Practical data skills for the farm animal vet

Robert Hyde, Martin Green, Chris Hudson, John Remnant

Abstract

Whilst farm animal veterinarians receive extensive training in the diagnosis and treatment of individual animals, population medicine or “herd health” have also become commonplace in modern production animal veterinary medicine, with clinicians using routinely collected farm data to identify epidemiological patterns to implement preventive management changes. Production animal veterinarians are in a unique position having to large volumes of farm data available and having the means to enact necessary changes on farms based on data analysis. Whilst data handling and statistical techniques can initially be challenging to learn, a relatively small investment in time spent learning can result in dramatically shorter times analysing farm outcomes, and a useful evidence base on which to base on-farm management decisions. This article aims to provide a practical guide to data skills for the production animal veterinarian, briefly covering basic descriptive statistics and graphical representations of data, with further examples of more advanced techniques such as statistical modelling.

Introduction

Farm animal veterinarians are frequently faced with disease or productivity issues that go far beyond the clinical treatment of a single animal. There are often many potential factors influencing these issues, and often the relative importance of these factors are unknown. Whilst farm animal veterinarians receive extensive training in the diagnosis and treatment of individual animals, population medicine or “herd health” have also become commonplace in modern production animal veterinary medicine, with clinicians using routinely collected farm data to identify epidemiological patterns to implement preventive management changes.

The use of statistical analysis and data science have become more common, particularly as computing power has increased the availability of statistical methods. Whilst farm animal veterinarians are unlikely to have received extensive training in statistical methods, they will often be utilising many statistical principles already within their clinical work. Determining whether a change in treatment protocol has altered clinical cure rates, a change in calf milk powder has altered growth rates or a new fertility protocol has improved fertility efficiency, all require an understanding of basic statistics. Many modern farmers now require this information to be analysed and communicated to a high level from their routine veterinarian. This has been identified as important role for farm vets in the future (Woodward et al., 2019).

Despite a large volume of medical research exploring advanced analytical techniques, examples of translation into an effect in a clinical setting are relatively rare. The use of big data in farm animal medicine is increasingly being discussed (Hudson et al., 2018), and production animal veterinarians are in a unique position having both access to large volumes of farm data and the means to enact necessary changes on farms based on the data analysis. This article aims to introduce the reader to some of the underlying concepts used in data analysis as well as providing some practical tips to apply these concepts in a clinical setting. The principles outlined may also be of value when conducting research in practice or when interpreting the results of published research. Whilst extensive training statistical techniques would be beyond the scope of a single article, the purpose of this article is to highlight a selection of useful methods, and provide a basic overview of statistical techniques, and an introduction to statistical coding that will have practical value to the farm animal veterinary clinician.

Sources and types of data

In addition to on-farm records of reproductive management, treatment records and disease incidences, many other sources of electronic data are available. These include milk recordings, milking parlour software and environmental or animal sensors. It is rare to find a solution that combines all these data into an easily accessible format for all analyses. Where more analysis is required beyond those available in proprietary herd health software it will usually be necessary to export sensor or on-farm records into a common format (such as a spreadsheet or comma separated values file) for analysis. The detail of how to do this is beyond the scope of this article, however sensor and machinery manufacturers may be willing to provide advice in individual cases. One downside to data analysis is the manual entry of data. Paper based recording obviously necessitates manual data entry into electronic form which can be time consuming. If the veterinarian is expected to manually type up paper-based records this should be encompassed within the consultancy charge.

Once data has been collated/exported for analysis, it is important to understand what type of data is being analysed. This will inform the statistical or data science methods that can be used (Figure 1).

As well as understanding the source and type of data, the data should also be critically assessed for potential flaws. The same is true of any metrics or key performance indicators (KPIs) calculated from the data. A detailed description of interpreting metrics is provided by Hermans and others (2018), and lag, momentum, bias and variation may all affect interpretation (Overton, 2009).

Data analysis overview

Summary statistics

Simple descriptive statistics are often a sensible starting point in data analysis and can provide a useful overview of your data. Descriptive statistics often provide some sort of measurement of central tendency (Figure 2) and some measurement of the spread of data (e.g. variance/standard deviation or range, see Box 3 and Figure 7). Data can also be tabulated to show the number or proportions falling within different categories.

There are several pitfalls to avoid when interpreting basic statistics. Consider a 100-cow dairy herd, where you are asked to interpret results from a milk recording. Figure 3 shows the somatic cell count (SCC) results for each cow. A relatively high mean of 208,000 cells/ml might seem relatively concerning at a herd level. On further analysis however, most cows have a relatively low SCC, and the mean is being largely influenced by a small number of cows with a particularly high SCC. This is an example of a right “tail” or “skew”, as the mean (sum of all scores divided by number of samples) is higher than the median (“middle” score when ranked) SCC is far lower at 115,000 cells/ml.

This example highlights the downsides of utilising single descriptive metrics for data analysis in isolation. A simple histogram (Figure 3) shows immediately that there is a skew to the data, allowing the clinician to focus any interventions in a more targeted area.

Visualising data

Graphical representation can be extremely powerful in data analytics, and even the most advanced statistical techniques often start with a simple visualisation of the raw data distributions. Carefully organised graphs can be extremely powerful knowledge exchange tools, taking relatively dull data and transforming it into something intuitive and engaging. Whilst here are a great number of statistical tests available to assess the structure of data, graphical representation of data can be one of the easiest methods of getting to know the dataset. Visualising data can often be more informative than calculating summary statistics and can also be particularly useful when assessing data quality.

There are an enormous number of potential graphical software options to choose from, and a wide range of techniques available depending on whether you are comparing one, two, three or even more variables. Graphs can be created in Excel, and pivot charts can be useful to quickly investigate relationships between many variables. The use of Excel based graphs however are often relatively time consuming to repeat when analysing many farms compared with code-based approaches that can be replicated for multiple farms in a fraction of a second, presuming data is recorded in the same format. This example dataset involves weights taken from preweaned calves at a range of ages. Graphs shown in Figure 4 were created using the ggplot2 package (Wickham, 2016) in R (R core team, 2020) (see *Software*). A tutorial in downloading and getting started in R is available in *Appendix i*, and code for the creation of these graphs is available in *Appendix ii*.

Box 1: Graphical representations of data (Figure 4)

Scatter: A scatter plot is a simple way of examining two continuous variables, in this case age (d) vs weight (kg). As age increases weight unsurprisingly follows, and there seems to be a strong pattern. By visualising an entire population in one graph however, more subtle patterns may be missed.

Regression line: Adding a “line of best fit” effectively demonstrates the association between age and weight, i.e. for each increase in 1d, how much does weight increase. By adding an equation to the graph, we can see that the formula for weight is best represented by:

```
weight = 42 + (0.8 x age)
```

This means that the average daily liveweight gain is 0.8kg/d, with a y-intercept of 42 (the weight at 0d of age – an estimation of birthweight). This is useful to determine a rough daily liveweight gain (DLWG) for a farm, however estimations from this method will be inaccurate if a small number of weights are used. A more accurate method for calculating DLWG would be to take birthweights and weaning weights for each individual animal.

Colour: By adding a grouping variable “sex” in this case, it now becomes apparent that whilst the overall DLWG may be around 0.8kg/d there are clearly differences between sex in terms of growth rate. Whilst heifers (blue) are generally born lighter than bulls (red), their DLWG appears to be higher than the bulls, which may be suggestive of management differences between the groups to be explored in more detail on the next farm visit.

Histogram: By “binning” observations into chunks it is possible to visualise distributions of single continuous variables. This graph shows that whilst most calves are weaned around 100kg, there are a small population of calves that are being weaned at 110-120kg. It is possible these animals could have been weaned earlier and would be worth investigating why these animals were this heavy by the time they were weaned.

Boxplot: If it is possible to calculate DLWG for each calf from a birthweight and a weaning weight, DLWG can be expressed in boxplots to indicate the median (middle line), and interquartile range (50% of calf DLWGs are within the box, see Box 3) and range (top to bottom of lines) to allow visualisation of spread in data. Whilst it is important to achieve a high DLWG, it is also important to do so consistently, and it appears in this case that heifers achieve both a higher and more consistent DLWG than bull calves.

Violin plots: A version of the boxplot, where the width of the “violin” demonstrates the number of observations at a given value. Whilst there are a small number of bull calves that have a higher DLWG than heifer calves, they are relatively inconsistent, with heifer calves being much more consistently close to 1kg/d DLWG.

Regression

There are various types of statistical “model”, which are methods of exploring relationships between variables that we might use in a clinical veterinary environment. Regression models are often useful in both clinical and research settings. Regression allows the strength of the association between two variables to be measured. Regression also has the benefit that multiple variables can be investigated simultaneously (multiple regression). It is also possible to carry out regression with a binary outcome (logistic regression), where the binary variable is transformed to a probability-based scale. Linear regression is a commonly used statistical method within research but is relatively easy to perform in practice to determine effects at a farm level.

Consider a dairy farm that manages their beef and dairy calves separately and wish to analyse the performance of both groups with their veterinarian. The farm has been weighing calves >21d, and after plotting the ages and weights of calves, it appears there may be some difference between the two groups. A simple regression model can put a numeric figure on this difference, which may aid in clinical decisions.

It is vital for regression models that all data points are independent (the *assumption of independence*). For example, if the only Belgian blue bull calf were weighed 5 separate times, the effect of being a beef calf might be exaggerated through multiple inclusions of a single fast-growing animal. For these models it is important that each data point comes from an independent unit (in this case a single calf weighed once) with as few biases between groups (such as breed, sex or housing) as possible. The failure to achieve independence of observations may lead to incorrect model results, and whilst in research this can be achieved by prospective random allocation to groups, this will not always be practical in a clinical setting. “Mixed-effects” models can be used to overcome issues with independence and can account for example multiple measurements per calf, or multiple calves per farm within the model to account for the effect of “calf” or “farm” on the outcome of interest. Even with appropriate study design and model analysis, correlation does not necessarily imply causation, and it is important to fully understand the dataset through graphical analysis as well as statistical modelling.

Box 2: Linear regression

To analyse the weights of beef and dairy calves, a clinician might first examine the weight of calves by age. This can easily be done in Excel by highlighting the "Age" and "Weight" columns and inserting a scatter plot. A trendline can be added to the graph, and a linear regression equation can be added to the line. In this example, calves gain on average 0.85kg for every 1d increase in age (a DLWG of 0.85kg), and the y-intercept (weight at age 0d) is 31.9kg on average. An R^2 of 0.87 means that around 87% of the variation in weight is explained by the predictor variable in the linear regression model.

Whilst plotting data in Excel (Figure 5) allows the analysis of weight by age, it would not be easy to add more variables for analysis. For example, we might be interested in whether beef or dairy calves grow faster (Figure 6). By using coding software, it is extremely easy to add or remove as many variables as required as shown below.

The formula for a regression model is shown below. This model will attempt to model weight, by using the variables age (d) and breed type (either beef or dairy), which would have been challenging to analyse in Excel.

```
weight ~ intercept + age + breed_type
```

Variable	Coefficient	95% confidence interval	P value
Intercept	40.62	37.22-44.03	<0.001
Age	0.80	0.76-0.85	<0.001
Breed: Beef	7.15	4.52-9.78	<0.001

Table 1: Results from regression model exploring effect of age (d) and breed type (beef or dairy) on weight (kg).

Results from this model are shown in table 1. An "*Intercept*" of 40.62 suggests that at day zero (birth), calves weigh on average 40.62kg (and we are 95% sure that this value lies between 37.22 and 44.03kg). A coefficient of 0.80 for variable "*age*" means that for every 1d increase in age, calves can be expected to grow 0.80kg (effectively a DLWG of 0.80kg/d). A 95% confidence interval effectively means that whilst our best guess is 0.80kg/d, we are 95% certain that this figure is between 0.76-0.85. Whilst model outcomes suggest an association between age and weight, this does not always indicate causality. In this model every 1d increase in age is associated with a 0.8kg increase in weight, however the increase in weight is not directly caused by the increase in age. A coefficient of 7.15 for the categorical variable "*Breed: Beef*" suggests that beef calves are 7.15kg heavier at any given time point compared with dairy calves (the default or "Reference" category "Breed: Dairy"). Again, our confidence interval shows that we are 95% sure beef calves are between 4.52-9.78kg heavier than dairy calves. In this example, it appears we could be quite confident that beef calves are significantly heavier than dairy calves. Whilst this model is likely to provide a robust estimate of calf weight within the age ranges shown, it is important to be cautious when extrapolating model predictions to areas with no data. In this example we have data from ~21-120d of age and the effects of age and breed type on weight reported in the regression model be accurate if applied earlier or later in life. The R^2 for this model is 94%, which indicates more of the variation in weight is explained by the predictor variables when age and breed are accounted for than the previous Excel model using only age. Multiple variables can be added to regression models such as the housing type, feeding and other management factors, and regression models provide a powerful method of determining the impact of multiple factors on an outcome of interest such as DLWG. The concept of "overfitting" in model building is largely beyond the scope of this article and describes analyses that correspond too closely to a particular set of data but failing to fit with additional data. For example, by including "Calf ID" in our model, we might well be able to predict DLWG perfectly, but the model would not be very useful for the prediction of DLWG from future calves, or different farms. Graphical analysis of data in conjunction with model outcomes should help prevent errors in the over interpretation of model results in relatively small datasets.

Box 3: Statistical definitions

Correlation: The degree of association between two variables.

Interquartile range: A measure of the variability or "spread" of data representing the "middle" 50% of the data; between the 25th (1st quartile) and 75th (3rd quartile) percentiles.

Variance and standard deviation: Measures of the variation around the mean. A lower variance or standard deviation suggests datapoints are closer to the mean value (see Figure 7). Standard deviation is calculated as the square root of the variance, and variance as standard deviation squared.

Linear regression: A statistical modelling technique that can be used to analyse the strength and nature of an association between single or multiple variables and a continuous outcome such as weight.

Logistic regression: A statistical modelling technique that can be used to analyse the strength and nature of an association between single or multiple variables and a binary outcome such as conception to a given insemination.

Multivariable regression: Either linear or logistic regression where there are more than one predictor variables included such as the effect of both powder type and age on weight.

Machine learning: A set of algorithms that can be used to classify or predict an outcome by "learning" patterns from data, often outperforming more traditional approaches such as regression in terms of predictive performance.

Black-box: Where the effect of individual variables on the outcome are challenging to interpret (often used to describe types of machine learning algorithms)

Box 4 – Useful spreadsheet functions

Spreadsheets can be a useful and accessible way to store, analyse and visualise data. The internet is full of discussion boards and blogs giving advice on functions and techniques for spreadsheets and other analysis software. Using a search engine to search for what you want to achieve is often the quickest way of finding an effective approach. The help features in the software and manufacturers websites and instructions also usually provide details examples of using functions. The following examples from Microsoft Excel (Table 2) are functions or features that the authors have found valuable when reviewing data, equivalent functions will be available in other software packages.

Function or feature	Description	Example usage
"=" or "Fx"	Used to enter a function into a cell, you can type "=" at the start or click on the *Fx* symbol in the formula bar. Using the formula bar allows you to search for functions and see a description of the correct usage more easily.	To enter a formula or perform a calculation in any cell. The Fx bar can be used to search for a formula and check how to enter it.
Fill handle	The small box in the bottom right corner of the cell can be dragged down or across to fill the following cells. Double clicking the box fills the column downwards to the end of the data. Excel recognises a range of options here from days of the week or dates to sequences of numbers or letters. The fill handle will also allow you to drag a formula across lots of rows.	To drag a formula down across multiple rows of data to perform the same calculation on lots of lines of data.
Static and dynamic references	Dragging a column down in Excel usually updates the "reference" (the other cells the formula uses) with each cell i.e. they are dynamic (cell A1+ cell B1 in cell C1 will change to A2+B2 in C2 etc). You can use a $ sign (or press F4) to make them static (cell A$1+ cell B$1 in C1 will still be A$1+B$1 in C2 etc when dragged down).	To fix a cell in a formula so that when you drag and fill the formula this part doesn't change. For example, a mating start date or average milk yield value that will be used lots of times on different rows of data.
Custom sort	The custom sort option allows data to be sorted by multiple columns.	To sort calf weight data by eartag number and then date so that all weights for the same calf are listed, in date order, together
IF	Returns different values or text depending on whether a	To recode continuous data (e.g. Milk somatic cell count)

	logical statement is true or false e.g. =IF(logical argument, value if true, value if false). If functions can be "nested" within each other to allow further logical arguments to be used.	as a category (e.g. infected or not infected). To apply a formula only if certain criteria are met e.g. if this eartag number matches the one above calculate the difference in weight, if they are not the same leave the cell blank.
COUNTIF	This counts the number of cells in a given range that meet a specific condition e.g. =COUNTIF(range, criteria)	To count the number of cows with a somatic cell count of over 200,000.
AVERAGE	The AVERAGE function calculates the arithmetic mean of a selection of data, other measures of central tendency are available and include MEDIAN, MODE. Other summary statistics are also available for example MIN and MAX.	To calculate the average age at first calving from a list of calculated age at first calving dates.
VLOOKUP	VLOOKUP can be used to lookup data from another worksheet using an identifier common to both sheets e.g =VLOOKUP(lookup_value, table_array, col_index_num, range_lookoup). The lookup_value is the cell in the first table that is being referred to, the table array is the second table, the col_index_num is the column number of the second table that contains the data of interest. range_lookup is either TRUE (for an approximate match) or FALSE (for an exact match). INDEX and MATCH formulas are more advanced but can be used to provide greater flexibility than VLOOKUP.	To add the date of birth of calves to a table containing weight data. The calf eartag number can be used as the identifier and a column added to a table of weights that uses VLOOKUP to find the date of birth in a different table containing calf eartag numbers and birth dates. This would allow the age at each weight recording to be calculated.
Pivot tables	Pivot tables are a way to summarise findings from longer tables and lists of data. There are lots of ways of using	To summarise individual cow data to produce a summary plot of average milk yield by calendar year or to summarise milk recording data by

	and combining data in to a pivot table (or pivot chart).	displaying counts of infected cows by calendar month.

Table 2: There are a number of useful tools and functions available in Excel that can be used to quickly analyse and report farm data.

P-values, hypothesis testing and sample sizes

P-values are usually calculated for hypothesis testing. This is when a research or clinical question is rephrased into a hypothesis – a statement that can be tested. The alternative hypothesis is usually the statement that would be true if there is a difference, for example "calves fed milk replacer A will have a higher growth rate than calves fed milk replacer B" or "Cows artificially inseminated by operator X are more likely to become pregnant than those served by operator Y". The null hypothesis is the opposite statement, in this case that there is no difference between milk replacer A and B or operator X and Y. There are lots of hypothesis tests available, an extensive description of all possible tests is beyond the scope of this article. When using statistical tests on data, it will often be necessary to decide whether to use parametric or non-parametric tests. Data that are parametric follow a typical distribution that can be defined with "parameters" that follow certain patterns. The most common parametric distribution described is the "normal" or "gaussian" distribution. A normal distribution can be described by its mean and standard deviation, we know that 95% of the data will be spread symmetrically 1.96 standard deviations either side of the mean value (Figure 7).

Many parametric statistical tests will have an "assumption of normality". Whilst the software may give a value, if the assumptions are not met the results may be inaccurate. If in doubt, non-parametric tests are usually considered a more conservative approach, being less likely to overestimate statistical significance.

Common hypothesis tests and the type of data they are used on are shown in table 3.

Type of data	Hypothesis test	Example
Comparing two continuous, parametric, normally distributed groups	T-test	To test for differences in daily live weight gain between calves fed two different milk replacers
Comparing continuous non parametric outcomes between two groups	Mann-Whitney test	To test for differences in age at first calving between heifers reared on two different diets
Comparing multiple groups with a continuous outcome	ANOVA	To test for differences in daily live weight gain between three different breeds of calf
Comparing categorical data e.g. proportions	Chi-squared tests	To test for differences in the proportion of cows becoming pregnant following insemination by two different operators

Table 3: Common hypothesis tests for various data types.

In a research setting, P-values being <0.05 is a conventionally accepted threshold to indicate the statistical significance of a result. The definition of a p-value is the likelihood that the results would have been obtained if the null hypothesis was true (for example that there was no difference between two interventions being tested). In simple terms, a p-value of 0.05 effectively means that if a study was repeated 100 times, we might expect this result to occur purely by chance five times, even if there was no difference. The somewhat arbitrary cut-off of $p<0.05$ suggesting *statistical* significance does not guarantee *clinical* importance, and similarly $p>=0.05$ does not mean there is *no* clinical importance. P-values do not describe effect size, and clinicians must consider the likelihood of a true difference between interventions, likely effect sizes and the cost of potential interventions.

For example, if a fertility treatment were shown to increase pregnancy rate with a p-value of 0.04, would it be worth the clinician using the treatment purely because the result was "statistically significant"? This result may have occurred by chance, although it is much less likely than if the p-value were 0.4. There may be a true difference between treatment and no treatment, but it is also worth considering the effect size and cost of the intervention. If the intervention cost £10 and resulted in only a 1% increase in conception rate this may no longer be worth considering even though there was a "statistically significant" difference. Similarly, an intervention costing £10, with a 20% increase in conception rate, but with a p-value of 0.07 might still be deemed clinically worthwhile, (if the sample size of animals tested was larger, a true difference might have been found).

This concept can be applied to clinical examination of animals; if examining a dairy cow that is clinically normal with a temperature 0.1^{o}C higher than the clinicians' "normal" range, this would might constitute a statistically significant finding (being outside of the normal 95% coverage range for normal individuals) but the clinician may not consider this clinically significant in terms of deciding on treatment. Conversely, a clinician detecting temperature at the upper end of "normal" in an animal demonstrating clinical signs suggestive of an infectious disease might place less 'clinical weighting' on temperature in terms of clinical decision making

There has been a move away from p-values within scientific literature, to end the string of hyped claims and dismissal of potentially crucial effects due to this arbitrary threshold that is commonly referred to (Amrhein et al., 2019; Ioannidis, 2005). Sample size calculations (also known as power calculations) are essential to determine the number of animals required to have a reasonable chance of detecting a statistically significant difference. This largely depends on the size of the expected difference between two groups. For example, more calves would be required to detect a difference of 0.05kg/d in growth rate between two groups of calves than to detect a 0.1kg/d difference in growth rates. Typically, sample size calculations are used to calculate the numbers of animals needed to have an 80% chance of detecting a true difference between groups (the "power") at a significance level of 0.05. For example, to have an 80% chance of detecting a statistically significant difference of 0.05kg/d between two groups of calves, we would require at least 63 calves to be present in each of the two groups. If we have fewer calves than this and do not detect a statistically significant difference then it does not mean there is no difference between the group, as this study would be underpowered to detect a difference.

When analysing whether results from analysis are likely to be clinically important, it is worth considering the effect size, p-value and cost/benefit of a given intervention. It is important veterinary clinicians are involved in data analysis and interpretation as these complex interactions of client expectation and aspiration, uncertainty, cost-benefit and practicality go far beyond a simple (or even a complex) mathematical equation, even before interventions are planned and implemented. It is important to understand the various pitfalls of p-values, and not place too much reliance on a somewhat arbitrary threshold, particularly if the study is underpowered.

Box 5: Epitools

Whilst statistical significance and sample size calculations can be performed in Excel to calculate statistical significance, it can often be complex to arrange. Epitools is one of a number of online tools where a number of statistical tests can be performed and is freely available. For example, whether a fertility intervention would be statistically significant could be calculated with using the following 2x2 table in figure 7 (https://epitools.ausvet.com.au/twobytwotable).

Using Epitools, a 2x2 table can be filled out to calculate statistical significance between two treatment groups (Figure 8). In this example, 18/50 cows were pregnant in the treatment group, with only 10/50 cows being pregnant in the control group. This resulted in a p-value of 0.07, and the clinician will have to decide whether this result is clinically significant, even if not "statistically significant" by conventional thresholds of $p<0.05$.

A sample size calculation can also be performed using Epitools using the sample size calculator (https://epitools.ausvet.com.au/twomeansone)

Using Epitools, a sample size can be calculated (Figure 9) to see how many animals would be required to have an 80% (power 0.8) chance of detect a difference in growth rates of 0.8kg/d and 0.85kg/d at a significance level of $p<0.05$ (confidence level 0.95). This resulted in a requirement of 63 calves per group (126 calves in total), and any fewer calves means that a difference of 0.8kg/d and 0.85kg/d might not be detected.

Software

Even with limited experience, the calculation of basic statistics and the creation of powerful visualisations are relatively straightforward and can be valuable tools for on farm investigations and routine herd health analysis. A wide range of data analysis tools are available, and some of the most powerful are free to download. Most veterinarians will be familiar with Microsoft Excel, which can certainly provide adequate graphs and basic statistical analysis for most situations. Microsoft Excel or other spreadsheet packages can also be useful for collecting and storing data for analysis where this is not possible in on farm software. Learning a coding language provides a far more powerful data analysis toolkit to the farm animal vet, which ultimately will result in higher quality, faster reports for farms. Proprietary on-farm or vet software will often provide tools to create graphs and reports, but these are often limited in their options and do not provide the flexibility that Excel or coding languages provide.

Whilst coding requires more effort to learn initially, there are enormous benefits in the long term, in terms of flexibility and repeatability. Analysis that might take 10 minutes per farm to create in a spreadsheet can be replicated almost instantly for any number of farms once code has been written. There is also vastly improved flexibility with coding, with almost limitless opportunities compared with the more limited analytical potential of spreadsheet software. Whilst a simple graph can be easily created with either option, the potential for more advanced analytics such as regression or even advanced machine learning techniques are greatly limited in spreadsheet software, compared with enormous potential within coding environments. A range of coding languages available such as R and Python, and whilst most languages allow for powerful data analysis, some can be easier to learn than others. Whilst coding is initially more difficult, an investment in learning how to write efficient and powerful code will pay dividends in the long term for a clinician keen to develop more data skills (Figure 10).

There is also the advantage that both R and Python are freely available, with no licensing costs at all, in contrast to many proprietary statistics packages and spreadsheet programs, where the software must be purchased. There are also freely available online tools such as the Nottingham Herd Health Toolkit (www.nottingham.ac.uk/herdhealthtoolkit), which allows upload of farm-based data with a predefined output for those less keen to write their own code or taking their first steps in this area.

Box 6: Tidy data

Data often comes to veterinarians in a variety of formats. The principles of "Tidy data" (Grolemund and Wickham, 2016) is extremely useful in data analysis, as it ensures that analytical techniques are easily repeatable between datasets, and aids in creating an efficient data analysis workflow.

"Tidy data" essentially describes a dataset that involves one row per observation, and one column per variable. Data being recorded on farm will often come in a wide range of formats, and it is not often tidy. A sensible first step in data analysis is to ensure your data fits the "Tidy data" framework before attempting more advanced analytics.

Box 7: Calf growth rate graphs in Excel

Monitoring growth rates in calves can be extremely valuable, in benchmarking farms and identifying performance issues to ensure maximum efficiency in terms of health and productivity.

Creating box plots is relatively straightforward in Excel (Figure 12). The disadvantage is that to repeat this analysis for a farm you generally have to go through all these steps every single time. By using a coding language, you simply re-run the code you have already written to create a bespoke graph in a fraction of a second. Note also that this dataset is not in the “Tidy data” format (See box 6, figure 11) and would require a lot of manual work in Excel to reorganise to create other graphs, for example scatter plots.

Box 8: Calf growth rate graphs in R

Whilst there is not much difference between the plot in Figure 13 and the excel plot in Figure 12, the real advantage comes when this process needs to be repeated for future analysis. Next time you wish to plot this graph for a farm, all that is needed is to change the file path to the new .csv file and rerun the code. Whilst the excel graph would need to be constructed from scratch for a new set of data, this code takes under a second to import data, calculate DLWG and produce the graph shown. A small amount of effort learning how to write 13 lines of code has resulted in a graph that can be produced for any farm that records calf weights in less than a second.

This graph can be pasted into a word document to add comments etc before sending as part of a report. There is also the option in R to create an automated report that includes whichever graphs and data analytics you require, enabling a practice to write detailed reports for multiple farms, or even also benchmark multiple farms against each other in a matter of seconds once the code has initially been written. Further coding examples are provided in *Appendix i.*

In this example, it seems that bull calves generally grow faster than heifers. If there are differences in management between bull and heifer calves this is worth exploring more on farm.

Summary

Data skills are becoming increasingly important in farm animal population medicine. A wide range of "big data" sources are available on farms, and the ability to utilise this data to improve animal health outcomes can be a useful tool to the modern production animal veterinarian.

Whilst data skills can be initially be challenging to learn, a relatively small investment in time spent learning can result in dramatically shorter times analysing farm outcomes and writing reports in the future when analysing data for multiple farms. The statistical and analytic techniques discussed in this article are freely available and should provide veterinarians with new tools to prevent disease at a population level.

Further resources

The following resources are available for free:

For a tutorial in downloading and getting started in R see *Appendix i*

Further coding examples for common herd health problems are provided in *Appendix ii*

Epitools epitools.ausvet.com.au

Nottingham herd health toolkit nottingham.ac.uk/herdhealthtoolkit

R for data science book r4ds.had.co.nz YaRrr! The Pirate's Guide to R bookdown.org/ndphillips/YaRrr

Teacups Giraffes and Statistics tinystats.github.io/teacups-giraffes-and-statistics

Stack overflow stackoverflow.com

Quick-R statmethods.net

References

Amrhein, V., Greenland, S., McShane, B., 2019. Scientists rise up against statistical significance. Nature. https://doi.org/10.1038/d41586-019-00857-9

Grolemund, G., Wickham, H., 2016. R for Data Science. O'Reilly.

Hermans, K., Opsomer, G., Waegeman, W., Moerman, S., De Koster, J., Eetvelde, V., Ranst, V., Hostens, M., 2018. Interpretation and visualisation of data from dairy herds 40, 195–203. https://doi.org/10.1136/inp.k2166

Hudson, C., Kaler, J., Down, P., 2018. Using big data in cattle practice. In Pract. 40, 396–410. https://doi.org/10.1136/inp.k4328

Ioannidis, J., 2005. Why Most Published Research Findings Are False. PLoS Med. 2, e124. https://doi.org/10.1371/journal.pmed.0020124.t004

Overton, M.W., 2009. Using Reproductive Records: Basics of Monitoring.

Team, R.C., 2020. R: A Language and Environment for Statistical Computing.

Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis.

Woodward, H., Cobb, K., Remnant, J., 2019. The future of cattle veterinary practice: Insights from a qualitative study. Vet. Rec. 185, 205. https://doi.org/10.1136/vr.105321

Figure legends:

Figure 1: Types of data commonly encountered in farm animal data

Figure 2: Commonly used summary statistics

Figure 3: Somatic cell counts for 100 dairy cows showing a “right tail” with a small number of very high cell count animals resulting in an increased mean relative to the median.

Figure 4: A variety of graphical options to describe the calf performance

Figure 5: Plotting data in Excel allows for the simple analysis of weight by age.

Figure 6. By colouring by breed type, it is clear that in addition to the effect of age, there is also an effect of breed type on weight.

Figure 7: Daily liveweight gain of 500 calves from two farms. This dataset has a “normal” or “gaussian” distribution, with a mean growth rate of 0.8kg/d (dashed line), however the calves from farm A (red) have a far more consistent growth rate (standard deviation 0.1kg/d) compared with the wider spread in growth rates of farm B (blue, standard deviation 0.2kg/d). 95% of the data is spread between 1.96 standard deviations either side of the mean for both farms (0.196kg/d for farm A and 0.392kg/d for farm B).

Figure 8: Epitools can be used to calculate statistical significance between two treatment groups

Figure 9: Epitools can be used to perform sample size calculations. Variance is the standard deviation squared (see Box 3).

Figure 10: Whilst statistical coding using R/Python initially takes longer, there are significant productivity benefits to more advanced users in both analytical performance and speed.

Figure 11: An example of “Tidy data”

Figure 12: Excel allows relatively straightforward plotting of calf performance by sex.

Figure 13: An example of calf performance analysis using a programming language such as R.

Questions

What statistical test could be used to compare two continuous parametric groups with normal distribution?

- T-test
- Mann-Whitney test
- ANOVA
- Chi-squared test

With normally distributed data how many standard deviations either side of the mean would contain 95% of the data?

- 0.96
- 1.96
- 2
- 3.96

What type of model might be used to analyse the effect of age on calf weight?

- Logistic regression
- Linear regression
- ANOVA
- Mann-Whitney

In a regression model, what term describes the percentage of variation explained by the model?

- R^2
- MAE
- RMSE
- Coefficient

What advantages do statistical coding languages provide over spreadsheets when performing data analysis?

- Speed
- Repeatability
- Performance
- All of the above