Statistics Modeling Examples
using AS-EASY-AS v1.6 for Win95/98/NT

By

Paris Karahalios, TRIUS, Inc.

P.O. Box 249, N. Andover, MA 01845-0249, Tel: 978-794-9377, Fax: 978-688-6312
Web: http://www.triusinc.com, e-mail: info@triusinc.com

All material herein is provided on an as-is basis. Neither TRIUS nor any of its employees and/or agents provide any warranty as to the applicability of this information to specific situations. Any such applicability has to be determined solely by the user.

TABLE OF CONTENTS

This document is not meant to be an exhaustive treatise in statistics and probability, but rather to help users of AS-EASY-AS for Win 98/NT understand some basic concepts so that they can assess the validity of results produced by the built-in statistical and probability functions.

1. PRINCIPLES OF SETS

Set is a collection of objects referred to as members or elements. A set may also be called **class**, **aggregate** or **collection**. In general, sets are denoted by an uppercase letter while elements of a set with lower case ones. Set notation is curly brackets, e.g., {1,2,3,4,....n} is the set of integers 1 to n. Here is a brief listing of some other important set notation and rule.

(1) Element *a* belongs in set B

$$a \in B$$

For example, if B = {1,2,3,4,5,6,7,8} then we have

$$2 \in B,\ \ 3 \in B,\ \ 5 \in B,\ \text{etc.}$$

(2) Element d does NOT belong in B

$$d \notin B$$

Using the same set, as in (1) above, we have

$$0 \notin B,\ \ 9 \notin B,\ \text{etc.}$$

(3) Each element of set A belongs to set B, i.e., A is a subset of B

$$A \subset B$$

For example, if A = {1,2} and B = {1,2,3} then since ALL the elements of A belong to B, we have

$$A \subset B$$

(4) Set A and B do NOT have exactly the same element

$$A \neq B$$

For example, if A = {1,2} and B = {1,2,3} then A is a subset of B but they do not have exactly the same elements, therefore

$$A \neq B$$

(5) Given three sets A, B, and C

$$\text{if}\ \ A \subset B\ \text{ and }\ B \subset C,\ \text{then}\ A \subset C$$

For example, if A = {1,2,3} B = {1,2,3,4} and C = {1,2,3,4,5,6,7} then since each element of A also belongs to B we have,

$A \subset B$, and since every element of B also belongs to C we have $B \subset C$ but we can also clearly see that every element of A also belongs to C, therefore $A \subset C$.

(6) The set of all elements that belong to either A or B, or both is called the union of A and B, denoted by

$A \cup B$

For example, if A = {1,2,3} and B = {2,3,4} then,

$A \cup B = \{1,2,3,4\}$

(7) The set of elements that belong to both A and B, i.e., they are common, is called the intersection, denoted by

$A \cap B$

For example, if A = {1,2,3,4,5,6} and B = {3,4,5,6,7,8} then

$A \cap B = \{3,4,5,6\}$

(8) The set of all elements of A which do not belong to B is called the difference of A and B, denoted by

$A - B$

Using the sets A and B defined in (7) above, we have,

$A - B = \{1,2\}$

(9) If $B \subset A$ then $A - B$ is called the **complement** of B relative to A. The complement of $A \cup B$ is denoted by

$(A \cup B)'$

For example, consider the sets A = {1,2,3,4,5} and B = {1,2,3}. Based on this rule we have,

$$B \subset A$$

$$A \bigcup B = \{1,2,3,4,5\}$$

$$A - B = \left(A \bigcup B\right)' = \{4,5\}$$

And here are some additional important theorems involving sets relationships.

$$A \bigcup B = B \bigcup A$$

$$A \bigcup \left(B \bigcup C\right) = \left(A \bigcup B\right) \bigcup C = A \bigcup B \bigcup C$$

$$A \bigcap B = B \bigcap A$$

$$A \bigcap \left(B \bigcap C\right) = \left(A \bigcap B\right) \bigcap C = A \bigcap B \bigcap C$$

$$A \bigcap \left(B \bigcup C\right) = \left(A \bigcap B\right) \bigcup \left(A \bigcap C\right)$$

$$A \bigcup \left(B \bigcap C\right) = \left(A \bigcup B\right) \bigcap \left(A \bigcup C\right)$$

$$A - B = A \bigcap B'$$

2. PROBABILITIES

In a random experiment, there is always an uncertainty as to whether a particular even will occur or not. As a measure of the chance or probability with which we expect the event to occur, we assign a number between 0 and 1. If we are certain that the event will occur, then we say that the probability is 100% (or 1). If we are certain the event will not occur, we say its probability is zero. For example, if a probability of an event is 1/4, we say that there is a 25% chance of it occurring, and a 75% chance that it would not occur. We can also state that the odds against it occurring are 75% to 25%, or 3 to 1.

There are two basic approaches with which we can calculate probabilities. One is the classical (a priori) approach, and the other is the frequency (a posteriori) approach.

The ***classical approach*** states that, if an event can occur in *h* different ways out of a total number of *n* possible ways, all of which are equally likely, then the probability of the event is *h/n.*

The ***frequency approach*** states that, if after *n* repetitions of an experiment, where *n* is very large, an event is observed to occur in *h* of these, then the probability of the even is *h/n*.

If we have a class of events *C* and we associate a probability *P* to each event *A* in the class, such that *P(A)* is the probability of the event A, then the basic rules for probabilities are:

For any event A in class C,

$$P(A) \geq 0$$

or more specifically,

$$0 \leq P(A) \leq 1$$

For a number of mutually exclusive events A1, A2, A3,... in class C,

$$P(A_1 \cup A_2 \cup A_3 \cdots) = P(A_1) + P(A_2) + P(A_3) + \cdots$$

Which says that the probability that either A1, or A3, or A3,... will occur is equal to the sum of the probabilities of each one of them individually occurring.

Let us examine some additional relationships between the probabilities of various events.

If $A_1 \subset A_2$ then $P(A_1) \leq P(A_2)$ and $P(A_2 - A_1) = P(A_2) - P(A_1)$

If A' is the complement of A then $P(A') = 1 - P(A)$

If $A = A_1 \cup A_2 \cup A_3 \cdots \cup A_n$ where $A_1, A_2, A_3, \ldots, A_n$ are mutually exclusive, then $P(A) = P(A_1) + P(A_2) + P(A_3) + \cdots + P(A_n)$

For three events A1, A2, and A3

$$\begin{aligned} P(A_1 \cup A_2 \cup A_3) &= P(A_1) + P(A_2) + P(A_3) \\ &\quad - P(A_1 \cap A_2) - P(A_2 \cap A_3) - P(A_3 \cap A_1) \\ &\quad + P(A_1 \cap A_2 \cap A_3) \end{aligned}$$

For any events A and B,

$$P(A) = P(A \cap B) + P(A \cap B')$$

To see how useful these rules are in understanding probabilities, If we have a sample space S, consisting of the elementary events A1, A2, A3,...An then, according to the equations above, the probability for all events A is,

$$P(A_1) + P(A_2) + P(A_3) + \cdots + P(A_n) = P(A)$$

and if we assume equal probabilities, then

$$P(A_i) = \frac{1}{n} \qquad i = 1,2,3,\ldots,n$$

and if A is any event made up of *h* such simple events, we have

$$P(A) = \frac{h}{n}$$

2.1 CONDITIONAL PROBABILITIES

Let A and B two events, in event space S, such that P(A)>0. P(B|A) denotes the probability of B occurring given that A has already occurred. Since A has already happened, A becomes the new sample space replacing S. This leads to the simple definition,

$$P(B \mid A) \equiv \frac{P(A \cap B)}{P(A)}$$

or

$$P(A \cap B) \equiv P(A)P(B \mid A)$$

Which says that the probability that both A and B occur is equal to the probability that A occurs times the probability that B occurs given A has already occurred.

Conditional Probability rules

For any A1, A2, A3 we have,

$$P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1 \cap A_2)$$

If the occurrence of one event B is not affected by the occurrence of another event A, then they are independent events, i.e.,

$$P(B \mid A) = P(B)$$
$$P(A \cap B) = P(A)P(B)$$

and if A1, A2, and A3 are to be independent, they must be pair wise independent, i.e.,

$$P(A_i \cap A_j) = P(A_i)P(A_j) \quad i \neq j, \text{where } i, j = 1,2,3 \text{ and}$$
$$P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3)$$

Finally, although we will deal with this later, it's important to state Bayes' Theorem which allows us to find the probabilities of the various events A1, A2,...An which can cause A to occur.

$$P(A_k \mid A) = \frac{P(A_k)P(A \mid A_k)}{\sum_{k=1}^{n} P(A_k)P(A \mid A_k)}$$

3. COUNTING

If a certain something can be accomplished n_1 different ways, and after that a second thing can be accomplished n_2 different ways, and so on until an m-th thing can be accomplished n_m different ways, then all m things can be accomplished in $n_1 n_2 n_3 \ldots n_m$ different ways. For example, if a man has 3 shirts and 2 pairs of pants, then he has 2*3=6 ways of choosing a shirt and then a pair of pants.

Permutations of ***n*** objects taken ***r*** at a time are denoted by,

$$ {}_nP_r = n(n-1)(n-2)\cdots(n-r+1) $$

and in the special case of ***n=r***, the above equation yields,

$$ {}_nP_r = n(n-1)(n-2)\cdots 1 = n! $$

where $n!$ is called the n factorial and the generic equation can be written as

$$ {}_nP_r = \frac{n!}{(n-r)!} $$

For example, the number of different arrangements, or permutations consisting of 2 letters each that can be formed from the 5 letters A, B, C, D, E is

$$ {}_5P_2 = \frac{5!}{(5-2)!} = \frac{1\cdot 2\cdot 3\cdot 4\cdot 5}{1\cdot 2\cdot 3} = \frac{120}{6} = 20 $$

The worksheet file perm1.wks has a few examples of using the built-in function to calculate permutations as well as their details calculation, as above.

When we talk about permutations, we are interested in the particular order of the objects, for example the permutation ABC would be different than the permutation BCA. However, some times we are only interested only in selecting certain objects, without regard to their order. Such selections are called combinations. For example, combinations ABC, BCA, CBA and BAC are equivalent.

The total number of combinations of n objects taken r at a time is given by,

$$ \binom{n}{r} = {}_nC_r = \frac{n!}{r!(n-r)!} $$

or

$$ \binom{n}{r} = \frac{n(n-1)(n-2)\cdots(n-r+1)}{r!} = \frac{{}_nP_r}{r!} $$

and it can be shown that,

$$ \binom{n}{r} = \binom{n}{n-r} = {}_nC_r = {}_nC_{n-r} $$

Example of using Combinatorial mathematics to solve probability problems.

A jar contains 8 red, 3 white and 9 blue marbles. If 3 balls are drawn at random without replacement, determine the probability that all 3 are red.

$$\text{required probability} = \frac{\text{number of selections of 4 out of 7 red marbles}}{\text{number of selections of 3 out of 19 marbles}} = \frac{{}_{7}C_{4}}{{}_{20}C_{4}} =$$

$$= \frac{\dfrac{7!}{4!(7-4)!}}{\dfrac{20!}{4!(20-4)!}} = \frac{\dfrac{1\cdot 2\cdot 3\cdot 4\cdot 5\cdot 6\cdot 7}{(1\cdot 2\cdot 3\cdot 4)\cdot(1\cdot 2\cdot 3)}}{\dfrac{1\cdot 2\cdot 3\cdot 4\cdot 5\cdot 6\cdot 7\cdot 8\cdot 9\cdot 10\cdot 11\cdot 12\cdot 13\cdot 14\cdot 15\cdot 16\cdot 17\cdot 18\cdot 19\cdot 20}{(1\cdot 2\cdot 3)\cdot(1\cdot 2\cdot 3\cdot 4\cdot 5\cdot 6\cdot 7\cdot 8\cdot 9\cdot 10\cdot 11\cdot 12\cdot 13\cdot 14\cdot 15\cdot 16\cdot 17)}} =$$

$$= \frac{\dfrac{5\cdot 6\cdot 7}{1\cdot 2\cdot 3}}{\dfrac{18\cdot 19\cdot 20}{1\cdot 2\cdot 3}} = \frac{5\cdot 6\cdot 7}{18\cdot 19\cdot 20} = \frac{7}{228}$$

4. STATISTICS

Mathematical expectation (expected value or expectation) of a random variable, for a discrete random variable X having the possible values x1, x2, x3,….xn is defines as:

$$E(X) = x_1 P(X = x_1) + \cdots + x_n P(X = x_n) = \sum_{i=1}^{n} x_i P(X = x_i)$$

And, if we set

$$P(X = x_i) = f(x_i)$$

then the above equation can be written as:

$$E(X) = x_1 f(x_1) + \cdots + x_n f(x_n) = \sum_{i=1}^{n} x_i f(x_i) = \sum x f(x)$$

where the last summation is considered over all appropriate values of x. If all the probabilities are equal, then we have a special case where,

$$E(X) = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} = \bar{x}$$

which you may recognize as the arithmetic mean, or simply the mean.

Some special theorems of expectation, that would help understand more complex topics later on, are:

(1) If c is any constant, then

$$E(cX) = cE(X)$$

(2) If X and Y are random variables, then

$$E(X + Y) = E(X) + E(Y)$$

If X and Y are independent random variables, then

$$E(XY) = E(X)E(Y)$$

The expectation of a random variable X is usually called the mean (*u* or *m*). Another important statistic is the variance (or the square root of the variance, the standard deviation), which is really a measure of the dispersion, of scatter of the values of a random variable about the mean. If the values of the random variable tend to concentrate near the mean, then the variance is small. IF the values tend to be distributed far from the mean then the variance is large.

The variance is defined as

$$Var(X) = E\left[(X - \mu)^2\right]$$

and the standard deviation is given by the equation

$$\sigma_x = \sqrt{Var(X)} = \sqrt{E\left[(X-\mu)^2\right]}$$

Usually, the standard deviation is denoted by σ without the subscript, and the variance by s or σ^2. If X is a discrete function having probability function f(x), then variance is given by the equation

$$\sigma_x^2 = E\left[(X-\mu)^2\right] = \sum_{i=1}^{n}(x_i-\mu)^2 f(x_i) = \sum (x-\mu)^2 f(x)$$

And in the special case where all the probabilities for each random variable value are equal we have,

$$\sigma_x^2 = \left[(x_1-\mu)^2 + (x_2-\mu)^2 + \cdots + (x_n-\mu)^2\right]$$

Similar to the expectation, there are a number of simple theorems regarding the variance.

(1) if $\mu = E(X)$ then

$$\sigma^2 = E(X-\mu)^2 = E\left(X^2\right) - \mu^2 = E\left(X^2\right) - [E(X)]^2$$

(2) If c is any constant, then

$$Var(cX) = c^2 Var(X)$$

(3) The quantity $E\left[(X-a)^2\right]$ is a minimum when $a = \mu = E(X)$

(4) If X and Y are independent random variables, then

$$Var(X+Y) = Var(X) + Var(Y) \text{ or } \sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2$$

(4) If X and Y are independent random variables, then

$$Var(X-Y) = Var(X) + Var(Y) \text{ or } \sigma_{X-Y}^2 = \sigma_X^2 + \sigma_Y^2$$

5. MOMENTS

Although a full dissertation about moments is clearly beyond the scope of this sample, it is important that we at least give a basic overview of *"moments"* as they are important statistics of a sample. The ***i-th*** moment of a random variable ***X*** about the mean ***u***, also referred to as the ***i-th*** central moment, is given by:

$$\mu_i = E\left[(X-\mu)^i\right]$$

from this, it follows that,

$$\mu_0 = 1$$

$$\mu_1 = 0$$

$$\mu_2 = \sigma^2$$

i.e., the second moment is also the variance of the sample. For discrete variables, the generic moments equation becomes,

$$\mu_i = \sum_{j=1}^{n} \left(x_j - \mu\right)^i f\left(x_j\right) = \sum (x-\mu)^i f(x)$$

and the ***i-th*** moment of a random variable ***X*** about the origin is given by:

$$\mu_r^{'} = E\left(X^r\right)$$

if we use the special case,

$$\mu_1^{'} = \mu$$

$$\mu_0^{'} = 1$$

The moments about the mean and about the origin are related as follows:

$$\mu_2 = \mu_2^{'} - \mu^2$$

$$\mu_3 = \mu_3^{'} - 3\mu_2^{'}\mu + 2\mu^3$$

$$\mu_4 = \mu_4^{'} - 4\mu_3^{'}\mu + 6\mu_2^{'}\mu^2 - 3\mu^4$$

6. NORMAL DISTRIBUTION

An important continuous probability distribution in the field of statistics is the normal, or Gaussian distribution. Its graph, called the normal curve, is a bell-shaped curve that describes many populations that occur in nature.

The shape of the bell curve, whether it's tall and thin or short and fat, is defined by the mean of the distribution, u, and the standard deviation, d. Knowing the values for u and d, we can calculate the value of the normal distribution at any point x using @GAUSS(x,u,d).

The curve of any continuous probability distribution is constructed so that the area under the curve bounded by the two ordinates x=X0 and x=x1 equals the probability that a measurement selected at random from the given population will fall between x=X0 and x=x1. The AS-EASY-AS function @INTG is used to evaluate the integral of a curve between the limits X0 and X1. The general syntax is @INTG("@x…",X0,X1) where "@x…" is the function to be evaluated, X0 is the start value of @x, and X1 is the end value of @x.

By combining the @INTG and @GAUSS functions, AS-EASY-AS can be used to calculate the area under the normal curve between x=X0 and x=x1, which is the probability of an event occurring between X0 and X1. The resulting formula is :

```
@INTG("@gauss(@x,u,d)",X0,X1).
```

This is best illustrated with a few examples.

6.1 Normal Distribution Example #1

A type of storage battery lasts on the average 3.0 years, with a standard deviation of 0.5 year. Assuming the battery lives are normally distributed, find the probability that a given battery will last less than 2.3 years.

From above, we know the mean is 3 years, the standard deviation is 0.5years, and we want to evaluate the normal distribution from 0 (X0) to 2.3 (X1). Plugging the values into our formula, we get:

```
@INTG("@gauss(@x,3,0.5)",0,2.3)
```

The probability the battery will last less than 2.3 years is 0.081, or 8.1%.

In ***NORM1.WKS***, the values for mean, standard deviation, X0, and X1 have been entered into cells C10..C13. The formula in cell C15 is @INTG("@gauss(@x,c10,c11)",C12,C13), which shows a result of .081.

You can vary the input in cells C10..C13, and see the impact on the result in C15.

6.2 Normal Distribution Example #2

A light bulb has a lifetime that is normally distributed with mean equal to 800 hours and a standard deviation of 40 hours. What is the probability that a bulb burns between 778 and 834 hours?

From above, we know the mean is 800 hours, the standard deviation is 40 hours, and we want to evaluate the normal distribution from 778 (X0) to 834 (X1). Plugging the values into our formula, we get:

```
@INTG("@gauss(@x,800,40)",778,834)
```

The probability the bulb burns between 778 and 834 hours is 0.5111, or 51.1%.

In ***NORM2.WKS***, the values for mean, standard deviation, X0, and X1 have been entered into cells C10..C13. The formula in cell C15 is @INTG("@gauss(@x,c10,c11)",C12,C13), which shows a result of .0.511.

You can vary the input in cells C10..C13, and see the impact on the result in C15. For example, if the standard deviation was only 40 hours, which would suggest a more narrow bell curve, the probability of a light bulb lasting between 778 and 834 hours increases to 0.820, or 82.0%.

6.3 Normal Distribution Example #3

The quality grade-point averages of 300 college freshmen approximately follow a normal distribution with a mean of 2.1 and a standard deviation of 1.2. How many of these freshmen would you expect to have a score between 2.5 and 3.5 inclusive, if the point averages are computed to the nearest tenth?

Since the scores are recorded to the nearest tenth, we require the area between X0 = 2.45 and X1 = 3.55. Our formula can be written as:

```
@INTG("@gauss(@x,2.1,1.2)",2.45,3.55) = 0.272
```

Therefore, 27.2%, or approximately 82 of the 300 freshmen, should have a score between 2.5 and 3.5 inclusive.

In ***NORM3.WKS***, the values for mean, standard deviation, X0, and X1 have been entered into cells C10..C13. The formula in cell C15 is @INTG("@gauss(@x,c10,c11)",C12,C13), which shows a result of .0272.

You can vary the input in cells C10..C13, and see the impact on the result in C15. For example, if the mean grade increased to 2.3, the probability of a freshman having a score between 2.5 and 3.5 increases to 0.301, or approximately 91 students.

6.4 Normal Distribution Example #4

On an examination, the average grade was 74 and the standard deviation was 7. If 12% of the class are given As, and the grades are curved to follow a normal distribution, what is the lowest possible A and the highest possible B?

Calculating the area under the curve between X0 and X1 solved the previous examples. In this problem, we know X1 (100) and the area under the curve (0.12), but we need to calculate X0. This can by done by using another AS-EASY-AS command, Goal Seek.

The Data Goal Seek command allows you to search for the input that would result in the desired output from the model. It involves first making an educated guess at the expected result. AS-EASY-AS then modifies the input cell by using the Newton-Raphson convergence technique until the specified output is reached.

We know the mean grade is 74, the standard deviation is 7, and the evaluation of the normal distribution from (X0) to100 (X1) is 0.12. The Data Goal Seek requires we make an initial guess for X0, let's use 80. Plugging the values into our formula, we get:

@INTG("@gauss(@x,74,7)",80,100), which returns a value of 0.196.

Lets go to the sample worksheet file, ***NORM4.WKS***. In order to allow Data Goal Seek to modify the input cell, we need to set up the formula such that it contains cell references, rather than values, and we'll place the values in the referenced cells. In our worksheet, mean is cell F10, standard deviation is F11, X0 is F12, and X1 is F13. Our formula, in F15 now should read:

@INTG("@gauss(@x,F10,F11)",F12,F13), which still returns a value of 0.196.

Now, select the Data Goal Seek command. The dialog first prompts for the Input Cell. This is the cell whose value will be modified to obtain the desired goal. In our case, it is F12, which contains our guess of 80. The next prompt is for the Output Cell. This is the cell that contains the formula for the desired goal. In this case, it is F15. The next prompt is for the Desire. In our example, it is 0.12, which is the known area under the curve. For the tolerance value, you can leave it at its default value. Clicking on OK results in AS-EASY-AS calculating the Input cell value X0 that results in a result of 0.12.

You should see F12 change to display 82.22. Therefore, the lowest A is 83, and the highest B is 82.

7. POISSON DISTRIBUTION

The Poisson distribution is a function that is generally used to approximate binomial distributions, when evaluating point values, in particular when the sample (N) is large and the product N*p (p=individual probability) is moderately small. For example, if we had 500 students attending a seminar, and each student had a probability p = 0.00002 of breaking his/her pencil while taking notes in the seminar, the calculation of *i student pencils being broken* could very well be approximated using a Poisson distribution function, analytically,

$$P\left(x=i\right)=e^{-Np}\,\frac{Np^{t}}{i!}$$

AS-EASY-AS for Win95/NT makes the calculation much simpler by providing a built-in function for it. In the above simple example, if we want to evaluate the probability that 2 students would break their pencils, then entering the formula @POISSON(2,500*0.00002,1) in a cell and pressing enter, would yield the correct result of 4.95E-05.

It is many times hard to decide when a Poisson distribution should be used. There is no "absolute" guideline for doing so, other than "N has to be very large, and p has to be moderately small". Some situations where a Poisson distribution may be applicable are:

-- Number of movies to gross over 50 million USD in a year

-- Number of High School senior students that don't graduate in a given year

-- Number of days with more that 1 inch of rain in Boston over the last 10 years.

The worksheet file ***poisson1.wks*** contains the solution to the following example:

Suppose that 300 misprints have been randomly distributed throughout a report of 500 pages. What is the probability that a given page will contain 2 misprints? What is the probability that it will contain 2 or more misprints?

8. BAYESIAN THEOREM

Most of us are familiar with the concept of conditional probabilities, i.e., the probability of the occurrence of an event given the occurrence of an earlier event. Many times, however, it is useful to look at it in reverse, i.e., find the probability of an *earlier* event conditional on the occurrence of a *later* event.

In theoretical terms, the Bayesian Theorem states:

Let A_1, A_2, A_3,....An be n mutually exclusive events whose union is the sample space S. Let E be an arbitrary event in S such that P(E)<>0. Then,

$$P(A_1 \mid E) = \frac{P(E_1 \cap E)}{P(E)}$$

$$P(A_1 \mid E) = \frac{P(E_1 \cap E)}{P(E_1 \cap E) + P(E_2 \cap E) + P(E_3 \cap E)}$$

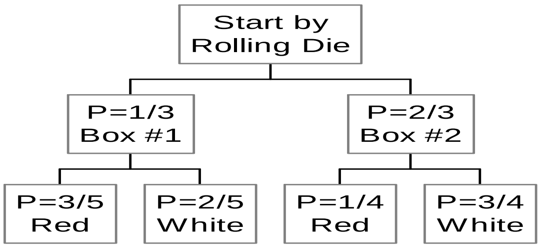
Let's investigate this concept by first looking at a couple of simple problems.

8.1 Bayesian Theorem Example #1

One box has three red and two white socks. A second box has one red and three white socks. A single die is rolled and if it comes up 1 or 6, a sock is taken out of the first box, otherwise a sock is taken out of the second box. If a red sock has just been taken out, what is the probability that it came out of the first box?

Let's formulate this problem. B1 stands for the first box, and B2 for the second. R represents a red sock and W represents a white one. Then, we assign probabilities to the appropriate outcome branches, as shown below:

Bayesian Theorem
Example #1



On the roll of the die, P(B1)=1/3 (since two out of six possibilities will result in picking from the first box), and P(B2)=2/3. Similarly, if 1 or 6 has been rolled and the sock is being taken out of the first box, the probability that it will be a red sock is 3/5, and that of being white is 3/5.

We are interested in finding P(B1|R), that is "the probability that the sock came out of the first box, given that the sock is red."

That can be written as:

$$P(B_1 \mid R) = \frac{P(B_1 \cap R)}{P(R)}$$

$$P(R) = P(B_1 \cap R) + P(B_2 \cap R)$$

$$P(B_1 \mid R) = \frac{P(B_1 \cap R)}{P(B_1 \cap R) + P(B_2 \cap R)}$$

$$P(B_1 \cap R) = P(B_1)P(R \mid B_1)$$

$$P(B_2 \cap R) = P(B_2)P(R \mid B_2)$$

$$P(B_1 \mid R) = \frac{\frac{1}{3} \cdot \frac{3}{5}}{\frac{1}{3} \cdot \frac{3}{5} + \frac{2}{3} \cdot \frac{2}{4}} = 0.55$$

The formulation of this simple example using Bayesian Theory is contained in the *AS-EASY-AS for Win95/98/NT* worksheet ***bayes1.wks***. In that spreadsheet, you can change the number of socks in each box and see the new conditional probability calculated automatically. For example, it may be interesting to see that if you change the number of socks in each box to 3 red and 3 white, and if you decided that a die roll of 1,2,3 would mean pick from box#1, while a roll of 4,5,6 would mean box#2, then the calculated probability should be 0.50. It might also be interesting to observe that if the number of white and red socks in each box is the same, the probability we are looking for is simply based on the number of rolls we assign for each box.

8.2 Bayesian Theorem Example #2

Now, let's look at a slightly more complex problem. A new inexpensive test is being developed for detecting tuberculosis. In order for the government agencies to evaluate the effectiveness of the test before it's put into general use, a medical team selects a random sample of 1000 people. Using precise, but significantly more expensive methods already available, it is determined that 8% of the 1000 people in the sample tested have tuberculosis. Now, each of the 1000 subjects is given the new skin test and the following results are observed. The new test detects tuberculosis in 96% of the test subjects who indeed have it (according to the more expensive test), and it finds tuberculosis in 2% of the people who do not have it (false positive). Based on these findings, what is the probability of a randomly chosen person having tuberculosis, if the skin test detects the disease?

Let's start by defining a number of parameters and forming a probability tree.

TB = Percent of the 1000 patients who were found to have tuberculosis using the expensive, already existing test.

NTB = Percent of the 1000 patients who were found to *NOT* have tuberculosis using the expensive, already existing test.

TBS = Percent of the patients who were found to have tuberculosis using new the skin test, out of those who were determine to have tuberculosis using the expensive, already existing test (confirming results).

TBNS = Percent of the patients who were found to *NOT* have tuberculosis using new the skin test, out of those who were determine to have tuberculosis using the expensive, already existing test (false negative).

NTBS = Percent of the patients who were found to have tuberculosis using new the skin test, out of those who were determine to *NOT* have tuberculosis using the expensive, already existing test (false positive).

NTBNS = Percent of the patients who were found *NOT* to have tuberculosis using new the skin test, out of those who were determine *NOT* to have tuberculosis using the expensive, already existing test (confirming results).

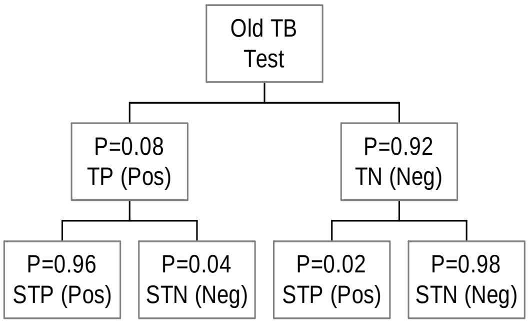
TP = Old (Original) Test positive.

TN = Old (Original) Test negative.

STP = New skin test positive.

STN = New skin test negative.

Bayesian Theorem
Example #2



We are looking for P(TP|STP), i.e., the probability of a person having tuberculosis, if the skin test indicates the disease.

$$P(TP \mid STP) = \frac{P(TP \cap STP)}{P(TP \cap STP) + P(TN \cap STP)}$$

$$P(TP \mid STP) = \frac{(0.08)(0.96)}{(0.08)(0.96) + (0.92)(0.02)} = 0.80$$

The formulation of this simple example using Bayesian Theory is contained in the *AS-EASY-AS for Win95/98/NT* worksheet ***bayes2.wks***. In that spreadsheet, you can change the results of each test and see the new conditional probability calculated automatically. For example, it may be interesting to see that if you change the "false positive" of the new skin test from 2% to 4%, the probability of the person having tuberculosis decreases from 80% to 67%.

Furthermore, if you change the "false positive" of the new skin test from 2% to 15%, the probability of the person having tuberculosis decreases from 80% to 35%.

8.3 Bayesian Theorem Example #3

A company produces 10,000 printed circuit boards a year for computers in 3 different manufacturing plants in the US. Plant A produces 3500 boards, plant B produces 2500 boards and plant C produces 4000 boards a year. Detailed production records indicate that 5% of the boards produced at plant A are defective, 3% of those produced at plant B will be defective and 7% of those produced at plant C will be defective. All the boards are shipped to a central warehouse, before being distributed. If a board at the warehouse is found to be defective, what is the probability that it was manufactured at plant A?

First, let's define some variables and construct a probability tree.

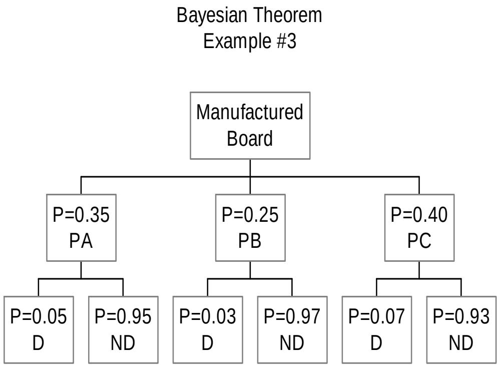
PA = Produced in plant A.

PB = Produced in plant B.

PC = Produced in plant C.

D = Defective unit.

ND = Not Defective Unite.



We are looking for P(PA|D). Using the Bayesian theorem we have:

$$P(PA \mid D) = \frac{P(D \cap A)}{P(D \cap A) + P(D \cap B) + P(D \cap C)}$$

$$P(PA \mid D) = \frac{(0.35)(0.05)}{(0.35 \cdot 0.05) + (0.25 \cdot 0.03) + (0.4 \cdot 0.07)}$$

The formulation of this simple example using Bayesian Theory is contained in the *AS-EASY-AS for Win95/98/NT* worksheet ***bayes3.wks***. In that spreadsheet, you can change the fractions produced at each plant, or the fraction of defective boards found from each plant, and see the new conditional probability calculated automatically. For example, it may be interesting to see that if you change the fraction of defective boards from plant A from 5% to 15%, the calculated probability changes from 0.33 to 0.59.

9. SAMPLE STATISTICS

If a sample is very large, it is difficult to determine the various characteristics or compute statistics such as mean, standard deviation, etc. in particular when we are trying to determine such statistics using manual calculations. For example, let's us assume that our sample consists of the heights of 100 male students at a local college. The data is shown below in Table-1.

Table-1
Male Student Height in inches

69	70	65	69	63
69	69	68	70	66
66	70	64	71	67
66	66	67	69	71
66	63	67	73	67
68	69	67	66	71
67	64	73	65	65
64	65	67	66	70
61	69	68	67	64
66	66	71	72	73
66	69	69	66	67
74	74	67	69	67
60	68	65	69	72
65	72	68	67	70
60	63	66	63	66
66	66	66	69	70
71	61	69	66	71
67	67	65	71	67
63	67	66	67	64
64	62	68	69	67

It would be fairly tedious to calculate statistics with all these numbers, and it would also be prone to errors in transcription, etc. Instead, we can arrange the data in categories or classes and determine the number of students in each class, usually referred to as the class frequency. For example, we can decide to arrange the data in groups of height, 60-62", 63-65", 66-68", 69-71", and 72-74". We could do that manually, but the built-in Data, Bin command makes the job easier. The worksheet ***dist1.wks*** contains the raw data and the results of the Data, Bin operation. Note that column H in the worksheet contains the median of each class, e.g., the median for the class 60-62" is 61". This is labeled as **ClMed** for Class Median. Furthermore, note the definition of the data bins in AS-EASY-AS. The first bin value is 63, that indicating that the bin will contain the count of data items with values less than 63", i.e., 60-62". The frequency table shown can be re-calculated by pressing Control-A, which executes the macro command shown in cell K13, or by selecting Data, Bin and specifying A3.E23 as the input range, and I4..J8 as the output range.

Now that we have created the frequency table using the Data, Bin command sequence, we can use it to calculate the mean by simply using the formula:

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{\sum fx}{n}$$

= (305 + 1152 + 2814 + 1890 + 584)/100

= 6745/100

= 67.45

The result of that calculation is shown in cell J12. When using manual calculation, however, even something like that gets involved by the sheer magnitude of the numbers being used. There is an even further simplification that can be made, provided the data meets a couple of simple criteria. If we take a look at the data labeled "ClMed" (column H in worksheet ***dist2.wks***), we see that the interval between each ClMed value is equal to 3, i.e., constant interval c=3, and 67 is the median of all the values in ClMed. If we code the value 67 as zero, then we can code 64 as -1 (being one interval below the coded median of 67), we can code 61 as -2 (two intervals below 67), we can code 70 as 1 (one interval above 67), and so on. The coded values are shown in column N of the worksheet ***dist2.wks***. Column M, of the same worksheet, contains the product of the coded value for each class (u) and the sample count in the class (f). So the important consideration in developing the coded values is that we make a transformation from the class mark ***x*** to a corresponding integer ***u*** given by,

where ***a*** is an arbitrary chosen class mark corresponding to ***u=0***

The mean may now be calculated from this coded data using, the coded values ***u*** and the formula:

$$\bar{x} = a + \frac{c}{n}\sum fu = a + \frac{\sum fu}{n}c = a + c\bar{u}$$

```
  where:  a = 67  (median coded value)
          c = 3   (constant interval)
          n = 100 (total sample count)

  Mean = 67*(15/100)*3

       = 67.45
```

Note that this calculation does NOT involve large values like the earlier calculation of the mean. Finally, cell J16 in the worksheet file ***dist2.wks*** contains the average (mean) calculated using the built-in function. You may note the difference in this value and those calculated using the call and the coded approaches, which lose some of the data definition due to grouping.

Since we are discussing calculations using coded values, the variance calculation using coded values is given by the equation,

$$s^2 = c^2\left[\frac{\sum fu^2}{n} - \left(\frac{\sum fu}{n}\right)^2\right] = c^2\left(\overline{u^2} - \bar{u}^2\right)$$

Getting back to analyzing the sample at hand, and to simplify the notation a bit, the r-th moments about the mean and about the origin, in our discrete variable data, are given by:

$$m_r = \frac{f_1\left(x_1 - \overline{x}\right)^r + \ldots + f_k\left(x_k - \overline{x}\right)^r}{n} = \frac{\sum f\left(x - \overline{x}\right)^r}{n}$$

$$m_r^{'} = \frac{f_1 x_1^r + \ldots + f_k x_k^r}{n} = \frac{\sum f x^r}{n}$$

and the two kinds of moments are related by,

$$m_1 = 0$$

$$m_2 = m_2^{'} - m_1^{'2}$$

$$m_3 = m_3^{1} - 3m_1^{'}m_2^{'} + 2m_1^{'3}$$

$$m_4 = m_4^{'} - 4m_1^{'}m_3^{'} + 6m_1^{'2}m_2^{'} - 3m_1^{'4}$$

Now, if we write

$$M_r = \frac{\sum f\left(u - \overline{u}\right)^r}{n}$$

$$M_r^{'} = \frac{\sum f u^r}{n}$$

then the above relations for ***m*** also hold true for ***M***. However, we can further simplify these relationships by using the coded variables as follows:

$$m_r = \frac{\sum f\left(x - \overline{x}\right)^r}{n} = \frac{\sum f\left[\left(a + cu\right) - \left(a + c\overline{u}\right)\right]^r}{n} = \frac{\sum f c^r\left(u - \overline{u}\right)^r}{n} = c^r M_r$$

and now, the expression for ***m*** can be re-written for coded variables as,

$$m_1 = 0$$

$$m_2 = c^2\left(M_2^{'} - M_1^{'2}\right)$$

$$m_3 = c^3\left(M_3^{'} - 3M_1^{'}M_2^{'} + 2M_1^{'3}\right)$$

$$m_4 = c^4\left(M_4^{'} - 4\mathrm{M}_1^{'}\mathrm{M}_3^{'} + 6\mathrm{M}_1^{'2}\mathrm{M}_2^{'} - 3\mathrm{M}_1^{'4}\right)$$

Now using these formulas, we can calculate the desired moments as illustrated in the worksheet file ***dist3.wks***. Note the intermediate values of **Mi** calculated in row 13 of the above worksheet, which makes the calculation of the 4 moments a bit simpler. It should also be noted that the value of the second moment is *m_2=8.5275 (using coded values).* Going back to the worksheet file ***dist1.wks***, we see that the calculated variance (based on all 100 samples), is 8.7724, which is in good agreement with the variance calculated using coded value (less resolution).

10. SKEWNESS and KURTOSIS

Often, a distribution is not symmetric about the maximum but instead has one of its "tails" longer than the other. If the longer tail occurs to the right of the maximum, the distribution is said to be "skewed to the right" . If the longer tail occurs to the left of the maximum, it is said to be "skewed to the left." The statistic describing this "asymmetry" is called coefficient of skewness of simply skewness. Its values are positive if the distribution is skewed to the right, and negative if it's skewed to the left. (This is one of the statistics calculated using the moments described earlier).

The coefficient of skewness is simply given by:

$$\alpha_3 = \frac{E\left[(X-\mu)^3\right]}{\sigma^3}$$

$$\alpha_3 = \frac{\mu_3}{\sigma^3}$$

and using coded variables and the new expressions we developed earlier,

$$a_3 = \frac{m_3}{s^3}$$

In some instances, the distribution may have its values concentrated near the mean (forming a large peak), or it may be relatively flat. The statistic describing this "peakness" of a distribution is the kurtosis. This statistic is meaningful when compared to the normal curve which has a kurtosis value of 3.

The coefficient of kurtosis, or simply kurtosis, is given by,

$$\alpha_4 = \frac{E\left[(X-\mu)^4\right]}{\sigma^4} = \frac{\mu_4}{\sigma^4}$$

and using coded variables and the new expressions we developed earlier,

$$a_4 = \frac{m_4}{s^4}$$

Using these two simple equations, we can now calculate the two valuable statistics as shown in the worksheet file ***dist4.wks***. Once again, the calculated values in this worksheet file agree very well with those calculated in the worksheet file ***dist1.wks*** using all 100 raw samples.