

Mining your Data to Make Money

Robert B. Morrison

Dept Clinical & Population Sciences, 385 Animal Science/Veterinary Medicine, 1988 Fitch Ave,
University of Minnesota, St Paul, MN 55108 USA; **Email:** BobM@UMN.Edu

■ Introduction

Data mining is a popular concept these days and refers to the process of extracting information from large databases. Data mining can be used to predict trends and find behaviour that may lie beyond the expectations of experts. Data mining is part of a larger process sometimes called knowledge discovery; specifically, advanced statistical analysis and modeling being applied to data to find useful patterns and relationships. For example, data mining is used in the banking industry to model and predict credit fraud (160 million payment card accounts are continuously monitored), evaluate risk, and perform trend analysis. In the finance industry, data mining is used to try and forecast stock prices, predict cross-selling opportunities for other products (e.g. home refinance and car loans) and predict the effectiveness of marketing messages. In the retail industry, a wealth of consumer purchasing data is mined to understand purchasing preferences of a target market of customers.

In the pig industry, we are starting to develop large databases where some of these same techniques will be useful. For example, in a diagnostic lab database, what is the trend for erysipelas infection? Or, what is the relationship between age of the pigs and *H. parasuis* diagnosis? In a pig herd or system of herds, what is the relationship between nursery and farrowing performance? And, are culling reasons consistent among herds within a large system?

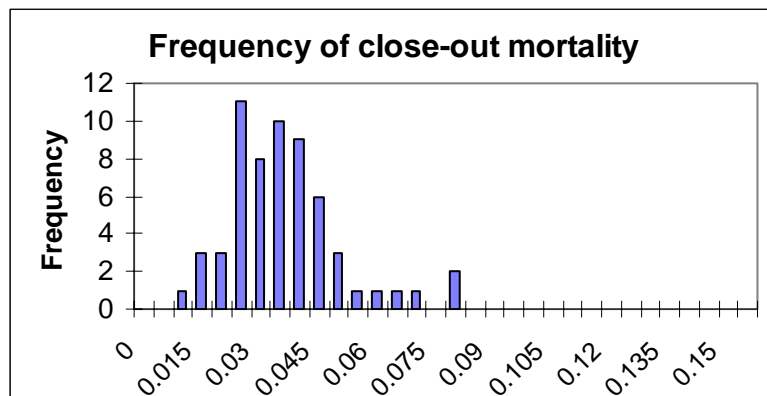
In this session, we will review some of the tools commonly used to assess datasets. Very complex analytic techniques exist, but the procedures described here will work for 95% of what we are called upon to look at. Some are simple and available with Excel, some are more complex and may require additional software, and a few may require expert assistance.

■ Describing Single Measurements

Before jumping in to calculate an average and make assumptions about the underlying distribution of the measurement, we should look at a picture of the data; a histogram (**Figure 1**). This will show us whether a measure is approximately bell shaped with one hump of data. Having done this, we can use measures of centrality (mean, median) to describe where the bulk of the data lie and measures of spread (range, standard deviation) to describe how variable the measure is.

- Bar or column,
 - histogram of finishing close-out mortality (**Figure 1**).
- Pie chart

Figure 1. Frequency distribution of close-out mortality in 59 groups (using Excel).



■ Two Measurements

We may be interested in breaking down a measure by categories. For example, how frequently do we see *H. parasuis* vs *S. suis* at a farm? Or, let's compare PRRS % seropositive among sites. Pivot charts and tables are extremely useful for helping us with this in Excel.

- Pivot charts
 - PigCHAMP database application and analysis of culling reasons (**Figure 2**),
 - *H. parasuis* occurrence at one system,
 - Flu serology from one system,

- Breakdown (using Statistica or PigCHAMP)
 - Pleuritis at slaughter and pull number (**Figure 3**),
 - site, complex, barn vs finishing mortality,
 - sire line and litter size,
 - gender, weight cutoff and performance of weaned pigs,

Figure 2. Pivot chart of culling reasons (using Excel).

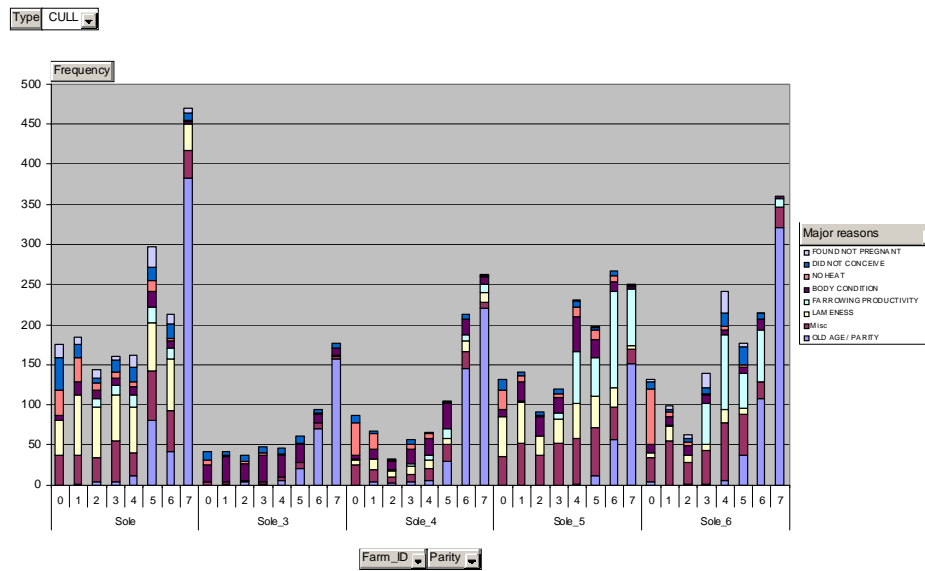
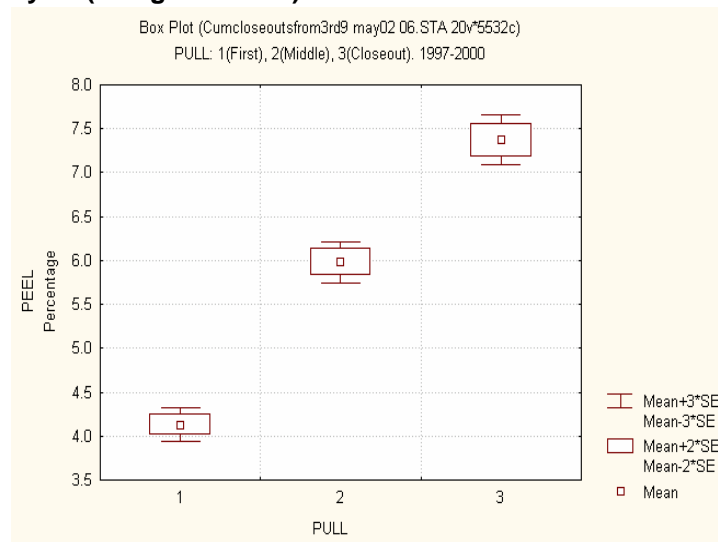


Figure 3. Breakdown of a continuous score into categories with statistical analysis (using Statistica).



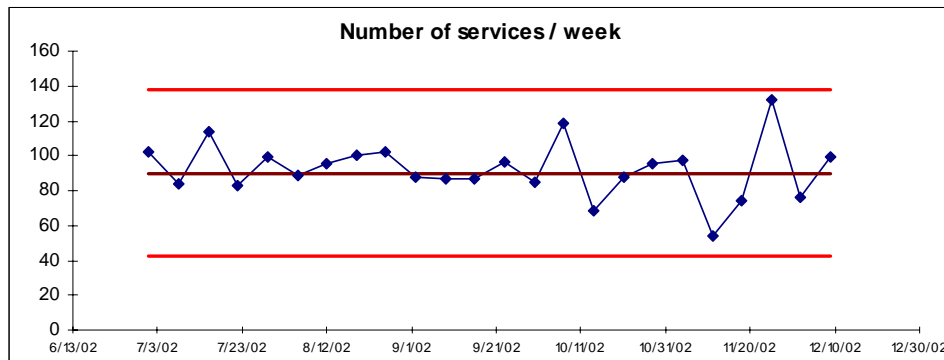
We commonly look at something over time as a line chart. Over time, two main uses for measurement are:

- to make data available as a source of ideas for improvement and,
- to check progress against a goal.

If progress is less than expected, revert to number 1 (Townsend and Gebhardt; 2002). When looking at data over time, we might plot a line graph. We should try to distinguish between random change and “real” change. Statistical process control helps us do this.

- Reproduction; litter size, services (**Figure 4**),
- Finishing mortality (watch out for seasonality in both examples),

Figure 4. Line chart with 99% confidence intervals (control limits) made with Excel.

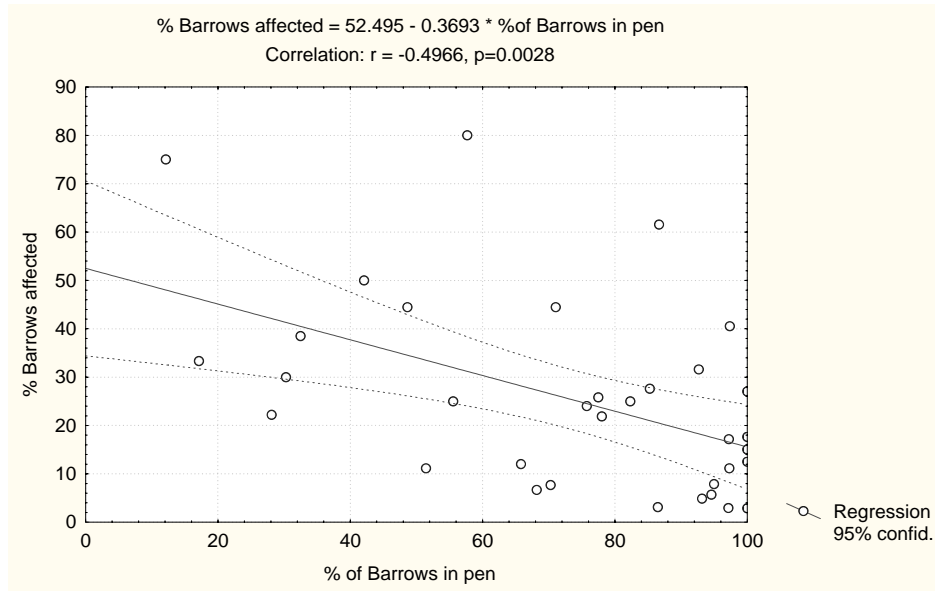


Forecasting is an extension of plotting a measure over time. It involves determining a predictive equation and projecting it forward.

We might be interested in the relationship between two continuous measures such as age and weight of growing pigs. The tool here is a scatterplot (correlation) and possibly a predictive equation if the relationship is linear (simple linear regression).

- Correlation between barrows with bitten tails and % barrows in the pen (**Figure 5**),
- chest girth & weight (or maybe height off floor and weight),
- nursery performance and reproductive performance

Figure 5. Correlation of percentage of barrows in the pen and the percentage of bitten barrows (in affected pens).



Three other analytic methods that are starting to be used are:

- Decision analysis – calculating break-even probability of an outcome with estimated cost and impact of a particular decision.
 - PRRS depopulation
 - Antibiotic group therapy or not
- Analysis of variability (or risk analysis),
 - @Risk modeling to examine impact of variability on optimum parity decision.
- Optimization routines
 - Use Excel Solver to determine optimum parity distribution for a herd.

■ References

Townsend, P, Gebhardt J. (2002) Simple quality for smaller organizations. Quality Progress; Oct 2002. pp76-80.