INTRODUCTION TO DATA ANALYSIS WORKSHOP

Notes and Workbook

5th SEPTEMBER 2002
Springfield, Dominica

Prepared by Dr Chris Magin
Senior Protected Areas Specialist
Fauna & Flora International

A National Strategy for Sustainable Wildlife Use, Commonwealth of Dominica
Project Ref. 162 / 10 / 010

INTRODUCTION SESSION

Have two minutes to interview each other in pairs and discover name, position, years of experience with FWD and one humorous fact about other, then introduce each other to group.

DATA ANALYSIS AND INTERPRETATION

Introduction and overview

This workshop aims to inform participants about the types of analysis we will be using to analyse the data collected during agouti, crab and frog transects.

What is data?

Scientists and researchers use the term data to describe any information gathered in a systematic or structured way.

Why analyse and interpret data?

Almost self-evident. If you have collected data, you need to make some kind of sense of it. Unanalysed data is a meaningless collection of numbers. You need to analyse your data to answer your questions.

What questions will you want to ask about your transect data?

Ask participants for e.g.s:
Relative abundance of agoutis / crabs / frogs / pigeons
Seasonality (of population structure of agoutis, of numbers of pigeons)
Differences between forest districts (Central, Southern, Eastern, Northern)
Relation between numbers and other factors (e.g. hunting, habitat)

But analysed data also needs to be interpreted, since uninterpreted data is of very little use to anyone, including the people who collected it. So once your questions are answered, they need to be placed in the biological and environmental context.

E.g. Find fewer agoutis in one district than another. Could be hunting, or habitat, or distribution, or time of day data collected etc. etc..

So, we need to think intelligently about what the analysed data shows.

Types of data:

Data can be measured – that is put into categories, which doesn't necessarily have to include assigning numbers - in three main ways:

Nominal or classificatory – observations are classified into several categories by their attributes.
E.g. gender (male or female); age class (adult or juvenile e.g. agouti)

Ordinal – observations are ranked (that is put in order) and the categories are related
E.g. scoring individuals in a group of animals according to social dominance, i.e. Chicken A is more aggressive than Chicken B etc.

Quantitative – observations can be represented by numbers.
E.g. temperature, length, weight. They can be continuous e.g. length, or discrete e.g. counts of individuals (you can't have 1.67 agoutis).

What type of data are we collecting?
Answer – mainly nominal and quantitative

Types of analysis:

1. Visual analysis – tabular, graphical, mapping

Pie and bar charts (including stacked bar charts). Can indicate means and relative proportions of components.
Scatter plots
Histograms (for continuous variables)

2. Descriptive analysis – mathematical description of the data

Mean – the average
Median – the number which divides the frequency distribution into upper and lower halves
Mode – the peak of the frequency distribution, the most common occurrence.
Percentages
Proportions

3. Statistical analysis
Statistics are used to determine whether a result is likely to have arisen by chance, or is likely to be a real result. Chance effects can arise because of sampling (bias, error, non-representativeness of the sample etc.).

Session 1 Visual data analysis

A picture paints a thousand words. Human's primary sense is sight. Therefore visual presentation is extremely important to us. Think of impact of TV, photos etc.
Q. What would an octopus prefer (or a dog)?

Tables are necessary to compile and summarize data collected on data sheets. Can also be considered as first stage in visual analysis.
E.g. from Northern region our agouti data sightings from 23 data sheets can be summarized in one table:

NORTHERN REGION		**AGOUTI**				Of those seen:			
		Date	No. Seen	No. Heard	Total (seen and heard)	No. Adults	No. Juveniles	No. pregnant	Unknown
NFR 1	MARCH	7.3.02	1	2	3	1	0	0	0
NFR 1	APRIL	10.4.02	1	2	3	1	0	0	0
NFR 1	MAY	9.5.02	1	3	4	0	0	0	1
NFR 1	JUNE	4.6.02	0	3	3	0	0	0	0
NFR 1	JULY	2.7.02	2	1	3	0	0	0	2
NFR 1	AUGUST	15.8.02	1	0	1	1	0	0	0
NFR 2	MARCH	6.3.02	0	0	0	0	0	0	0
NFR 2	APRIL	4.4.02	0	0	0	0	0	0	0
NFR 2	MAY	9.5.02	0	0	0	0	0	0	0
NFR 2	JUNE	4.6.02	0	0	0	0	0	0	0
NFR 2	JULY	4.7.02	1	0	1	0	0	1	0
NFR 3	MARCH	14.3.02	1	2	3	1	0	0	0
NFR 3	APRIL	10.4.02	0	0	0	0	0	0	0
NFR 3	MAY	7.5.02	0	0	0	0	0	0	0
NFR 3	JUNE	10.6.02	0	0	0	0	0	0	0
NFR 3	JULY	2.7.02	0	0	0	0	0	0	0
NFR 3	AUGUST	15.8.02	0	0	0	0	0	0	0
NFR 4	MARCH	12.3.02	1	0	1	0	0	0	1
NFR 4	APRIL	4.4.02	1	0	1	1	0	0	0
NFR 4	MAY	2.5.02	2	0	2	0	0	0	2
NFR 4	JUNE	20.6.02	2	1	3	2	0	0	0
NFR 4	JULY	4.7.02	1	2	3	0	0	0	1
NFR 4	AUGUST	8.8.02	0	0	0	0	0	0	0
Total (March to August)			**15**	**16**	**31**	**7**	**0**	**1**	**7**

Graphs = second stage in visual analysis. As a general rule, graphs are much more informative than tables or figures. Displaying your data in a graph is one of the most fruitful types of exploratory data analysis, and it is always wise to plot your data and inspect them visually before plunging into confirmatory data analysis.

Many different types. Commonest (and most useful to us) are:

Bar chart / column bar chart
Compiled by plotting categories of observation along the x axis, and numbers along the y.
E.g. Plot Months and total number of agoutis seen – do with participants calling out numbers from above data.

Stacked bar chart
Compiled by plotting related categories one on top of the other (i.e. cumulatively) on the same bar
E.g. Plot Months and total numbers of agoutis heard on top of previous graph. Gives total of seen and heard

Pie chart – shows relative proportions of observations.
Technique is to divide a circle of 360 degrees into the same number of sectors that you have types of observation. E.g. agouti data have adult, juvenile, pregnant, unknown.
The angle of each sector is equivalent to its proportion of the total observations.
i.e. Angle of sector A = A x 360 / Total A+B+C etc.

In e.g. above, 15 agoutis were seen, 7 adult, 1 pregnant and 7 unknown. Adults and unknowns make up 7/15 = 46.7%, pregnant 1/15 = 6.7%
In a pie chart adult and unknown sectors would therefore make up 168 degrees

Scatter plot
Two independent variables are plotted on x and y axis
Can help to determine if there is any relation between the two.

CENTRAL REGION – data			**AGOUTI**	**RED-NECKED PIGEON**
		Date	Total (seen and heard)	Total (seen and heard)
CER 1	MARCH	5.3.02	0	10
CER 1	APRIL	9.4.02	0	0
CER 1	MAY	2.5.02	0	2
CER 1	JUNE	6.6.02	1	17
CER 1	JULY	2.7.02	1	14
CER 1	AUGUST	2.8.02	4	12

No reason as yet to believe there is a correlation.

EXERCISE 1.

Individuals / groups to plot

i) Bar chart of monthly average against month
ii) Stack bar chart showing average numbers in each region per month
iii) Pie chart showing overall proportion of pigeons seen in each district

Red necked Pigeon					
	CENTRAL	EASTERN	SOUTHERN	NORTHERN	**MONTHLY AVERAGE**
MARCH	6.0	4.0	8.5	3.0	**5.4**
APRIL	6.5	2.0	4.3	4.8	**4.4**
MAY	9.5	1.0	4.8	2.8	**4.5**
JUNE	12.8	2.3	9.3	6.0	**7.6**
JULY	11.3	2.3	2.8	6.3	**5.6**
AUGUST	5.3	3.3	2.8	1.3	**3.2**
Overall average	8.6	2.5	5.4	4.0	**5.1**

Session 2 Descriptive data analysis

In addition to graphing your data, it is also helpful to summarize data in the form of summary descriptive statistics such as means or medians or ranges before statistical analysis is carried out.

Mean – the average of a population or a sample. X bar = Sum xi / n
Where n = no. in population or sample and Sum xi = sum of all the observations
E.g. 4 + 6 + 8 + 10. Mean = 28/4 = 7

Median – the number which divides the frequency distribution into upper and lower halves. I.e. The number which lies in the centre of a set of values. To find the median arrange the data values in increasing order of size. The median is the value having as many observations above it as below
E.g. the numbers 2, 5, 6, 7, 8, 8, 10 have median of 7

If there is an even number of values in the sample then the median is the arithmetic mean of the two middle values.
E.g. the numbers 5, 7, 9, 10, 12, 12 have median (9 + 10)/2 = 9.5

Mode – the peak of the frequency distribution, the most common occurrence.
Q. What is the mode of the no. of agouti seen in transects in the Northern Forest district?

Proportions – the fraction (usually expressed as a decimal) of an observation or observations out of the total set.
E.g. 10 observations of animals, of which 3 frogs, 5 agoutis, 2 pigeons.
Proportion of X = X / (Sum xi), so for frogs = 3 / (3 + 5 + 2) = 3 /10 = 0.3 etc.
Proportions are 0.3, 0.5, 0.2

Proportions are usually expressed as percentages, which should total 100.
Percentage of X = 100 X / (Sum xi)

In the above example 30% of the observations were of frogs, 50% agoutis, 20% pigeons.

Range – the range is the difference between the highest and lowest values in a distribution

EXERCISE 2.

Calculate separately for Central and Southern regions from the red-necked pigeon data below

i) The mean totals
ii) The medians
iv) The range
i) The mode
v) The proportion seen and the proportion heard
vi) The percentage seen and the percentage heard

RED-NECKED PIGEON			Date	Number Seen	Number heard	Total (seen and heard)
D'leau Gommier	CER 1	MARCH	5.3.02	1	9	10
	CER 1	APRIL	9.4.02	0	0	0
	CER 1	MAY	2.5.02	0	2	2
	CER 1	JUNE	6.6.02	0	17	17
	CER 1	JULY	2.7.02	0	14	14
	CER 1	AUGUST	2.8.02	2	10	12
TMT	CER 2	MARCH	14.3.02	0	8	8
	CER 2	APRIL	16.4.02	0	6	6
	CER 2	MAY	14.5.02	0	6	6
	CER 2	JUNE	13.6.02	0	11	11
	CER 2	JULY	11.7.02	0	13	13
	CER 2	AUGUST	7.8.02	0	3	3
Gommier Ellick	CER 3	MARCH	12.3.02	0	1	1
	CER 3	APRIL	18.4.02	0	3	3
	CER 3	MAY	14.5.02	0	6	6
	CER 3	JUNE	4.6.02	0	3	3
	CER 3	JULY	4.7.02	0	5	5
	CER 3	AUGUST	1.8.02	0	8	8
	CER 4	APRIL	11.4.02	3	14	17
	CER 4	MAY	9.5.02	1	23	24
	CER 4	JUNE	11.6.02	2	18	20
	CER 4	JULY	9.7.02	0	13	13
	CER 4	AUGUST	13.8.02	0	1	1

Morne Raquette	SFR 1	MARCH	5.3.02	0	4	4
	SFR 1	APRIL	2.4.02	0	4	4
	SFR 1	MAY	7.5.02	0	5	5
	SFR 1	JUNE	4.6.02	0	4	4
	SFR 1	JULY	4.7.02	2	9	11
	SFR 1	AUGUST	16.8.02	0	6	6
Sylvania	SFR 2	MARCH	7.3.02	0	2	2
	SFR 2	APRIL	2.4.02	0	1	1
	SFR 2	MAY	9.5.02	0	1	1
	SFR 2	JUNE	5.6.02	0	0	0
	SFR 2	JULY	2.7.02	0	0	0
	SFR 2	AUGUST	13.8.02	0	5	5
Salisbury	SFR 3	MARCH	11.3.02	0	5	5
	SFR 3	APRIL	4.4.02	0	3	3
	SFR 3	MAY	9.5.02	3	7	10
	SFR 3	JUNE	11.6.02	1	9	10
	SFR 3	JULY	9.7.02	0	0	0
	SFR 3	AUGUST	15.8.02	0	0	0
Laudat	SFR 4	MARCH	11.3.02	7	16	23
	SFR 4	APRIL	4.4.02	1	8	9
	SFR 4	MAY	7.5.02	0	3	3
	SFR 4	JUNE	4.6.02	15	8	23
	SFR 4	JULY	2.7.02	0	0	0
	SFR 4	AUGUST	19.8.02	0	0	0

Session 3 Statistical data analysis

Statistics are used to determine whether a result is likely to have arisen by chance, or is likely to be a real result. Chance effects can arise because of sampling (bias, error, non-representativeness of the sample etc.).

Statistics determine the probability of an event occurring at random. If the probability is small, we accept that the effect observed is unlikely to have occurred by chance, i.e. the effect is a real one.

Conventionally, we say that if there is a less than one in twenty chance of an effect occurring, the difference or result is a real one.

A one in two chance (or a 50 : 50) chance such as getting a heads or tails when spinning a coin is called a $p = 0.5$
One in twenty is a $p = 0.05$ chance.

There are two main types of statistical tests

Parametric tests are used for quantitative data that are normally distributed. That is, when the data are plotted in a frequency histogram the resulting curve is bell-shaped, symmetrical about the mean and approximately 2/3 of all values fall within +/- 1 standard deviation of the mean

Draw Curve

Non-parametric tests can be used for nominal (= classificatory) and ordinal data. They can also be used for quantitative data, and do not require that the data are normally distributed. They are also usually quick and easy to use. Many occurrence data contain a great many 0's so the curve would not be normal.

Almost all statistical tests use a null hypothesis, H_0. This is a hypothesis of "no effect" or "no difference", and is usually proposed for the express purpose of being rejected. If it is rejected, the alternative hypothesis, H_1 is supported.

E.g. When comparing the means of two samples, H_0 would be that the samples were drawn from the same population and that therefore there should be no difference between them. If H_0 is rejected, we conclude that the opposite hypothesis, that there is a difference between the populations, must be correct.

Most statistical tests have what at first sight appear to be complicated formulas. However, if you work through them step by step and are careful, they are actually straightforward. The difficult part of statistics is choosing the correct test to use, because there are literally dozens of them.

We will only look at two, non-parametric tests.

The Mann-Whitney U test can be used to estimate the statistical difference between the medians of two unrelated groups of numerical observations. In other words we test whether the samples are drawn from similar groups, or are statistically different.

DO EXAMPLE WITH CARD SAMPLING
Emphasise that the packs are our populations, and we are sampling them.

Get groups to draw 7 cards at random from each of two packs. Average score for each normal pack should be 7

See my statistics notes
Run through test with group. Hopefully no difference!

Then re-distribute cards with Ace – three swapped with Jack - King

Average for two packs now 9.3 and 4.7. Redo Hopefully difference!

EXERCISE 3.

Do a Mann-Whiney U test to compare numbers of pigeons seen during transects in CER 1 and SFR 1.

The Spearman rank correlation test measures the degree of association between two variables X and Y grouped in N pairs (X, Y). A correlation is used when the variables are numerical scores. When one or both of them is just a qualitative classification into a few classes a different test, the Chi-squared test, is used.

Example

Village	No. of hunters (X)	No. of agouti hunted / year (Y)
A	5	25
B	7	50
C	6	30
D	8	45
E	3	5
F	10	70

Graph – shows that there is an apparent correlation. But is it significant?

Then rank X's and Y's in order of increasing magnitude, and find the difference between their ranks

Village	X	Rank X	Y	Rank Y	Difference = X rank – Y rank
A	5	2	25	2	0
B	7	4	50	5	-1
C	6	3	30	3	0
D	8	5	45	4	1
E	3	1	5	1	0
F	10	6	70	6	0

$Rs = 1 – 6 (\text{sum } D^2) / N(N^2-1)$

$Rs = 1 – 6x2/ 6(36-1) = 1-(12 / 210) = 0.94$

Value of Rs at p=0.05 for N = 6 is (by looking at tables) 0.829

Therefore there is a positive correlation, significant at p=0.05

EXERCISE 4.

Calculate Rs for the data given