# LAB 2 INSTRUCTIONS

PROBABILITY DISTRIBUTIONS IN EXCEL

There is a wide range of probability distributions (both discrete and continuous) available in Excel. They can be accessed through the **Insert Function** button in the formula bar or through the **Formulas** menu. The most common two applications for any specific probability distribution are those that return the cumulative probability or return the value that produces a given cumulative probability.

In this lab, we will discuss some of the above applications for binomial, Poisson and normal distributions. Examples are provided to illustrate how to use the tools in simple problems.

1.   Binomial Distribution

The distribution of the count X of successes in n independent observations, each with the same probability of success p, is called the binomial distribution with parameters n and p. The binomial probabilities in Excel can be obtained by the BINOMDIST function.

The function is accessible in the *Statistical* category of the *Insert Function* (*Lab1 Instructions*, page 11).
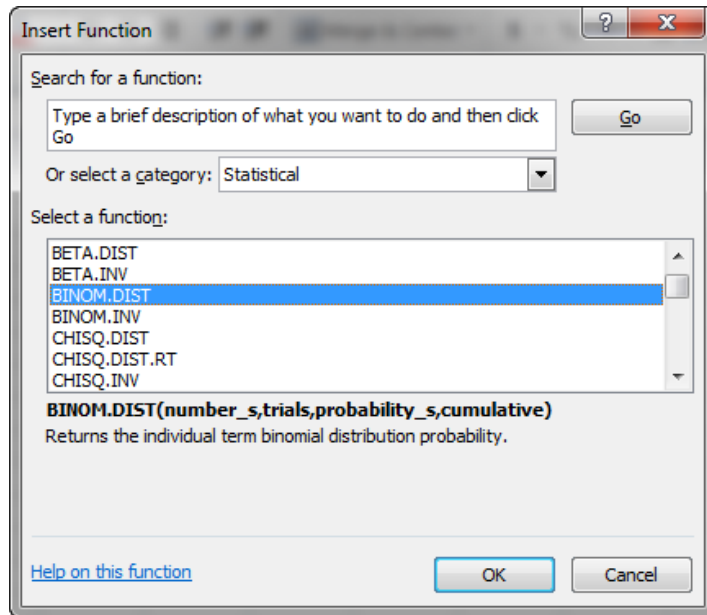
The BINOMDIST function takes four arguments: the number of successes x, the number of independent trials n, the probability of success p on each trial, and the logical variable *cumulative* that takes on the values TRUE or FALSE. When *cumulative* = TRUE, the BINOMDIST(x, n, p, *cumulative*) function returns the probability of x or fewer successes in n independent trials (cumulative probability). When *cumulative* = FALSE, BINOMDIST returns the probability of exactly x successes (probability mass function).

The binomial probability mass function is calculated in Excel as

$$BINOMDIST(x, n, p, FALSE) = \binom{n}{x} p^x \cdot (1-p)^{n-x}.$$

The arguments in the BINOMDIST function must satisfy the following conditions: x is a non-negative integer, n is a positive integer (n greater or equal to x), the probability p is between 0 and 1, and cumulative is either FALSE or TRUE.

For example, in order to calculate the probability of obtaining exactly x=10 successes (correct answers) in n=20 independent trials (multiple-choice test consisting of 20 multiple-choice questions, each with five possible answers) with the probability of a success p=0.20 (assuming a student is guessing answers randomly), enter the following four parameters 10, 20, 0.2, and FALSE into the above dialog box:

Once the function arguments are entered into the appropriate entry boxes in the *Function Arguments* dialog box, the computed value is displayed in the dialog box. Clicking on OK enters the computed value into the Excel active cell.

Notice that if you wish to calculate the probability that a student will guess at least 11 answers in the multiple-choice test, you will have to use the following relationship

$$P(X \geq 11) = 1 - P(X \leq 10) = 1 - BINOMDIST(10, 20, 0.20, TRUE).$$
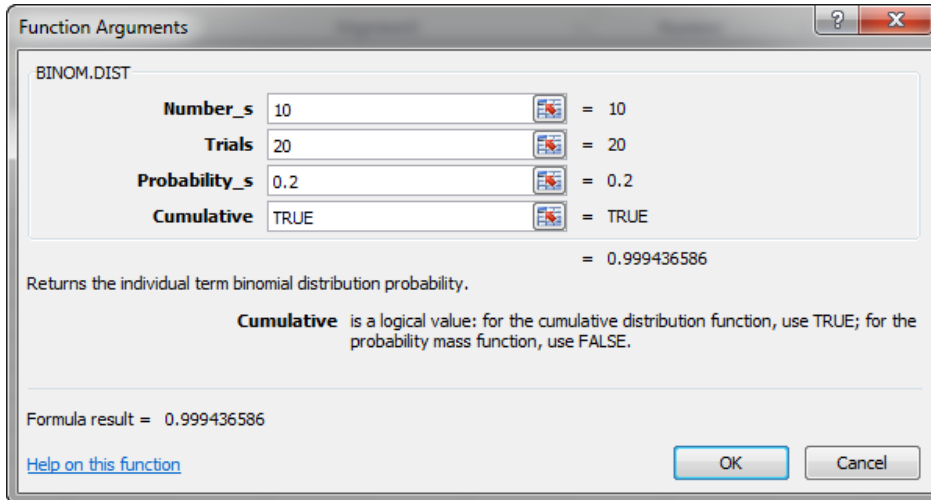

Thus the probability of obtaining at least 11 correct answers is 1-0.999436586≈0.000563.

The interactive template *Binomial* available in the Excel file *lab2.xls* that can be downloaded on Stat 235 Labs web site allows you to calculate the binomial probabilities without using the function directly. The only thing you will have to do is to enter the parameters of the binomial distribution. The binomial probabilities and cumulative binomial probabilities will be calculated automatically and displayed in your worksheet.
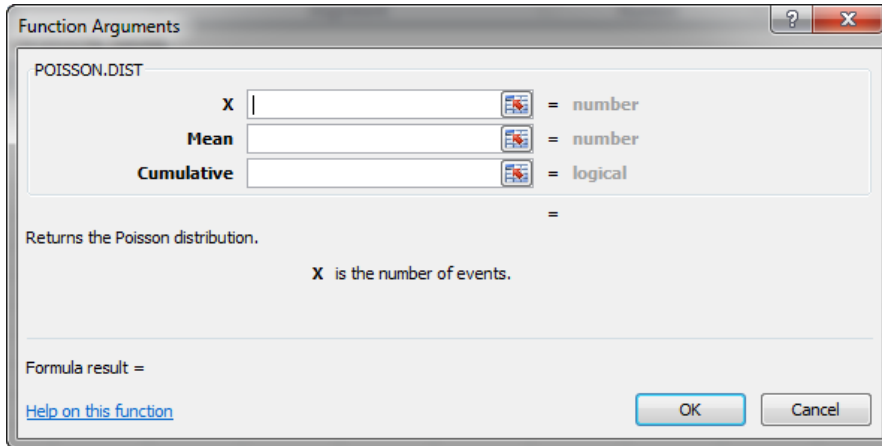
2.   Poisson Distribution

Number of vehicles passing a specified point on a highway, number of arrivals of customers per hour, or number of flaws in a glass sheet is often described by a Poisson distribution. In general, a Poisson random variable represents the number of counts in some interval.

The function POISSON is accessible in the *Statistical* category of the *Insert Function*.
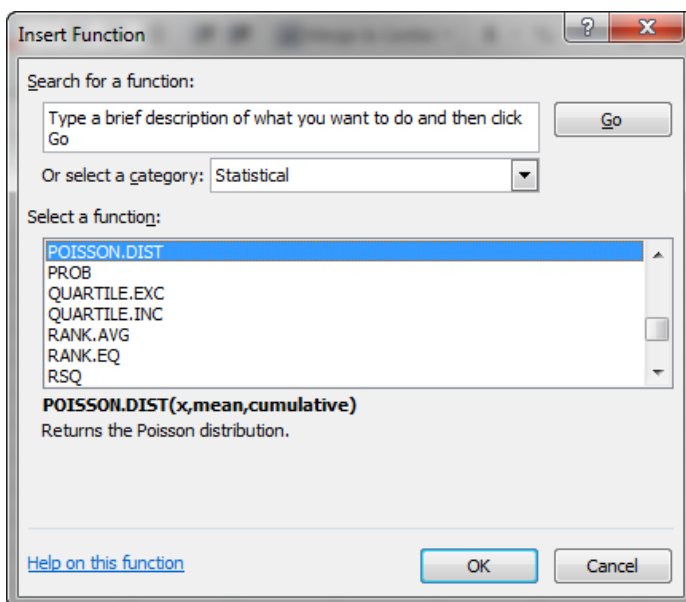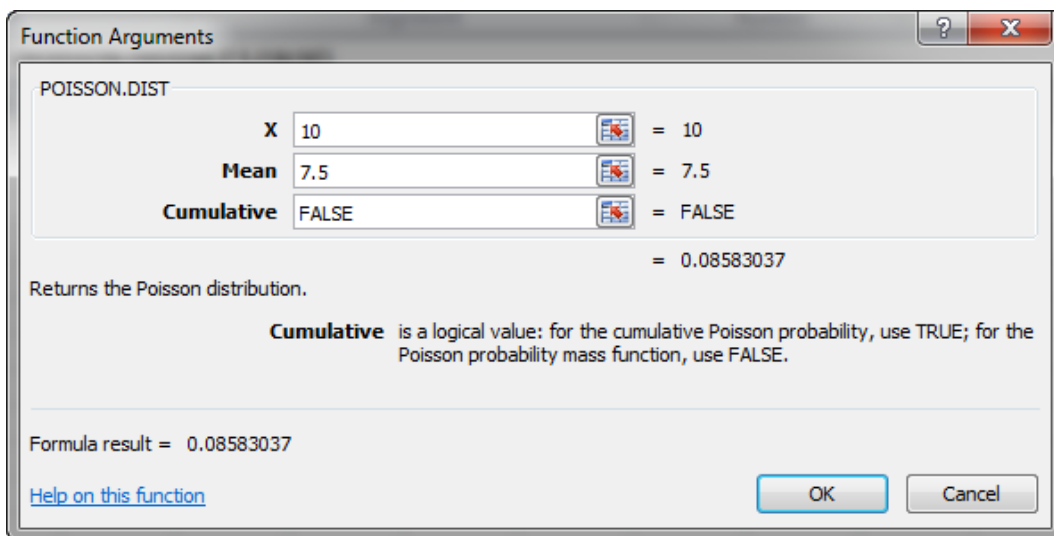
The function POISSON(x, *mu*, cumulative) takes three arguments: the number x, the mean mu, and the logical variable *cumulative* that takes on the values TRUE or FALSE.

When cumulative = TRUE, the function POISSON(x, *mu*, cumulative) returns the probability that a POISSON random variable with mean *mu* takes on a value less than or equal to x. When cumulative = FALSE, POISSON returns the probability that such a random variable takes on a value exactly equal to x.

In order to illustrate the Poisson distribution, suppose vehicles arrive at an intersection at a rate of 10 per minute. A traffic light cycle lasts 45 seconds. Then, the number of vehicles that arrive at the intersection follows a Poisson distribution with the mean *mu*= 10 * 0.75 = 7.5 because 10 vehicles arrive per minute on average, and 45 seconds is 0.75 minutes. The probability that exactly 10 vehicles will arrive at the intersection at a randomly chosen cycle can be obtained by entering the dialog box below as follows:

The template *Poisson* in the Excel file lab3.xls enables you to calculate Poisson probabilities and cumulative Poisson probabilities. The only thing you will have to do is to enter the parameter λ of the distribution into the worksheet. The parameter describes the mean number of counts in a unit of time or space.

### 3.	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, the standard deviation σ measures the spread of the distribution about its mean.

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

Standard Normal Distribution Density Curve
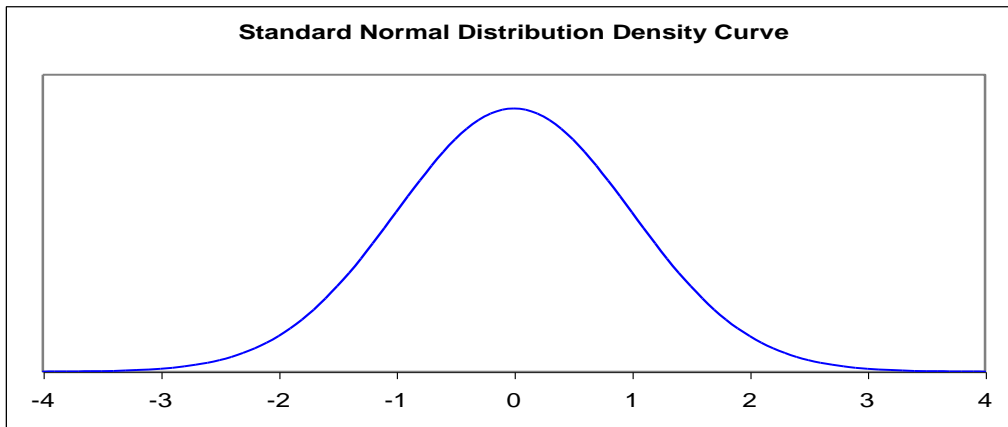
The four basic functions for normal distributions available in EXCEL are NORMDIST, NORMSDIST, NORMINV and NORMSINV. The are described in detail below.

NORMDIST Function Syntax: NORMDIST (x, mean, standard deviation, cumulative).
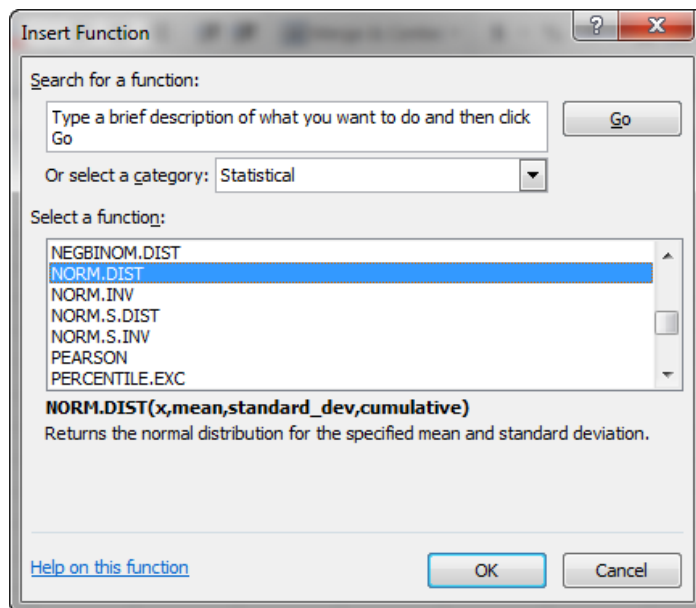
If the *cumulative* argument is FALSE, the function returns the height of the normal density function at x. If the cumulative argument is TRUE, the function returns the cumulative relative frequency that the normal variable X is less than or equal to x (the area under the density curve to the left of

x). The relative frequency that a variable X assumes values not exceeding a given number x will be denoted here by P(X<x).

Notice that you can calculate $P(2 < X < 3)$, where X is a variable following a normal distribution with a mean of 5 and standard deviation of 9 by entering the formula:

$$= NORMDIST (3, 5, 9, TRUE)- NORMDIST(2, 5, 9, TRUE).$$

You should get .042629.

NORMSDIST Function Syntax: NORMSDIST(z).

The function provides a cumulative relative frequency that $Z < z$, where Z is a variable following a standard normal distribution and z is a given real number. This value is the area under the standard normal density curve to the left of z.

To calculate the relative frequency that -1<Z<1, enter NORMSDIST(1) - NORMSDIST(-1). You will get 0.682689.

NORMINV Function Syntax: NORMINV(p, mean, standard deviation).

The function returns the value of x such that the relative frequency P(X<x)=p, where X is a variable that follows the normal distribution and p is a given number between 0 and 1. Thus NORMINV returns the 100pth percentile or the pth quantile of the normal distribution.

The first quartile of the normal distribution with the mean 100 and the standard deviation 20 can be calculated by entering the formula: NORMINV(.25, 100, 20). Excel returns the value of 86.51019.

NORMSINV Function Syntax: NORMSINV(p).

The function returns the 100pth percentile of the standard normal distribution, where p is a given number between 0 and 1. For example, NORMSINV(.05) returns the value of -1.6448530.

4.   Using Excel to Generate Random Numbers

Excel includes the *Random Number Generation* tool that fills a range of a worksheet with random numbers from one of six probability distributions: the uniform, normal, Bernoulli, binomial, Poisson, and discrete.

In order to access the tool, choose the **Data** tab, the **Analysis** group and click on **Data Analysis.** Excel opens the **Data Analysis** dialog box. To use the tool choose the *Random Number Generation* option in the dialog box and click OK.
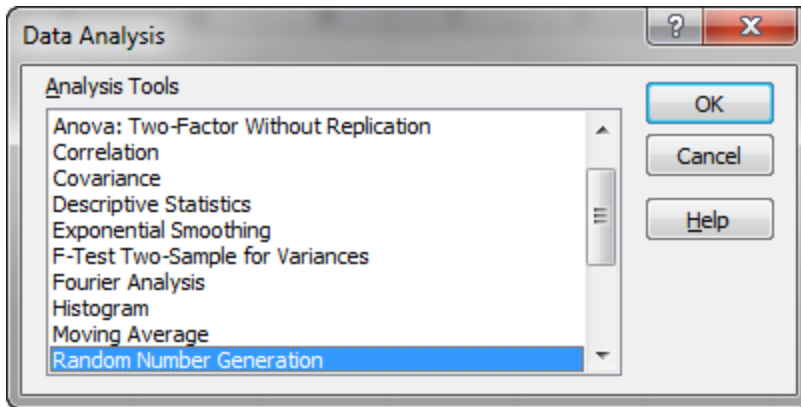
The *Random Number Generation* dialog box will appear.
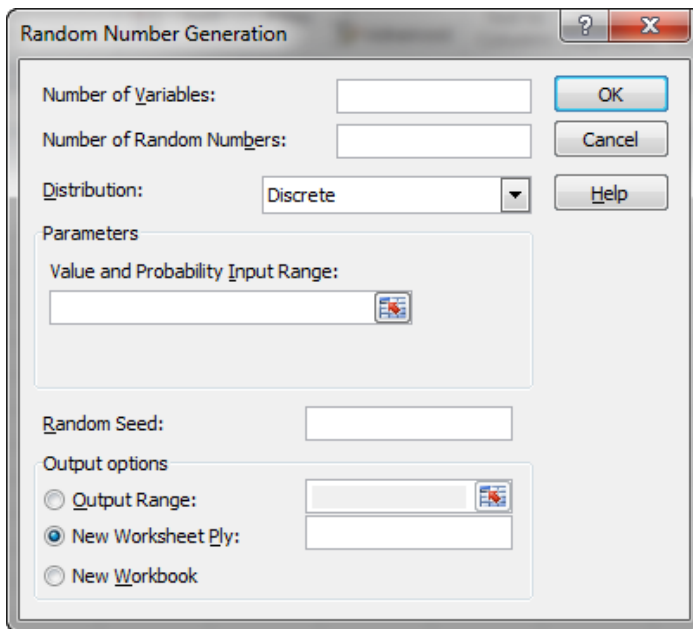
If you want one column of random numbers, type 1 in the *Number of Variables* box , then press *Tab.* Type the number of random observations you want in the *Number of Random Numbers* box. Then click *Normal* in the *Distribution* drop-down list. Enter the values of the mean, standard deviation, and the output range.

5.   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 5.1 and 5.2 can be only used to detect substantial deviations from normality. Normal probability plot described in 5.3 is the most reliable method to verify the normality assumption.

5.1   Examining a histogram of the data

A first step in determining whether a distribution is normal is to look for obvious nonnormality in a histogram of the data. Look for skewness and asymmetry. Look for gaps in the distribution - intervals with no observations. However, remember 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.

5.2   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 (text, page 123, Figure 4-12). 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 Appendix.

5.3   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 template *Normal Probability   Plot* in the file *lab2.xls* allows to verify the assumption of normality for the data in your lab assignment.
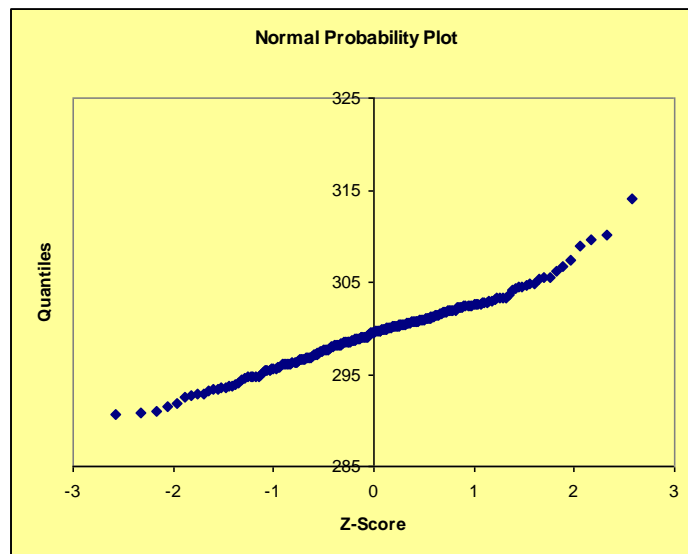
The above normal probability plot supports the assumption of normality for the data.

6. Appendix: COUNTIF Function

The COUNTIF function is used to count the number of cells in a given range that meet a single criterion. The function is accessible either from the *Insert Function* dialog box in the *Statistical* function category or by entering the following formula in a blank cell on the worksheet:

=COUNTIF(*range*, *criteria*).

The function has two arguments: *range* and *criteria*. The range argument is the cell addresses you want Excel to evaluate, and *criteria* is the value you want counted or the conditon to apply to the range. For example, to count all cells that contain the label *NO* in the range A1:A100, enter the formula =COUNTIF(A1:A100, "NO").

To count all cells in the range A1:A100 with the entries exceeding 10, you can use the formula =COUNTIF(A1:A100,">10").

To provide a count of all cells in the range A1:A100 with the entries identical to the contents of the cell C1 with an absolute address, enter the formula =COUNTIF(A1:A100, C1).

To count all cells in the range A1:A100 with the entries from the interval [1,2], you can use the formula =COUNTIF(A1:A100,"<=2") - COUNTIF(A1:A100,"<=1").

To count all cells outside of the interval [1,2] in the same range, you can use the formula =COUNTIF(A1:A100,"<1") + COUNTIF(A1:A100,">2").