distributed with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.

The larger the sample size n, the more normally distributed the distribution will be and the closer the distribution will converge about the mean μ.

Although statisticians and texts suggest n to be at least 30 for $\bar{X}$ to approach normality, you can see from the plots on this page that the histograms are displaying a normal shape for $n < 30$. The sample size required for convergence to normality depends on the shape of the original distribution. Fairly flat or symmetric distributions tend to have normal $\bar{X}$ plots for smaller sample sizes n compared to distributions that are highly skewed.

Advantages of Normally Distributed $\bar{X}$'s

If we can assume that a distribution, such as $\bar{X}$, is normally distributed and we know the original distribution's mean and standard deviation, then we can determine probabilities for $\bar{X}$ using a standard normal table or Minitab. In essence, we repeat the technique used in the ***Normal Distribution*** lesson, except that we replace the random variable X with its average $\bar{X}$ for sample size n.

Example 3

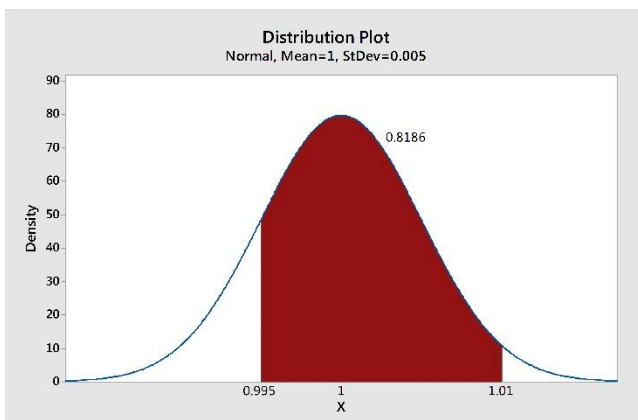
Let $X_1, X_2, X_3, \ldots, X_{100}$ denote the weights of 100 independent and identically distributed bags of candy corn. If the mean weight of each bag is 1 lb. and its standard deviation is 0.05 lb., determine the probability that the average of 100 bags weighs between 0.995 lb. and 1.01 lb.

According to the Central Limit Theorem, the distribution for $\bar{X}$ for n = 100 is approximately normal, regardless of the distribution of each X_i.

Using the formulas that we verified in this lesson, the mean and standard deviation of $\bar{X}$, respectively, are 1 lb. and $0.05/\sqrt{100} = 0.005$ lb. Since $\bar{X}$ is approximately normal, then we can use the standard normal table to compute the probability. Recall that Z represents a standard normal random variable.

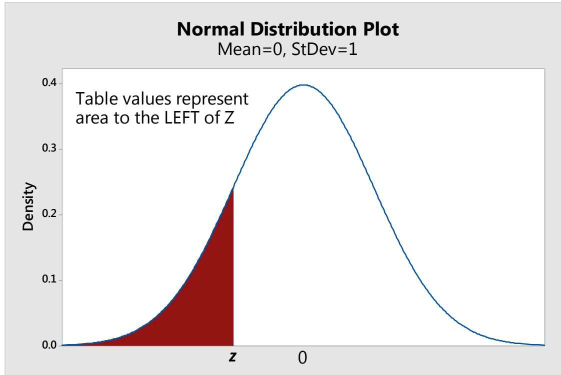
$$P(0.995 < \bar{X} < 1.01) = P\left(\frac{0.995-1}{0.005} < Z < \frac{1.01-1}{0.005}\right) \cong P(-1 < Z < 2) = 0.97725 - 0.15866 = \mathbf{0.81859}$$

We can determine this probability in Minitab as well, as explained in the ***Normal Distribution*** lesson.



Using the Standard Normal Distribution Table

The graph below depicts how to interpret the standard normal distribution tables provided on the following pages.



STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.9	.00005	.00005	.00004	.00004	.00004	.00004	.00004	.00004	.00003	.00003
-3.8	.00007	.00007	.00007	.00006	.00006	.00006	.00006	.00005	.00005	.00005
-3.7	.00011	.00010	.00010	.00010	.00009	.00009	.00008	.00008	.00008	.00008
-3.6	.00016	.00015	.00015	.00014	.00014	.00013	.00013	.00012	.00012	.00011
-3.5	.00023	.00022	.00022	.00021	.00020	.00019	.00019	.00018	.00017	.00017
-3.4	.00034	.00032	.00031	.00030	.00029	.00028	.00027	.00026	.00025	.00024
-3.3	.00048	.00047	.00045	.00043	.00042	.00040	.00039	.00038	.00036	.00035
-3.2	.00069	.00066	.00064	.00062	.00060	.00058	.00056	.00054	.00052	.00050
-3.1	.00097	.00094	.00090	.00087	.00084	.00082	.00079	.00076	.00074	.00071
-3.0	.00135	.00131	.00126	.00122	.00118	.00114	.00111	.00107	.00104	.00100
-2.9	.00187	.00181	.00175	.00169	.00164	.00159	.00154	.00149	.00144	.00139
-2.8	.00256	.00248	.00240	.00233	.00226	.00219	.00212	.00205	.00199	.00193
-2.7	.00347	.00336	.00326	.00317	.00307	.00298	.00289	.00280	.00272	.00264
-2.6	.00466	.00453	.00440	.00427	.00415	.00402	.00391	.00379	.00368	.00357
-2.5	.00621	.00604	.00587	.00570	.00554	.00539	.00523	.00508	.00494	.00480
-2.4	.00820	.00798	.00776	.00755	.00734	.00714	.00695	.00676	.00657	.00639
-2.3	.01072	.01044	.01017	.00990	.00964	.00939	.00914	.00889	.00866	.00842
-2.2	.01390	.01355	.01321	.01287	.01255	.01222	.01191	.01160	.01130	.01101
-2.1	.01786	.01743	.01700	.01659	.01618	.01578	.01539	.01500	.01463	.01426
-2.0	.02275	.02222	.02169	.02118	.02068	.02018	.01970	.01923	.01876	.01831
-1.9	.02872	.02807	.02743	.02680	.02619	.02559	.02500	.02442	.02385	.02330
-1.8	.03593	.03515	.03438	.03362	.03288	.03216	.03144	.03074	.03005	.02938
-1.7	.04457	.04363	.04272	.04182	.04093	.04006	.03920	.03836	.03754	.03673
-1.6	.05480	.05370	.05262	.05155	.05050	.04947	.04846	.04746	.04648	.04551
-1.5	.06681	.06552	.06426	.06301	.06178	.06057	.05938	.05821	.05705	.05592
-1.4	.08076	.07927	.07780	.07636	.07493	.07353	.07215	.07078	.06944	.06811
-1.3	.09680	.09510	.09342	.09176	.09012	.08851	.08691	.08534	.08379	.08226
-1.2	.11507	.11314	.11123	.10935	.10749	.10565	.10383	.10204	.10027	.09853
-1.1	.13567	.13350	.13136	.12924	.12714	.12507	.12302	.12100	.11900	.11702
-1.0	.15866	.15625	.15386	.15151	.14917	.14686	.14457	.14231	.14007	.13786
-0.9	.18406	.18141	.17879	.17619	.17361	.17106	.16853	.16602	.16354	.16109
-0.8	.21186	.20897	.20611	.20327	.20045	.19766	.19489	.19215	.18943	.18673
-0.7	.24196	.23885	.23576	.23270	.22965	.22663	.22363	.22065	.21770	.21476
-0.6	.27425	.27093	.26763	.26435	.26109	.25785	.25463	.25143	.24825	.24510
-0.5	.30854	.30503	.30153	.29806	.29460	.29116	.28774	.28434	.28096	.27760
-0.4	.34458	.34090	.33724	.33360	.32997	.32636	.32276	.31918	.31561	.31207
-0.3	.38209	.37828	.37448	.37070	.36693	.36317	.35942	.35569	.35197	.34827
-0.2	.42074	.41683	.41294	.40905	.40517	.40129	.39743	.39358	.38974	.38591
-0.1	.46017	.45620	.45224	.44828	.44433	.44038	.43644	.43251	.42858	.42465
-0.0	.50000	.49601	.49202	.48803	.48405	.48006	.47608	.47210	.46812	.46414

STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
0.1	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56749	.57142	.57535
0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
1.0	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91309	.91466	.91621	.91774
1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189
1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327
1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670
2.0	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169
2.1	.98214	.98257	.98300	.98341	.98382	.98422	.98461	.98500	.98537	.98574
2.2	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899
2.3	.98928	.98956	.98983	.99010	.99036	.99061	.99086	.99111	.99134	.99158
2.4	.99180	.99202	.99224	.99245	.99266	.99286	.99305	.99324	.99343	.99361
2.5	.99379	.99396	.99413	.99430	.99446	.99461	.99477	.99492	.99506	.99520
2.6	.99534	.99547	.99560	.99573	.99585	.99598	.99609	.99621	.99632	.99643
2.7	.99653	.99664	.99674	.99683	.99693	.99702	.99711	.99720	.99728	.99736
2.8	.99744	.99752	.99760	.99767	.99774	.99781	.99788	.99795	.99801	.99807
2.9	.99813	.99819	.99825	.99831	.99836	.99841	.99846	.99851	.99856	.99861
3.0	.99865	.99869	.99874	.99878	.99882	.99886	.99889	.99893	.99896	.99900
3.1	.99903	.99906	.99910	.99913	.99916	.99918	.99921	.99924	.99926	.99929
3.2	.99931	.99934	.99936	.99938	.99940	.99942	.99944	.99946	.99948	.99950
3.3	.99952	.99953	.99955	.99957	.99958	.99960	.99961	.99962	.99964	.99965
3.4	.99966	.99968	.99969	.99970	.99971	.99972	.99973	.99974	.99975	.99976
3.5	.99977	.99978	.99978	.99979	.99980	.99981	.99981	.99982	.99983	.99983
3.6	.99984	.99985	.99985	.99986	.99986	.99987	.99987	.99988	.99988	.99989
3.7	.99989	.99990	.99990	.99990	.99991	.99991	.99992	.99992	.99992	.99992
3.8	.99993	.99993	.99993	.99994	.99994	.99994	.99994	.99995	.99995	.99995
3.9	.99995	.99995	.99996	.99996	.99996	.99996	.99996	.99996	.99997	.99997