Iran. J. Plant Path., Vol. 48, No. 4, 2013: 161-168

FDD – A USEFUL MS EXCEL *ADD-IN* FOR FITTING DISCRETE DISTRIBUTIONS TO DISEASE INCIDENCE, WEED, INSECT AND NEMATODE FREQUENCY DATA

H. HAMZEHZARGHANI[1]* and A. TAHAVVOR[2]

(Received : 21. 10. 2011; Accepted : 14. 7. 2012)

Abstract

An *add-in* for MS Excel 2007 was developed to fit discrete distributions to the frequency of disease incidence, weed, insect and nematode in the form of count per sampling unit data. Probability distributions such as binomial, Poisson, negative binomial and beta binomial are discrete distributions, which are respectively appropriate for describing uniform, random and aggregated or clustered binary data such as disease incidence or insect/nematode/weed count per quadrate observations. The *fit discrete distribution* (FDD) program is a very user-friendly and flexible *add-in* and can take data in the form of either Table of raw observations or a frequency Table. The program estimates distribution parameters and their standard errors using a maximum likelihood procedure and determine the expected (theoretical) values of the distribution for any distribution from its list. It also calculates both chi-square and log-likelihood ratio goodness of fit statistics to test the null hypothesis of goodness of fit and plot the expected frequencies according to all used distributions against observed values in one single bar graph for comparison purposes. Application of the program in studying spatial pattern of plant pests is discussed.

Keywords: Spatial pattern, Discrete distribution fit, Negative binomial, Binomial, Beta binomial, Poisson.

*: Corresponding Author, Email: zarghani@shirazu.ac.ir
1. Assis. Prof. of Plant Pathol., College of Agriculture, Shiraz University., Shiraz, Iran.
2. Department of Agricultural Economics, College of Agriculture, Shiraz University, Shiraz, Iran.
A copy of the FDD program can be via email Corresponding from author.

Introduction

Sampling is fundamental in plant pathology, weed science and agricultural entomology. In plant pathology sampling is necessary to improve our understanding of spatio-temporal dynamics of epidemics. Sampling is also of crucial importance when information is used to inform disease management decisions (Binns*et al*. 2000; Pedigo and Buntin 1994). The analysis of spatial patterns of diseased plants, weeds, plant parasitic nematodes and insect pests is of prime importance in decision making for their management. Information on the spatial patterns of diseased individuals for instance may be used to transform data to meet statistical assumption for assessing treatment effect and also to calculate minimum sample size for development of sampling protocols that support accurate and precise estimates of the mean disease intensity (Madden et al., 2007).Estimation of population parameters such as mean and variance of disease intensity/weed or pest density in a field is critical for their reliable estimation. Such estimates might be used for prediction of yield losses caused by the disease/weed or pest and hence are extremely important for management of economically important plant diseases, weeds and pests. As sampling needs resources, it is important that the data obtained by sampling meet reliability requirements which requires careful planning based on statistical descriptions. Among many approaches used to characterize spatial patterns, a popular one is fitting discrete probability distributions to the data (Kish, 1995; Perry, 1994; Cochran, 1977). The Poisson and negative binomial distributions can be fitted to counts of lesions, diseased plants, insects, nematodes and weeds per sampling unit. Based on an appropriate measure of goodness of fit, if the negative binomial distribution is the best fit to the observations, there is evidence of a clustered (or aggregated) pattern where the degree of aggregation in this case can be measured by the estimated k parameter of the distribution. If Poisson distribution is the best fit, then this is an indication of random pattern (Madden *et al*. 2007). While Poisson and negative binomial distributions may be considered as suitable for analyzing count data of insects, nematodes and weeds per sampling unit, some believe that fitting such distributions to disease incidence data, considering the binary nature of disease incidence, may be misleading (Hughs and Madden 1993). For binary random variables such as disease incidence, binomial is an appropriate probability distribution provided the pattern of the studied items is random. In most cases, however, because of the clustered nature of disease incidence data at sampling unit level, the beta-binomial distribution can be a good alternative (Roumagnac *et al.* 2004; Griffiths 1973). The equations and parameters for the Poisson, binomial, negative binomial and beta-binomial distributions are shown in Table 1. In all equations 1-4 in Table 1, Pr(Y) denotes the probability that a sampling unit contains Y entries (diseased plants, nematodes, insects or weeds). Poisson and binomial distributions have one parameter, probability of occurrence of an entry (diseased plant, insect or weed) per sampling unit (μ and p , respectively) with n=the total number of entries (plants or plant units, insects, weeds, etc…) being observed in a sampling unit. Negative binomial has three terms: μ and Y are as described before and k is an aggregation parameter ($k > 0$ where aggregation at the sampling unit increases as k decreases).

Gates and Ethridge (1970) developed a FORTRAN computer program to fit the Poisson and negative binomial distributions to data which its new version was called DISCRETE. This program can fit binomial distribution; but it fails to fit beta-binomial distribution to data. To estimate the parameters of beta-binomial distribution using maximum likelihood method, a FORTRAN subroutine was published by Smith (1983).These DOS-based programs could not be controlled by the user without revising the FORTRAN source codes and were unable to calculate the expected frequencies. Madden and Hughes (1993) wrote a MINITAB macro to use Smith algorithm and input n and estimates of θ and p to compute the beta-binomial expected frequencies and also a χ^2 goodness of fit test. Recent software such as EasyFit®(mathwave) and BestFit® (Palisade Product #1006), which provides user- friendly and easy to use interfaces, have been developed that require the user to buy the software from the publisher. Common statistical software like SAS does not offer a special procedure to fit discrete distributions and calculate their expected frequencies along with acceptable goodness of fit tests and requires the user to know writing long

Table 1. Probability distributions and their parameter(s) used to study the spatial aspects of disease epidemics and dispersion pattern of insects, plant parasitic nematodes and weeds at the sampling unit level

Probability Distribution	Equation of probability generating function		Parameter(s)
Poisson	$\Pr(Y) = \frac{\mu^Y e^{-\mu}}{Y!} \quad Y = 0,1,2,...$	(Eq.1)	μ
Binomial	$\Pr(Y) = \binom{n}{Y} p^Y (1-p)^{n-Y}, \binom{n}{Y} = \frac{n!}{Y!(n-Y)!}, \; Y = 0,1,..,n$	(Eq.2)	p
Negative binomial	$\Pr(Y) = \left(\frac{\Gamma(k+Y)}{\Gamma(k)\Gamma(Y+1)}\right)\left(\frac{\mu}{k}\right)^Y \left(1+\frac{\mu}{k}\right)^{-(k+Y)}$	(Eq.3)	μ and k
Beta-binomial	$\Pr(Y) = \binom{n}{Y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+Y)\Gamma(\beta+n-Y)}{\Gamma(\alpha+\beta+n)}$	(Eq.4)	α and β

macros which is impractical for plant pathologists, entomologists and weed scientists. To facilitate the analysis of most commonly used discrete distributions in agricultural research and providing a summary of outputs on parameter estimation, calculation of expected values, goodness of fit tests and production of accessory bar graph of the observed versus theoretical frequencies (Figure 4), we wrote an MS-Excel-based macro that can be added to Excel as *add-in* and easily used to fit discrete distributions. The purpose of this article is to describe the data input and output and generally how a user with simple basic knowledge on MS-Excel can install the *add-in* with an example of the use of the macro.

]

Fit Discrete Distribution (FDD) macro

Microsoft Excel is one of the powerful spreadsheet software that can program with VBA (Visual Basic for Applications).This program is written in Microsoft Excel VBA and can be added to Excel by *add-in*. One needs office 2007 or higher version to use this program.

The form of *binomial distribution* used for fitting purposes is presented in Table 1 Eq.2.In binomial distribution three methods are defined for estimation of *n* and *p* parameters. The first method is MLE that uses the highest observed number as an acceptable guesstimate of *n* and *p* equals $mean/n$. The method of maximum likelihood is based on the principle that the best estimate of the population parameters is the estimate which maximizes the probability of obtaining the observed sample. In the second method, *n* is defined by the user and again $p = mean/n$. The third technique or the method of moments computes least square estimation of parameters of binomial distribution (i.e. *n* and *p*) and is not recommended by statisticians.

Parameter of *Poisson distribution* (Eq.1 in Table 1) or μ can also be estimated via two ways, either using the MLE method where μ is considered the mean of the data or by least square technique as stated for binomial distribution.

The *negative binomial distribution* is a mixture of the Poisson and the Gamma distributions with a basic form of:

$$\Pr(Y) = \binom{Y+k-1}{k-1} p^k (1-p)^Y$$

If distribution parameter p is reparameterized as $k/(\mu+k)$, the form of negative binomial distribution presented as Eq.3 in Table 1 is derived. In this form, μ refer to the mean of the data and *k* is often referred to as an over-dispersion parameter. Now it will be possible to estimate μ and *k*. A MLE estimation of μ is the mean of observations and an estimate of *k* can be obtained by Newton Raphson method (Park & Lord 2008).

The *Beta-Binomial distribution* is also a mixed

	A
1	0
2	0
60	0
61	1
85	1
86	2
98	2
99	3
100	3
101	3
102	4
103	4
104	5
105	6
106	8

	A	B
1	X	Frequency
2	0	60
3	1	25
4	2	13
5	3	3
6	4	2
7	5	1
8	6	1
9	7	0
10	8	1

Fig. 1. Example input data sets for use with FDD program. Left set: data on individual sampling units recorded in one column of an Excel worksheet. To save space observations between rows 61 and 85 and also between rows 86 and 98 were omitted. Right set: Frequency of observations mode which is recorded in two columns. Heather of column A in the left set could be the number of items (diseased plants, weeds, or insects) per sampling unit (quadrate/plot). In the right set "X" in column A is the numbers of items per sampling unit and in column B (labeled with "Frequency") the frequencies are entered, for example 60 sampling units contained no diseased plants.

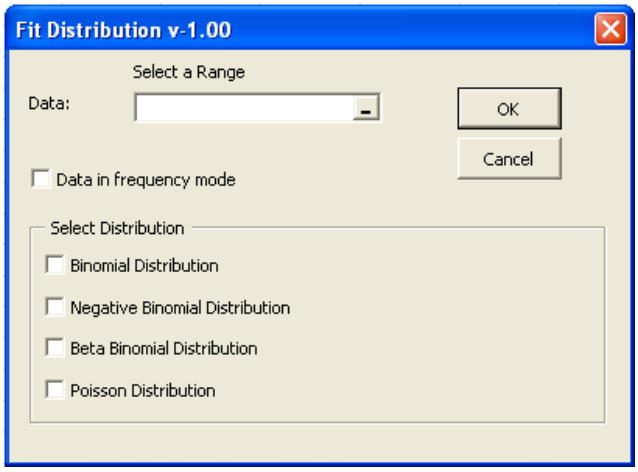
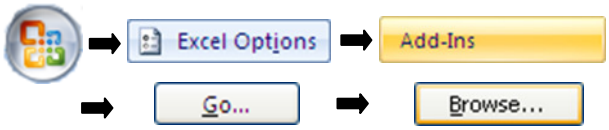


Fig. 2. Fit Distribution Input data and options dialog box. The range of the data is selected using the mouse by pressing the button on the left of the "Select a Range" data box if it is in raw data format or directly entering the addresses of the cells that contain the data. If data are arranged in a frequency table, the box on the right of "Data in frequency mode" must be ticked to activate a second data box called "Frequency" where the address of the cells that contain frequencies must be entered in this box. In the "Select Distribution" section one can choose as many as probability distributions as required.

distribution of the Beta and the Binomial distributions. The probability function of the Beta-binomial distribution is shown in Table 1 (Eq.4). It is convenient to parameterize $\mu = \alpha/(\alpha+\beta)$ and $\theta = 1/(\alpha+\beta)$ because parameters μ and θ are more meaningful. Then μ is the mean of the Binomial parameter p and θ is a scale parameter which measures the variation of p. There are two main approaches to estimate the parameters μ and θ for the Beta-Binomial distribution. One approach involves moments and the other involves maximum likelihood (ML). ML estimation is a more efficient parameter estimation method than moment estimation (Ennis and Bi 1998). Following parameter estimation, expected values of distribution are computed and used to calculate a χ^2 and/or log-likelihood ratio statistics to test the goodness of fit between the observed and expected frequencies. The significant level of the χ^2 and log-likelihood ratio value may then be calculated by checking it against a chi-square distribution with appropriate degrees of freedom.

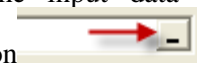
Running the FDD* add-in *in MS Excel and Simulation results

To run FDD, first the *add-in* must be added to Microsoft Excel, to do that, one needs a copy of a file named*FitDistribution.xlam* which contains all the source codes and algorithms. In the second step, one needs to follow the route shown below to add FDD to MS Excel:



and paste a copy of *FitDistribution.xlam* file after selecting and finish installation by pressing OK button. Finally in the *add-in* window, check mark *FitDistribution* and click OK button. The program will now be ready for use. The input data can be read or entered in either of the two possible formats; a table of raw observation or a table of frequency data (Figure**Error! Reference source not found.** 1).

To run FDD, one needs simply to press Ctrl+Shift+D or run it from Developer tab → Macros and run *FitDistribution*, and then the FDD dialog box appears (2).

The input data is selected by clicking on button  and then selecting the range of cells that contain data. If the data is arranged in frequency mode, first cells showing the number of observations must be selected followed by putting a check mark besides the "Data in frequency mode" option (

) and then selecting cells that contain frequency of each observation.

The next step in the "Select Distribution" section of the dialog box is choosing the distribution(s) of interest to fit to the data. One can choose any of the three methods for Binomial distribution for Poisson distribution and either of the two ways to estimate parameters of the distribution of interest as described before. After selecting the distribution(s) of interest, just one click is required to run the program. If no error occurs during processing, output will appear in a new worksheet (Figure 3).

In the first two columns of the output sheet, FDD reproduces the observed data in the frequency table mode regardless of the type of input data used (Figure 1). In the next couple of columns FDD gives the expected values (relative and absolute) based on the probability distributions selected and finally a graphical representation of this table is also plotted in the form of a bar graph (Figure 4). In the middle rows, FDD gives the parameter estimates and their standard errors and finally tests goodness of fit based on chi-square distribution and log-likelihood ratio in the bottom rows of the output worksheet (Figure 3).

DISCUSSION

We found the FDD program very useful in fitting Poisson, negative binomial, binomial and beta binomial discrete distributions. As the program can be easily added to Microsoft Excel 2007 as an *add-in*, it is expected to provide a relatively easy tool for researchers in the field of agriculture who do not want to be involved in a lot of programming for fitting distributions including parameter estimation and goodness of fit statistics and test of randomness or aggregation. In a recent study on sampling optimization for root lesion nematodes [*Pratylenchus neglectus* (Rensch) Filipjev et Schuurmans Stekhoven. and *P. thornei* Sher et Allen.] in the irrigated wheat fields of Marvdasht region, FDD program was used to Fit Poisson and negative binomial discrete distributions to the

	A	B	C	D	E	F	G	H	I	J	K
1	X	Observed Value	Relative Observed Value	Relative Expected Binomial	Absolut Expected Binomial	Relative Expected Poisson	Absolut Expected Poisson	Relative Expected Negative Binomial	Absolut Expected Negative Binomial	Relative Expected Beta Binomial	Absolut Expected Beta Binomial
2	0	60	0.5660	0.4206	44.59	0.4401	46.65	0.5706	60.49	0.5779	61.25
3	1	25	0.2358	0.3847	40.78	0.3612	38.29	0.2284	24.21	0.2058	21.81
4	2	13	0.1226	0.1539	16.32	0.1482	15.71	0.1042	11.05	0.1053	11.16
5	3	3	0.0283	0.0352	3.73	0.0406	4.30	0.0495	5.25	0.0568	6.02
6	4	2	0.0189	0.0050	0.53	0.0083	0.88	0.0240	2.54	0.0301	3.19
7	5	1	0.0094	0.0005	0.05	0.0014	0.14	0.0117	1.25	0.0149	1.58
8	6	1	0.0094	0.0000	0.00	0.0002	0.02	0.0058	0.61	0.0065	0.69
9	7	0	0.0000	0.0000	0.00	0.0000	0.00	0.0029	0.31	0.0023	0.24
10	8	1	0.0094	0.0000	0.00	0.0000	0.00	0.0014	0.15	0.0005	0.05
11											
12		Mean=	0.8208	mean=	0.8208	mean=	0.8208	mean=	0.8208	mean=	0.8407
13		Variance=	1.8057	Variance=	0.7365	Variance=	0.8208	Variance=	1.6832	Variance=	1.6645
14		n=	106								
15				Probility=	0.1026	Mean=	0.8208	Mu=	0.8208	n=	8
16				n=	8			S.E. Mu=	0.1260	Mu=	0.1051
17								k=	0.7811	S.E. Mu	0.0158
18								S.E. k=	0.2663	Teta=	0.2095
19										S.E. Teta=	0.0673
20			Test H0: Goodness of fit								
21			chi square		769001.2126		4259.4446		6.7762		20.0868
22			Prob > Chi-Square		0.0000		0.0000		0.3420		0.0012
23											
24			Log Likelihood		54.1440		33.4667		4.8616		7.9461
25			Prob > Chi-Square		0.0000		0.0000		0.5617		0.1592
26											
27			degrees of Freedom		6		7		6		5

Fig. 3. Example of output worksheet from FDD program. Data are an example in Radjabi, 2009 on counts of insects per quadrate. For more details on the output please see the text. In this example according to Log Likelihood GOF statistics both negative binomial distribution and beta binomial distribution show good fit (see *p>0.05* in cells I25 and K25), however according to Chi Square GOF statistics only negative binomial had good fit to the data (see *p>0.05* in cell I22). As the data are in the count of insects and not disease incidence with binomial nature, negative binomial is best distribution to describe the observations. Absolute expected values of the number of insects per quadrate are in cells I02 to I10.

number of nematodes per 100 g soil using FDD (Ghaderi *et al*., 2012). The result showed that neither model did fit to the data indicating deviation of spatial pattern of the nematodes from both random and aggregated at the field level. It was concluded that presumably tillage operations that are frequently done for seed-bed preparation and also movement of equipment through the field, may have caused repeated redistribute of root lesion nematodes leading to an intermediate status between uniform, random and aggregated distributions. The *add-in* has also been used to study the spatial pattern of Fababean plants infected to Faba bean necrotic yellows virus (FBNYV) in Fababean fields near Shiraz (unpublished data). Preliminary results showed that the spatial pattern of infected plants changed from aggregated to random towards the end of the season that may suggest dependency of spatial pattern on the behavior of the vector. These findings were verified applying Taylor power law model to the data and interpreting the estimated parameters of the model. In another survey the spatial pattern of lime trees infected to *Candidatus* Phytoplasma aurantifolia

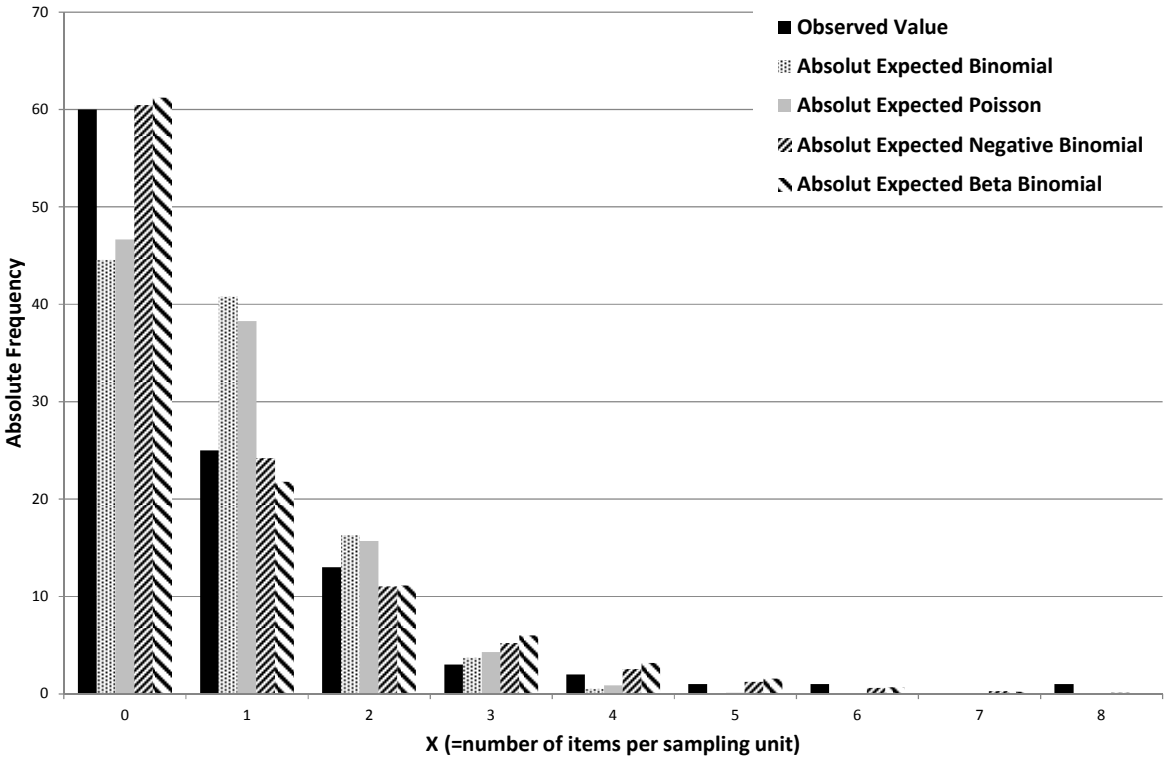


Fig. 4. A sample Bar chart for observed and expected frequencies of binomial, Poisson, Negative binomial and Beta binomial distributions outputted with FDD. The third bars show the negative binomial expected values which show excellent agreement with the first bars (the observed values). For details please refer to the text.

causal agent of Witches-broom disease of lime (WBDL) in orchards across Hormozgan province was investigated (unpublished data). Fitting statistical probability distributions (binomial and Beta-binomial) to the frequency of symptomatic trees per quadrate proved a very good fit of Beta-binomial distribution to the data which supports the aggregative nature of WBDL spatial pattern.

FDD provides a helpful and easy to use interface making a research project more productive and efficient. The automatic fit distribution to the data in seconds is brought about by a powerful parameter estimation algorithm. Presently FDD supports the discrete distributions that are most commonly used in the fields of plant pathology, entomology and weed sciences and can be easily added to Microsoft Excel 2007. It is possible to extend the features of FDD in the future including adding more discrete and continuous probability distributions and specifying the custom parameters manually. The goodness of fit (GOF) tests measures the compatibility of a random sample with a theoretical probability distribution function which is an indication of how well the selected distribution fits to the data. FDD supports the Chi-Squared and Log-likelihood GOF tests.

Reference

BINNS, MR., NYROP, JP. and VAN DER WERF, W. 2000. **Sampling and Monitoring in Crop Protection: The Theoretical Basis for Developing Practical Decision Guides**. CABI Pub., Oxon, UK.

CAMPBELL, CL. and L.V. MADDEN.1990. **Introduction to Plant Disease Epidemiology.** John Wiley and Sons, Inc., New York.

COCHRAN, WG. 1977. **Sampling Techniques**. 3rd ed., John Wiley and Sons Inc., New York.

ENNIS, D.M. and BI J. 1998. The Beta-binomial model: accounting for inter-trial variation in replicated difference and preference tests. **J. Sensory Stud.** 13: 389-412.

GATES, CE. and ETHERIDGE, F. G. 1970. A generalized set of discrete frequency distributions with FORTRAN program. **Math. Geol.** 4:1–24.

GRIFFITHS, D. A. 1973. Maximum likelihood estimation for the beta-binomial distribution and an application to the household distribution of the total number of cases of a disease. **Biometrics** 29: 637-648.

HUGHES, G. and MADDEN, L.V. 1993. Using beta-binomial distribution to describe aggregated patterns of disease incidence. **Phytopathology** 83: 759-763.

JULIUS, S. 2004. Letter to the editors of Biometrics. **Biometrics** 60:285.

KARANDINOS, MG. 1977. 1976. Optimum sample size and comments on some published formulae. **Bull. Entomol. Soc. Amer.** 22: 417-422.

KISH, L. 1995. **Survey Sampling**. John Wiley and Sons Inc., New York.

KREBS, C.J. 1989. **Ecological Methodology**. 2nd ed., HarperCollins Pub. Inc., New York.

MADDEN, L.V., HUGHS, G. and VAN DEN BOSCH, F. 2007. **The Study of Plant Disease Epidemics.** APS Press, St. Paul, MN.

MADDEN, LV. and HUGHS, G. 1993.BBD-Computer Software for Fitting the Beta Binomial Distribution to Incidence Data. **Plant Dis.** 78(5): 536-540.

MADDEN, LV. and HUGHS, G. 1999. Sampling for plant disease incidence. **Phytopathology** 89: 1088-1103.

PARK, B. J. and LORD D. 2008. Adjustment for maximum likelihood estimate of the negative binomial dispersion parameter. **J. Trans. Res. Board** 2061: 9-19.

PEDIGO, LP.and BUNTIN, GD. 1994. **Handbook for Sampling Methods for Arthropods in Agriculture**.CRC Press, Boca Raton, FL.

PERRY, J. 1994. Sampling and applied statistics for pests and diseases. **Asp. Appl. Biol.** 37: 1-14.

RADJABI, GH. 2009. **Insect Ecology, Applied and Considering the Condition of Iran**. 2nd ed., Agricultural Research, Education and Extension Organization (AREEO), 648pp.

ROUMAGNAC, P. PRUVOST, O. CHIROLEU, F. and HUGHES, G. 2004. Spatial and temporal analyses of bacterial blight of onion caused by *Xanthomonas axonopodis* pv. *allii.* **Phytopathology** 94:138-146.

RUESSINK, WG. 1980. Introduction to sampling theory. Pp.61-78. *In:* M.Kogan and D.C. Herzog (Ed.), **Sampling Methods in Soybean Entomology**. Springer Pub., New York,

SIMON, L.J. 1961. Fitting negative binomial distributions by the method of maximum likelihood. **Proc. Casual Actuary Soc.** 48: 45-53

SMITH, D. M. 1983. Maximum likelihood estimation of the parameters of the beta binomial distribution. **Appl. Stat.** 32: 196-204.

SOUTHWOOD, T.R.E. 1978. **Ecological Methods**. 2nd ed., Chapman and Hall Pub., London.

STRANDBERG, J. 1973. Spatial distribution of cabbage black rot and the estimation of diseased populations. **Phytopathology** 63: 998-1003.