**Statistical Testing**

Table of Contents

[Preliminary Reading 2](#_Toc329771238)

[Introduction 3](#_Toc329771239)

[Definition of a Statistic 4](#_Toc329771240)

[Some Often-Used Statistical Tests 4](#_Toc329771241)

[Chi-Squared () Test 4](#_Toc329771242)

[F-test 5](#_Toc329771243)

[T-test 5](#_Toc329771244)

[Linear Regression and Pearson’s Correlation Coefficient 5](#_Toc329771245)

[Anova 5](#_Toc329771246)

[Probability Distributions 6](#_Toc329771247)

[The Chi Squared () Test 10](#_Toc329771248)

[Expected Value and Mean 16](#_Toc329771249)

[Variance in Written as an Expected Value 16](#_Toc329771250)

[F-Test 16](#_Toc329771251)

[Pearson’s Correlation Coefficient 17](#_Toc329771252)

[Transforming Distributions 19](#_Toc329771253)

[One-Tailed and Two-Tailed Tests 20](#_Toc329771254)

[The Paired T-Test 22](#_Toc329771255)

[One-Sample and Two-Sample T-tests 23](#_Toc329771256)

[The Bonferroni Correction 24](#_Toc329771257)

[ANOVA (to be continued) 24](#_Toc329771258)

[Confidence Intervals 24](#_Toc329771259)

[Confidence Intervals for Estimation of a Variable (N >> 1) 24](#_Toc329771260)

[Confidence Intervals for Estimation of a Variable (N 1) 26](#_Toc329771261)

[Confidence Intervals for Engineering Design 28](#_Toc329771262)

[Conclusion 29](#_Toc329771263)

[Appendix 30](#_Toc329771264)

[Use of the Excel ERF(x) routine to calculate the cumulative normal distribution 30](#_Toc329771265)

[Calculation of the inverse error function from the inverse of the normal distribution 32](#_Toc329771266)

[Calculation of the inverse error function from the gamma distribution 32](#_Toc329771267)

List of Exercises

[**Exercise 1:** Nomenclature 3](#_Toc327536463)

[**Exercise 2:** Probability Density Function for a Coin Toss 8](#_Toc327536464)

[**Exercise 3:** Show that the sum of two Gaussian numbers has a Gaussian distribution 9](#_Toc327536465)

[**Exercise 4:** Running a Student’s t-test in Excel 20](#_Toc327536466)

[**Exercise 5:** Transforming uniform random numbers into Gaussian random numbers 20](#_Toc327536467)

[**Exercise 6:** Comparing t-tests on Gaussian and uniform random numbers 20](#_Toc327536468)

[**Exercise 7:** The paired t-test 24](#_Toc327536469)

[**Exercise 8:** Confidence intervals for different percentages. 28](#_Toc327536470)

List of Examples

[**Example 1:** Illustration of the Central Limit Theorem 6](#_Toc327536416)

[**Example 2:** Test 10](#_Toc327536417)

[**Example 3:** Test for a Uniform Distribution 15](#_Toc327536418)

[**Example 4:** F-Test 17](#_Toc327536419)

[**Example 5:** Pearson’s Correlation Coefficient 17](#_Toc327536420)

[**Example 6:** Generating Gaussian (Normal) Random Numbers 19](#_Toc327536421)

[**Example 7:** Generating Gaussian Variables with a Given Mean and Standard Deviation 20](#_Toc327536422)

[**Example 8:** Application of a One-Tailed T-Tests 21](#_Toc327536423)

[**Example 9:** Application of a One-Tailed T-Tests 21](#_Toc327536424)

[**Example 10:** Correct interpretation of the one-tailed t-test 22](#_Toc327536425)

[**Example 11:** Interpretation of the two-tailed t-test 22](#_Toc327536426)

[**Example 12:** Is my redfish too big? 23](#_Toc327536427)

[**Example 13:** Calculation of a 95% confidence interval 26](#_Toc327536428)

[**Example 14:** Calculation of an arbitrary confidence interval 26](#_Toc327536429)

[**Example 15:** Use Excel to compare the 95% confidence intervals for t and normal distributions. 27](#_Toc327536430)

# Preliminary Reading

This document assumes that the student is familiar with the mathematical definitions of mean, standard deviation and variance. Before reading this document, the student should read the articles in “Dictionary of Statistics” on the following topics:

Population

Sample Population

Distribution

Uniform Distribution

Normal Distribution

Rayleigh Distribution

Hypothesis

T-Test

F-Test

Chi-Squared Test

Correlation Coefficient

Anova

Confidence Intervals

T-distribution

Sample Mean

**Preliminary Question 1:** Let  be a random variable. Why is it not possible for a new random variable, defined as  to have a normal distribution?

**Preliminary Question 2:** Let  be a probability density function. For example, the probability density function for a normal distribution is , where  and  are the standard deviation and mean of , respectively. What are the equivalent expressions for a uniform distribution, a chi-squared distribution, and a Rayleigh distribution?

**Preliminary Question 3:** You propose that crows are larger than blackbirds. What are the populations implied by this proposal?

**Preliminary Question 4:** Imagine that you wish to test the hypothesis that crows are larger than blackbirds. How might your sample populations differ from the populations you identified in Question 3?

**Preliminary Question 5:** What is the relationship between a correlation coefficient and a standard deviation?

# Introduction

Statistical tests are used to determine how confident one can be in reaching conclusions from a data set. They are highly important when data sets lead to wide variability, as in biological experiments.

A *population* is a group under study. For example if you are interested in comparing men to women, men would be one population and women would be another.

There are several types of statistical testing. The test chosen depends on the hypothesis you are testing. For example, the *student’s T test* is used to determine whether, on average, the mean value of some *variable of interest* (e.g. height, age, temperature) in one population is different from the mean value of the same variable in another. For example, examine the question “On average, are men taller than women?” Here the variable of interest is height, the populations are men and women, and the *statistic of interest* is the average height.

Each statistical test yields a *p value* (short for probability value) that represents the probability that the *null hypothesis* is correct. The null hypothesis is generally the opposite of what you are trying to prove. For example, you could formulate the hypothesis that Biomedical Engineers perform better on the FE exam than Industrial Engineers. The null hypothesis is:

Biomedical Engineers do not perform better on the FE exam than Industrial Engineers.

**Exercise 1:** Nomenclature

Identify the population, the variable of interest and the statistic of interest implied by the above null hypothesis.

If you do a T-test and obtain a p value of 0.05, it means that:

“Given the standard deviation of these data and the number of data points, there is a 5% probability that we would obtain a difference in the means this large or larger if the performance of Biomedical and Industrial Engineers were exactly the same.”

In other words, there is a 95% chance that you are right to say that Biomedical Engineers perform better, or equivalently, given this data set, we have only a 1 in 20 chance of being wrong if we claim that Biomedical Engineers perform better on the FE exam than Industrial Engineers.

Be careful in interpreting statistical tests. *Students tend to incorrectly believe that if their p-value is less than the designated value (in biological applications this is usually taken as 0.05) then their hypothesis is true.* Some dangers are:

1. If you do enough statistical tests on something, the odds are that the t-test will show significance on something even though significance is not there. For example, if p=0.05 is taken as the cutoff point, then 1 time out of 20 you will get significance when the underlying distributions are the same. Thus, if you perform 20 t-tests, odds are that one of them will show significance even though no significance exists.
2. If the p value exceeds 0.05, *it does not prove the null hypothesis*. Indeed you can never prove the null hypothesis. If your hypothesis is that Burmese cats weigh more than Siamese cats and you find no significance (p > 0.05), it does not prove that Burmese cats and Siamese cats weigh the same. It only means that there is not enough evidence in your data set to state with confidence that they have different weights.

# Definition of a Statistic

A statistic is a numerical value that is derived from a set of data. Examples are the sample mean () and sample standard deviation (). These statistics are typically differ from the true mean () and true standard deviation (), and much of statistical analysis is performed to determine the relationships between sampled values and true values. For example, one might be interested in the average weight of the adult human brain. One can weight 20 human brains and take the average of the weights to obtain a sample mean. That value us likely to differ from the true mean, a value that could be obtained only by weighing the brain of every adult human on the planet.

# Some Often-Used Statistical Tests

Chi-Squared () Test

The test is used to test the hypothesis that the data you are working with fits a given distribution. For example, if you want to determine whether the times of occurrence of meteorites during the Leonid meteor shower are inconsistent with a Poisson distribution, you could formulate the null hypothesis that the arrival times follow such distribution and test whether the data contradict this null hypothesis.

A test is typically the first test you would like to perform on your data because the underlying probability distribution determines how you will perform the statistical tests. Note, however, that you cannot prove that the data follow a given distribution. You can only show that there is a strong probability that the data do not follow the distribution.

F-test

You choose two cases of something and formulate the hypothesis that the variances of the variable of interest for populations are different. For example, assume that you have two tools to measure height and you want to know if one leads to more consistent results than the other. You could collect repeated measurements of some item from both of these tools and then apply an F-test. (The two populations in this case are 1. measurements taken with the first tool and 2. measurements taken with the second tool). Note that in the T-test it matters whether the variances of your two data sets are different. Therefore, it is a good idea to perform an F-test on your data before you perform a T-test.

T-test

This test is probably the most widely known of all the statistical tests. You choose two populations and formulate the hypothesis that they are different. For example, if you would like to know if Altase (a blood pressure medicine) reduces blood pressure, you could form the hypothesis that “People who are given Altase (population 1) will have lower blood pressure than people who are given a placebo (population 2).

Linear Regression and Pearson’s Correlation Coefficient

Another hypothesis might be that one variable is correlated with another. For example, “Blood pressure is correlated with the number of cigarettes smoked per day.” In this case you would do a linear regression of the blood pressure vs number of cigarettes smoked and examine the p-value for this regression. This test is different from the T-test in that you are looking at a functional relationship between two quantitative values rather than a difference in means between two cases. The p value depends on the r value (which is Pearson’s Correlation Coefficient) for goodness of fit of the regression and the number of data points used in the regression. When you perform a least squares fit in Excel, one of the parameters that the software provides in the output is the p value.

Anova

The Anova examines the variance within each population and compares this to the variance between populations. The simplest case is where there are three populations, and you wish to determine whether some statistic varies from population to population. If you were interested in determining whether FE exam scores differed for Biomedical Engineering, Industrial Engineering and Mechanical Engineering students, this would be the test to use. It can also be used for cases where you do not expect a linear correlation but do expect some effect of a given variable. Weight, for example generally increases as one ages, but then typically diminishes in old age. The trend is not linear, but it certainly exists. For example, look at the variability of blood pressure as a function of age. The categories are obtained by dividing the subjects into specific age groups, such as 20-30, 30-40, 40-50, 50-60, 60-70, and 70-80 years old.

More details of each statistical test are provided later in this document.

# Probability Distributions

We denote the probability distribution of a random number by . F-Tests and T-Tests assume that the probability distribution of the noise in the data follows a Gaussian (or normal) distribution, according to Equation 1:

|  |  |  |
| --- | --- | --- |
|  |  | Eq.  |

The rand() function in Excel generates a uniformly distributed random variable between 0 and 1. This definition means that the number is just as likely to fall between 0.2 and 0.3 as it is to fall between 0.3 and 0.4, or between 0.9 and 1. The Gaussian distribution and uniform distribution are shown in Figure 1. The area under both curves must equal 1, which means that it is assured that the value of a given experiment will be somewhere in the possible range. For example, if the experiment is the roll of a die, the result must be one of 1, 2, 3, 4, 5, or 6. Hence, the probability of the result being 1, 2, 3, 4, 5, or 6 is 1.

**Figure 1:** Gaussian (left) and uniform (right) distributions. For the Gaussian, the dashed line is the mean and the dotted lines are one standard deviation from the mean. In both cases, integration over the entire range (- to for the Gaussian and 10 to 15 for the uniform distribution) yields a value 1, as required of any probability distribution.

The Gaussian distribution is important because many distributions are (at least approximately) Gaussian. The “Central Limit Theorem” states if one takes the average of samples from a population, regardless of the underlying distribution of the population, and if is sufficiently large, the distribution of this mean will be approximately Gaussian with a mean equal to the mean of the original distribution, and a standard deviation of approximately: .

**Example 1:** Illustration of the Central Limit Theorem

Show that when a new random variable is defined as “the sum of the values when a die is thrown three times,” the probability distribution begins to take on the shape of a Gaussian distribution.

**Solution:** First look at the probabilities for the sum of two dice. Anyone who has played Monopoly is aware that 2 or 12 occur with low probability, whereas a 7 is the most likely number to be thrown. Table 1 demonstrates all possible combinations of Throw 1 and Throw 2. Note that there is one way to obtain a “2,” 2 ways to obtain a “3,” 3 ways to obtain a “4,” 4 ways to obtain a “5,” 5 ways to obtain a “6,” 6 ways to obtain a “7,” 5 ways to obtain an “8,” 4 ways to obtain a “9,” 3 ways to obtain a “10,” 2 ways to obtain an “11,” and 1 way to obtain a “12.”

**Table 1:**  All possible outcomes for the throwing of a die twice. For example, the “6” in the 4th row, 2nd column is obtained from throwing a “4” on Throw 1 and a “2” on Throw 2, and there are 5 ways of throwing a “6.” Thus, the probability of a “6” is 5 out of 36 (number of rows × number of columns = 36).

It follows that the distribution for 2 rolls of a die is trianglular in shape. Table 2 builds on this result. On the left of the table are the possible outcomes for Throw 3, and at the top of the table are the possible outcomes for the combination of throws 1 and 2. At the bottom of the table, the row marked “Frequencies” shows the frequency for each outcome. For example, the 6 at the bottom indicates that there are 6 different ways to obtain 7 from the roll of 2 dice.

**Table 2:**  All possible outcomes for the throwing of a die twice followed by the throwing of a die a third time. The row labeled “Frequencies” is the number of ways that the given outcome from the 2-throw experiment can occur. For example, the “12” in row 5, column 7 means that the sum of the first two throws was 7 (achieved in one of 6 ways) and the outcome of the third throw was “5,” making a total of 12.

To obtain the number of combinations for each possible result, it is necessary to multiply the number of times a given number occurs in each column by the frequency for that column and then sum over all columns. For example, the number of possible 8’s 1+2+3+4+5+6 = 21. The total number of possible combinations is 63 = 216, so the odds of obtaining an 8 are 21/216. Table 3 shows all combinations that can occur for 3 throws of a die and the number of times they can occur.

The probability density for the 3 rolls of a die are obtained by taking the frequency values in Table 3 and dividing by the total possible number of combinations (256). These values are plotted in Figure 2 along with the probability density for the Gaussian. Even when the number of values in the sum is as small as 3, close agreement is found with a Gaussian distribution.

**Figure 2:** Probability density for the sum of 3 die throws as compared to the Gaussian distribution. The agreement is close, even thought the probability density for a single die is uniformly distributed and only 3 values are summed to obtain the final result.

**Table 3:**  All possible values that can be obtained from the sum of three throws of a die and the frequencies with which they can be obtained. For example, a value of “4” (second column) can be obtained from 3 combinations, 1, 1, 2; 1, 2, 1; and 2, 1, 1.

**Exercise 2:** Probability Density Function for a Coin Toss

Define a random number as the number of times a coin comes up heads when tossed 20 times. For example, if the outcome is T, T, T, H, T, H, H, H, T, H, T, H, T, H, H, H, T, T, T, T, there are 9 heads and 11 tails, so the random number’s value is 9. This value is the same as would be obtained by defining H as 1 and T as zero and defining a new random variable as the sum results from all 20 tosses. Find the probability density function for this new random variable and compare it directly to a Gaussian distribution. (Hint: for 1 toss the probability density is 0.5 at 0 and 0.5 at 1. For 2 tosses, there is one way to obtain a value of 0 (two tails), two ways to obtain a value of 1 (H, T and T, H) and 1 way to obtain a value of 2 (two heads). The density is 0.25 at 0 and 2 and 0.5 at 1. For 3 tosses, there is a 50% chance of all values remaining the same (the 3rd toss is tails) and a 50% chance of them increasing by 1. Thus, the possibilities are given by Table 4:

**Table 4:** The possible number of ways to obtain different values from three tosses of a coin. For two tosses we already know that there is one way to obtain 0 (two tails), two ways to obtain 1 (Heads + Tails or Tails + Heads) and one way to obtain 2 (two heads). The table then considers how many ways there are to obtain each new variable (which can have a value from 0 to 3) when the third coin is tossed. It considers first the case where the third toss is Tails and then the case where the third toss is Heads.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **New Value** | 0 | 1 | 2 | 3 |
| **Ways of obtaining if 3rd toss is Tails** | 1 | 2 | 1 |  |
| **Ways of obtaining if 3rd toss is Heads** |  | 1 | 2 | 1 |
| **Total Possible Ways of Obtaining New Value** | 1 | 3 | 3 | 1 |

Table 4 can be continued as in Table 5. One takes the probability distribution from the previous toss, shifts it to the right and sums. This pattern is easy to implement in Excel. The astute student will notice that the process is equivalent to convolving each successive probability distribution with the probability distribution for a single coin. The pattern is not unexpected. In general, when forming a new random number as the sum of random numbers from two distributions, the probability density of the new random number is the convolution of the distributions from the two original distributions.

**Table 5:** Row 1 shows the number of ways that 0 and 1 can be obtained from a single coin toss (there is one way to obtain tails (0) and one way to obtain heads (1). In row 2, the pattern is shifted to the right by one. Row 3 is the sum of rows 1 and 2, indicating the number of ways that 0, 1 and 2 can be obtained from two tosses. Row 4 is row 3 shifted to the right. Row 5 is the sum of rows 3 and 4, indicating the number of ways that 0, 1, 2 and 3 can be obtained from 3 coins. The process is continued until the total number of coin tosses is reached. The process is equivalent to a convolution of probability densities. To obtain probability densities, the values must be divided by the total possible outcomes. For example, the probability of obtaining “4” from 8 tosses is 70/(1+8+28+56+70+56+28+8+1) = 70/256 = 0.2734375.



**Exercise 3:** Show that the sum of two Gaussian numbers has a Gaussian distribution

The probability distribution for the sum of two random numbers is the convolution of the two numbers’ probability distribution. I.e., if , and if and are Gaussian, then the distribution for will be the convolution of the distributions and . Show that the convolution of two Gaussian distributions is Gaussian and that therefore that a random number formed as the sum of two Gaussian random numbers is still Gaussian. Recall that from the definition of convolution:

where

so that

**Hint:** The shortcut for performing the convolution is to use the convolution theorem of Fourier analysis. First, take the Fourier transforms of the two distributions, and , then multiply them together, and then take the inverse transform of that product.

# The Chi Squared () Test

It is important to know the distribution of the data you are looking at because the statistical tests assume a specific distribution, and if your data do not follow that distribution, the test will be invalid.

For the Chi-Squared test, the probability distribution is divided into a set of bins and the number of expected numbers in each bin is determined. For example, if the distribution is uniform from 0 to 5, one can divide it into 5 bins (0 to 1, 1 to 2, 2 to 3, 3 to 4, and 4 to 5). If 60 random numbers are obtained in the data set, then it is expected that, on average, one should obtain 60/5, or 12 data points per bin. One then examines the data to determine how many points do occur in each bin and forms the statistic:

|  |  |  |
| --- | --- | --- |
|  |  | Eq.  |

Where is the observed number of values in bin *i* and is the expected number of values in bin *i*. One then compares this Chi-Squared statistic to a table of significance.

**Example 2:** Test

 Use a  test on the set of data in Table 6 to determine whether it is consistent with a *Gaussian* distribution with a mean of 2 and standard deviation of 1.

**Table 6:** Data set for Example 2.

|  |
| --- |
| **Data Values** |
| -3.14 | -0.72 | -0.41 | 1.24 | 2.02 |
| -3.07 | -0.54 | 0.48 | 1.31 | 2.29 |
| -2.73 | -0.47 | 0.51 | 1.35 | 2.68 |
| -2.56 | -0.17 | 0.59 | 1.5 | 3.61 |
| -1.91 | -0.06 | 0.63 | 1.56 | 4.12 |
| -1.63 | 0.01 | 1.03 | 1.71 | 4.43 |
| -1.15 | 0.16 | 1.09 | 1.9 | 4.6 |
| -0.93 | 0.16 | 1.09 | 1.9 | 4.6 |
| -0.93 | 0.16 | 1.1 | 1.97 | 4.87 |
| -0.8 | 0.33 | 1.19 | 1.99 | 6.74 |

**Solution:** Because the test depends on the assumed probability distribution, the first step is to define that distribution. From the statement of the problem, a Gaussian distribution is to be used, which is defined as:

|  |  |  |
| --- | --- | --- |
|  |  | Eq.  |

The Gaussian depends only on the parameters and . These parameters were not given in the statement of the problem, but they can be approximated from the data. Use the mean of the data for , and use the standard deviation of the data for (obtained with AVERAGE() and STDEV() in Excel). Specifically, and . The resulting distribution is plotted in Figure 3.



**Figure 3:** Gaussian distribution with mean and standard deviation determined from the data in Table 6.

Next, the bins will be defined. One would like to have few enough bins that at least five data values fall in each bin. The number of data points in a bin between values and is:

where is the total number of points in the data. For example, there are 50 data values in the given data set, and all 50 values must be somewhere between and , which is consistent with the result that .

To define the bins, calculate the running integral of the probability density function. The first few lines of an Excel spreadsheet that accomplishes this task are shown in Figure 4. The number of points with a value less than is plotted as a function of in Figure 5. The rule of thumb for the Chi-Squared test is to have at least 5 expected data points in each bin. For the sake of this problem we will choose 10 expected data points per bin. The horizontal lines in Figure 5 mark increments of 10 points, and the vertical lines indicate where the accumulated number of points crosses each of the horizontal lines. Thus, we expect approximately 10 points with values between and -0.9, another 10 with values between -0.9 and 0.3, another 10 with values between 0.3 and 1.4, another 10 with values between 1.4 and 2.6, and another 10 with values between 2.6 and .



**Figure 4:** The first few lines of an Excel spreadsheet that calculates the Gaussian distribution (Column B), integrates it (Column C), and calculates the expected number of points (of 50) less than each x value (Column D). Cell B1 has been given the name “xbar,” and cell B2 has been given the name “sig.”



**Figure 5:** The integral of the Gaussian, multiplied by the number of points and partitioned such that 10 points, on average, should occur in each bin.

The values -0.9, 0.3, 1.4 and 2.6 are approximate locations of the intersection of the horizontal lines and the accumulated probability curve, so the expected number of points in each interval will be slightly different from 10. For example, the number of data points between -0.9 and 0.3 will be:

The integral can be calculated numerically, as in Figure 4, or, more precisely, with the use of the Excel function NORMDIST(xval, xmean, stdev, cumulative). If the parameter cumulative is 0, his function calculates the normal (Gaussian) distribution:

If the parameter cumulative is 1, the function calculates the integral of that distribution from to xlimit.

Consequently, you can determine the expected number of points in the range of to with the Excel command “=NORMDIST(x2, xmean, sigma, 1) = NORMDIST(x1, xmean, sigma, 1),” where x1, x2, xmean and sigma are named cells that contain the lower limit, the upper limit, the mean and the standard deviation, respectively. The cumulative normal distribution is closely related to the error function (which frequently arises in problems that involve mass and heat transfer). The appendix describes the use of Excel’s error function routine to calculate the cumulative normal distribution.

Table 7 shows the integrated probability density function, , and the expected number of data points in each bin with the given upper limit value.

**Table 7:** Data pertaining to each bin in the solution to Example 2.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Bin Uppler Limit Value |  -0.9 |  0.3 | 1.4 | 2.6 |  |
|  | 0.201 | 0.192 | 0.205 | 0.195 | 0.207 |
| Expected *n* in that bin | 10.04 | 9.6 | 10.26 | 9.73 | 10.36 |
| *n* calculated from the histogram | 9 | 11 | 13 | 9 | 8 |

The last row of Table 7 gives the values of for Equation 2. The values of are obtained from a histogram of the original data set. If the data and the bins are arranged on an Excel spreadsheet as shown in Figure 6, then the dialog box in Figure 7 will generate the Bin/Frequency data in range **H8:I12**.



**Figure 6:** Layout of the data and the bins to generate the histogram.



**Figure 7:** Dialog box to generate the histogram data shown in Figure 6.

The values from the histogram are given in the last row of Table 7. Equation 2 can now be evaluated as shown in Figure 8.



**Figure 8:** Spreadsheet commands to calculate the value.

The value in cell E8 is 1.64. This number must now be converted into a p value.

The probability depends on the number of degrees of freedom (). In this example, the number of degrees of freedom is the number of bins minus 3, i.e. it is 2. We subtract 1 from the number of bins because once we know the number of data points in 5 of the bins, we know the number in the final bin because we know the total number of points. In addition, we subtract 1 because we used a value for that was calculated from the data, and another 1 because we used a value for that was calculated from the data.

The p value can be obtained from Excel with the function CHIDIST(value, degree). For example, “=CHIDIST(1.64, 2)” will calculate the p value for a value of 1.62 and 2 degrees of freedom. Table 8 shows p values for Chi-Squared with 1, 2, 3, and 4 degrees of freedom.

**Table 8:** Chi-Squared values for various values of ν.

|  |  |
| --- | --- |
|  | ***p*** |
| **df** | **0.995** | **0.99** | **0.975** | **0.95** | **0.9** | **...** | **0.1** | **0.05** | **0.025** | **0.01** | **0.005** |
| **1** | --- | --- | 0.001 | 0.004 | 0.016 | ... | 2.71 | 3.84 | 5.02 | 6.64 | 7.88 |
| **2** | 0.01 | 0.02 | 0.051 | 0.103 | 0.211 | ... | 4.61 | 5.99 | 7.38 | 9.21 | 10.60 |
| **3** | 0.072 | 0.115 | 0.216 | 0.352 | 0.584 | ... | 6.25 | 7.82 | 9.35 | 11.35 | 12.84 |
| **4** | 0.207 | 0.297 | 0.484 | 0.711 | 1.064 | ... | 7.78 | 9.49 | 11.14 | 13.28 | 14.86 |

Because the value is smaller than 5.99, the p value is greater than 0.05 and hence the null hypothesis, that the two distributions are equal, cannot be rejected. We thus accept that the data could have come from the proposed Gaussian distribution.

One is suspicious that the distribution is not correct when the p value is low, but one can also be concerned if the p value is too high. Recall that p value indicates the probability of obtaining the data given the underlying probability distribution. It is unlikely to obtain data that are far from the distribution, but it is also unlikely to obtain data that match the distribution highly well. Thus, if the value had been less than 0.103, there would have been concern that the fit to the data was “too good,” perhaps indicating that the data had been faked.

If the problem statement had provided the mean and standard deviation, it would not have been necessary to calculate these parameters from the data set, and the number of degrees of freedom would have been increased by 2, to become 4.

The student may be interested to find that the data for this exercise were generated by transforming uniform random variables to Gaussian random variables via the Box Muller procedure described below. Thus, the data truly were generated from the proposed underlying Gaussian distribution.

**Example 3:** Test for a Uniform Distribution

Use a Chi-Squared test on the data set in Table 6 to determine whether it is consistent with a *uniform* distribution.

# Expected Value and Mean

The expected value of a random variable is defined as the first moment of its distribution. Specifically,



One should notice that the expected value of a Gaussian distribution is located at the peak of the distribution and is equivalent to the distribution’s “mean.” Often the terms “mean” and expected value are used interchangeably. One may also speak of a “sample mean,” which is the average of a number of random variables taken from a distribution. For example, the average of the data in Table 6 is 0.79 even though they are generated from a distribution with a mean of 1. Typically, if one is able to obtain an infinite number of data points from the distribution, one will find that the mean of the data approaches the expected value. This behavior does not always hold, but it is an intuitive result and when it does hold, the random variable is said to be *ergodic*.

# Variance Written as an Expected Value

The variance of a random variable is defined to be:

One can verify that this definition for expected value provides a value of σ that is equal to the parameter σ in the definition of the Gaussian distribution.

**Exercise 4:** Standard Deviation of a Uniform Random Variable

Use the expression for variance in terms of expected value to find the standard deviation of a set of uniformly distributed random numbers between and . Express your answer in terms of and and then rewrite the answer in terms of and , where and (i.e. is the halfway point between and ).

**Answer:**

The function will have a value of zero if or . Between and it will have a value of . This value ensures that the integral of the probability density function from to is 1. Therefore, the equation for can be rewritten as:

This integral can be evaluated directly, but the answer will be more clear if a u substitution is used, where . It follows that and the limits are from to .

In terms of and ,

For example, if a uniformly distributed random variable is needed with a mean value of 7 and a standard deviation of 2, then the variance is , and we require , or . The variable therefore needs to be uniformly distributed between and , i.e. between 3.536 and 10.464.

# F-Test

The F-test is used to compare variances. One may use it to either determine whether two data sets come from the same distribution or whether a single data set matches a known underlying distribution. The F statistic is:

,

where  is the standard deviation calculated from one data set and  is the standard deviation calculated from the other data set.

|  |
| --- |
| **Table 9:** Pressure drop measurements taken by two residents. |
|  |

**Example 4:** F-Test

A physician has both of his two new residents measure the pressure drop in an arterial segment of a patient. He notices that the standard deviation of the first intern’s data is 0.93 and that the standard deviation of the second intern’s data is 1.32. Do the data, shown in Table 9, support the hypothesis that Resident 1 is able to take data with less scatter than Resident 2?

|  |
| --- |
| **Table 10:** F-test output for the data in Table 9. |
|  |

**Solution:** The output from Excel is shown in Table 10. There are 8 degrees of freedom for data set 1 and 7 for data set 2. In this case, the degrees of freedom is the number of points minus 1. The F statistic is 0.490, and the probability of obtaining this if the two distributions are identical is 0.17. Therefore, the null hypothesis cannot be rejected and we cannot say that one intern is better than the other at making this measurement.

In the old fashioned way of doing this test, one would look up the p-values in tables. The tables were typically written for F values greater than 1, which meant that one would have to use the data set with the larger variance as data set 1 (to obtain ).

# Pearson’s Correlation Coefficient

The student is probably familiar with the r value from a least squares fit. The r value measures how will the data points fit the given line, but it does not directly state how likely it is that the line has significance. If there are only 2 data points, the r value must be 1, regardless of how valid the data are. However, if there are 100 points and each point fits the line perfectly, then one can state that the least squares fit is probably a good model for the underlying data. The Pearson’s correlation coefficient takes into account the number of data points used in the fit and provides the probability of obtaining the given set of if there is no correlation between the two variables in the underlying physics of the problem.

**Example 5:** Pearson’s Correlation Coefficient

John Q. Researcher proposes that a person’s blood pressure is linearly proportional to the person’s car’s gas mileage. He surveys 10 people and collects the data in Table 11. Is this survey consistent with the hypothesis within the p < 0.01 range?

|  |
| --- |
| **Table 11:** Data used to determine whether blood pressure is related to gas mileage. |
|  |

**Solution:** The easy way to do this problem is to input the data values into Excel and perform a linear regression. Select “Tools | Data Analysis” and then click on “linear regression.” (If the Data Analysis menu does not appear see the Help menu under “regression” for instructions on how to get it to appear). Fill in the requested data (cells for y-range, cells, for x-range, and cells for output) and hit “OK.” The output should look like the data in Tables 12, 13 and 14 (although not quite as pretty):

The relevant statistic here is the P-value for “X Variable 1” in Table 14. In this case the value is 0.18, which is much larger than 0.01, so the null hypothesis cannot be rejected, and it appears that the fit is not good.

|  |  |
| --- | --- |
| **Table 12:** Regression Statistics for the data in Table 11. | **Table 13:** Anova for the data in Table 11. |
|  |  |

|  |
| --- |
| **Table 14:** Coefficients, statistics and confidence intervals for the data in Table 11. |
|  |

To obtain further insight into this problem, one can generate a plot of mileage as a function of blood pressure. The data and the least squares fit appear in Figure 9. Although the data may appear to have a trend, the statistical test indicates no trend. As confirmation that there is no underlying pattern, the data were generated from the “rand” function of Excel. The example illustrates the danger of relying on one’s perception in making conclusions from this kind of data.



**Figure 9:** Random data with the linear least-squares model superimposed. Although there is no correlation between the two variables, the eye is tricked into thinking that there might be.

# Transforming Distributions

It is possible to transform random variables that have one distribution to random numbers with another distribution. For example, if you wanted to generate Gaussian random variables with the rand() function, you could use the Box-Muller procedure. In this method, two random numbers, *x*1 and *x*2, that are uniform between 0 and 1 are generated. The Gaussian numbers *y*1 and *y*2 are then generated as follows:



These numbers will have a mean of zero and a standard deviation of 1.

**Example 6:** Generating Gaussian (Normal) Random Numbers

Show how you could generate 6 normal random numbers in Excel.

**Solution:** Generate the following spreadsheet.

|  |  |  |  |
| --- | --- | --- | --- |
|  | A | B | C |
| 1 | =rand() | =sqrt(-2.\*ln(A1)) \* cos(2\*pi()\*A2) |  |
| 2 | =rand() | =sqrt(-2.\*ln(A1)) \* sin(2\*pi()\*A2) |  |
| 3 | =rand() | =sqrt(-2.\*ln(A3)) \* cos(2\*pi()\*A4) |  |
| 4 | =rand() | =sqrt(-2.\*ln(A3)) \* sin(2\*pi()\*A4) |  |
| 5 | =rand() | =sqrt(-2.\*ln(A5)) \* cos(2\*pi()\*A6) |  |
| 6 | =rand() | =sqrt(-2.\*ln(A5)) \* sin(2\*pi()\*A6) |  |
| 7 |  |  |  |

Column B will have 6 independent normal random with mean 0 and standard deviation of 1 numbers in it.

**Example 7:** Generating Gaussian Variables with a Given Mean and Standard Deviation

How would you generate normal random numbers that had a mean of 21 and a standard deviation of 5?

**Solution:** One needs to multiply the zero-mean normal random variables by the standard deviation and add the mean, as in Column C.

|  |  |  |  |
| --- | --- | --- | --- |
|  | A | B | C |
| 1 | =rand() | =sqrt(-2.\*ln(A1)) \* cos(2\*pi()\*A2) | =B1\*5 + 21 |
| 2 | =rand() | =sqrt(-2.\*ln(A1)) \* sin(2\*pi()\*A2) | =B2\*5 + 21 |
| 3 | =rand() | =sqrt(-2.\*ln(A3)) \* cos(2\*pi()\*A4) | =B3\*5 + 21 |
| 4 | =rand() | =sqrt(-2.\*ln(A3)) \* sin(2\*pi()\*A4) | =B4\*5 + 21 |
| 5 | =rand() | =sqrt(-2.\*ln(A5)) \* cos(2\*pi()\*A6) | =B5\*5 + 21 |
| 6 | =rand() | =sqrt(-2.\*ln(A5)) \* sin(2\*pi()\*A6) | =B6\*5 + 21 |
| 7 |  |  |  |

**Exercise 4:** Running a Student’s t-test in Excel

Use Excel to generate 40 sets of pairs of 10 random numbers having a uniform distribution between 0 and 1. Because these two have the same distribution, the t-test should show that there is no statistically significant difference in their means. Perform a 2-sample, equal variance, one-tailed t-test on each set and examine the p-values. Are any of them significant to the p < 0.05 level? Note, you can use the ttest() function in Excel. For example, if data set 1 is in a1:a10 and data set 2 is in b1:b10, you can write “=ttest(A1:A10, B1:B10, 1, 2)” in cell C1 to get the 1 tailed, equal variance test. If the next set is in columns A11:A20 and B11:B20, you then need only copy cell C1 to cell C11.

**Exercise 5:** Transforming uniform random numbers into Gaussian random numbers

Use the Box-Muller method to transform the pairs of numbers generated in Exercise 5 to Gaussian numbers. Check that the distributions of the normal and Gaussian random numbers are reasonable by doing a historgram (found under “Tools | Data Analysis” in Excel) on the data and plotting the results. On the same plot show the probability density of the corresponding data set. Do the distributions look reasonable?

**Exercise 6:** Comparing t-tests on Gaussian and uniform random numbers

Now repeat the T-test on each set. Do any of the T-tests show significance to the 0.05 level? Note the meaning of the p < 0.05 statistic: “This is the probability that these data could be generated by two distributions that are exactly identical.” Is the result you obtained consistent with this statement? Why or why not?

# One-Tailed and Two-Tailed Tests

The null hypothesis is typically a statement that two entities are equal. For example, means are proposed to be equal for the T-test, variances are assumed to be equal for the F-Test, and probability distributions are assumed to be equal for the Chi Squared-test. To use a statistical test, the “alternative hypothesis” must be specified. For the T-test, for example, one can propose that the mean for variable 1 is greater than that for variable 2, or one can propose that the two means are different. If one proposes that variable 1 is greater than variable 2, one is being more restrictive (taking a greater risk of being wrong) than if one proposes simply that the two variables are different. The reward for taking the extra risk is that one need only examine one tail of the T-test distribution. If the alternate hypothesis is simply that the two means are different, both tails of the T-test must be examined.

Use a 1 tailed T-test if your alternative hypothesis is that one of the variables is greater than the other. Use a 2 tailed test if your alternative hypothesis is that the two variables are different. How do you determine which alternative hypothesis to make? It depends on the circumstances.

**Example 8:** Application of a One-Tailed T-Tests

A coin is tossed 7 times and comes up “heads” every time. Is the coin biased towards heads at the p < 0.01 level?

**Solution:** The probability of a coin toss yielding *N* heads in a row is . Therefore, the probability of having 7 heads in a row is 1/128, or 0.00781, which is less than 0.01. Therefore it is concluded that the coin is biased toward heads at the 0.01 level.

**Example 9:** Application of a One-Tailed T-Tests

A coin is tossed 7 times and comes up “heads” every time. Is the coin biased at the p < 0.01 level?

**Solution:** In asking if the coin is biased, one must look at all outcomes that would make one conclude that the coin was biased. One of these outcomes is 7 heads in a row, but 7 tails in a row would provide an equivalent conclusion. Therefore, one must add the probabilities in both tails of the distribution. The result is thus twice 0.00781, or 0.0156, which is greater than 0.01. Therefore it cannot be concluded that the coin is biased at the 0.01 level.

The two examples above form a paradox. How is it possible to conclude that the coin is biased in a certain direction and yet not be able to conclude that it is biased? After all, if the coin is biased in a particular direction, it must be biased, right?

The resolution of the paradox lies in how the alternative hypothesis is formed. If one proposes that the coin tosses will generate 7 heads in a row and then obtains 7 heads in a row, the initial prediction is correct and all onlookers are impressed. If, on the other hand, he makes the same prediction and obtains 7 tails in a row, the initial prediction is incorrect and nobody is impressed. On the contrary, the process may generate gales of laughter from the audience. However, if one proposes that the coin will generate the same side 7 times in a row, and it comes up with either 7 heads or 7 tails, everyone is still impressed. It must be remembered that the experimenter is not allowed to look at the data before he/she formulates the alternative hypothesis. Thus, one is not simultaneously concluding that “the coin is almost certainly biased toward heads but is not certain to be biased.” Rather, one is measuring how closely the data match the original prediction, which was either that the coin was biased towards heads *or* that the coin was biased in one direction or the other.

|  |
| --- |
| **Table 15:** Data used in a two-tailed t-test. |
|  |

**Example 10:** Correct interpretation of the one-tailed t-test

If you initially proposed that the means of Event 1 and Event 2 in Table 16 were different, what would you conclude at the p < 0.01 level?

**Example 11:** Interpretation of the two-tailed t-test

For the same set of data, what would you have concluded if you had initially proposed that the mean of data set 1 was different from the mean of data set 2?

# The Paired T-Test

When comparing means of data, there are often relationships between the individual points in each data set. For example, assume that you wish to determine whether, on average, the left kidney weighs less than the right kidney. It does not make sense to pool all left kidneys in one group and all right kidneys in the other. A better approach is to compare left and right kidneys from each individual. Consider the data set in Table 16.

**Table 16:** A comparison of left and right kidney weight.

|  |  |  |
| --- | --- | --- |
| **Patient** | **Left Kidney Weight (oz)** | **Right Kidney Weight (oz)** |
| Abrams | 6.3 | 6.5 |
| Bradley | 5.7 | 5.9 |
| Dillard | 7.1 | 7.5 |
| Prudhomme | 6.2 | 6.3 |
| Richland | 5.1 | 5.7 |
| Saunders | 4.8 | 5.1 |
| Waltham | 6.5 | 6.8 |

The means and standard deviations of the two columns are similar. The p value of 0.247 indicates that there is no significant difference. However, a second look at the data shows that each value for the left kidney is smaller than the corresponding value for the right kidney. A possible explanation is that each person has a different weight, and the kidney’s weight may scale to the patient’s weight. The paired T-test takes into account the possibility that each pair of numbers in the data set may have some innate connection. In these cases you can use the paired t-test in Excel. There are obvious cases where the paired t-test would not be of value, however. For example, if the left and right kidneys did not come from the same patient there would be no grounds for pairing. One who designs experiments should be aware of cases where this kind of pairing can be taken advantage of.

# One-Sample and Two-Sample T-tests

In most cases, the two-sample t-test is used to compare the mean values of two data sets. However, it is also possible to run a one-sample t-test, where one wishes to determine whether the data suggest that the mean value for a single data set is significantly different from a fixed value. Excel’s Data Analysis package does not include a one-sample t-test, but the test can be performed relatively easily with the following steps:

1. Calculate the sample mean () and sample standard deviation () of the data set.
2. Calculate the t-statistic as:

where is the expected mean and is the number of data points.

1. Use Excel’s tdist(x, deg\_freedom, tails) function to calculate the p value. The value *x* is the absolute value of the t statistic. The number of degrees of freedom (deg\_freedom) is . The value of tails will be 2 if the alternative hypothesis is that the sample mean is different from the fixed value, and it will be 1 if the alternative hypothesis is that the sample mean is greater than (or less than) the fixed value.

**Example 12:** Is my redfish too big?

Florida state law prohibits a sportsman from keeping a redfish (Figure 10) that is longer than 27 inches. Assume that you have caught a redfish, and the Fish and Game officer tells you that it is over the limit. You agree to have five independent observers measure the fish and to pay the fine if the measurements indicate that the fish length is statistically larger than the 27 inch limit at the 5% level (p < 0.05). The measurements are 27.4, 27.3, 26.9, 27.0, and 26.8. The results of the three steps are as follows:

1. The mean is 27.08 and the standard deviation is 0.259.
2. The t-statistic is:
3. The p-value is tdist(1.54, 4, 1) = 0.099. The number of tails is 1 because you are testing whether the length of your fish is greater than the limit.

Because the p value is greater than 0.05, you and the officer agree that there is insufficient evidence to suggest that the fish is larger than the limit. Consequently, you do not have to pay the fine, and you get to go home and have a delicious redfish for dinner.



**Figure 10:** A redfish that could be used as a sample for a one-tailed (no pun intended) Student’s t-test.

**Exercise 7:** The paired t-test

Perform a paired t-test on the data in the table above. Do the data support the hypothesis that the left kidney weighs less than the right kidney to the p < 0.01 level?

# The Bonferroni Correction

As was noted earlier, if you perform a Student’s T-Test at the *p* = 0.05 level, and find significance, there is a 5% chance that the conclusion you make is wrong. It follows that if you perform two tests (e.g. you perform one test to determine whether a person’s typing speed is

# ANOVA (to be continued)

# Confidence Intervals

A confidence interval is the range in which one is confident that the true value of a variable lies, within the given percent. For example, if the calculated mean of a variable is 3.1, and the 95% confidence interval is from 2.8 to 3.4, then there is a 95% probability that the true value of that variable is somewhere between 2.8 and 3.4. Conversely, there is a 5% probability that the true mean value is outside of that interval.

Confidence Intervals for Estimation of a Variable (N >> 1)

For the simplest example of a confidence interval, consider an experiment in which a given variable is measured times, and assume that the distribution of the error in the measurement is Gaussian. The method described here will work if is sufficiently large. It is known that the average of Gaussian variables with mean and standard deviation has a Gaussian distribution with the same mean and a standard deviation of . An example of the distribution and the cumulative probability is shown in Figure 11. The standard deviation for this example is 0.5. The distribution (probability density) shows that a large probability exists in the range , and only a small amount of probability lies outside of that range. The confidence limits are ends of the range in which most of the probability lies. The 95% confidence limits, for example, surround a range within which 95% of the area under the probability density curve exists. Because the Gaussian density is symmetric, the two limits will be equally spaced from the mean. The upper tail of the distribution will contain 2.5% of the probability, and the lower tail will contain another 2.5%. The upper limit lies at a value of where the cumulative probability is equal to 0.975, which is the intersection of the dashed vertical line with the solid horizontal line.



**Figure 11:** Probability density function for a measurement with mean zero and standard deviation 0.5. 0.025% of the probability is in the upper tail of , where . Consequently, crosses the value 0.975 when .

If we approximate by the sample mean, , and the standard deviation by the sample standard deviation, then we need to find the interval for which the area under the distribution is 95%. I.e., we need the value of such that:

The confidence interval will then be . To simplify the calculation, the probability density can be shifted so that the mean is zero, and the integral can be taken over only half of the (symmetric) range and then multiplied by two. The value of can then be calculated more readily as:

This integral can be evaluated for a known value of in Excel as NORMDIST(xi, xmean, stdev,1) – 0.5, where xmean is and stdev is . However, we need to invert it to find , so we can write

NORMDIST(xi, xmean, stdev,1) – 0.5 = 0.475.

This equation gives

NORMDIST(xi, xmean, stdev,1) = 0.975.

And we can invert both sides to obtain:

xi = NORMINV(0.975, xmean, stdev);

The NORMINV function inverts the cumulative normal distribution (not the normal density function) and does not require the fourth parameter that NORMDIST uses. For a 95% probability, the value of will always be equal to 1.959964.

**Example 13:** Calculation of a 95% confidence interval

A set of 234 numbers has a mean of 18.4 and a standard deviation of 6.5. Find the 95% confidence interval.

**Answer:** The confidence interval will be , where Thus, 17.57 < < 19.23.

**Example 14:** Calculation of an arbitrary confidence interval

For the data used in , calculate the 98% confidence interval

**Answer:** For this case, is NORMINV(0.99, 0, 6.5/sqrt(234)) = 0.9885, so 17.41 < < 19.35.

The above method works when a large number of data points have been collected. When a small number of data points have been collected, the mean value will follow the t distribution instead of the normal distribution.

Confidence Intervals for Estimation of a Variable (N 1)

The Gaussian distribution will work well if the standard deviation of the variable is accurately known, as it will be if is large. In a more general case, where is not large, the Gaussian probability distribution must be replaced with the t distribution. The two distributions are similar in shape, and are compared in Figure 12. The cumulative probability is compared in Figure 13.



**Figure 13:** Comparison of the probability density function for the Gaussian and the t distribution with various values of . The amount of probability in the tails increases as decreases for the t distribution.



**Figure 13:** Comparison of the cumulative probability for the Gaussian and the t distribution with various values of .

**Example 15:** Use Excel to compare the 95% confidence intervals for t and normal distributions.

**Answer:** Excel has a built-in function TDIST(value, dof, tails), which calculates the p-value for a t statistic (value) with dof degrees of freedom and tails tails. If we represent the density of the t distribution as , then TDIST(value, dof, tails), with tails set to 2 will calculate:

However, we seek the cumulative probability, which is

The inverse function, TINV(p, dof, tails), calculates the value of that corresponds to a given value of for a two-tailed test. Therefore,

Take the inverse of both sides to obtain

For example, to obtain the 95% confidence interval (where ) with seven degrees of freedom, use:

xi = TINV(2\*(1-0.975), 7)

Figure 14 shows how the parameter varies with the number of degrees of freedom for the t distribution. The value for the Gaussian distribution (the dashed line) is always 1.96. The value for the t distribution rapidly converges to this value.



**Figure 14:** Value of for the t distribution and the Gaussian distribution.

**Exercise 8:** Confidence intervals for different percentages.

Ues Excel to calculate the value of for confidence intervals of 60%, 70%, 80%, 85%, 90% and 95% for (1) the Gaussian distribution, (2) the t distribution with , and (3) the t distribution with . Plot the results as a function of confidence interval percentage. I.e. you will have one plot of as a function of percent with three curves shown.

Confidence Intervals for Engineering Design

In the design of a device, it is important to know the variability in certain device parameters. For example, if you designed a low-pass filter, you would want to tell your customers that the cutoff frequency was known to within a given amount of error (e.g. 1 kHz 1 Hz). If you can construct prototypes and measure the parameter of interest on each one, then you can use the technique of the previous section to determine the confidence intervals. In many cases, however, it is not practical to construct prototypes. For example, a bridge across a large waterway may have unique design requirements, so only one such bridge would be built. Important factors in the construction will nonetheless be subject to variability, so you would like to determine, in advance, how confident you are that the bridge will be able to support a given load or withstand the force of a given wind speed. The methods described in the handout on measurement error can help you to estimate the standard deviation of the final design parameters. From the standard deviation, one can then use the methods of the previous section to determine the confidence limits.

# Conclusion

Several statistical tests are available to test certain hypotheses. The appropriate test depends on the statistic of interest (mean, standard deviation, probability distribution, etc.), the null hypothesis, and the alternative hypothesis. A good book on statistics is a valuable tool for anyone who needs to design experiments or interpret experimental data. Generally, for a given test it will be necessary to calculate a statistic (e.g. T statistic, F statistic, Chi-Squared statistic, or Pearson coefficent) from the data and determine a probability value based on the number of degrees of freedom. A wide variety of software is available to perform statistical tests. Although Excel is not specifically designed for statistical tests, it has several of them build in and is therefore convenient for some of the more common statistical tests.

To be valid, a statistical test must be proposed before the experiment is performed. If an experimenter looks at the data before forming a hypothesis, the validity of the test is contaminated. For example, one cannot first notice that the mean of one set of data is larger than the mean of the other and then perform the t-test. If one notices such a trend, it is necessary to collect a new set of data to be completely unbiased.

# Appendix

Use of the Excel ERF(x) routine to calculate the cumulative normal distribution

Although the Excel function NORMDIST, which includes all three of the parameters x, xmean and stdev, simplifies the evaluation of the cumulative normal distribution, it is often necessary to use functions of a single variable to evaluate a function of multiple variables. The method is illustrated here to evaluate the normal distribution from the Excel error function routine, ERF(x). Excel defines the error function as:

|  |  |  |
| --- | --- | --- |
|  |  | Eq. A1 |

which is similar to the integral of Eq. 3, but does not account for the mean and standard deviation. Substitute into Eq. 3:

so that

The limits of integration must also be changed:

These relations can be used in Equation A1 to obtain:

The integrand is is now the same as that in Eq. 3, but we need the lower and upper limits to be and , respectively. Therefore, we set:

It follows that:

So we can calculate the integral as:

|  |  |  |
| --- | --- | --- |
|  |  | Eq. A2 |

Unfortunately, the Excel function is limited in that it does not allow negative values of the two arguments. One workaround is to write a Visual Basic macro that uses the Excel-defined error function to obtain the integral. Another workaround is to break up the integral into sections for which the limits are always positive. Three cases arise:

**Case 1:** and are both negative.

Because is symmetric,

Therefore,

**Case 2:** is negative, and is positive.

But

Therefore,

**Case 3:** and are both positive.

Therefore,

All three of these cases can be summarized with the equation:

where the function returns the sign of its argument (i.e. ).

So in Excel one can calculate the integral from to as:

|  |  |  |
| --- | --- | --- |
|  | = SIGN(*b*)\*ERF(0,ABS(*b*))- SIGN(*a*)\*ERF(0,ABS(*a*)) | Eq. A3 |

If is , then the second ERF() in Eq. A3 is 1, and SIGN(*a*) is 1, so the Excel command becomes

|  |  |  |
| --- | --- | --- |
|  | = SIGN(*b*)\*ERF(0,ABS(*b*)) + 1 | Eq. A4 |

Calculation of the inverse error function from the inverse of the normal distribution

Although Excel does not provide the inverse error function, it does provide two related functions, NORMDIST(x, mu, sigma, cumulative) and NORMINV(x, mu, sigma), where mu and sigma are the mean and standard deviation, respectively, and cumulative is a parameter that determines whether the probability density or the cumulative probability is calculated. If cumulative is 0, then the probability density is calculated, i.e.

If cumulative is 1, then the cumulative probability is calculated, i.e.

If and , then the cumulative probability becomes

Because the integral of the normal distribution from to 0 gives half of the probability (when the mean is zero),

But the integral term in the above expression is half the error function, so

It follows that ERF(x) = 2\*NORMDIST( x, 0, 1/sqrt(2) ) − 1

Calculation of the inverse error function from the gamma distribution

The inverse error function can be calculated from the inverse gamma distribution as:

erf -1(xi) = SQRT(GAMMAINV(xi,0.5,1))

This method works because the gamma distribution is defined as

Therefore, the gamma distribution of , with and becomes:

The cumulative distribution of is

Let

Now let so that (and so that when , )

But . Thus, the gamma distribution of is the same as the Gaussian distribution of , i.e. . The Excel routine GAMMAINV takes the inverse of the gamma distribution, so:

In other words, if then , so .

Although the function is not included in Excel, the natural log of this function, is included. Therefore, to evaluate, for example, , you need to use