# Lies, Dammed Lies and Statistics

# or

# How to Prove that Storks Bring Babies

## Statistics and Rehabilitation

**Author**: J Hopper
Wildlife Assistance & Information Foundation Inc
judy@waif.org.au
0412 945 659

*Judith Hallinan fell in love with bats when she had her first microbat to care for at the age of nine. Then she read raising Archie and discovered the entire world of rehabilitation. From then on bats, particularly flying-foxes, became the passion of her life.*

*She has been rehabilitating for nearly 10 years and has been a Head Carer for a large part of that time.*

*Concerned at the number of flying-foxes routinely euthanised and a lack of knowledge about treatments she started a master of science by research under Dr. Gemma O'Brien at the University of New England. This is now nearing completion and she would prefer that nobody mentions the words "thesis submission".*

*Seven years ago Judy and a friend started WAIF (Wildlife Assistance & Information Foundation Inc). WAIF now runs a weekly vet clinic at St Ives and employs two of the few vaccinated vets. Not surprisingly WAIF treats a lot of flying-foxes.*

*In her spare time Judy works as a business consultant in project and quality management but would like to start her own fabric printing business to make Australian designs for quilting, embroidering and sewing.*

We are moving into a more scientific era of rehabilitation where rehabilitators are extending the types of treatment they do and attempting to work out which ones might be best in a wide variety of situations. This is a wonderful trend. Unfortunately most rehabilitators are not trained in research or statistics. So how do we assess claims made for various treatments and ensure that we are not simply creating more "myths"?

There are some, I believe quite normal, people who actually like the study of statistics and I have even met one or two of them. But most of us find the subject at best boring and at worst totally incomprehensible. Further we don't have time to become statisticians. On the other hand if we are to assess the information being given to us (and not overreact to our own data) we need to have some way to determine what the data is really telling us, i.e. some understanding of statistics. Note I say understanding not technical knowledge.

You can run statistics on just about any set of data you like but it is very much a case of garbage in and garbage out. If the results are to be of any use then the data has to be "good" data and the test done on it have to be the appropriate tests (biological systems require special tests). Statistical analysis is usually run on data specifically collected for that purpose, in other words data from scientific research. This research follows a stringent protocol (set of procedures) and the data collection also follows a stringent protocol. The aim of scientific research stringency is to remove any bias from the research method, the person applying the research method, the recording of the outcomes, and the deductions drawn. There are also other forms of data which may be used such as rehabilitator records, Census information, etc but, by their very nature, this data is more likely to have bias in its recording or in its interpretation, not to mention the methodology underlying its collection. So before getting tangled up in the statistical analysis side it is important to consider whether the underlying data is/was worth the effort to analyse. In other words were the activities which generated the data and the recording of the data well thought out, stringently applied, and not influenced by the people involved.

The first thing to be aware of is that in Australia everyone requires a scientific licence in order to carry out research on Australian native wildlife (this includes any "normal" activity if it is being used to record data for scientific purposes, e.g. spotlighting to estimate populations rather than enjoy an evening walk in the bush). This holds even if the wildlife in question has come in to a rehabilitation group or a veterinary clinic and is undergoing normal veterinary or rehabilitation activities. The first question to ask, therefore, is "what is the scientific licence number". In New South Wales this is issued by Department of Environment, Climate Change and Water, in Victoria it is the Department of Sustainability and Environment, in South Australia it is the Department of Environment and Heritage and so on. All of these also require that the research has approval by an appropriate animal ethics committee. Animal ethics committees are associated with all research organisations such as CSIRO and the Universities. On rare occasions an individual may obtain a research licence without an affiliated research organisation but they still require ACEC approval and this is usually through their state government ACEC. While neither a scientific licence nor animal ethics approval guarantee the *usefulness* of the research the processes to be followed usually mean that independent experienced people have reviewed the methods involved (suitable protocol) and the numbers of animals/samples involved (sampling techniques) and the required aim (outcome to be tested) with the intention of determining whether or not the research can meet the aim (in a welfare manner). If the researcher does not have a scientific licence and they are not affiliated with a research establishment than it raises a question over the validity of the research and therefore the statistical analysis.

The second question is clear: "is the researcher trained or experienced in biostatistics?". Biostatistics is a specialised branch of statistics because biological systems have some unique and headache making problems. If the researcher does not have this training (and the average vet, like the rest of us, does not) or has not involved a biostatistician then their entire statistical analysis is a very open to question.

A negative answer to the above two questions does not mean that useful and predictive data has not been collected, it merely raises a question.

There is an important distinction between data and information. Data is an actual measurable fact for example the bat weighs 3 g. Information is a judgement or conclusion drawn from data, e.g. it is a very small bat. Of course to extract information from data it must first be processed in some form. For scientific research this processing is statistical analysis. In the previous example it was processed through my personal experience with bats, but note that the bulk of my experience has been with one kilogram flying-foxes. I may be considered to be biased or just plain wrong although not deliberately lying. If I had used 14g would you have considered this bat small or would you have considered it average or even large?

To obtain clear measurable data which can be used for scientific analysis very stringent protocols (methods) must be followed. This is rarely the case for normal rehabilitation practices.

One of the stringent requirements for scientific protocols is that the protocol must be followed exactly. For example I was trialling a film dressing on a wing wound. While I was away for two days an assistant accidentally used oil and not a dressing. In most rehabilitation situations this would not be a problem, and in fact would probably not even get recorded, however for scientific analysis that injury treatment combination could no longer be considered in the data set. This raises an important issue in the collection of data for scientific analysis. The protocol of what is recorded and when it is recorded is as important as the protocols for the research itself. This is one reason why conclusions drawn from an analysis of rehabilitation records must be carefully considered. This is particularly the case for retrospective data, that is data which was not originally recorded for the purpose now being analysed. This is so whether the data was recorded purely as a rehabilitation record or whether it was recorded as part of another research project. In my previous example, if that bat had originally been part of another project, such as response to stress, it is quite possible that the two days of incorrect treatment may not have been noted. If, however, I later (retrospectively) reviewed the data for success/failure of treatments and recorded a failure against dressings I could quite easily be doing so incorrectly. The failure, if one occurred, may have been due to the lack of a dressing for two days. Further, in rehabilitation practice it is very difficult to maintain the treatment in the face of a worsening condition in an animal. The normal rehabilitator instinct is to try other treatments. This is fine as a welfare response but not suitable as a scientific protocol.

What is done to one must be done to all - exactly, to the letter. If this isn't followed then the data from that animal can not be included. This is a stringent requirement for scientific protocols: only one variable is changed at a time. This is rarely done in rehabilitation practice and may not be well recorded if the original research was for another reason. This sort of thing can be looked for in the “small print” of some reporting papers and worried about in any paper which is short on detail in the methods section. Such throwaway statements as “other normal nursing practices followed”, “normal veterinary care as required” and so on can hide a multitude of statistical sins. For instance, in the above example, the use of antibiotics, anti-inflammatories, massage and other treatment activities may in fact be the contributing cause to the success or failure of the dressing. Another hiding place to check for these types of issues is the “standard practice” of the group. Continuing the previous treatment analogy, standard practice can be Metacam for three days, dex injection (despite research showing it is contra indicated), antibiotics and so on. These are often not recorded just because they are so standard and therefore taken for granted, e.g. some commercial feeds have a coccidiostat in them which is obviously going to impact coccidiosis results in macropods.

Another point to check is where the paper, if there is one, is published (if there isn’t a paper get very nervous). The reason papers are published in scientific journals is because they are "peer-reviewed". This means that a committee of people with some understanding of the topic and some understanding of statistical analysis have already done a filtering on papers and rejected those with poor processes and analysis. It isn't foolproof but the more respected the journal in which the paper is published the higher the chance that the work is done correctly. So you can start to trust the conclusions they are drawing.

A word of warning on papers. Be careful of basing your belief in the “claim“ being made because of the writer has “claimed“ certain results achieved by other researchers. Always, always, always check the references yourself before believing them. And this is regardless of where the paper was published. I have known some truly terrible misreporting not to mention blindly stupid interpretation of other people’s work reported in scientific papers.

Okay - so at this point you are prepared to give the processes leading up to the analysis the benefit of the doubt. Now we need to consider the statistical analysis itself

To understand statistics

- you need to know what it is (and isn’t),
- be able to determine whether or not the statistic is a “good” statistic, and,
- recognise whether that statistic is useful in the conclusion being drawn from it.

This is not difficult for rehabilitators as we do it all the time without realising it. We just need to be a little more stringent in what we believe.

I'm going to use plovers as an example as I go along. The plovers are to prove that I do know of the existence of wildlife which are not bats. So try to imagine that for some reason I have taken it into my head to that plovers need to increase their reproductive success and I am going to treat them.

What is a statistic? (In some practical definition which may be useful)

Well it is usually a number and it needs to carry some sort of information. The plover has five eggs is a number - this is our data. But unless you are estimating the size of the incubator needed to hatch them it does not carry much additional information (and even then you need to now the size of the eggs). It is in fact a single point of data - a datum. If I tell you that this is a 25% increase over last year then that is a statistic. Furthermore, given my initial desire, whatever treatment I did appears to have worked. Right? Maybe, maybe not. It brings us to our second consideration “is it a good statistic?”.

What defines a good statistic?

This is the essence of statistical analysis. And where we get into all the horrible tests and degrees of freedom and n equals something and p equals something else and log transforms and normal distributions etc. Use the wrong tests, carry out the wrong transforms on the data, include the wrong data points and the statistic may look good but be useless and misleading. This is when an expert is needed.

So, without a statistics degree how are we going to recognise good from bad? Well first of all don't throw your normal intelligence out the window because somebody has waved a statistic at you. “Baffling with bullshit” is an old art. Bigger is not necessarily better.

If I tell you (possum rehabilitators) that my possums eat (toxic) oleander leaves are you going to rush out and feed all your possums on oleander? I hope not. I hope you will ask me a few (pointed) questions. How many possums ate the oleander leaves and survived (and how many died) would be a good start. Starting with my competency to make this decision might be better, i.e. can I tell the difference between a dead and a live possum, or possibly do I know an oleander if I fall over one. If you get past those two then looking at the research protocol comes next: did I feed raw leaves or had I processed them, how did I ensure/determine that the possums ate the oleander leaves, what percentage of the browse constituted oleander leaves and so on.

One of the big areas of concern is the sample size. There is a whole area of statistics on sampling theory. The aim is to have a sample size big enough to represent the normal population but as small as possible to reduce impact on the population and general/random enough to represent the population and not be a biased sample. This is a good area for exercising commonsense. If a person tells you they have “treated” 250 animals in the last year (or even three years) then you might, reasonably, get suspicious. Very few people could manage that level of animal intake. And even fewer are going to have that number of animals with the injury/illness to be treated/tested. You may discover that the 250 were all the animals who came into care over that period. Only a small percentage actually presented with the appropriate illness/injury and these were treated by 10 different people. So n does not equal 250! For example I sourced flying-foxes from all over NSW for my Master’s wing treatment research. Approximately 150 flying-foxes were treated, by me, for injuries during the period of the study. Of these nearly 140 had wing injuries (10 had other injuries and so weren’t included). Of these flying-foxes, 90 were submitted to the research program (the rest were eliminated due to: problems with treatment pre arrival, complications requiring multiple treatments, arrived with healing already too advanced, wrong type of injury - e.g. pulled muscle, etc). 75

flying foxes were ultimately included (although 7 of these did not develop injuries). The injuries of the 68 flying-foxes, totalled 300 abrasions, bruises and blisters. So you can see that n may appear to be 150 but is really 68. A big difference. Actually n is 300 because I was treating injuries not animals. So keep focussed and don’t get react to numbers based on size - take a close look at the number.

Who carried out the treatment is also significant (no pun intended). The experience of the rehabilitator has a major impact on survival of the animal. Taronga zoo clinic discovered that with inexperienced nurses (everything else the same) they had a survival rate of 25%. Over time this rate soared. Then the nursing staff changed and it plummeted again. The same effect of experience can be seen in the success of treating oiled sea birds. It improves with experienced staff. This can be particularly important in rehabilitator data. The mere fact that some-one has decided to focus on a particular treatment may mean that all those animals are now being handled by only experienced people, the rehabilitators received special training, and/or they are handling far more than they used to do and so are rapidly increasing their experience. Which may mean that improvement is due to experience and treatment is irrelevant. Not to mention that some people understand instructions better than others. “Use honey and don’t give it anything else” being variously interpreted as: “only put honey on the wound”, “use mac oil until you feel like using honey”, “give it Metacam for 3 days (with or without the honey)” and so on. None of these (other than the first) could have been included in the research collection, which is one reason I went from 140 to 90 in the above example.

So apply your commonsense to a statistic to determine significance (i.e. what was done brought about the result). If I have five eggs (back to the plovers) and a 25% increase then I must have had four eggs last year (I mean the plovers of course). So I have one extra egg this year. If you know plovers then this doesn't sound very exciting. Any pair of plovers might have four eggs one time, five eggs next time and then three eggs and so on. This variation in numbers is called normal variation, noise in the system, expected fluctuation, etc. In other words it is <u>not statistically significant</u>. It barely rates as an indication that my treatment might be having an effect. This is the first place that statistics are <u>misused, deliberately or otherwise</u>. If I have applied this treatment several times and to several pairs of plovers and every time I get a 25% increase (reproducability) then perhaps I have statistical significance. Reproducibility is one of the key points of research based statistical analysis. It is important to distinguish a genuine improvement or change from normal variation or noise.

The next thing to look at is what else I was doing with the plovers at the time. This can have a serious impact on the outcome and is often hidden in the small print or the throwaway line. Particularly the one “applied normal veterinary procedures otherwise”, as previously mentioned. When you enquire into that one you tend to find that the entire so-called statistical proof isn't worth the paper it is written on. For instance you may find that my treated plovers were also nesting in laboratory controlled (warm, quiet, safe, high nutritional diet) conditions as compared to the untreated plovers who were trying to produce eggs in a paddock in the middle of a five-year drought. It is near enough to impossible to attribute an outcome (such as improved healing or increase in plovers eggs) to an activity if more than one activity has been applied (there are some truly horrible statistical techniques for this but let's not go there). This is why, as I said before, scientific experimentation is so stringent in what can or can't be done during the experiment. The aim of the stringency is to ensure that only one variable changes and therefore if an outcome changes it is due to that variable. So if I put my plovers on a healthy diet and treat them with catfish extract and reduce all their stress I can hardly claim that the catfish extract caused that extra egg. The need for stringency is why it is very hard to use data from veterinary or rehabilitation records because in general too many other things vary; such as accompanying treatments, conditions during rehabilitation, conditions during injury etc. A real-life example is arnica which I trialled on my flying-foxes with high hopes. These were completely dashed to the ground. Further investigation suggests that the people who were/are telling me that arnica works wonderfully for bruises were also using cold compresses, limb elevation, massage, antibiotics, anti-inflammatories or all of the above.

Another thing to look at is when the researcher decided to use the data for the analysis they are presenting. If I had originally decided to use catfish extract for fertility than I designed the observation/measurement process to match (or should have done). If I designed it for something else and decided half way through that fertility was improving, or even decided to examine previous results from another study (retrospective analysis) then bias, inaccurate observations and other problems are likely to be introduced. If I was actually giving catfish extract to improve plumage I may have kept a less than accurate record of all eggs laid. For example I may not have recorded broken eggs. In fact I may have recorded chicks hatched, or birds released, and be extrapolating the number of eggs laid.

It isn't always easy to interpret presented results for stringency and accuracy. This is why I recommend starting with the research and finding out if they have a scientific licence etc. Getting a scientific licence and animal ethics approval is frequently a nuisance but does have the advantage that somebody with statistical knowledge or experimental understanding has done at least a preliminary review of the procedures to be followed in the gathering and measurement of useful data. They would have pointed out that the number of eggs hatched doesn't equal the number of eggs laid. And that in a lab loss of eggs to predators won't occur so "laid" may be closer to "hatched" in the lab than in the wild. In fact plovers in the wild may lay 6 or 7 eggs and catfish extract decreases fertility.

Even if everything seems scientific and appropriate etc you still need to consider the interpretation being made, i.e. the conclusions being drawn. I was at the international Wildlife Disease Association conference a couple of years ago when some Japanese scientists reported on their findings of disease in whales. It looked like a good justification for whale hunting if one identified disease. One person stood up and said "you have identified a range of potentially disease causing organisms but were there any signs of clinical disease – were the whales in fact sick?". The answer was "no". In other words they could statistically prove the presence of organisms but the whales were fine. Good statistical analysis but poor implied conclusion or interpretation. I can, and have (as an undergraduate exercise in first-year statistics), easily shown a direct, positive relationship between the birth rate of humans in England and the number of storks, therefore storks clearly bring babies. QED!

I don't want to discourage the gathering of veterinary and rehabilitation data and the development of information from it. I am a strong believer in the value of knowledge collected by rehabilitators. I do want to raise a warning that analysing and interpreting data is a rich source of myth development. Until such data (and the collection protocol) has been analysed by a biostatistician the conclusion should not be accepted as gospel and become part of standard practice. Such conclusions need to be reviewed by:

- people who understand research protocols and statistics (e.g. researchers and statisticians).
- people who understand the underlying processes involved (e.g. biologists, vets, physiologists), and,
- people who know the animal involved (e.g. rehabilitators, wildlife vets, ecologists).

Maybe the NWRC should put together review panels to analyse presented results!

Nor should the baby be thrown out with the bathwater. Even if proven to be not statistically significant it can often point a strong indicator. Better still if you intend to investigate a treatment ask a good biostats person for some help on how to collect useful data BEFORE you start. And develop good and standardised recording protocols for rehabilitator records so that the mass of data we can all collect can be used!