Beta Probability Distribution

Animal Health Risk Analysis
Canadian Food Inspection Agency
3851 Fallowfield Road
Ottawa, Ontario, Canada K2H 8P9

Analyse des risques de la santé des animaux
Agence canadienne d'inspection des aliments
3851, chemin Fallowfield
Ottawa (Ontario) Canada K2H 8P9

Canadian Food Inspection Agency (CFIA)

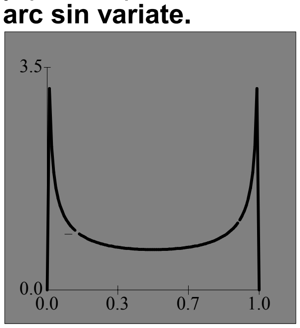
Agence canadienne d'inspection des aliments (ACIA)

Properties of the Beta Distribution

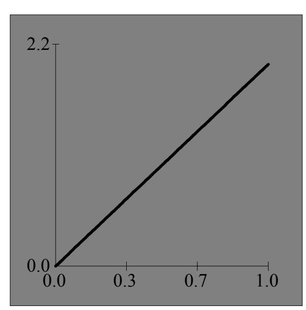
The beta distribution belongs to the family of probability densities of continuous random variables taking on values in the interval (0,1). It is useful to model the uncertainty about the unknown probability *p* of some event. The most typical application is with the use of the binomial distribution for which the knowledge or uncertainty about *p* is represented by assuming that *p* is random with a beta distribution.

The distribution can be readily transformed to a four-parameter distribution in which the additional parameters represent the range endpoints. The flexibility of the distribution encourages its empirical use in a wide range of applications. It is useful in Bayesian statistics because the beta can easily be updated to account for new data while retaining the prior information.

The beta variate β (0.5, 0.5) is an arc sin variate.

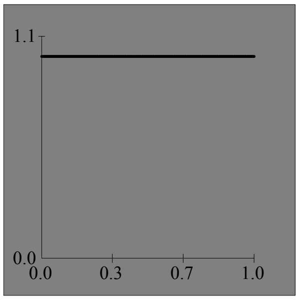


The beta variate β (2, 1) is a power function variate and is equivalent to the triangular distribution where the minimum value is 0, the most likely is 1 and the maximum is 1 (Triang(0, 1, 1).

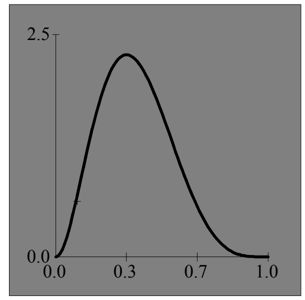


Properties of the Beta Distribution

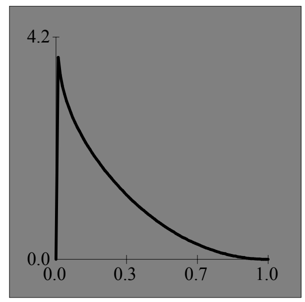
The beta variate β (1, 1) is a uniform or rectangular variate, i.e.,Uniform(0,1).

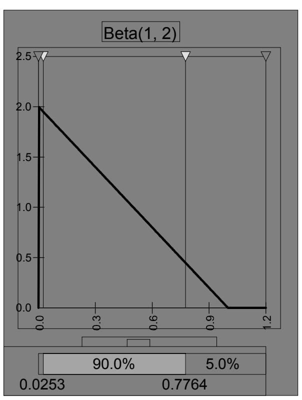


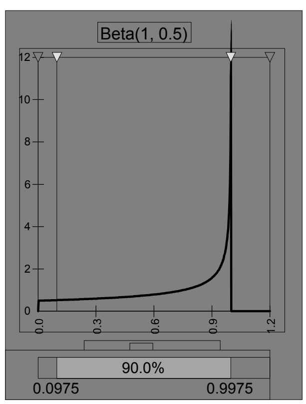
The beta variate where $\alpha_1 > 1$, $\alpha_2 > 1$ the beta density takes the shape such as that of β (3, 5).

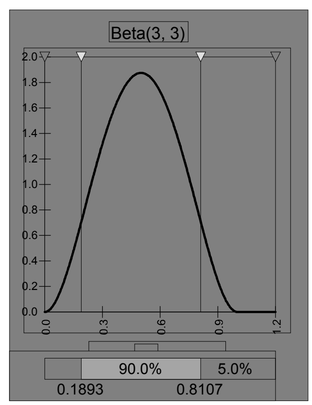


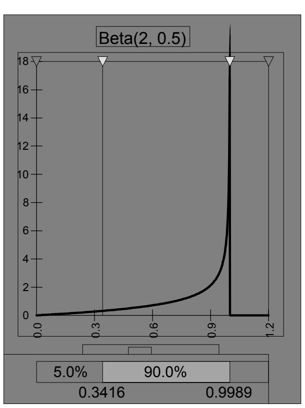
The beta variate where $\alpha_1 < 1$, $\alpha_2 > 1$ the beta density takes the shape such as that of β (0.8, 3).

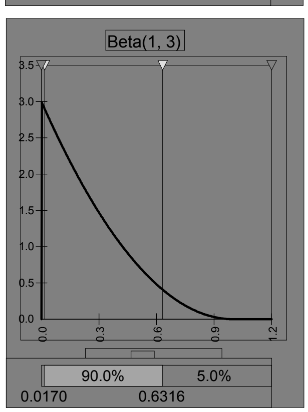


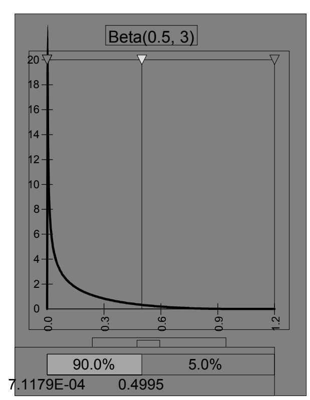








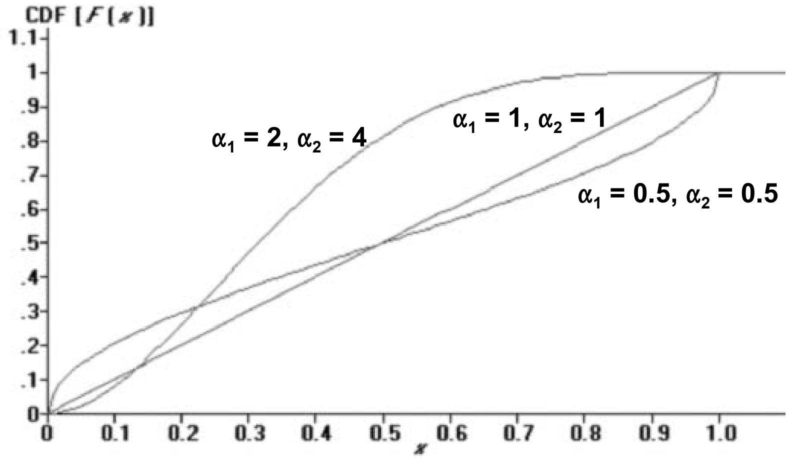




Properties of the Beta Distribution

Cumulative Distribution Function Curves

Distribution function curves for the beta variate $\beta(\alpha_1, \alpha_2)$.



Properties of the Beta Distribution

Probability density function:

$$f(x) = \frac{x^{\alpha_1 - 1}(1-x)^{\alpha_2 - 1}}{B(\alpha_1, \alpha_2)} \; where \; B(\alpha_1, \alpha_2) = \int_0^1 t^{\alpha_1 - 1}(1-t)^{\alpha_2 - 1} dt$$

Cumulative distribution function: no closed form (tables and numerical methods exist)

Parameters: $\alpha_1 > 0$, $\alpha_2 > 0$. Both are shape parameters.

Domain: $0 \leq x \leq 1$

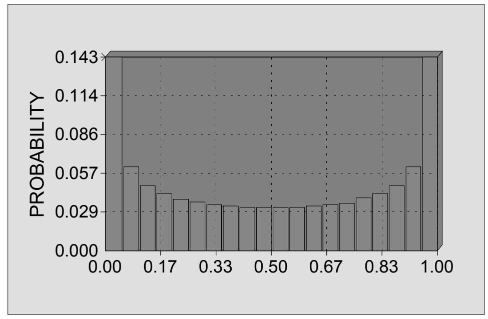
Mean (μ): $\alpha_1/(\alpha_1 + \alpha_2)$

Variance (σ^2): $(\alpha_1 \alpha_2)/[(\alpha_1 + \alpha_2)^2(\alpha_1 + \alpha_2 + 1)]$

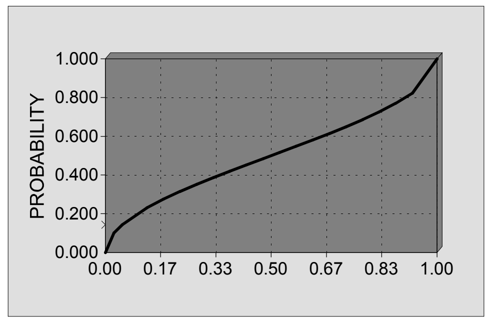
Mode: $(\alpha_1 - 1)/(\alpha_1 + \alpha_2 - 2)$ if $\alpha_1 > 1$, $\alpha_2 > 1$
0 and 1, if $\alpha_1 < 1$, $\alpha_2 < 1$
0, if $\alpha_1 < 1$, $\alpha_2 \geq 1$ or if $\alpha_1 = 1$, $\alpha_2 > 1$
1, if $\alpha_1 \geq 1$, $\alpha_2 < 1$ or if $\alpha_1 > 1$, $\alpha_2 = 1$
does not uniquely exist if $\alpha_1 = 1$, $\alpha_2 = 1$

Properties of the Beta Distribution

Probability density function for Beta (0.5, 0.5) as an output of @RISK simulation



Cumulative distribution function for Beta (0.5, 0.5) as an output of @RISK simulation



Properties of the Beta Subjective Distribution

Probability density function:

$$f(x) = f_B(x', \alpha_1, \alpha_2)$$

$$where\ x' = \frac{(x - \min imum}{(\max imum - \min imum)}$$

$$f_B = density\ of\ a\ beta\ distribution$$

$$\alpha_1 = \frac{(\mu - \min imum)(2 \times most\ likely - \min imum - \max imum)}{(most\ likely - \mu)(\max imum - \min imum)}$$

$$\alpha_2 = \alpha_1 \frac{(\max imum - \mu)}{(\mu - \min imum)}$$

Cumulative distribution function:

$$F(x) = F_B(x', \alpha_1, \alpha_2)$$

$$where\ x' = \frac{(x - \min)}{(\max - \min)}, F_B = density\ of\ a\ \ beta\ \ distribution$$

Parameters:

$$\min imum < most\ likely < \mu < \max imum, if\ most\ likey < \frac{(\min imum + \max imum)}{2}$$

$$\min imum < \mu < most\ likely < \max imum,\ if\ most\ likely > \frac{(\min imum + \max imum)}{2}$$

$$\min imum < most\ likely = \mu < \max imum,\ if\ most\ likely = \frac{(\min imum + \max imum)}{2}$$

Domain: minimum<x<maximum

Mean (μ):

Variance (σ^2): $((\alpha_1\ \alpha_2)/[(\alpha_1 + \alpha_2)^2(\alpha_1 + \alpha_2 + 1)])\times(\text{maximum-minimum})^2$

Properties of the Beta-Pert Distribution

Probability density function:

$$f(x) = f_B(x', \alpha_1, \alpha_2)$$
$$where\ x' = \frac{(x - \min imum}{(\max imum - \min imum)}$$
$$f_B = density\ of\ a\ beta\ distribution$$
$$\alpha_1 = \frac{(\mu - \min imum)(2 \times most\ likely - \min imum - \max imum)}{(most\ likely - \mu)(\max imum - \min imum)}$$
$$\alpha_2 = \alpha_1 \frac{(\max imum - \mu)}{(\mu - \min imum)}$$
$$\mu = \frac{\min imum + 4 \times most\ likely + \max imum}{6}$$

Cumulative distribution function:

Parameters: minimum<most likely<maximum

Domain: minimum<x<maximum

Mean (μ):

Variance (σ^2): $((\alpha_1\ \alpha_2)/[(\alpha_1 + \alpha_2)^2(\alpha_1 + \alpha_2 + 1)])\times(\text{maximum-minimum})^2$

Sun Rise Example

One of the questions posed by the mathematician Pierre-Simon Laplace was: What is the probability that the sun will rise tomorrow? The question can be rephrased as: Suppose that some event A occurred in all of a large number N of trials. What is the probability that it will occur again on the next trial?

The question can be answered by assuming the Bernoulli model of independent trials and equally likely structure (Laplace's Rule of Succession). It says that if some particular event has occurred in *n* consecutive trials, the probability is approximately (*n*+1)/(*n*+2) that it will occur on the very next trial. Part of the derivation is set out below.

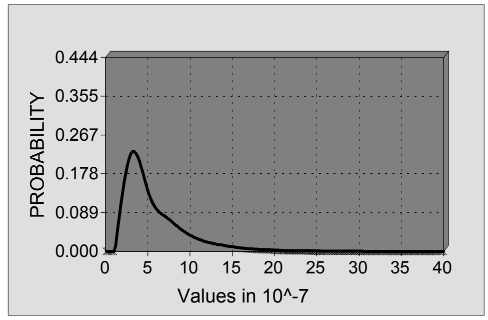
$$P(B) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{i}{N}\right)^{n}$$
$$P(A \cap B) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{i}{N}\right)^{n+1}$$
$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{i}{N}\right)^{n+1}}{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{i}{N}\right)^{n}}$$
$$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{i}{N}\right)^{n+1} \approx \int_0^1 x^{n+1}dx = \frac{1}{n+2}$$
$$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{i}{N}\right)^{n} \approx \int_0^1 x^{n}dx = \frac{1}{n+1}$$
$$P(A|B \approx \frac{1/(n+2)}{1/(n+1)} = \frac{n+1}{n+2}$$

Sun Rise Example

A rather conservative estimate of the time when we know that the sun was rising every day is about 7,000 years. Maybe it is possible that before the first written records, the sun did not rise but was (say) switched on suddenly in the sky or operated under some other principles. Thus letting n = 365 × 7,000 = 2,555,000 sunrises, then (n+1)/(n+2) ≈ 0.9999996.

An approach using the beta distribution is to estimate the beta shape parameters as $\alpha_1 = x+1$, $\alpha_2 = n+1-x$., where x = 0 and n = 2,555,000 for the probability distribution of failure. This beta distribution could be considered input fraction f_1. This distribution has a mean value of 3.91×10^{-7}. Therefore, the probability of interest is $(1 - f_1)$ which is the complement of the failure probability.

The graph at right depicts the beta distribution (@RISK simulation output) of the failure of the sun rising based on the evidence of success over 7,000 years.



Probability of failure

Dry Cured Ham and HCV Example

In the Parma ham-hog cholera virus (HCV) study conducted by McKercher et al. (1987) 12 pigs were inoculated intramuscularly with a 1 ml of 1:10 suspension of spleen and blood (105.3 PFU/g) in both a US and an Italian conducted experiment. There were 5 uninfected controls in the US experiment and 1 control in the Italian experiment. The pigs were slaughtered 5 days after inoculation. A 1 g sample of muscle, bone marrow and fat from each of 3 hams from inoculated pigs and from one of the control pigs was tested in triplicate at each sampling time. Each pig had been identified and each ham marked with the numbers 1 or 2. The method of identification was used so that no two hams from the same pig would be tested at a respective sampling period. In samples in which negative results were obtained, the fluids from the initial tests were subpassaged two additional times. Thus, negative samples would have been tested three times before being considered negative for the virus. When all 9 samples from inoculated pigs were negative, then portions of the samples were pooled for inoculation into two 20-30 kg pigs. When two tests performed at two consecutive time intervals were negative in vitro and confirmed negative in vivo in pigs, the experiment was considered terminated and the ham considered not to contain any infective virus after that respective time period. This amounted to 9 samples (one of muscle, bone and fat taken from each of 3 infected hams) at two time periods in both the US and Italian experiments, i.e., 36 samples tested in triplicate for the in vitro tests. These same samples were pooled and inoculated into two pigs in both the US and Italian experiments, i.e., 18 samples pooled and inoculated into each of two pigs for the two time periods in both the US and Italian experiments. The US phase of the experiment showed the loss of infectivity to pigs between 188-313 days while the Italian phase showed that the infectivity of HCV disappeared between 112-189 days.

Dry Cured Ham and HCV Example

<u>Question</u>: Can these experimental study results be used to estimate the probability of survival of hog cholera virus in the dry cured hams following processing?

Yes, the studies provide evidence of the survivability of hog cholera virus in the dry cured hams in the form of *x* successes in *n* trials. The probability of HCV survival can be modelled using the beta distribution.

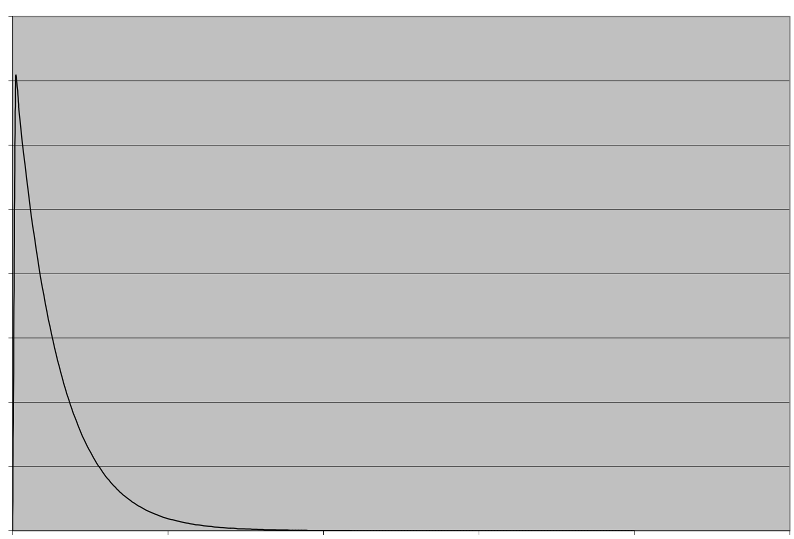
The experimental study of both the US and Italian phases combined thus indicated that n = 72 samples and x = 0. The two detection systems, *in vitro* and *in vivo* (pig inoculation), are sufficiently different to exclude any intra-test correlation in repeated tests. The sensitivity of each test was considered to be 100%. Parameters of the beta distribution can be estimated as follows:

$\alpha_1 = x+1$

$\alpha_2 = n+1-x$.

The probability of HCV survival in the dry cured ham is thus $\beta(1, 73)$.

Dry Cured Ham and HCV Example



Beta distribution curve β (1, 73) of the survival of hog cholera virus in dry cured hams.

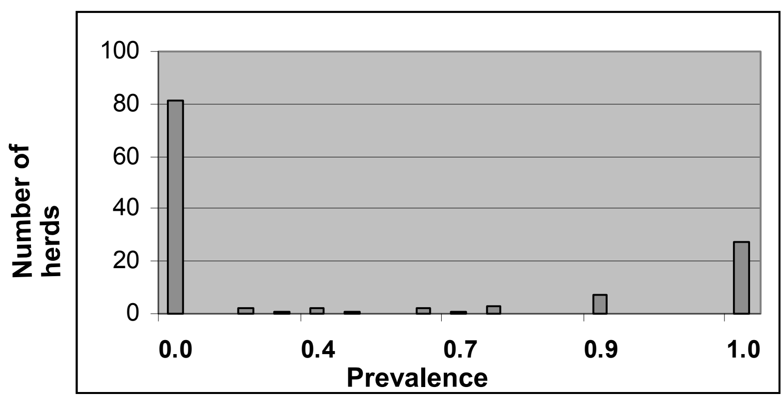
Pseudorabies Within-herd Prevalence Example

Estimates of the prevalence of the within-herd prevalence of pseudorabies in growing and finishing pigs in infected herds were obtained from serological studies. The table at right presents the frequency of 127 herds as to prevalence. None of the pigs had been vaccinated against pseudorabies.

For a risk assessment on the importation of finished pigs from a pseudorabies infected country, what probability distribution can be used to model the within-herd prevalence of exposure to pseudorabies virus?

Frequency	Prevalence
81	0
1	0.03
1	0.1
1	0.3
2	0.4
1	0.5
1	0.57
1	0.6
1	0.7
1	0.8
1	0.82
1	0.83
6	0.9
1	0.93
1	0.95
2	0.97
24	1
127	

Pseudorabies Within-herd Prevalence Example



Graph of frequency distribution of 127 herds according to pseudorabies serological prevalence in growing and finishing pigs.

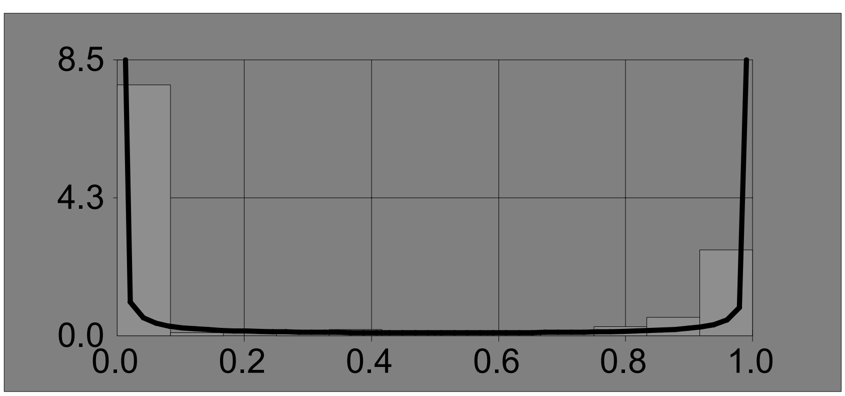
Pseudorabies Within-herd Prevalence Example

A bimodal distribution of prevalence is apparent from the frequency distribution of the seroprevalence in the 127 herds. The mean prevalence is 0.309213 and the standard deviation is 0.438373. A two parameter beta distribution is a useful probability distribution for representing a prevalence, probability or proportion bounded by 0 and 1. The shape parameters α_1 and α_2 were estimated with the following expressions:

$$\alpha_1 = \mu\left(\frac{\mu(1-\mu}{\sigma^2} - 1\right) \qquad \alpha_2 = \left(1-\mu\right)\left(\frac{\mu(1-\mu}{\sigma^2} - 1\right)$$

The within-herd prevalence of exposure to pseudorabies virus was modelled using the beta distribution β(0.034554, 0.077157).

Pseudorabies Within-herd Prevalence Example



Comparison of the input and beta β(0.034554, 0.077157) distributions of within-herd prevalence of exposure to pseudorabies virus.

Bayesian Update Example

The seroprevalence of some disease in a population of export animals is the parameter of interest. In this example, the prior probability distribution is based on no prior information. The unknown population proportion can take on any value between 0 and 1, hence, p is a continuous random variable. The probability density function for the beta distribution is as follows:

$$f(p) = \frac{(\alpha_1 + \alpha_2 - 1)!}{(\alpha_1 - 1)!(\alpha_2 - 1)!} p^{\alpha_1 - 1}(1 - p)^{\alpha_2 - 1}$$

With no information, the choice for α_1 and α_2 is to let both parameters equal 1. Substituting this value for the two parameters above equates to 1, and the shape of the distribution can be represented by a horizontal line. This is equivalent to Uniform(0, 1), implying that every value of p between 0 and 1 is equally likely. Although the probability that a continuous random variable takes on a specific value equals zero, one should only speak about the probability that p lies within intervals. The uniform distribution indicates that the probability that p falls in an interval of given length is the same no matter where the interval is located in the range from zero to one.

Bayesian Update Example

Bayes' Rule specifies how the prior probability distribution should be updated to produce a posterior probabiltiy distribution that accounts for the additional data. If P(A) is the prior probability of the population prevalence, P(E) is the probability that the evidence E will be observed and P(E|A) is the conditional probability that the evidence will be oberved given that P(A) is correct, then Bayes' Rule states that the posterior probability that A is true given that E has been observed is:

Prior

Posterior

$$P(A|E) = P(A) \times \left[\frac{P(E|A)}{P(E)} \right]$$

Correction factor

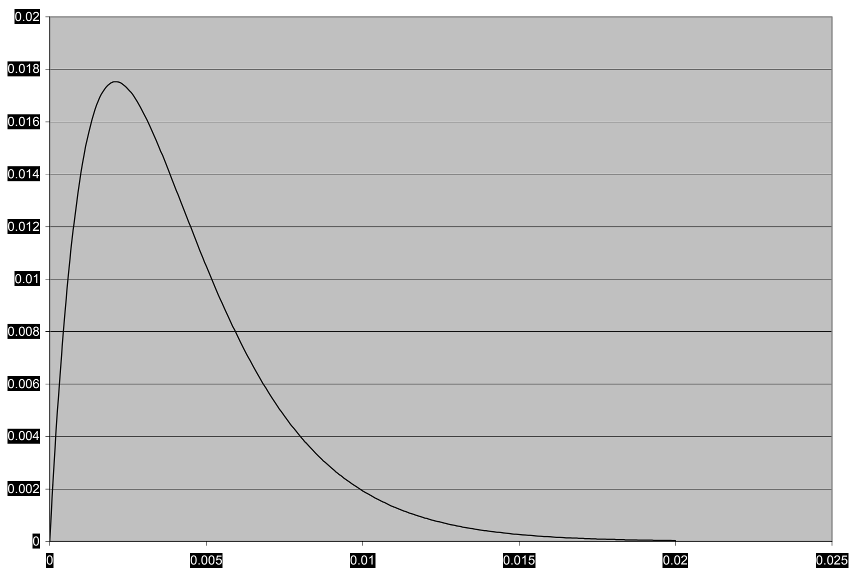
The correction factor comprises P(E|A) which is often called the likelihood since it specifies the likelihood of obtaining the evidence given the prior information. The denominator of the correction factor is regarded as a normalizing factor, because the posterior probabilities must sum to unity. Thus Bayes' Rule states that the posterior probability is proportional to the prior probability and the likelihood of the evidence. In this example, the additional evidence is a sample of the export animal population in which 1 test of 500 animals tested is seropositive. In the second column of the table on the following slide, the prior probability P(A) is indicated. P(E|A) is derived from the binomial distribution in which $x = 1$ and $n = 500$. P(A)P(E|A) is simply the product of the prior and conditional probability. The posterior probability is obtained by dividing P(A)P(E|A) by the sum so that the posterior probabilities sum to unity.

Bayesian Update Example

Microsoft Excel function BINOMDIST(1, 500, *p*(Se)+(1-*p*)(1-Sp),0) where Se = 0.95 and Sp = 1.00 and *p* is the discretized prevalence.

Prevalence	P(A) Prior Distribution	P(E|A)	P(A)P(E|A)	P(A|E)
0	1	0	0	0
0.0001	1	0.045300699	0.045300699	0.002156
0.0002	1	0.086406066	0.086406066	0.004113
0.0003	1	0.123606945	0.123606945	0.005884
0.0004	1	0.157176289	0.157176289	0.007482
0.0005	1	0.187370195	0.187370195	0.00892
0.0006	1	0.214428873	0.214428873	0.010208
0.0007	1	0.238577565	0.238577565	0.011357
0.0008	1	0.260027415	0.260027415	0.012378
0.0009	1	0.278976289	0.278976289	0.01328
0.001	1	0.295609548	0.295609548	0.014072

Bayesian Update Example



The posterior probability P(A|E) is represented as a beta distribution as portrayed above.