

## The Analysis of Data from Studies of Crop-Raiding

Richard F. W. Barnes

*Division of Biological Sciences, University of California at San Diego, La Jolla, USA*

### Introduction

Do elephants raid fields at random? This is a fundamental question for wildlife managers. If the answer is “yes,” then the manager must find ways to discourage elephants from every field. But if the answer is “No, elephants select certain fields for special attention while ignoring others,” then cost-effective management becomes possible. For example, we can concentrate protection on fields with the risk factor we have identified, or we can work with farmers to reduce that risk factor to make their fields less attractive to elephants.

In an earlier paper I described the methods for collecting data to answer this question (Barnes 2008). In the present paper I describe simple methods for analyzing those data. This paper is written with the needs of graduate students in mind. Many graduate students live in remote villages to collect data in order to help the farmers reduce their losses from crop-raiding elephants. The lack of funds, the absence of electricity or the humidity means that often they cannot use even a laptop. So first I present simple methods that require only a calculator. They will enable students to get answers while still in the village or while living in a tent. Then I move to the logistic model that requires access to a computer. In a subsequent paper I will deal with more complex multivariate methods of analysis that make more use of the data.

Before sitting down to start your data analysis, you must remind yourself of the original goals of the study, and you must have a clear idea of the questions you are expecting to answer.

### The odds ratio

Let us start with a simple question: does the probability of a farm being raided depend upon the presence of a particular crop? In other words,

do farms with this crop have a greater risk of being visited by elephants compared to farms without the crop? This question arose on our Kakum study site (Barnes *et al.* 2005) with respect to plantains: it seemed to us that elephants were more likely to raid farms with plantains. At certain times of year elephants sought out the plantains because of their fleshy stems and then stayed on the farm to eat other crops. The null hypothesis we are testing is therefore: there is no difference in risk between farms with plantains and those without. We select the 0.05 level of significance.

The risk that a farm will be raided is a probability. The odds compare the probability of an event (such as a farm being raided) occurring relative to the probability that the event does not occur.

$$\text{Odds} = \frac{\text{Probability of being raided}}{\text{Prob. of not being raided}} = \frac{P}{1 - P}$$

For example, if there is a 20% probability ( $P = 0.2$ ) that my farm will be raided during the growing season, then there is an 80% probability ( $1 - P = 0.8$ ) that it will not. The odds that my farm will be raided are  $0.2/0.8=0.25$ .

Imagine that you set out to test the hypothesis that it is the presence of plantains on a farm that attracts elephants. You can draw up a  $2 \times 2$  table to compare the frequencies of raided and intact farms against the frequencies of farms with and without the risk factor of interest, which in this case is plantains.

		Farms raided	
		Yes	No
Plantains	Yes	<i>a</i>	<i>b</i>
	No	<i>c</i>	<i>d</i>

In this example *a* farms had plantains and were raided while *b* farms that also had plantains were untouched. On the other hand, there were *c* farms without plantains that were raided compared to *d*

that had no plantains and remained unscathed.

The odds that a farm with plantains was raided is  $a/b$ . Similarly, the odds that a farm without plantains was raided is  $c/d$ . The odds ratio, OR, is then:

$$OR = (a/b)/(c/d) = ad/bc$$

The odds ratio is a measure of how more likely it is for an event to occur for farms with plantains than farms without plantains. An odds ratio of 1 means no association: the odds of being raided are independent of whether or not plantains are present. An odds ratio of greater than 1 indicates an association between plantains and the probability of crop damage: elephants are attracted to fields with plantains. An odds ratio less than 1 would suggest that elephants avoid farms with plantains. In theory the odds ratio can have values from zero to infinity. Usually you will find it varies from 0.2 to about 20 in extreme cases.

Below is a real example of 203 farms, some of which had plantains. We suspected that plantains were a major risk factor that drew elephants on to those farms. The null hypothesis says there was no association between plantains and raids.

		Farms raided	
		Yes	No
Plantains	Yes	14	63
	No	15	111

A superficial glance might suggest that it doesn't matter whether or not you have plantains: almost equal numbers of damaged farms fell into the plantains present/absent categories. However, it is a cardinal rule in crop raiding studies that one must also look at the farms that were not raided, those in the right hand column.

We first calculate the odds that a farm with plantains was raided. The odds =  $P_1/(1-P_1) = 14/63 = 0.2222$ . The odds that a farm without plantains was raided =  $P_0/(1-P_0) = 15/111 = 0.1351$ .

The odds ratio is therefore:

$$OR = \frac{P_1/(1-P_1)}{P_0/(1-P_0)} = \frac{14/63}{15/111} = \frac{0.2222}{0.1351} = 1.64$$

Thus farms with plantains were 1.64 times more likely to be raided than those without plantains. Is this value significantly different from unity, the "null value" indicating no association? If the confidence limits lie outside 1, then we can conclude that the estimated OR is significant. The OR is not normally distributed, whereas its natural logarithm,  $\ln OR$ , is more likely to be normally distributed (Hosmer & Lemeshow 2000). The standard error (SE) of the natural log of the odds ratio is:

$$SE = \sqrt{(1/a + 1/b + 1/c + 1/d)}$$

The 95% confidence interval for the log of the odds ratio ( $\ln OR$ ) is given by:

$$\ln OR \pm 1.96 \times SE$$

The lower confidence interval for the odds ratio is therefore  $\exp[\ln OR - 1.96 \times SE]$  and the upper confidence interval is  $\exp[\ln OR + 1.96 \times SE]$ . The 95% confidence interval for the plantain example is from 0.75 to 3.63 (the calculations are shown in Appendix 1). Note that the confidence interval is asymmetric. We present the results in this manner:

$$\text{Estimated odds ratio} = 1.64 (0.75, 3.63)$$

Our interpretation of this result is: the odds that a farm with plantains will be raided by elephants are 1.64 times greater than for a farm without plantains. However, since the confidence interval includes 1, this association is not significant at the 0.05 level. Therefore we cannot reject the null hypothesis; we conclude that there is no association between elephant raids and the presence of plantains. Our suspicion that plantains attracted elephants is not supported by the data.

The odds ratio is a very simple method for determining whether a particular crop attracts elephants onto farms in your area. These calculations can be done on a calculator while sitting in your tent.

If your data are stratified, for example you have fields at different distances from the river or from the national park boundary, then the Mantel-Haenszel test can be used. This estimates the association between raids/no raids and a dichotomous variable (like plantains/no

plantains) after controlling for strata (e.g. >2 km, 2-4 km and >4 km from the river). It can also be used to estimate the association between raids/no raids and a dichotomous variable (like plantains/no plantains) after controlling for another categorical variable, such as rice/no rice. The calculations can be done on a calculator, but are lengthy, and programming a spreadsheet would avoid arithmetical errors. Details and worked examples are given in Rosner (2006: 651-661).

### Logistic regression

The method above describes the association between raided farms and a categorical variable. There are two disadvantages. First, you can only evaluate one crop at a time. In practice, you will often want to evaluate the effect of one crop in the presence of a second crop or several other variables. Second, you can only use it with categorical variables (e.g. plantains/no plantains). In practice we often have independent variables that are continuous, for example the number of crop types grown on each farm, or the size of rice fields. Does the risk of raiding increase with the size of your rice fields? Logistic regression is a type of model that describes the relationship between a dichotomous dependent variable (like “raided” or “not raided”) and one or more independent variables (e.g.  $X_1$ ,  $X_2$ , and  $X_3$ ). The independent variables may be dichotomous (like “plantain” or “no plantain”) or continuous, such as number of different crop types or distance to a national park boundary.

Consider a continuous variable  $X_1$  that we suspect of influencing elephant behaviour. The probability,  $P$ , that an event ( e.g. a raid by elephants) will occur is:

$$P = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 X_1))}$$

Where  $\beta_0$  and  $\beta_1$  are regression coefficients.

Note that there is a negative sign in front of the expression in brackets:  $-(\beta_0 + \beta_1 X_1)$ . This can be re-arranged so that the odds that the event will occur are:

$$P/(1 - P) = \exp(\beta_0 + \beta_1 X_1)$$

Another tweak gives us the expression in the form of logits (log of the odds):

$$\ln(P/1 - P) = \beta_0 + \beta_1 X_1$$

This is a univariate logistic regression model that describes the relationship between the odds of the event occurring and the risk factor  $X_1$ .

If  $X_1$  is a binary variable (“yes/no” or “present/absent”) then  $\exp(\beta_1)$  is the simple odds ratio for  $X_1$ . It is identical to the odds ratio calculated from the  $2 \times 2$  contingency table. Thus the method of association from the  $2 \times 2$  table and logistic regression are closely related (Hosmer & Lemeshow 2000).

If  $X_1$  is a continuous variable, such as the number of food crops, then  $\exp(\beta_1)$  ---often written as  $e^{\beta_1}$  ---expresses the change in odds for each unit change in  $X_1$ . See Hosmer & Lemeshow (2000) or Woodward (2005) for the derivation of these equations.

As an example, at Kakum we needed to test whether the number of crop types was a major risk factor. The fitted logistic regression model is shown in Table 1. The risk  $P$  of being raided was:

$$P = \frac{1}{1 + \exp(-(-2.9076 + 0.3739X_1))}$$

Which can be re-written as:

$$P/(1 - P) = \exp(-2.9076 + 0.3739X_1)$$

The odds ratio for the number of crops is the exponentiated value of the regression coefficient, i.e.  $\exp(\beta_1)$ , which in this case is  $\exp(0.3739) =$

**Table 1.** Logistic regression model where the dependent variable is crop raids (raids/no raids) and the independent variable is the number of crops on each farm. Test for the overall fit of the model:  $\chi^2 = 6.50$ ,  $df = 1$ ,  $p = 0.0108$ .

Parameter	Estimated Coefficient $\beta$	SE ( $\beta$ )	Odds ratio	Confidence interval for odds ratio	p
Intercept	-2.9076	0.5116			<0.0001
No. of Crops	0.3739	0.1463	1.453	1.091, 1.936	0.0106

1.453 with a 95% confidence interval of (1.091, 1.936). This means that each increase in the number of crops, for example from 4 to 5 crops, raised the odds of being raided by 1.45 times. Since the confidence interval did not include 1, this association is significant ( $P < 0.05$ ).

In practice there are likely to be several risk factors that determine the probability that a farm will be raided. The simple logistic equation can then be generalized to:

$$P = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3))}$$

In this case there are three risk factor,  $X_1$ ,  $X_2$  and  $X_3$ , while  $\beta_0$ ,  $\beta_1$ ,  $\beta_2$ , and  $\beta_3$  are regression coefficients. The odds that the event will occur are:

$$P/(1 - P) = \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3)$$

Or in the logit form:

$$\ln(P/1 - P) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

There may be one particular variable (for example,  $X_1$ ) that we are interested in testing, perhaps because we know that we can persuade farmers to manage that particular variable. We need to know the importance of  $X_1$  after controlling for other risk factors. The odds ratio for  $X_1$  after controlling for the other two variables is  $\exp(\beta_1)$ . For example, number of crops is clearly important, but is it a significant risk factor after controlling for distance to the national park boundary? After all, we know that the risk of crop raiding falls steeply as you move further away from the national park. Table 2 shows the estimates from the fitted logistic regression model: after adjusting for distance to the park boundary, the number of crop types is still a significant risk factor for crop raids ( $p = 0.0228$ ). After adjusting for distance, this model tells us that a farmer increases his risk by a factor of 1.417 each time he plants an extra crop on his farm. Thus at a given distance, a

farmer with 4 crops on his farm will have  $1.417^2 = 2.01$  times the risk of a farmer who grows only two crop types.

In this example, both distance to park boundary and number of crops are significant contributors to risk. Distance was measured in metres, so the odds ratio ( $OR = 0.997$ ) means that for each metre one moves away from the park boundary there will be a reduction in risk of 0.997. Thus, a farm 500 metres away will have a risk that is  $0.997^{500} = 0.22$  times as bad as one directly on the park boundary.

What practical value is all this? Can it help the park manager to reduce the problem of crop damage by elephants? In this case the answer is clearly “Yes”: a farmer can reduce his risk dramatically by placing his fields at least 500 metres from the park boundary. If the pressures on land do not allow him the luxury of deciding where his field must go, then he can halve his risk by growing two crops instead of four.

We can now return to the question of plantains. After adjusting for distance to the park and for the number of crops, are plantains an important risk factor? We add plantains to the logistic regression model (Table 3) and the result shows that the odds ratio for plantains, after adjusting for the other two variables, is 1.01 (0.39, 2.66). The odds ratio is one: it does not matter whether or not farmers have plantains on their farms.

Logistic regression is discussed in many statistics textbooks, e.g. Hosmer & Lemeshow (2000), Kleinbaum *et al.* (2008), Woodward (2005). All statistical software packages include modules for logistic regression.

All the calculations described above can be done with *Epi-Info*, an analysis package provided absolutely free by the United States government

**Table 2.** Logistic regression model where the dependent variable is crop raids (raids/no raids) and the independent variables are the number of crops on each farm and distance to the national park boundary. Test for the overall fit of the model:  $\chi^2 = 19.27$ ,  $df = 2$ ,  $p < 0.0001$ .

Parameter	Estimated coefficient $\beta$	SE ( $\beta$ )	Odds ratio	Confidence interval for odds ratio	p
Intercept	-1.4164	0.6535			0.0302
No. of crops	0.3485	0.1530	1.417	1.050, 1.913	0.0228
Distance	-0.0030	0.0009	0.997	0.995, 0.999	0.0009

**Table 3.** Logistic regression model where the dependent variable is crop raids (raids/no raids) and the independent variables are number of crops on each farm, its distance from the park boundary, and presence/absence of plantains. Test for the overall fit of the model:  $\chi^2 = 19.27$ ,  $df = 3$ ,  $p = 0.0002$ .

Parameter	Estimated coefficient $\beta$	SE ( $\beta$ )	Odds ratio	Confidence interval for odds ratio	p
Intercept	-1.4157	0.6540			
Crops	0.3463	0.1739	1.414	1.005, 1.988	0.0465
Distance	-0.0030	0.0009	0.997	0.995, 0.999	0.0009
Plantains	0.0131	0.4921	1.013	0.386, 2.658	0.9788

(www.cdc.gov). The odds ratios and Mantel-Haenszel test are in the *Utilities\StatCalc* drop-down menu.

## Discussion

We cannot properly understand the phenomenon of crop raiding by elephants if we do not quantify it. But we must never forget that each event represents a farmer's loss, sometimes catastrophic, and our goal must always be to minimize his suffering.

The logistic regression model illustrates how the risk declines steeply with distance from the park boundary. In some sites it may be distance from a river or a road, or some other feature, that is of interest. These models can help the agricultural extension agents or the park manager to work with local farmers to decide on the number of crops to cultivate on their farms (Barnes *et al.* 2005). By growing fewer crops they will obviously reduce food production, but that will be balanced by less risk of loss from elephants.

The logistic model described here makes an important assumption: that the farms are all independent of each other, i.e. they are randomly distributed about the study area. In practice, this is unlikely. Our study farms are often in groups, for example we may have 30 farms, with 10 farms in each of 3 villages. This clustering presents a complication for the diligent analyst. In a subsequent paper I will describe how one should deal with clustered farms.

## References

Barnes, R.F.W. (2008) The design of crop-raiding studies. *Gajah* **28**: 4-7.

Barnes, R.F.W., Hema, E.M., Nandjui, A.,

Manford, M., Dubiure, U.F., Danquah, E.K.A. & Boafo, Y. (2005) Risk of crop-raiding by elephants around the Kakum Conservation Area, Ghana. *Pachyderm* **39**: 19-25.

Hosmer, D.W. & Lemeshow, S. (2000) *Applied Logistic Regression*. John Wiley & Sons, New York.

Kleinbaum, D.G., Kupper, L.L., Nizam, A. & Muller, K.E. (2008) *Applied Regression Analysis and Other Multivariable Methods*. Thomson Brooks/Cole, Belmont, California.

Long, J.S. (1997) *Regression Models for Categorical and Limited Dependent Variables*. Sage Publications, Thousand Oaks, California.

Rosner, B. (2006) *Fundamentals of Biostatistics, 6<sup>th</sup> Edition*. Thomson Brooks/Cole, Belmont.

Woodward, M. (2005) *Epidemiology: Study Design and Data Analysis*. Chapman & Hall, Boca Raton.

## Appendix 1

At Kakum we believed that plantains attracted elephants that then stayed on the farm to wreak damage on other crops. The  $2 \times 2$  table is shown in Table 2. The odds ratio is:

$$\begin{aligned} \text{OR} &= (14 \times 111) / (15 \times 63) = 1.6444 \\ \text{SE of lnOR} &= \sqrt{(1/14 + 1/63 + 1/15 + 1/111)} \\ &= \sqrt{(0.0714 + 0.0159 + 0.0667 + 0.0090)} \\ &= 0.4037 \end{aligned}$$

The 95% confidence limits of lnOR  
 $= 0.4974 \pm 1.96 \times 0.4037 = 0.4974 \pm 0.7913$   
 The 95% confidence limits of OR =  $\exp(-0.2939)$   
 and  $\exp(1.2887) = 0.7454$  and  $3.6281$

*Author's e-mail: rfwbarnes@znet.com*