LAB 3 INSTRUCTIONS

NORMAL DISTRIBUTION & SAMPLING DISTRIBUTIONS

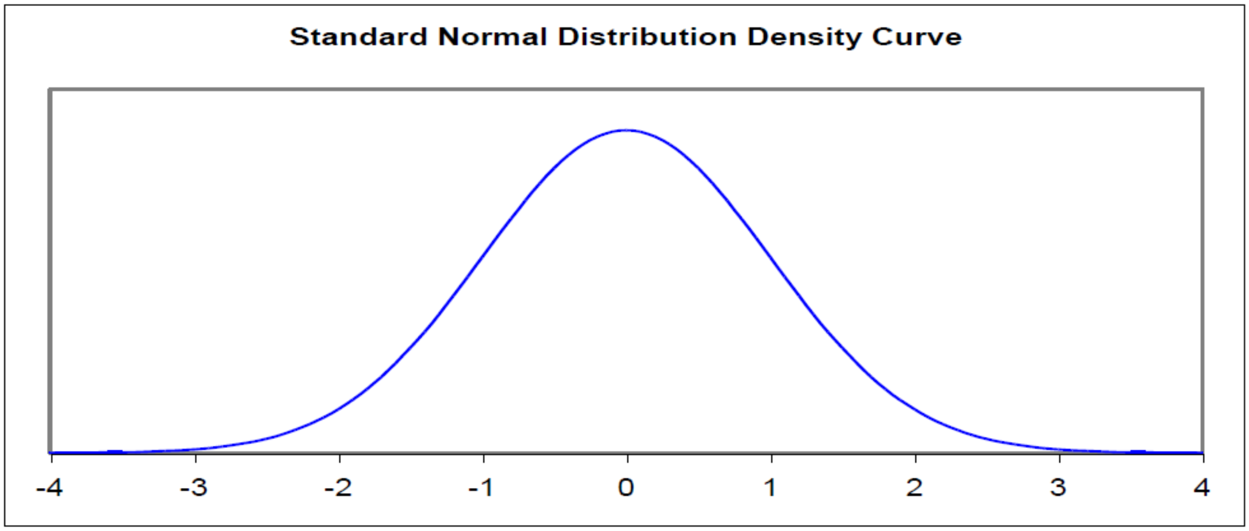
In the lab instructions, we will review the basic properties of the normal distribution and the sampling distribution of a sample mean. In particular, we will use Excel to demonstrate the Central Limit Theorem. For **Using Excel to Generate Random Numbers** and the **COUNTIF Function**, see the **Lab 2 instructions**.

For **Activating the Data Analysis Add-In** or **Inserting Excel Output into a Word Document**, see the **Lab 1 instructions**.

1. Normal Distribution

Any normal distribution is described by a symmetric bell-shaped density curve. The total area under the curve is 1. An area under the density curve gives the proportion of observations that fall in a range of values. Any normal distribution is specified by two parameters: its mean (μ) and standard deviation (σ). The mean (μ) is located at the center of the density curve while the standard deviation (σ) measures the spread of the distribution about its mean.

If a variable X follows a normal distribution with a mean of μ and a standard deviation of σ, then the standardized variable $Z = (X - \mu) / \sigma$, has the standard normal distribution with a mean of 0 and a standard deviation of 1.



1.1 NORM.DIST and NORMDIST functions

The normal distribution probabilities in Excel can be obtained by the NORM.DIST or NORMDIST functions. Either version takes three arguments as described below or via Microsoft.

NORM.DIST(x,mean,standard_dev,cumulative)

The NORM.DIST function syntax has the following arguments:

x	The value for which you want the distribution.
mean	The arithmetic mean of the distribution.
standard_dev	The standard deviation of the distribution.
cumulative	A logical value that determines the form of the function. If cumulative is TRUE, NORM.DIST returns the cumulative distribution function; if FALSE, it returns the probability density function.

The arguments in the NORM.DIST function must satisfy the following conditions: **standard_dev** is a number greater than zero and **cumulative** is either TRUE or FALSE.

Examples:

(a) If temperature in Edmonton in March follows a normal distribution with a mean of 5°C and a standard deviation of 9°C, then what is the probability the temperature on a random day will be less than 2°C?

If $X \sim N(\mu = 5, \sigma = 9)$, then $P(X < 2) = P(X \leq 2)$ = NORM.DIST(2, 5, 9, TRUE) = 0.369441

(b) Find the probability that the temperature on a random day is between 2°C and 3°C.

$P(2 \leq X \leq 3) = P(X \leq 3) - P(X \leq 2)$ = NORM.DIST(3, 5, 9, TRUE) – NORM.DIST(2, 5, 9, TRUE) = 0.042629

1.2 NORM.S.DIST and NORMSDIST functions

Alternatively, the NORM.S.DIST or NORMSDIST functions can be used. The first function has two arguments while the second has only one, yet they will produce the same answer. Examples are provided below or via Microsoft.

NORM.S.DIST(z,cumulative)

The NORM.S.DIST function syntax has the following arguments:

z The value for which you want the distribution.

cumulative Cumulative is a logical value that determines the form of the function. If cumulative is TRUE, NORMS.DIST returns the cumulative distribution function; if FALSE, it returns the probability mass function.

The arguments in the NORM.S.DIST function must satisfy the following conditions: **cumulative** is either TRUE or FALSE.

Examples:

(a) If temperature in Edmonton in March follows a normal distribution with a mean of 5°C and a standard deviation of 9°C, then what is the probability the temperature on a random day will be less than 2°C?

If $X \sim N(\mu = 5, \sigma = 9)$, then $P(X < 2) = P\left(\frac{X-\mu}{\sigma} < \frac{2-5}{9}\right) = P(Z < \text{-}1/3)$ = NORM.S.DIST(-1/3, TRUE) = 0.369441

OR, $P(Z < \text{-}1/3)$ = NORMSDIST(-1/3) = 0.369441

(b) Find the probability that the temperature on a random day is between 2°C and 3°C.

$P(2 \leq X \leq 3) = P\left(\frac{2-5}{9} \leq \frac{X-\mu}{\sigma} \leq \frac{3-5}{9}\right) = P(\text{-}1/3 \leq Z \leq \text{-}2/9)$

= NORM.S.DIST(-1/3, TRUE) – NORM.S.DIST(-2/9, TRUE) = 0.042629

OR, $P(\text{-}1/3 \leq Z \leq \text{-}2/9)$ = NORMSDIST(-1/3) – NORMSDIST(-2/9) = 0.042629

1.3 NORM.INV and NORMINV functions

Finding *x* from a provided probability can use the NORM.INV or NORMINV functions. Either version takes three arguments as described below or via Microsoft.

NORM.INV(probability,mean,standard_dev)

The NORM.INV function syntax has the following arguments:

probability A probability corresponding to the normal distribution.
mean The arithmetic mean of the distribution.
standard_dev The standard deviation of the distribution.

The arguments in the NORM.INV function must satisfy the following conditions: **probability** is a number between 0 and 1 while **standard_dev** is a number greater than zero.

Example:

If temperature in Edmonton in March follows a normal distribution with a mean of 5°C and a standard deviation of 9°C, then 36.9441% of random days will be less than what temperature value?

If $X \sim N(\mu = 5, \sigma = 9)$, then $P(X < x) = 0.369441$. Thus, x = NORM.INV(0.369441, 5, 9) = 1.999992 ≈ 2.

1.4 NORM.S.INV and NORMSINV functions

Alternatively, the NORM.S.INV or NORMSINV functions can be used. Both functions use one argument. Examples are provided below or via Microsoft.

NORM.S.INV(probability)

The NORM.S.INV function syntax has the following arguments:

probability A probability corresponding to the normal distribution.

The arguments in the NORM.S.INV function must satisfy the following conditions: **probability** is a number between 0 and 1.

Example:

If temperature in Edmonton in March follows a normal distribution with a mean of 5°C and a standard deviation of 9°C, then 36.9441% of random days will be less than what temperature value?

If $X \sim N(\mu = 5, \sigma = 9)$, then $P(X < x) = 0.369441$ and $P(Z < z) = 0.369441$.

Thus, z = NORM.S.INV(0.369441) = -0.333333… = -1/3 and $x = \mu + z\sigma = 5 + (-1/3)(9) = 2$

2. Assessing Normality

In this section, some statistical tools will be presented to check whether a given set of data is normally distributed. The methods described in 2.1 and 2.2 can be only used to detect substantial deviations from normality. Normal probability plot described in 2.3 is the most reliable method to verify the normality assumption.

2.1 Examining a histogram of the data

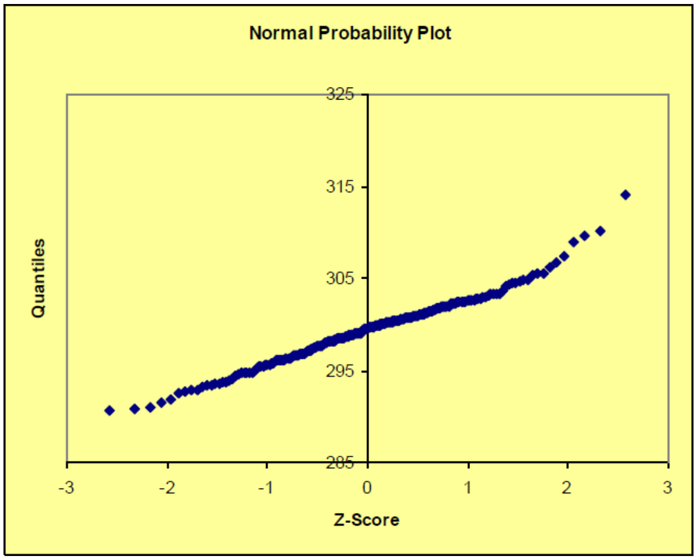
A first step in determining whether a distribution is normal is to look for obvious non-normality in a histogram of the data. Look for skewness and asymmetry. Look for gaps in the distribution (in other words, intervals with no observations). Remember, however, that normality requires more than just symmetry; the fact that the histogram is symmetric does not mean that the data come from a normal distribution.

2.2 Normal Counts

Another way to detect deviations from normality is to count the number of observations within 1, 2, and 3 standard deviations of the mean and compare the results with what is expected for a normal distribution in the 68-95-99.7 rule (in other words, the Empirical Rule). According to the rule, 68% of the observations lie within one standard deviation of the mean, 95% of observations within two standard deviations of the mean, and 99.7% of observations within three standard deviations of the mean. To count the number of observations in an Excel column, you may sort the data in ascending order and use another column of successive integer numbers to count the number of observations in each interval. You can also use the COUNTIF function described in the **Lab 2 Instructions**.

2.3 Normal Probability Plot

The plot can be obtained by plotting the standardized normal scores against ordered observations. If the data come from a normal distribution, the plotted points will fall approximately along a straight line. If the points deviate significantly from a straight line, the assumption of normality is not feasible. The *Normal Probability Plot* template in the *lab3.xlsx* file allows one to verify the assumption of normality for the data in the lab assignment. The normal probability plot below generally supports the assumption of normality for that dataset.



3. Sampling Distribution of a Sample Mean

A statistic is any function of the observations in a random sample. For example, a sample mean or a sample proportion are examples of statistics. A sampling distribution is the probability distribution of a given statistic based on a random sample of size *n*. It may be considered as the distribution of the statistic for *all possible samples* of a given size. The sampling distribution depends on the underlying distribution of the population (also called the parent population), the statistic being considered, and the sample size used.

For example, consider a normal population with a mean of μ and a variance of σ^2. Assume we repeatedly take random samples of size *n* from this population and obtain the average of observations for each sample. The distribution of these averages is called the sampling distribution of the sample mean.

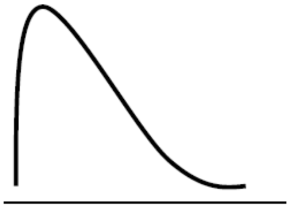
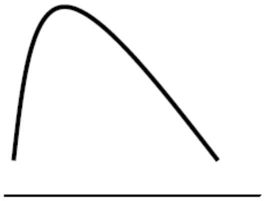
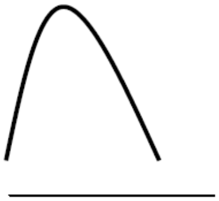
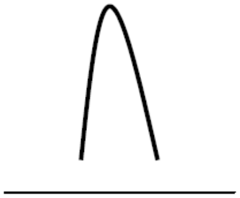
If a population has a mean of μ and a standard deviation of σ, then the sampling distribution of the sample mean $\overline{x}$ has a mean of $\mu_{\overline{x}}$ and a standard deviation of $\sigma_{\overline{x}}$ defined by the following formulas.

$$\mu_{\overline{x}} = \mu \qquad \sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$$

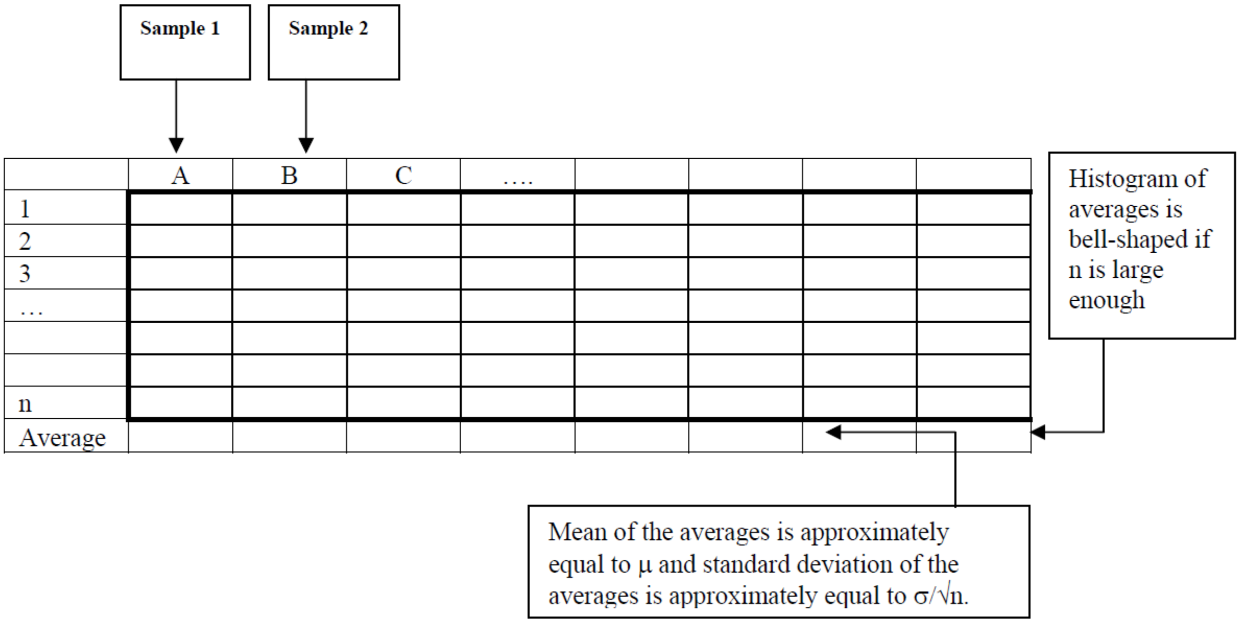
If the population from which the samples are drawn (parent population) is normal, then the distribution of the sample means will be normal regardless of the sample size. Based on the above equations, the sample means are centered at the population mean for any sample size and the spread of the sampling distribution of the sample mean decreases as the sample size increases.

4. The Central Limit Theorem

The Central Limit Theorem states that if the sample size is large enough ($n \geq 30$), then the distribution of the means of those samples (the sampling distribution of the sample mean) is approximately normal regardless of the population distribution. The larger the sample size *n*, the better the normal approximation.

Parent Population	Sampling Distribution of $\overline{x}$		
	n=5	n=10	n=30
			

The Central Limit Theorem can be demonstrated in Excel by first obtaining a large number of samples of size *n* from the population, where *n* is large enough. Each sample corresponds to one column in an Excel worksheet and consists of *n* observations (*n* rows in the worksheet). The samples can be easily obtained by using the **Random Number Generation** feature (see **Lab 2 Instructions**). Then, the averages for each column (sample) are calculated using the **Insert Function** feature. This is demonstrated in the drawing on the next page.



Histogram of averages is bell-shaped if n is large enough

Mean of the averages is approximately equal to μ and standard deviation of the averages is approximately equal to σ/√n.

If the sample size *n* is large enough, the histogram of the averages should resemble a normal curve. The **Insert Function** feature applied to the last row in the table allows you to calculate the mean and standard deviation of the averages. The Law of Large Numbers states that the actually observed sample mean of a large number of observations must approach the mean of the population.