Ann. Zool. Fennici 46: 82–92 ISSN 0003-455X (print), ISSN 1797-2450 (online)
Helsinki 30 April 2009

The importance of a research data statement and how to develop one

E. David Ford

College of Forest Resources, P.O. Box 352100, University of Washington, Seattle, WA 98195-2100, USA (e-mail: edford@u.washington.edu)

Received 18 Feb. 2008, revised version received 17 June 2008, accepted 22 July 2008

Ford, E. D. 2009: The importance of a research data statement and how to develop one. — *Ann. Zool. Fennici* 46: 82–92.

A research plan must answer four questions: (i) What is the scientific question that the research seeks to answer? (ii) What type of investigation will the researcher conduct? (iii) What measurements will be made? (iv) What type of data analysis will be used, and will there be sufficient statistical power in the data to give an effective answer to the question? I illustrate how each of these questions can be answered. The first requires conceptual and propositional analysis to refine concepts and postulates so that an important scientific question is developed and one that is possible to answer. Questions ii, iii and iv together form a Data Statement that must answer why a particular type of investigation is appropriate and how that question will be answered. It must define the extent to which measurements represent the concepts in question and the accuracy and precision of those measurements. And, most important, the Data Statement must define the type of data analysis that will be used, including a definition of the statistical power necessary to answer the question effectively. The development of a Data Statement will usually involve exploratory analysis of the scientific question to explore the effectiveness of proposed measurements and to enable calculation of a sampling regimen that will provide adequate statistical power.

Introduction

Graduate students who spend considerable time both in planning and conducting their first research are sometimes dissatisfied. Not infrequently they wish they had used different experimental treatments or had measured additional things and sometimes they fail to achieve results that are statistically significant yet are indicative of the effects they investigated. Research supervisors may attribute these problems simply to "lack of experience" but there is a more informative explanation — that the student's research planning has not been sufficiently precise in defining what must be done to answer a scientific question.

Research planning should be treated as a distinct and detailed activity that determines precisely how investigations will be conducted. A research data statement should be produced before starting the measurement phase of research that:

a. defines the scientific procedure to be used in investigating a scientific question,
b. specifies the measurements to be made, and

c. defines the requirement of the data for any statistical test that is to be applied.

Producing a research data statement should lead to answers to two questions: *What exactly is the scientific question*? and *Can the question be answered with precision*?

Defining the exact scientific question requires making a detailed conceptual and propositional analysis to organize what is already known and identify what should be done. I illustrate this process in the first section. This procedure can reduce the problem of not doing or measuring the right thing. Answering the second question frequently requires an exploratory investigation or pilot study that can help to ensure a more satisfactory use of statistical inference. In the second section of this paper I show why exploratory investigations are so important, discuss why they are frequently overlooked, and make some suggestions about the relationship between statistical and scientific inference.

What exactly is the scientific question?

When a student joins a research group he or she needs help from them in defining what might be an important research question with a reasonable prospect of being answered, and in learning the measurement and analytical techniques of the research field. However, it is quite typical for research groups to develop accepted ideas about the important questions that need to be answered and the sets of measurement and analytical techniques and their sampling protocols that should be used. But for each new specific question there is the possibility that these ideas have to be modified more than expected. This may occur when new papers are published on the topic or the variance associated with a particular type of measurement is different for a new problem, so that a new sampling procedure is required.

When starting a field or laboratory research project we must consider the *exact* scientific question to be investigated and not embark on an investigation until that is clear. The question that can be asked and the measurements that might be made are, of course, inter-related. The available measurements can determine the scope of the scientific question and may limit it in ways that can be underestimated. Working scientists seem to have a map inside their minds that puts together what is known, what needs to be discovered, and how it might be discovered. When first coming into research a student is unlikely to have such a map. Conceptual and propositional analysis (Ford 2000) can help you make one.

Figure 1 represents an overview diagram of the type of map of a scientific problem that conceptual and propositional analysis can help you produce. Analysis of the scientific literature is required to organize what is known in terms of axioms. Notice that each axiom is represented by the concepts that define it.

An axiom is a proposition assumed to be true on the basis of previous research, observations, or information, and is used in defining the working part of the theory that is the foundation for the research. Typically, an axiom specifies that something does or does not occur, or that one thing does or does not influence another; or it defines a mathematical relationship (Ford 2000).

We often start research with a broad question that provides an initial direction for the research as a whole but cannot be answered by a single investigation. For example: *How has climate change affected animal distribution*? is obviously of great topical interest but is too broad to be answered by a single investigation. Particular components of *climate change* must be defined, the *animal* or *animals* must be specified and what is actually meant by *affected* and *distribution* must be clarified. Each italicized word is a concept that requires definition. The process of making a definition should lead to refining the concept so that its meaning is more precise and, as each concept is refined, then the precise meaning of the question changes.

Conceptual and propositional analysis has two related objectives: to define what will be accepted in the investigation as established knowledge and to define the question. This helps to define the most important gaps in knowledge or where there is greatest potential for advance. The exact question we can ask depends upon what knowledge can be accepted as sufficiently well known to build upon.

The first component of conceptual and propo-

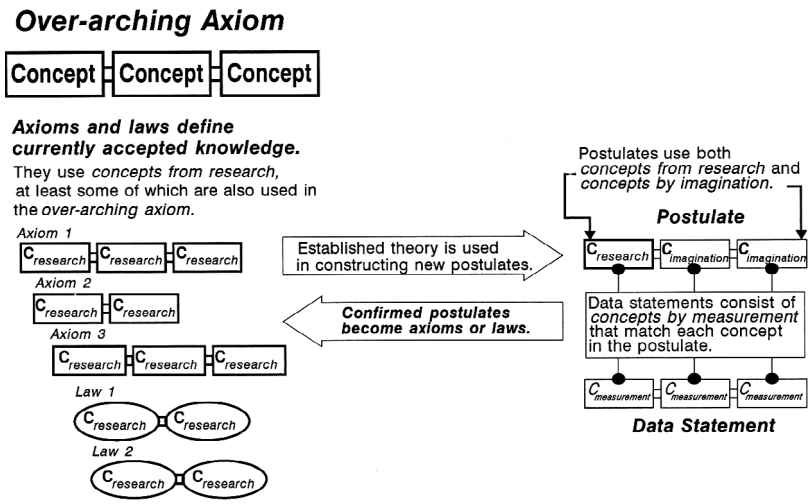

Fig. 1. A schematic view of the interaction between theory and postulates. The established body of the theory, i.e., those propositions that have been confirmed by research investigations or are accepted as necessary for further studies (left hand side) are represented by axioms and, where appropriate by laws. Each axiom is constructed of concepts from research, $C_{research}$. The theory as a whole has an over-arching axiom. In some theories there is an overarching postulate rather than overarching axiom. A postulate is constructed using at least one $C_{research}$ but at least one concept is new, termed here a concept by $C_{imagination}$. In order to construct a data statement a measurement must be developed for each concept used in the postulate. This process invariably involves making some compromise in precision or accuracy in measurement or sampling and it is important to define such problems when writing a data statement (Reprinted with permission from Ford 2000, Cambridge University Press).

sitional analysis is to change from using questions to using propositions. So, instead of asking "How has climate change affected animal distribution?" we form it as a proposition: *Climate change affects animal distribution* and ask if this proposition is true or false or is something we can confirm or reject. The scientific literature shows that a number of studies have been made, different types of climate change have been considered, and different effects reported. So, to decide on this proposition we see that it all depends upon what is meant by the concepts used.

The analytical process of conceptual and propositional analysis can be illustrated through a study about possible climate change effects published in this journal. Brommer (2004) studied change in the distribution of bird species in Finland over a 12-year period (1974–1979 to 1986–1989). He notes that: "The period studied here coincides with the period of the earth's most rapid climate warming in the last 10 000 years which started in 1976 (McCarthy *et al.* 2001)" and "Climate warming induces a poleward shift of isotherms (in Europe ca. 120 km in the last century)." Isotherms are lines of equal temperature and are usually calculated for individual months or seasons by smoothing between mean temperatures for separate weather station records. These statements provide two axioms:

Axiom 1: Rapid climate warming has occurred since 1976.

Axiom 2: Isotherms in Europe have moved northward during the period of rapid climate warming at a rate of some 120 km in 100 years.

Brommer made no direct measurement of temperature change over the study period and Axiom 1 and Axiom 2 are used to infer direct axioms for the research:

Axiom 3: Mean annual temperature in Finland increased between 1974–1979 and 1986–1989.

Using Axiom 2 the rate of northward movement can be calculated as around 13.2 km over

the 11 years of the study although Axiom 1 suggests this would be an underestimate. Brommer presents no calculations but writes " … and organisms are expected to follow this change (Parmesan *et al*. 1999, McCarthy *et al*. 2001)." The research of Parmesan *et al*. is specific to butterflies and Brommer extends that to birds. Extending axioms established in one piece of work to act as the foundation for research in a different type of investigation is quite frequent in ecology. Brommer cites a number of studies in Finland reporting climatic effects on different aspects of bird's behavior and life cycles to support this extension. Brommer also cites a study (Thomas and Lennon 1999) in the south of the United Kingdom indicating that bird species shifted their range margins polewards. So a further axiom of the research is:

Axiom 4: Some European bird species respond to northward isotherm movement with a poleward change in their range margins.

The overall conjecture in this research is that there has been a northward movement of bird species in Finland. This must be defined as a postulate.

A postulate is a conjecture written in the form of a proposition. It is untested, or considered sufficiently uncertain to be the subject of further direct investigation.

The straightforward postulate might seem to be: *There has been a northward movement of bird species in Finland*. We can consider this as the over-arching postulate, a general question, stated in propositional form, that motivates the research but that might not be completely answered by a single investigation. Two considerations make us carefully consider the details of the postulate to be investigated. The type of measurement available that can be used and the possibility that other factors than climate change may cause changes in bird distribution.

Range margins of bird distributions are calculated from atlases of breeding birds. Brommer (2004) used two Finnish atlases of bird species distribution, one constructed from data collected 1974–1979 and one 1986–1989. These atlases use a grid of 3813 10×10 km^2 grid cells and they record, for each cell, whether a bird species bred in the cell or not. No estimate is made of the number of birds breeding in a grid square. If such a number were available, a more comprehensive indication of movement might be possible. The available presence/absence data require that the study focuses on changes in range margins and so the postulate must reflect that. Change in range margins may occur due to other influences than as a direct response to climate change. Thomas and Lennon (1999) note such change may be part of overall population changes: "Overall increases in distribution size, whatever the cause, would be expected to cause species restricted to the south of Britain […] to move northwards at their northernmost boundary, and species restricted to the north […] to move southwards" while "a general decline in a species would be expected to cause northern species […] to retract northwards, and southern species […] to retract southwards towards their cores." Such a pattern may occur "because the most marginal records are likely to be further from the distributional core." Note that this alternative cause of movement also has its own set of axioms that define a bird population as having a core and that contraction and expansion of the population takes place relative to that core. These axioms should be specified (but for space considerations will not be in this paper) and the possibility that such movement may occur needs to be considered. So the postulate must specify changes in range margins relative to whether the overall distribution is shrinking or expanding.

Postulate 1: Over the period 1974–1979 to 1986–1989 the range margins of bird species in Finland moved northwards relative to changes in the overall size of a species' distribution.

This postulate is clearly more restricted than the over-arching postulate but is closer to a question that can be investigated. However, the measurements that can be taken require a further restriction due to qualification in the bird species that can be considered and where in Finland they are located. A comparison is made between the movement of the southern margins of species with a northerly distribution and the movement of northern margins of species with a southerly distribution. The reasoning being that the cool

(northerly) margins of temperate species might be more responsive to temperature changes than the warmer (southerly) margins. Thomas and Lennon (1999) suggest that warm margins might respond more to rainfall, species interactions or longer-term climate-related changes in the vegetation. Both Thomas and Lennon (1999) and Brommer (2004) make a two part comparison, which for the Finnish case is defined as:

Postulate 1a: Over the period 1974–1979 to 1986–1989 the northerly range margins of bird species in Finland with a southerly distribution moved northwards relative to changes in the overall size of a species' distribution.

Postulate 1b: Over the period 1974–1979 to 1986–1989 the southerly range margins of bird species in Finland with a northerly distribution remained stationary relative to changes in the overall size of a species' distribution.

Postulate 1a is the major postulate as it relates directly to the question of northward movement. If Postulate 1b is found to be correct it might be taken to strengthen ideas about differences between north and south margin changes during a period of northerly margin change. Of course, from a technical point of view if southerly margins, as well as northerly margins, were found to move north then doubts would be raised about the over-arching postulate and/or the techniques being use.

The concepts used in this study have normal common use definitions. However, in some studies, much of the work required in propositional analysis is careful analysis of concept definitions. This is particularly so when integrative concepts, such as "species diversity" are used in overarching postulates and have to be translated into measurable concepts.

Recall that a research data statement must be produced that:

a. defines the scientific procedure to be used in investigating a scientific question,
b. specifies the measurements to be made, and
c. defines the requirement of the data for any statistical test that is to be applied.

A really important issue here, and one that can be too easily glossed over, is that the measurements we make of something may not represent the complete meaning of the concept they are supposed to represent. In this work *distribution* and *range margins* will have to be defined by some metric and *bird species* will have to be restricted at least to those inhabiting only south Finland for Postulate 1a and only north Finland for Postulate 1b. Of course the species are necessarily different between north and south and what we may hope to find out using Postulate 1b may not relate directly to the species used to investigate Postulate 1a.

a. The scientific procedure is a comparison of change in species range margins between 1974–1979 and 1986–1989 relative to change in the size of species distributions using atlases of bird distribution for Finland. In this case the procedure is largely determined by the available data.
b. In this research the direct observations of whether birds were breeding or not were made during the bird surveys and it is assumed that these observations were correct. (This assumption is an axiom of measurement.)

The range margins of birds are calculated as the median latitude of the ten most marginal 10 × 10 km^2 grid cells. Distribution changes were calculated as the $\log_{10}$ of the proportion of the number of 10 × 10 km^2 grid cells of 1986–1989 over the number of grid cells occupied in 1974–1979. These definitions specify axioms of measurement, i.e., assumptions that are expected to hold true, so that the measurements used do actually represent the quantity in the postulate. For example, "The range margin of bird distributions can be defined as the median latitude of the ten most marginal 10 × 10 km^2 grid cells occupied by breeding birds." In this investigation no test was made of this axiom of measurement, e.g., by changing the numbers of cells used in the calculation or by using another metric than the median. In some investigations axioms of measurement can become critical and measurements and/or sampling schemes are found inadequate.

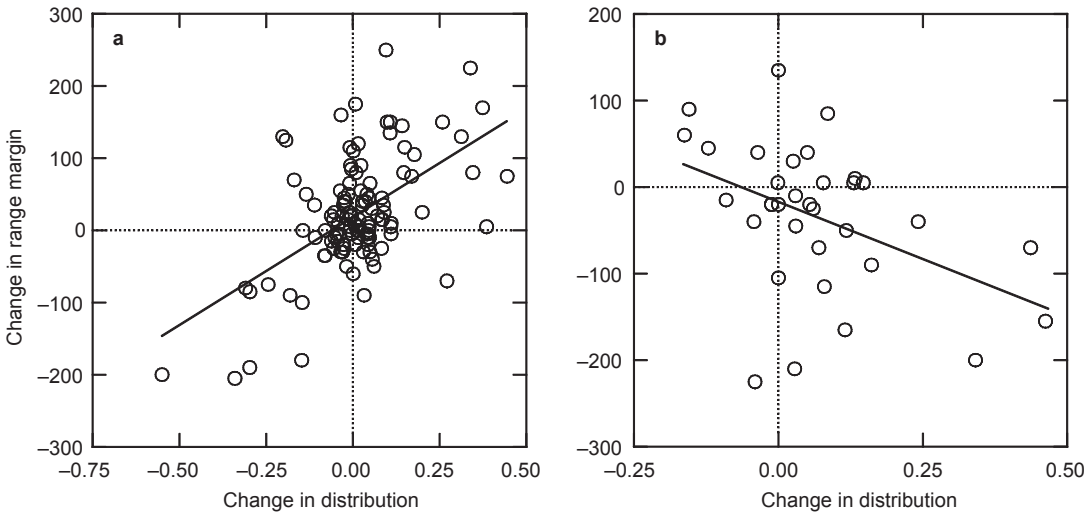

Fig. 2. The change in the latitude of range margins (in km) plotted against the change in distribution size for (**a**) 116 southerly species (regression: $F_{1,114} = 48.9$, $P < 0.001$, $R^2 = 29.4\%$) and (**b**) 34 northerly species (regression: $F_{1,32} = 7.4$, $P = 0.01$, $R^2 = 16.3\%$). The expected change in range margin if a species would not change in overall distribution is give by the intercept of the regression line (Brommer 2004).

The number of bird species that are used is restricted. First there is a classification into those with north and south distributions so the numbers used for Postulate 1a and 1b are restricted, and are actually different. Second, species distributed throughout the whole of the country are excluded as are species with disjunct distributions. Postulate 1a requires measurement of range margin changes and so there is also an axiom of the study that "a sufficient number of bird species in south Finland are sensitive to the rate of temperature change to affect the mean rate of range margin change." It is important to acknowledge this axiom because it specifies that we must detect a mean affect.

c. Statistical tests are based on null hypotheses. In this work the calculation is a regression of change in range margin on change in distribution area.

The null hypothesis for Postulate 1a is that the intercept, i.e., when change in distribution area is zero, on this regression line is not significantly significant from zero and Postulate 1a predicts that this null hypothesis will be rejected. Postulate 1b has the same null hypothesis, but applied to movement of the southern margin of northern species, and Postulate 1b predicts this null hypothesis will be accepted. Using the parameters of the regression calculated between change in range margin and change in area distribution, an estimate is made of the range margin change for the condition of no change in species area distribution (Fig. 2). Four assumptions are made in using linear regression:

i. that the relationship is linear,
ii. that the deviations from the true regression line are independent,
iii. that residuals are normally distributed, and
iv. the variance is constant across the relationship (homoscedasticity).

These assumptions were not tested by Brommer (2004) or by Thomas and Lennon (1999) but they should be specified in a research plan data statement that uses regression.

Thomas and Lennon (1999) do recognize that an important difficulty with linear regression is that points at extreme values can have a large effect on calculated regression slopes. In their work they recalculate one regression line after leaving two such points out and that did change

the slope. Brommer presents the intercept of the regression line at zero change in species area given in Fig. 1a as 18.8 ± 6.1 km with $P = 0.002$, in ~12 years. The equivalent value for northward movement of bird species in southern Britain was 18.9 km over the longer period 1968–1972 to 1988–1991, ~20 years. However, in both pieces of work the regression relationship actually accounts for only small amounts of variation, i.e., $R^2 = 29.4\%$ (Brommer 2004) and $R^2 = 33.7\%$ (Thomas and Lennon 1999). In regression analysis, relationships can be weak, e.g., accounting for less than 50% of the variability, but statistical significance can be high.

In a research plan, the data statement should describe both how the postulates will be answered and the possible limitations to the answer that might be given by following the data statement. The limitations are particularly important as one moves from statistical inference to the wider task of making a scientific inference. Clearly in this example the statistical significance indicates rejection of the null hypothesis for Postulate 1a and acceptance of the null hypothesis for Postulate 1b so that both postulates are accepted — but clearly there are limits to that acceptance. The second component of the data statement indicates that the species used were necessarily limited and neither paper indicates that the third component of the data statement is met, i.e., testing whether the assumptions of the statistics are met by the data. This becomes important if one wishes to interpret the parameters of the regression equation. Brommer (2004) mentions that the rate of northward movement in Finland is greater than that in Britain. However, deciding whether this is true or false requires that adequate precision is made in the parameter estimates of both papers. While there is high probability of a significant regression relationship in both cases, i.e., that interception of zero change in the distribution axis is itself non-zero, the actual values of these intercepts depend upon precise definition of the regression relationships. Because the assumptions of regression have not been tested, and because the amount of variance actually accounted for is low, the actual values for northward movement are best treated with caution.

Both papers are best considered as exploratory analysis. The mark of effective exploratory analysis is that it enables a question to be refined. Brommer acknowledges that the cause of northward movement is not revealed by his study, and that the apparent faster rate of change in Finland needs additional evidence. Both studies allow some classification of species as to the amount of their movement and a more detailed comparison between species could lead to more precise analysis. If this were to be attempted then species might be classified according to whether or not they showed northward movement of range margins relative to change in area distribution. This introduces a new concept of a *bird species sensitive in its range margins to the estimated temperature change*.

Can the question be answered with precision?

Developing a research data statement provides a focus for detailed research planning that *includes* exploratory empirical investigation of a problem, frequently called pilot studies. Unfortunately exploratory investigation is the least practiced aspect of research planning, yet it is frequently essential to determine how an investigation can resolve a hypothesis test. Before starting an investigation it is essential to know what type of measurements to make and how many samples must be taken. The purpose of exploratory investigation is to define how measurements should be made and sampling schemes constructed to estimate the variance of measurements.

There are a number of reasons for the lack of exploratory investigations. Obviously they require time and effort but also it may be that researchers may not appreciate their importance in calculating whether a Type I or Type II error may be committed, which is affected by the number of measurements relative to the variance of the observations.

Type I error: A true null hypothesis is rejected. The null hypothesis for Postulate 1b is that there is no significant difference in the southerly range margins of northern species over the time period of measurements. It would be a Type I error if we found there was a difference when in fact there is not. We generally

set an α level of 0.05 in judging statistics so that there is a 1 in 20 chance that we might reject a null hypothesis even when it is true.

Type II error: A false null hypothesis is accepted. The null hypothesis for Postulate 1a is that there is no significant difference in the northerly range margins of southern species over the time period of the measurements. It would be a Type II error if we accepted this null hypothesis when there was actually a difference.

The key to avoiding a Type II error is to have sufficient power in your analysis. The *power of a test* is $1 - \beta$; the probability that a null hypothesis will be rejected as false if it is false. Four things influence power (Table 1).

1. The larger the actual difference between the quantities then the greater power — the less likely that a Type II error will occur. Of course, we usually investigate situations where we do not know if there is a difference or not so, generally, the actual difference is likely to be small rather than large.
2. The chosen value of α determines the confidence in the test of a statistical hypothesis, which is $1 - \alpha$, and is the probability that a Type I error will not occur. So, for $\alpha = 0.05$, the confidence in the test is 0.95, there is a 1 in 20 chance of rejecting the null hypothesis of no difference when there is no difference. But if we make $\alpha = 0.01$ then the confidence in the test increases to 0.99, a 1 in 100 chance of falsely rejecting the null hypothesis. However, as α is decreased then there is a greater probability that a Type II error will occur — that the null hypothesis will be accepted rather than rejected. As α decreases then we make it more difficult to detect a real difference. The compromise value for α is usually 0.05.
3. As the variance of the data increases then power decreases. So, if we assume that there really is a difference, and we really should reject a null hypothesis, the chance of rejecting it decreases as the variance of the data increases. This can be a real difficulty in ecology. We may investigate things that are influenced by multiple effects. For example, the range margins of some of the bird distributions measured may be affected by things other than change in temperature, such as changes in land use. This is likely to increase the variance of the data.
4. As the number of observations increase then power increases. This is because we determine the mean response with more certainty. This is one thing that an investigator *can* usually do something about. Increasing sample size increases power without decreasing confidence in the test. Obviously increasing sample size can be costly. What is needed is a calculation of the number of samples required to achieve a specific power of test and for this calculation we must have an estimate of the variance. This is why exploratory investigations are so important — to obtain an estimate of the variance of the data so that the number of samples needed can be calculated. Perhaps the most disappointing thing to see is when students chose to increase the number of treatments made, or conditions investigated, and sacrifice the number of replicates in order to do that without making any calculations of how that may affect statistical power.

We can take an example to calculate numbers required in a sample from a study by Ratti *et al*.

Table 1. Factors that influence the power of a test and their effects.

Influence	Effect
The actual difference between the quantities being investigated.	As the actual difference increases then the power increases.
The chosen value of α.	Power increases as α increases.
The variance of the observations, s^2.	Power increases as the variance decreases.
The number of observations made in each sample.	Power increases as the number of observations increase.

(2006) into the concentration of selenium (Se) in bird eggs. Ratti *et al*. outline the debate about the Se levels at which negative impacts occur ranging from values of 6–7 ppm to 12–15 ppm for the EC10 value (Effect Concentration at which 10% of the organisms are affected). Ratti *et al*. (2006) sampled bird eggs in southeast Idaho where cases of Se toxicity had been identified in farm grazing animals. Using stratified random sampling for ecological site types they sampled a single egg per nest from an area with mining sites and an area with no mining sites, which they referred to as a reference site. In 1999, they collected 215 eggs (reference $n = 98$, mining $n = 117$) from 27 species and in 2000, 329 eggs (reference $n = 175$, mining $n = 154$) from 24 species. They found no overall difference in egg Se concentration between years and so pooled the data and found significant differences among species and a significant species by year interaction. They conclude that treatment effects must be evaluated separately for each species and report significant difference between reference and mining areas for 16 of 24 species and conclude that "The remaining species did not have large enough sample sizes for statistical tests." Of course tests had actually been made and a more appropriate wording might be that for some bird species there was inadequate power and that no significant difference was concluded when, in fact, sample size was small, and so there is possibility of a Type II error.

The data collected can be used to estimate the number of samples that might be required. For example, Ratti *et al*. (2006) report values for the killdeer (*Charadrius vociferous*) of $\bar{x} = 2.38$ dry weight concentration ppm, $n = 5$, SE = 0.31 for the reference area and $\bar{x} = 8.09$, $n = 5$, SE = 3.91 for the mining area. We can calculate the estimated standard deviations, SD, from the formula for the standard error of the mean, $SE = \hat{\sigma}/\sqrt{n}$, as reference = 0.6932, mining = 8.743. This large difference in SD between reference and mining sites is itself interesting and Ratti *et al*. show a tendency in mining data for there to be a few large values.

A number of computer programs are available that calculate required sample size given the expected difference between means and estimates of standard deviations. As an example we use GPOWER (Erdfelder *et al*. 1996; http://www.psycho.uni-duesseldorf.de/aap/projects/gpower/) that provides calculations for a number of statistical tests including a *t*-test, correlation, ANOVA, MANOVA and χ^2. If we consider a *t*-test for the difference in Se content of killdeer eggs between reference and mining areas then we first enter the means and SD values based on the small samples we have. The program first calculates a non-centrality parameter for the assumed normal distributions, i.e., how much the distributions depart from zero. Then for an $\alpha = 0.05$, and power $(1 - \beta) = 0.95$ and assuming we will allocate samples equally between the reference and mining areas (the GPOWER program allows for different distributions) the calculated sample size to achieve a critical *t* value is 27 eggs from both areas. The important factor influencing the size of this estimate is the large SD value for the mining area eggs. If, for the sake of demonstration, the estimated SD for the mining area was half of 8.743 then the number of samples required would be 8 from each area. The distribution of Se content in eggs from the mining area tends to be skewed, with many low values and a few large ones (Ratti *et al*. 2006, Table 2) and Ratti *et al*. conduct their statistical analyses using log transformations.

The relationship between the difference expected between means and the number in the sample required to detect the difference is markedly non-linear. In designing a green-house experiment to detect the effect of a mineral element on plant growth, exploratory investigation of the plant without the element produced a mean height of 5 cm with a standard deviation of 0.75. The total number of plants required for an investigation, with the element and without it, is plotted against the expected possible differences between the means (Fig. 3) (divided equally between treatment and control). For $\alpha = 0.05$ and $1 - \beta = 0.95$ then to detect a difference of 1.5 cm requires a sample of 14 plants but to detect a difference of 0.2 cm requires a sample of 612. A difference of 0.1 cm requires a sample of 2438 plants (not shown in Fig. 3). Such curves can be used in two ways: (a) To estimate the sample size needed to detect an effect of a particular magnitude. (b) If sample size is limited, by cost or some logistic factor, to estimate the maximum effect that can be detected.

The relationship between statistical inference and scientific inference

The way in which a data based investigation is set up to answer a scientific question influences how the result of that investigation should be interpreted. Postulates 1a and 1b of the bird range margins investigation illustrate two approaches. In Postulate 1a the null hypothesis is the opposite of what the researcher believes, it is a reject-support (RS) situation (StatSoft Inc. 2007). For Postulate 1b the opposite is true. The researcher believes there is no southward movement of northern species and so accepting the null hypothesis supports his hypothesis (AS). In the RS situation a Type I error represents a false positive. From the perspective of journal editors, and the scientific world in general, such false positives are undesirable as they could lead to wasted effort in pursuing unwarranted ideas. This is justification for journal editors keeping α low. In RS testing a Type II error is a tragedy for the researcher because a postulate that is true is, by mistake, not confirmed and a researcher should make it a top priority to keep β low.

In AS testing, a Type I error is a false negative for the researcher's postulate and a Type II error is a false positive. So, for example, maintaining a very low Type I error rate would be loading things in favor of the researcher's postulate. Falsely rejecting (Type I error) and falsely accepting (Type II error) a null hypothesis are equally to be avoided. Most important is that the power of a test should be reported.

The hypothesis testing approach is frequently a sham in scientific research when the result of a supposed statistical test is translated directly into inference in a scientific theory. A large number of decisions are made in establishing the "test" to be used, about measurements, species to be used, areas in which the investigation will be made, etc. and the result of the "test" is conditional on those decisions. Rather than using a supposed test, whether based on an RS or AS approach, it is frequently more appropriate to use confidence interval estimation (Läärä 2009). Confidence intervals contain information about precision of estimates not available in testing (StatSoft Inc. 2007) and particularly how precisely a mean has been estimated.

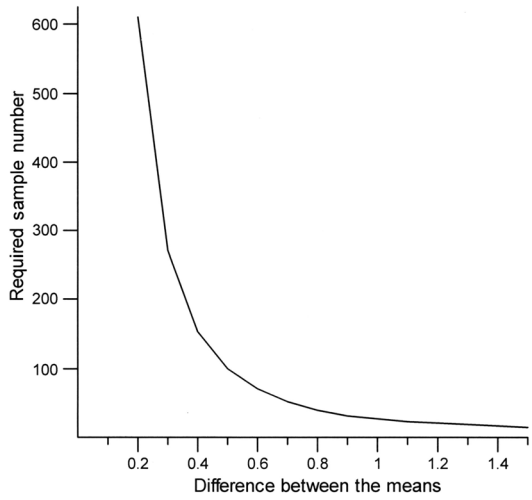

Fig. 3. The relationship between required sample number and difference between means that can be detected for α = 0.05 and 1 – β = 0.95 for an example with μ = 5 for the reference distribution and SD = 0.75.

The StatSoft manual (Statsoft Inc. 2007) notes that "Much research is exploratory. The fundamental questions in exploratory research are "What is our best guess for the size of the population effect?" and "How precisely have we determined the population effect size from our sample data?" Significance testing fails to answer these questions directly. Many a researcher, faced with an "overwhelming rejection" of a null hypothesis, cannot resist the temptation to report that it was "significant *well beyond* the 0.001 level." Yet it is widely agreed that a *p* level following a significance test can be a poor vehicle for conveying what we have learned about the strength of population effects (Läärä 2009).

A number of reasons have been suggested why interval estimates are less often reported (StatSoft Inc. 2007). These include: tradition, when testing is emphasized; pragmatism, for example when intervals are narrow but close to zero they might suggest a highly significant but trivial result and when they are wide they indicate a lack of precision; ignorance, many people are unaware of valuable estimation procedures; lack of availability, some procedures have not been implemented in statistical packages.

Further reading

The topics discussed in this paper are frequently aired in the literature. Some recent references are: Newman (2008) on the application of significance tests; Mcgarvey (2007) on problems in committing Type II errors in conservation biology; Strug *et al*. (2007) on sample size calculation; Garcia-Berthou (2001) on the analysis of residuals in regression analysis.

References

Brommer, J. E. 2004: The range margins of northern birds shift polewards. — *Annales Zoologicic Fennici* 41: 391–397.

Erdfelder, E., Faul, F. & Buchner, A. 1996: GROWER: A general power analysis program. — *Behavior Research Methods, Instruments & Computers* 28: 1–11.

Ford, E. D. 2000: *Scientific method for ecological research*. — Cambridge University Press, Cambridge.

Garcia-Berthou, E. 2001: On the misuse of residuals in ecology: testing regression residuals vs. the analysis of covariance. — *Journal of Animal Ecology* 70: 708–711.

Läärä, E. 2009: Statistics: reasoning on uncertainty, and the insignificance of testing null. — *Annales Zoologici Fennici* 46: 138–157.

Mcgarvey, D. J. 2007: Merging precaution with sound science under the endangered species act. — *BioScience* 57: 65–70.

Newman, M. C. 2008: "What exactly are you inferring?" A closer look at hypothesis testing. — *Environmental Toxicology and Chemistry* 27: 1013–1019.

Ratti, J. T., Moser, A. M., Garton, E. O. & Miller, R. 2006: Selenium levels in bird eggs and effects on avian reproduction. — *Journal of Wildlife Management* 70: 572–578.

StatSoft Inc. 2007: *Electronic statistics textbook*. — Available at http://www.statsoft.com/textbook/stathome.html.

Strug, L. J., Rohde, C. A. & Corey, P. N. 2007: An introduction to evidential sample size calculations. — *American Statistician* 61: 207–212.

Thomas, C. D. & Lennon, J. J. 1999: Birds extend their ranges northwards. — *Nature* 399: 213.