final report

Project Code: A.MFS.0169

Prepared by: Michelle Lorimer
Andreas Kiermeier
South Australian Research and Development Institute

Date published: January 2010

PUBLISHED BY
Meat and Livestock Australia Limited
Locked Bag 991
NORTH SYDNEY NSW 2059

National Microbiological Database Analysis Tool

Meat & Livestock Australia acknowledges the matching funds provided by the Australian Government and contributions from the Australian Meat Processor Corporation to support the research and development detailed in this publication.

This publication is published by Meat & Livestock Australia Limited ABN 39 081 678 364 (MLA). Care is taken to ensure the accuracy of the information contained in this publication. However MLA cannot accept responsibility for the accuracy or completeness of the information or opinions contained in the publication. You should make your own enquiries before making decisions concerning your interests. Reproduction in whole or in part of this publication is prohibited without prior written consent of MLA.

Abstract

The National Microbiological Database (NMD), also known as the ESAM (*Escherichia coli* and *Salmonella* Monitoring) database, was established to help Australia meet market access requirements to the US. All export establishments are required to collect and analyse carcase samples from all slaughter species for *E. coli* and *Salmonella*. Test results are forwarded to the Australian Quarantine and Inspection Service (AQIS) and entered into the NMD.

The aims of this project were to develop a reporting system for the data stored in the NMD, provide regular monthly reports about NMD results to each establishment, and provide training to industry and AQIS on the interpretation of those reports.

The reporting system that was developed is based on the R and LaTeX softwares. The R software is used to generate numerical and graphical data summaries, while the LaTeX software is used to generate the reports in PDF format which are emailed to establishments. To date, six of these reports, ending June to November 2009, have been emailed to establishments.

Executive Summary

The National Microbiological Database (NMD), also known as the ESAM (*Escherichia coli* and *Salmonella* Monitoring) database, was established to help Australia meet market access requirements to the US. All export establishments are required to collect and analyse carcase samples from all slaughter species for *E. coli* and *Salmonella*. Results are forwarded to the Australian Quarantine and Inspection Service (AQIS) on-plant vet who is responsible for entering the data into the NMD. This database provides useful information for benchmarking Australia's performance and can be used in market access negotiations.

The objective of this project was to develop a software-based reporting system so that individual establishments can be provided with regular monthly summaries of their own results and national benchmarks in the form of tables and graphs.

Data are obtained directly from AQIS each month. The reports are generated using a three-year moving window of data. This sampling period was chosen so that seasonal patterns could be identified.

The reporting system uses the R software to generate numerical and graphical data summaries. The LaTeX software is then used to generate the reports in PDF format.

A separate report is generated for each red meat species at each establishment, namely calves, cows/bulls, steers/heifers, sheep, lambs, goats skin-on and goats skin-off. Each report contains summary information on TVC prevalence and concentrations (where available), *E. coli* prevalence and concentrations, *Salmonella* prevalence and *Salmonella* serotype.

A series of training workshops were prepared and presented. A supporting document has been developed to aid establishments in the interpretation of their reports however, it requires review by MLA before distribution.

To date, establishments have received six reports (up to and including November 2009). Feedback has been positive and two establishments have detected errors in their data – these have now been corrected.

Other comments have related to providing modified data summaries, e.g. by calendar year or providing a summary of the previous month or three months. These potential modifications to existing reports should be investigated by MLA through a national survey of establishments. In addition, the inclusion of the most recent month's data summaries should be included immediately to allow better identification of data entry errors.

Contents

Page

1 Background 5
2 Project Objectives 5
3 Methodology 5
3.1 The data 5
3.2 Data Summaries 7
3.3 The Software 8
3.3.1 Report Templates 8
3.3.2 Generation of Data Summaries 8
3.3.3 Producing the PDF 9
3.3.4 Emailing the reports 10
4 Results and Discussion 10
4.1 Reporting System Rollout 10
4.2 Training Workshops 10
4.3 Report Progress to Date 10
4.4 Explanatory Guide 11
5 Success in Achieving Objectives 11
6 Impact on Meat and Livestock Industry – now and in five years time 11
7 Conclusions and Recommendations 11
8 Bibliography 12
Appendices 13
Appendix 1 – Explanatory Guide DRAFT 13
Appendix 2 – Example National Report 29

1 Background

The *E. coli* and *Salmonella* Monitoring (ESAM) program, now known as the National Microbiological Database (NMD), was established to help Australia meet market access requirements to the US. The program requires all export slaughter establishments to collect and analyse carcase samples from all slaughter species for *E. coli* and *Salmonella.* Data are entered into a National Database by AQIS on-plant vets and provide useful baseline information for benchmarking Australia's performance. These data, along with industry baseline data, have proven very useful in market access negotiations.

MLA, through their ESAM working group, developed statistical protocols for developing a reporting tool that could provide more in-depth analysis of the ESAM data to individual establishments on a regular basis. This enables them to compare their results against national benchmarks.

This report describes the development and implementation of the new reporting system and the response from industry so far.

2 Project Objectives

The project objectives were as follows:

1. Develop software capable of carrying out regular analysis of data from the ESAM database and provide regular reporting.
2. Provide a demonstration of the system along with training to ensure a fully functioning system.

3 Methodology

3.1 The data

Data from the NMD are exported by Paul Smith (AQIS) to a Comma Separated Values (CSV) file on a monthly basis and sent to SARDI. The reports are generated using a three-year moving window of data each month. The three-year reporting period was chosen so that seasonal patterns could be identified.

Each row in the CSV file contains the test result(s) for an individual sample. A detailed description of the columns in the CSV file is given below.

- `EstabId`: Number representing the establishment where the sample was taken.
- `TestId`: A number allocated by the database to the test at the time of entry into the database.
- `Species`: Type of animal the sample originated from. Possible values are **Calf**, Camels, **Cow/Bull**, Deer, Emu, Game Deer, Game Kangaroo, Game Pig, **Goat skin-off**, **Goat skin-on**, Horses, **Lamb**, Ostriches, Pig skin-off, Pig skin-on, **Sheep**, **Steer/Heifer**, Tripe – only those in **bold** are currently used in report preparation.
- `Dressing`: Method of dressing used within the establishment. Conventional, Gravity Rail, Inverted and Bed. At present, no differentiation is made between these during report generation.

- `Chain`: Indicates the slaughter chain and can take integer values 1-7. At present, no differentiation is made between these during report generation.
- `Shift`: Refers to the shift and can take values First, Second and Third corresponding to the Day, Afternoon of Night shift, respectively. At present, no differentiation is made between these during report generation.
- `Swabbed`: Refers to boning method, either Hot or Cold. At present, no differentiation is made between these during report generation.
- `Date`: The date and time the sample was collected.
- `Result`: This field has not been utilised in the NMD and contains only the value NO RESULT. It is ignored in the analysis.
- `TVCReading`: The TVC value (cfu/cm^2) for the sample as per the laboratory report. A zero (0) value denotes that no TVCs were detected, i.e. a negative sample.
- `TVCResultReg`: Indicates whether the TVC reading passed (P), failed (F), was marginal (M) or was not tested (N), where cut-off values, m and M, for marginal and fail are set by regulation. Not used for the reports.
- `TVCResultVol`: Indicates whether the TVC reading passed (P), failed (F), was marginal (M) or was not tested (N), where cut-off values, m and M, for marginal and fail are set voluntarily by the industry. Not used for the reports.
- `EcoliReading`: The *E. coli* value (cfu/cm^2) for the sample as per the laboratory report. A zero (0) value denotes that no *E. coli* were detected, i.e. a negative sample.
- `EcoliResultReg`; Indicates whether the *E. coli* reading passed (P), failed (F), was marginal (M) or was not tested (N), where cut-off values, m and M, for marginal and fail are set by regulation (Table 1) Not used for the reports.
- `EcoliResultVol`: Indicates whether the *E. coli* reading passed (P), failed (F), was marginal (M) or was not tested (N), where cut-off values, m and M, for marginal and fail are set voluntarily by the industry (Table 1) Not used for the reports.
- `SalmonellaTested`: Indicates whether the sample was tested (T) or not tested (N) for *Salmonella*.
- `SalmonellaResult`: Indicates whether a sample that was tested for *Salmonella* passed (P; negative for *Salmonella*) or failed (F; positive for *Salmonella*) the test.
- `SeroType`: The serotype, as identified by the laboratory, for samples that tested positive for *Salmonella*; blank otherwise.

Table 1: The regulatory limits for *E. coli* are presented in the following table (supplied by AQIS)

Species	*E. coli* Regulatory Limits		*E. coli* Voluntary Limits	
	m	*M*	*m*	*M*
Bovine Calf	5	100	5	100
Bovine Cow Bull	0	20	0	20
Bovine Steer Heifer	0	20	0	20
Ovine Lamb	5	100	5	100
Sheep	5	100	5	100
Goat	1	100	1	100
Pig	1	100	1	100
Horse	0	100	0	100
Deer	1	100	1	100
Emu Ostrich	1	10	1	10
Kangaroo	50	500	50	500
Camel	0	20	0	20
Game Pig	50	500	50	500

3.2 Data Summaries

A separate report is generated for each slaughter species processed at each establishment. Consequently, some plants receive only a single report while others receive several. Each report contains a self-contained glossary of terms. In addition, an Explanatory Guide was written to assist establishment staff to interpret the reports appropriately. It contains detailed explanations of the summary statistics found in the reports (Appendix 1).

Each report includes separate summary tables for the prevalence and concentration for TVC and *E. coli*. For *Salmonella*, the prevalence is summarised along with the corresponding serotypes.

Prevalence refers to the number of samples that were found to contain at least one colony of TVC (`TVCReading` > 0.0) or *E. coli* (`EcoliReading` > 0.0). For *Salmonella*, it refers to the proportion of samples that were positive for *Salmonella*, i.e. those with a failed result. The lower and upper bounds of a 95% Confidence Interval are also presented. This interval gives the range or the 'ballpark' of where the true prevalence may be. The width of the interval (difference between the lower and upper bounds) indicates the level of precision with which the prevalence is estimated – the smaller the better. However, to a large degree this interval depends on the number of samples used for the prevalence calculation.

For TVC and *E. coli* concentrations the data are summarised for positive samples only. Therefore, the data are restricted to include those samples where the concentrations are greater than or equal to the limit of detection. TVC and *E. coli* concentrations are both $\log_{10}$ transformed. Summary statistics include the minimum, first quartile, median, mean, third quartile, 90th percentile, 95th percentile, 99th percentile, maximum and the standard deviation.

The reports also contain two types of graphical representations of the data – box plots and time plots. In the box plots, the (positive) $\log_{10}$ concentrations are summarised on a monthly basis over the three-year period. In the time plots all individual observations (sample results) are used on the original (untransformed) scale, including negative samples (or those less than the limit of detection).

Box plots utilise the descriptive statistics described above, the minimum, first quartile, median, third quartile and maximum. Box plots were included to help identify differences in concentrations between months. They help identify seasonal trends and extreme or unexpected concentrations.

The time plot can be used to compare individual plant's level of *E. coli* compared to those found nationally over the same sampling period. This plot may also assist in summarising the number of 'alerts' a plant has had over the sampling period compared to those found nationally.

3.3 The Software

The reports are generated using two interlinked software systems – the statistical software R (R Development Core Team, 2008) and the document preparation system LaTeX (Lamport, 1994). Both of these systems are available for free download and commercial use.

3.3.1 Report Templates

Two reporting templates were developed, one for plants (`report.Rnw`) and one for national data (`reportNat.Rnw`). These templates are used by the R software[1] for each plant-by-species and national-by-species combination and populated with the appropriate data. This produces a separate LaTeX (TEX) file for each report. These LaTeX files contain text based commands, similar to HTML and need to be processed using the LaTeX software to produce the formatted PDF reports. An example of a national report is attached as Appendix 2 – for confidentiality reasons a plant report cannot be provided.

3.3.2 Generation of Data Summaries

The data manipulations and summaries are undertaken with the help of four scripts, thereby keeping the process modular. These scripts are as follows:

- The `SummaryFunctions.R` script contains all functions, date/time definitions and generates generic plot axes that are used in the other scripts. It is automatically sourced when running the `InputDataAndCheck.R` script.
- The `InputDataAndCheck.R` script performs data input and performs basic data integrity checks. The results of these checks are written to the log file InputDataAndCheck.log. Further information is given below.
- The `NationalGraphsAndSummaries.R` script is used to generate national data summaries and graphs. All output files are saved to the `national` subfolder (which must exist at the time of running the script). In particular, the reports contain a .tex extension and need to be processed with the LaTeX software.
- The `PlantGraphsAndSummaries.R` script is used to generate plant and species specific data summaries and graphs. All output files are saved to the `plants` subfolder (which must exist at the time of running the script).

These scripts are run, one at a time, directly from the R command prompt – either using the 'source' command or from the File menu.

[1] R uses the Sweave mechanism (Leisch, 2002) to combine the templates with the appropriate data.

Data checks

Once the data set has been imported into R and unnecessary observations have been removed – those which fall outside the three-year window, those that relate to other species and those that relate to plants that have nominated to not receive the reports – the following data checks are performed.

- `Dressing`: Check that the only values recorded are 'CONVENTIONAL', 'GRAVITY RAIL', 'INVERTED', 'BED' or 'NOT APPLICABLE'.
- `Chain`: Check that the only values recorded are integers 1 to 7.
- `Shift`: Check that only values recorded are 'First', 'Second' and 'Third'.
- `Swabbed`: Check that the only values recorded are 'Hot' or 'Cold' or that the value is blank (NA).
- `Date`: Check that the date is not missing.
- `Result`: This field has not been utilised in the NMD and contains only the value NO RESULT. It is ignored from the analysis.
- `TVCReading`: Check that any recorded value is greater than or equal to zero.
- `TVCResultReg`: Check that the value is 'N' (not tested) if `TVCReading` is blank or that the value is not equal to 'N' if the `TVCReading` is not blank.
- `EcoliReading`: Check that any recorded value is greater than or equal to zero. In addition, check that the value is a multiple of the sampling fraction (0.08 for cows, bulls, steers and heifers; 0.33 for calves, sheep, lambs and goats, with and without skin).
- `EcoliResultReg`; Check that the value is 'N' (not tested) if `EcoliReading` is blank or that the value is not equal to 'N' if the `EcoliReading` is not blank.
- Check `SalmonellaTested` (N=not tested, Z=not tested, T=tested) and `SalmonellaResult` (P=pass, F=fail) values
  - If `SalmonellaTested` is 'N' then there are no 'F' or 'P' in `SalmonellaResult`
  - If `SalmonellaTested` is 'T' then there are no 'Z' or 'N' in `SalmonellaResult`
  - If `SalmonellaTested` is 'Z' then there are no 'F' or 'P' in `SalmonellaResult`

The result of each test is written to the log file. If a test is failed then the problematic values along with the establishment and test ID's are written to the file for easy identification. These errors are inspected before proceeding further.

3.3.3 Producing the PDF

Once the data summaries (R scripts) have been run the TEX report files need to be compiled using the LaTeX typesetting program to produce the corresponding PDF files. This needs to be done from the command line (the 'DOS' window). To facilitate this two batch files, `runreports.bat` and `cleanup.bat`, have been written.

Running the `runreports.bat` batch file will result in all reports – national and plant based – to be compiled into PDF files. The reports for each plant are then compressed into ZIP files for emailing – each containing a plant's reports for all species processed under the same Establishment ID. In addition, a single ZIP file containing all reports is created for archiving.

Running the `cleanup.bat` batch file results in all files, including ZIP, PDF and log files, to be deleted so that the next month's reports can be created under a 'clean' environment.

3.3.4 Emailing the reports

Emailing ZIP files is facilitated via a macro in Microsoft Excel. This macro steps through a list of establishments and associated contact details, in some cases more than one person, and creates an email message for each establishment, with the ZIP file attached.

This has been developed to minimise the risk of sending an establishment's reports to the wrong person, which is much more likely if the process were manual.

4 Results and Discussion

4.1 Reporting System Rollout

The reporting system was initially developed and tested using de-identified data, supplied by AQIS. SARDI signed a confidentiality agreement with AQIS to enable SARDI to receive identified establishment codes.

To obtain the contact details of the individual establishments, a letter addressed to the QA Manager was forwarded to MINTRAC and then sent to each establishment. Establishments then contacted SARDI directly with the appropriate information including their Establishment Number so that reports could be matched accordingly. This information was also canvassed at four of the five training workshops.

Currently there are 19 establishments that have not yet received their reports as they have not provided their details. These establishments will be followed up in early 2010.

4.2 Training Workshops

A series of workshops were prepared and presented to establishment representatives and AQIS on-plant vets around the country as part of the AQIS/MLA/AMIC Industry Development Series. Participants were introduced to the reports and interpretation of the information contained was discussed in detail. The workshop was broken up into three sections. At the end of each section, participants were asked to work through a series of exercises based on the information discussed. The aim was to get establishment staff thinking about how to interpret the information provided in the tables and graphs of their reports.

4.3 Report Progress to Date

Paul Smith from AQIS has provided data up to 30 November 2009. So far, six reports have been sent to establishments for the following periods:

- July 2006 – June 2009
- August 2006 – July 2009
- September 2006 – August 2009
- October 2006 – September 2009
- November 2006 – October 2009
- December 2006 – November 2009

Feedback from establishments, where provided, has been positive. Two establishments have detected errors in their data, which have now been rectified. One of these was related to the

incorrect entry of two negative *E. coli* samples – entered as 0.08 (positive) instead of 0.0 (negative).

Swift Australia has requested an additional report where all their plants are combined into a single report. This report will be prepared in 2010 in collaboration with Swift and then included in the usual monthly reporting roll out.

Other comments include summarising the data in calendar years or providing a summary of the previous month or previous three months. Separating the data in terms of Hot and Cold swabbing has also been suggested. It is proposed to modify the reports early in 2010 to include the latest month's data (plant and national). Later in 2010 a survey should be undertaken to identify other required reporting options.

4.4 Explanatory Guide

A supporting document has been written to assist establishments in the interpretation of their reports. This follows on from the training workshops and provides detailed information on the data summaries provided in the reports. A draft copy is included in the Appendix 1. This document requires approval by MLA before being sent to establishments.

5 Success in Achieving Objectives

A fully functional software system has been implemented and regular reports are being provided to establishments on a monthly basis. A total of six reports have been provided to date.

Training was provided as part of AQIS/AMIC/MLA Industry Development workshops around Australia.

6 Impact on Meat and Livestock Industry – now and in five years time

Establishments are provided with regular reports so that they can compare their performance against the national baseline, enabling them to improve their process or benchmark their improvement over a period of time. Establishments will be able to demonstrate to exporters how "good" their product is compared to the national baseline.

In addition, establishments will be able to assess the validity of the data entered into the NMD and rectify any data entry errors. Consequently, data quality is expected to improve over time. This better quality data will assist processors, the industry and AQIS during market access negotiations.

7 Conclusions and Recommendations

The reporting system has been implemented successfully and feedback from industry has been positive.

Given the experiences gained from the development and implementation the following recommendations are made:

- At present only six reports have been issued to establishments. To ensure maximum industry uptake, it is recommended that MLA continue to fund the provision of reports in 2010 before handing project reporting over to industry, for example, as part of a subscription based service.

- A formal survey of establishments in 2010 would be beneficial in identifying how the reports could be modified to make them more useful to industry.
- The explanatory guide should be finalised in early 2010 and provided to establishments. This could be in the current format or be published by MLA.
- Only red meat species in the NMD are currently being reported on. Provision of reports to other species, e.g. pigs or kangaroos, could be implemented relatively easily. Consequently, co-investment by other industries should be considered favourably by MLA.
- Reports should be modified to include a summary of TVC, *E. coli* and *Salmonella* for the last month to allow easier identification of data entry errors.

8 Bibliography

Lamport, L. (1994) LaTeX: A Documentation Preparation System User's Guide and Reference Manual, Addison-Wesley Pub Co., 2nd edition.

Leisch, F. (2002) Sweave: Dynamic generation of statistical reports using literate data analysis. In Wolfgang Härdle and Bernd Rönz, editors, *Compstat 2002 - Proceedings in Computational Statistics*, pages 575-580. Physica Verlag, Heidelberg, ISBN 3-7908-1517-9.

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org

Appendices

Appendix 1 – Explanatory Guide DRAFT

Explanatory Guide for the National Microbiological Database Reporting System

Background

The National Microbiological Database (NMD), also known as the ESAM program, was established to help Australia meet market access requirements to the US. All export establishments are required to collect and analyse carcase samples from all slaughter species for *E. coli* and *Salmonella*. Results are forwarded to the AQIS on-plant vet who is responsible for entering the data into the NMD.

This database provides useful information for benchmarking Australia's performance and can be used in market access negotiations. Establishments are now provided with a regular (monthly) report so that they can compare their performance to that found nationally for the same reporting period.

This document has been developed to provide a more detailed explanatory guide to assist establishments in interpreting their monthly reports.

The Database

The NMD includes a combination of descriptive or **categorical** variables, such as Establishment ID and Species and **continuous** measurement variables such as *E. coli* Reading (cfu/cm^2) and TVC Reading (cfu/cm^2).

Categorical variables refer to those variables which have a countable number of values. Examples of categorical variables in the database are:

- Species – Calf, Cow/Bull, Steer/Heifer, Sheep, Lambs, Goat Skin on or Goat Skin Off
- Dressing – Bed, Conventional, Gravity Rail or Inverted
- Shift – First, Second or Third
- Swabbed – Hot or Cold
- *Salmonella* – Pass (Negative) or Fail (Positive)

At present the only categorical variables used are Species and *Salmonella*.

Continuous variables refer to those variables which can take on any value and are measured or counted. Examples are the concentration of *E. coli* and the Total Viable Count (cfu/cm^2).

The Reports

A separate report is generated for each of the species processed at each establishment. The species are: Calf, Cow/Bull, Steer/Heifer, Sheep, Lambs, Goat Skin on and Goat Skin off.

Summary tables for prevalence and positive concentrations are presented in the reports, along with graphs for the individual establishment and the national baseline.

A three-year moving window of data is used to generate the reports. Every month, the oldest month is dropped from the data and the latest month is added. The information presented in the tables is summarised over the given three-year period.

Prevalence Summaries

The *true prevalence* relates to the number of carcases that are positive for TVC, *E. coli* or *Salmonella*. However, this would involve testing every cm^2 of every carcase produced. Consequently, the best guess at this true prevalence is estimated from the NMD as the percentage of positive tests.

Prevalence = Percentage of positive tests

The first step in examining the ESAM data is to determine the prevalence of TVC, *E. coli* and *Salmonella*. Prevalence refers to the number of samples that were found to have at least **one** colony of TVC or *E. coli*. For *Salmonella* it is the number of samples that have recorded a **Fail** result, that is, the number of samples that were positive for *Salmonella*. This is then divided by the total number of tests that were performed during the sampling period and multiplied by 100 to give the percentage of positive tests.

Tests: The total number of TVC, *E. coli* or *Salmonella* tests recorded in the NMD during the reporting period.

Positives: The number of samples with positive concentrations (at least one colony) or a failed (positive) test.

Percent +ve: 100% × Positives/Tests

95% Confidence Interval

A Confidence Interval gives the range or the 'ballpark' of where the true prevalence may be. If nothing in the process changes in the future and we repeatedly sample then we can expect the prevalence to be contained in this range 95% of the time.

The **Upper** and **Lower Bound** describe the bounds of a 95% Confidence Interval.

A small range (or difference) between the lower and upper bounds indicates a more consistent estimate of prevalence. Wider confidence intervals may indicate inconsistency in the estimate of prevalence.

> *Consider two establishments at which the prevalence of* Salmonella *was found to be 5% each. In addition, assume that the 95% confidence intervals are 1.5% to 8.5% at Plant A and 4% to 6% at Plant B. This means we can be 95% confident that between 1.5 and 8.5% of carcases at Plant A are positive for* Salmonella *and between 4 and 6% at Plant B. Therefore we can conclude that the estimate for Plant B is more precise as the interval is narrower.*

A narrow confidence interval in relation to the prevalence estimate means that the estimated prevalence is more precisely estimated.

Note that the width of the confidence interval is influenced by the total number of tests performed during the sampling period. If only a small number of tests have been performed, it is likely that the confidence intervals will be wider (assuming the prevalence is the same). As such, it is expected that the width of the confidence interval for the national summaries are likely to be much smaller than those found at an individual establishment.

Example 1

An example of a prevalence summary obtained for *E. coli* for a particular export establishment and species is presented in Table 1.

Table 1: *E. coli* prevalence summary for this establishment and nationally

	Plant	National
Tests	891	16142
Positives	268	1135
Percent +ve	30.08	7.03
Lower Bound	27.08	6.64
Upper Bound	33.21	7.44

Interpretation

- The prevalence or percentage of positive *E. coli* samples at this plant was much higher than the national prevalence during the same period (30.08% versus 7.03%).
- Of the 1135 positive *E. coli* samples found nationally, 268 (268/1135 = 23.6%) were from this plant.
- Given the small range in the confidence interval for the national summary, we can be fairly confident that nationally around 7% of carcases of this species are positive for *E. coli*.
- At the plant level, the bounds of the confidence interval indicate that the *E. coli* prevalence could be as low as 27% and as high as 33%.

Example 2

An example of a prevalence summary obtained for *Salmonella* for a particular export establishment and species is presented in Table 2.

Table 2: *Salmonella* prevalence summary for this establishment and nationally

	Plant	National
Tests	980	5370
Positives	2	39
Percent +ve	0.204	0.726
Lower Bound	0.025	0.517
Upper Bound	0.735	0.991

Interpretation

- Nationally, the prevalence of *Salmonella* is generally low (0.7%).
- At the plant, the prevalence is somewhat lower than that found nationally (0.2%).
- Based on the bounds of the confidence intervals, it may be expected that few as 2 in 10,000 (0.02%) to as many as 7 in 1,000 (0.7%) carcases contain *Salmonella*.
- Nationally, the level of *Salmonella* is estimated to be between 5 in 1,000 (0.5%) to 1 in a 100 (0.99%).

TVC and *E. coli* Concentration Summaries

For TVC and *E. coli* concentrations, the data are summarised for positive samples only. This means that the data are restricted to include those samples where the concentrations are greater than or equal to the limit of detection.

In the NMD, concentrations less than the limit of detection should be recorded as 0. Hence data are restricted to concentrations > 0. TVC and *E. coli* concentrations are then transformed to the log_{10} scale.

Log_{10} Transformation

It is standard practice for microbiological concentration data to be transformed into logarithms with base 10, denoted by $\log_{10}$. After this transformation, the distribution of concentrations is made more symmetrical. For example, concentrations that range from 1 to 25,000,000 say, will after $\log_{10}$ transformation range from 0 to 7.4.

The effect of the $\log_{10}$ transformation on the concentration data is presented in Table 3.

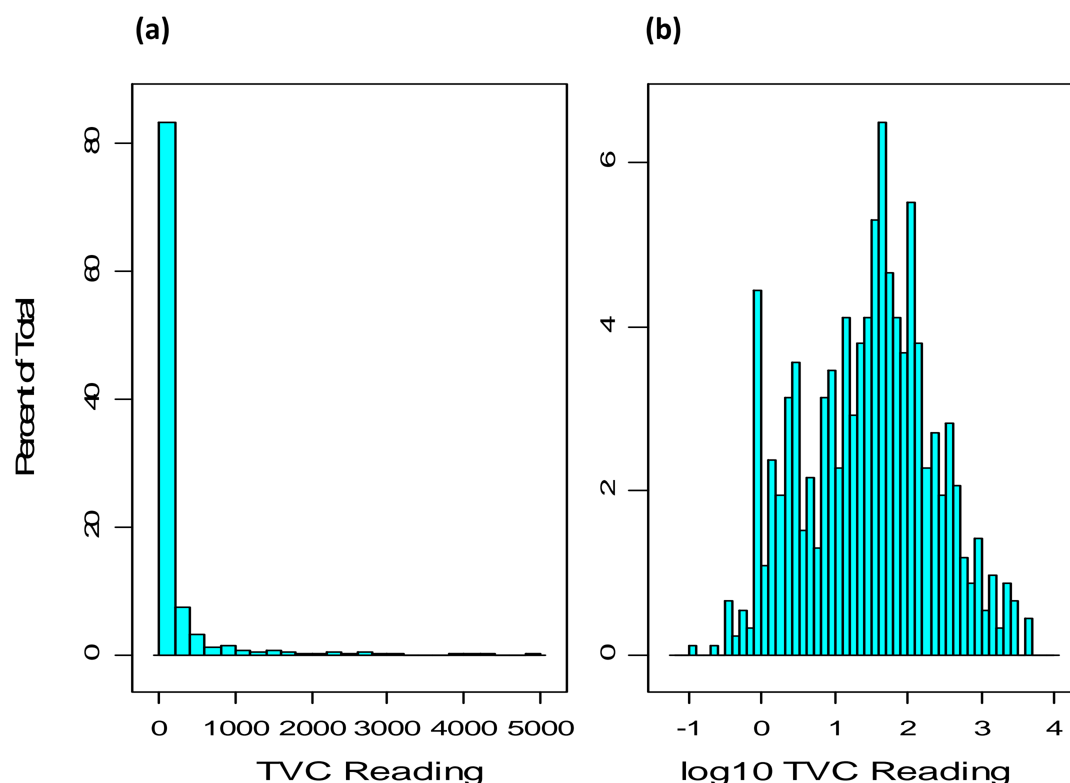
Table 3: Pattern of counts before and after $\log_{10}$ transformation

Count		$\log_{10}$
0.1	1×10^{-1}	-1
1	1×10^{0}	0
10	1×10^{1}	1
100	1×10^{2}	2
1,000	1×10^{3}	3
2,000	2×10^{3}	3.3
10,000	1×10^{4}	4
20,000	2×10^{4}	4.3
35,000	3.5×10^{4}	4.5
100,000	1×10^{5}	5
200,000	2×10^{5}	5.3
500,000	5×10^{5}	5.7

- A 1 $\log_{10}$ reduction equates to a 90% reduction in counts on the original scale (i.e. a change from 1,000 to 100 or 10,000 to 1,000 etc).
- A 2 $\log_{10}$ reduction equates to a 99% reduction in counts on the original scale (i.e. a change from 10,000 to 100 or 100,000 to 1,000 etc).
- A 3 $\log_{10}$ reduction equates to a 99.9% reduction in counts on the original scale (i.e. a change from 100,000 to 100 etc).

As an example, the distribution of TVC on the original scale is presented, along with the distribution of TVC after $\log_{10}$ transformation, in Figure 1. Figure 1a shows a large proportion of counts close to zero, with a “long tail” to the right as concentrations increase (few very large concentrations). The distribution of TVC after $\log_{10}$ transformation Figure 1b shows that the data is now more symmetrically distributed, that is, the left side of the distribution is an approximate mirror image of the right hand side.

Figure 1: Distribution of TVC concentration a) on the original scale and b) after log_{10} transformation



Summary statistics are calculated for both the individual plant and from the whole database (Nationally) for each species for the previous three-year period. To do this, the concentrations are initially ordered from lowest to highest.

Minimum Concentration

The AQIS Meat Notice (2003/6) includes procedures to convert concentrations to cfu/cm^2 of carcase surface. This conversion includes information regarding the number of colonies found, the appropriate dilution factor and the **sampling factor**. The sampling factor relates to the amount of area of swabbed surface that each ml of the undiluted sample represents.

For cows/bulls and steers/heifers the sampling factor is:

- 0.08 cfu/cm^2,
- On the transformed scale this is -1.097 log_{10} cfu/cm^2.

For sheep, lambs, calves and goats the sampling factor is:

- 0.33 cfu/cm^2,
- On the transformed scale this is -0.48 log_{10} cfu/cm^2.

For an undiluted sample with **no colonies** and a sampling factor of 0.08 the concentration is considered to be **less than** the limit of detection or <0.08 cfu/cm^2. These should be entered as 0.0.

Similarly for an undiluted sample with **no colonies** and a sampling factor of 0.33 the concentration is considered to be **less than** the limit of detection or <0.33 cfu/cm^2. These should be entered as 0.0.

For an undiluted sample with **one colony** and a sampling factor of 0.08 the concentration is considered to be **equal** to the limit of detection or 0.08 cfu/cm^2. These should be entered as 0.0.

Similarly, for an undiluted sample with **one colony** and a sampling factor of 0.33 the concentration is considered to be **equal** to the limit of detection or 0.33 cfu/cm^2. These should be entered as 0.0.

Therefore, it is expected that the absolute minimum concentration (i.e. 1 colony) that is summarised in the tables should be equal to the sampling factor for that species. In other words, the (absolute) minimum concentration (on the $\log_{10}$ scale) that can be expected in the tables should be:

- -1.097 $\log_{10}$ cfu/cm^2 for cows/bulls and steers/heifers and
- -0.48 $\log_{10}$ cfu/cm^2 for Sheep, lambs, calves and goats.

If values less than these have been entered into the database then establishment staff are encouraged to check the original data records for these observations to determine if it is an error or if the laboratory used an alternative approach (such as dilutions) to determine the concentration.

Measures of "Central Tendency"

The **median** and the **mean** are both measures which help to examine the central tendency of the data. In other words, they describe the "middle" or "expected" concentration in the data set.

The **median** describes the midpoint or middle concentration in the set of data. Therefore, 50% of the data are less than this value and 50% of the data are greater than this value. The median is also known as a resistant measure of centre because it is not influenced by extreme observations.

The term **mean** is just another word for average. It is calculated by adding all $\log_{10}$ concentrations in the sampling period and dividing by the number of positive tests. The mean is easily influenced by extreme observations.

Example 3

Consider the following $\log_{10}$ concentration data (11 samples) ordered from smallest to largest:

-0.22	0.16	0.23	0.30	0.48	0.69	0.75	0.97	1.10	1.32	1.57

The **median** or midpoint corresponds to the 6th largest sample and equals 0.69 $\log_{10}$ cfu/cm^2. Five samples are below the median, and five are above.

The **mean** is calculated by adding up all the concentrations and dividing by 11. For this example, the mean is 0.67 $\log_{10}$ cfu/cm^2.

Example 4

Let's now consider an example where the median and the mean are influenced by an extreme observation. The $\log_{10}$ concentration data (n=11) from Example 3 are again considered but the largest concentration was changed from 1.57 to 4.57 (e.g. a data entry error).

-0.22	0.16	0.23	0.30	0.48	0.69	0.75	0.97	1.10	1.32	4.57

The **median** or midpoint again corresponds to the 6th carcase sample and equals 0.69 $\log_{10}$ cfu/cm^2. Five samples are below this sample, and five are above. It is unaffected by the change in value of the largest sample value.

The **mean** is calculated by adding up all the concentrations and dividing by 11. For this example, the mean is 0.94 $\log_{10}$ cfu/cm^2. This shows how the mean can be influenced (shifted upwards in this case) by unusual (or extreme) observations.

Measures of variability

In addition to summarising the mean and median, we also need some indication of the variability or spread of the data. The simplest measure of variability is the **range** which is defined as the difference between the smallest (**minimum**) and largest (**maximum**) concentrations.

Consider the data in Example 4, we would calculate the range as 4.57 – (-0.22) = 4.79, suggesting a fairly wide spread. However, it is quite possible that the largest concentration may be incorrect (possibly a data entry error). Omitting the extreme concentration of 4.57 reduces the range to 1.54. Given the sensitivity to one extreme concentration, the range is not a very reliable measure of spread. Instead, measures of variability that are not so easily influenced by extreme or (unusual) observations are preferred, such as the **inter-quartile range** and **standard deviation**.

Inter-quartile range, Quartiles and Percentiles

To calculate the inter-quartile range, we need to define the quartiles. The definition of the quartiles is similar to that of the median. Where the median cuts the data into halves, the quartiles cut the data into quarters, each containing (as far as possible) an equal number of concentrations.

The inter-quartile range is the difference between the upper or third (Q3) and lower or first (Q1) **quartiles** and describes the range of the middle 50% of concentrations found in the data set. Because it uses only the middle 50% of data, it is not affected by unusual or extreme observations.

The lower quartile (**Q1**) is the value in which 25% of the data are below and 75% of the data are above.

The upper quartile (**Q3**) is the value in which 75% of the data are below and 25% of the data are above.

Inter-quartile Range (IQR) = Q3 – Q1 and describes the middle 50% of data.

Example 5

As an example, consider the $\log_{10}$ concentration (n=11) data used in Example 3 (reproduced below).

-0.22	0.16	0.23	0.30	0.48	0.69	0.75	0.97	1.10	1.32	1.57

The **median** was previously shown to be 0.69 $\log_{10}$ cfu/cm^2.

To determine the lower quartile (**Q1**), consider the five observations less than the median value of 0.69. Take the median or middle value from these five observations, giving **Q1**. This is 0.23 $\log_{10}$ cfu/cm^2.

Likewise, to determine the upper quartile (**Q3**), consider the five observations above the median value of 0.69. Take the median or middle value from these five observations, giving **Q3**. This is 1.10 cfu/cm^2.

The inter-quartile range is then defined as the difference between Q3 and Q1.

IQR = 1.10 – 0.23 = 0.87 cfu/cm^2.

The **percentiles** cut the data into hundredths. In the summaries of the data from the NMD the **90th, 95th and 99th percentiles** are presented. Examining the values at each of the percentiles will be helpful in detecting unusual or extreme observations in the database.

Percentiles are also useful for summarising the performance of the slaughter process. For example we can imply that (if all things being equal) 90% of carcases will have a concentration better (less) than this value.

90th Percentile: 90% of the data are less than this value, 10% are greater.

95th Percentile: 95% of the data are less than this value, 5% are greater.

99th Percentile: 99% of the data are less than this value, 1% are greater.

Standard deviation

The **standard deviation** is a measure of spread (or variability) about the mean.

To calculate the standard deviation:

- Subtract the mean from each observation (known as mean difference)
- Square the result
- Add them all up
- Divide by on less than the number of observations

- Then take the square root

For the mathematically minded, this is:

$$\text{Standard deviation} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Where
$i = 1, 2, ..., n$
n = the number of observations in the data set
x_i = individual concentrations
$\bar{x}$ = the mean concentration

This measure will be **large** if observations are widely spread about the mean.

This measure will be **small** if observations are close to the mean. A small standard deviation is indicative of more consistent carcase hygiene.

Example 6a

Consider the following $\log_{10}$ concentration data (the row labelled 'x') and calculate the standard deviation. Note that the **mean is 1**, for this data set.

x	0.8	0.9	1.0	1.1	1.2
(Mean difference)^2	$(0.8 – 1)^2 = 0.04$	0.01	0	0.01	0.04

Now add up the results.

0.04 + 0.01 + 0 + 0.01 + 0.04 = 0.1

Divide by n – 1, where n is 5 observations in this example

So 0.1/4 = 0.025

The **standard deviation** is the square root of this result and equals **0.158**.

Example 6b

Consider a second example of $\log_{10}$ concentration data, again calculating the standard deviation. Again the **mean is 1** for this data set.

x	0.4	0.7	1.0	1.3	1.6
(Mean difference)^2	$(0.4 – 1)^2 = 0.36$	0.09	0	0.09	0.36

Now add up the results.

0.36 + 0.09 + 0 + 0.09 + 0.36 = 0.90

Divide by n – 1, where n is 5 in this example

So 0.9/4 = 0.225

The **standard deviation** is the square root of this result and equals **0.47**.

In the first example (**Example 6a**), the standard deviation was smaller compared to that of the second example (**Example 6b**) although the means were the same. The standard deviation is larger because the points are more spread out.

Example 7

The table below is an example of a concentration summary obtained for $\log_{10}$ TVC for a particular export establishment and species.

Table 4: Total Viable Count ($\log_{10}$) cfu/cm^2) summary for this establishment and nationally

	Plant	National
Positives	1444	19246
Minimum	0	-1.00
Q1	1.66	0.70
Median	2.01	1.19
Mean (+ve)	1.98	1.21
Q3	2.35	1.70
90th Percentile	2.66	2.22
95th Percentile	2.92	2.51
99th Percentile	3.41	3.16
Maximum	4.13	5.57
SD	0.58	0.77

Interpretation

- TVCs were consistently higher than those found nationally at Q1, Median, Mean, Q3 and at the 90th, 95th and 99th percentile.
- But the maximum TVC at this plant is 1.4 $\log_{10}$ cfu/cm^2 less than was found nationally
- The standard deviation was smaller at this plant than that observed nationally for the same sampling period. This shouldn't come as a big surprise since the national standard deviation includes within and between plant variability.
- Overall, we can conclude that the levels of TVC are somewhat higher at this plant compared to those found nationally for the same sampling period, but the variability is less.

Example 8

The table below is an example of a concentration summary obtained for *E. coli* for a particular export establishment and species. Note that it is important to compare the concentration summary in the context of the prevalence table for *E. coli*, i.e. the number of positive *E. coli* samples need to be kept in mind. The corresponding prevalence table was discussed in Example 1.

Table 5: *E. coli* summary for this establishment and nationally

	Plant	National
Positives	268	1135
Minimum	-1.097	-1.097
Q1	-1.097	-1.097
Median	-1.097	-1.097
Mean (+ve)	-0.716	-0.799
Q3	-0.495	-0.620
90th Percentile	0.058	-0.097
95th Percentile	0.442	0.246
99th Percentile	1.120	1.186
Maximum	2.146	2.413
SD	0.553	0.525

Interpretation

- We previously showed (Example 1) that 30% of samples from this plant had positive *E. coli* samples, compared to only 7% nationally.
- However, it can be concluded from the summary of concentrations that although positive, the concentrations of *E. coli* were generally low.
- In fact, at this plant and also nationally, 50% of the data were recorded at the limit of detection (i.e. recorded as having only 1 colony).
- In addition, 95% of positive *E. coli* concentrations were less than $10^{0.442}$ = 2.77 organisms per cm^2.
- The rest of the concentrations were quite similar (percentiles, mean and standard deviation) to the national 'picture'.

Graphical representations of the data

Two graphical representations of the data are presented in the reports, box plots and time plots. In the box plots, the (positive) $\log_{10}$ concentrations are summarised on a monthly basis over the three-year period. In the time plots, all individual observations are used on the original scale, including negative samples (or those less than the limit of detection).

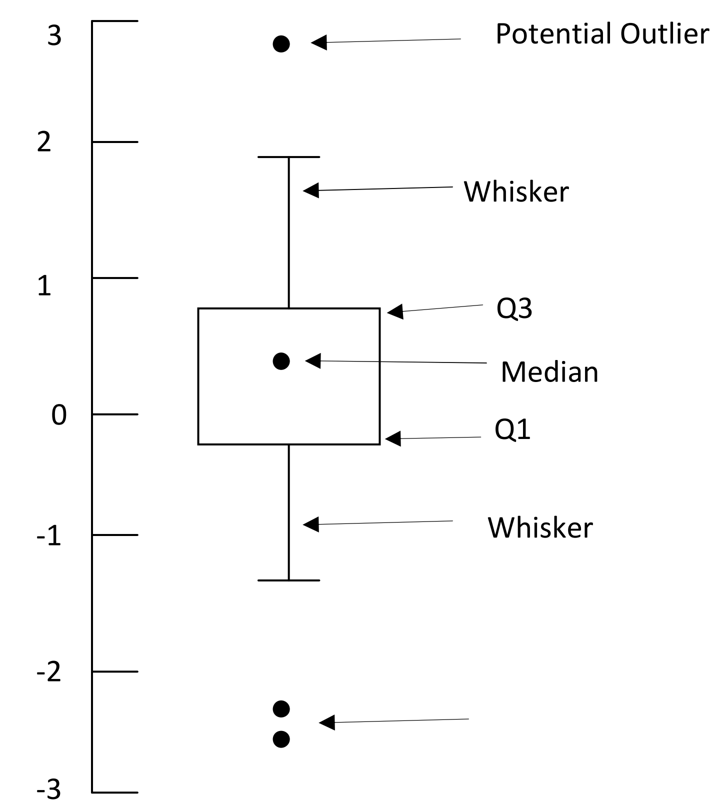
Box plot

A box plot is a convenient way to present groups of data. It utilises the descriptive statistics that we have explained above: the minimum, Q1, median, Q3 and maximum. Box plots help to identify differences between groups (in this case between months) of concentrations, showing the spread of the data within each sampling month. It may also help to identify seasonal trends and extreme or unexpected concentrations.

To construct a box plot:

- First draw a box from Q1 to Q3. Half of the data will fall within this box.
- Draw the median
- Draw the whiskers, the length of which is determined by 1.5*IQR and end on the closest observation to this defined value.
- Observations falling outside the extent of the whiskers are indicated separately. Values falling far outside the whiskers indicate potentially unusual or extreme observations (outliers). They should be investigated (as far as possible) to determine why they are so different.

Figure 2: Example of box plot

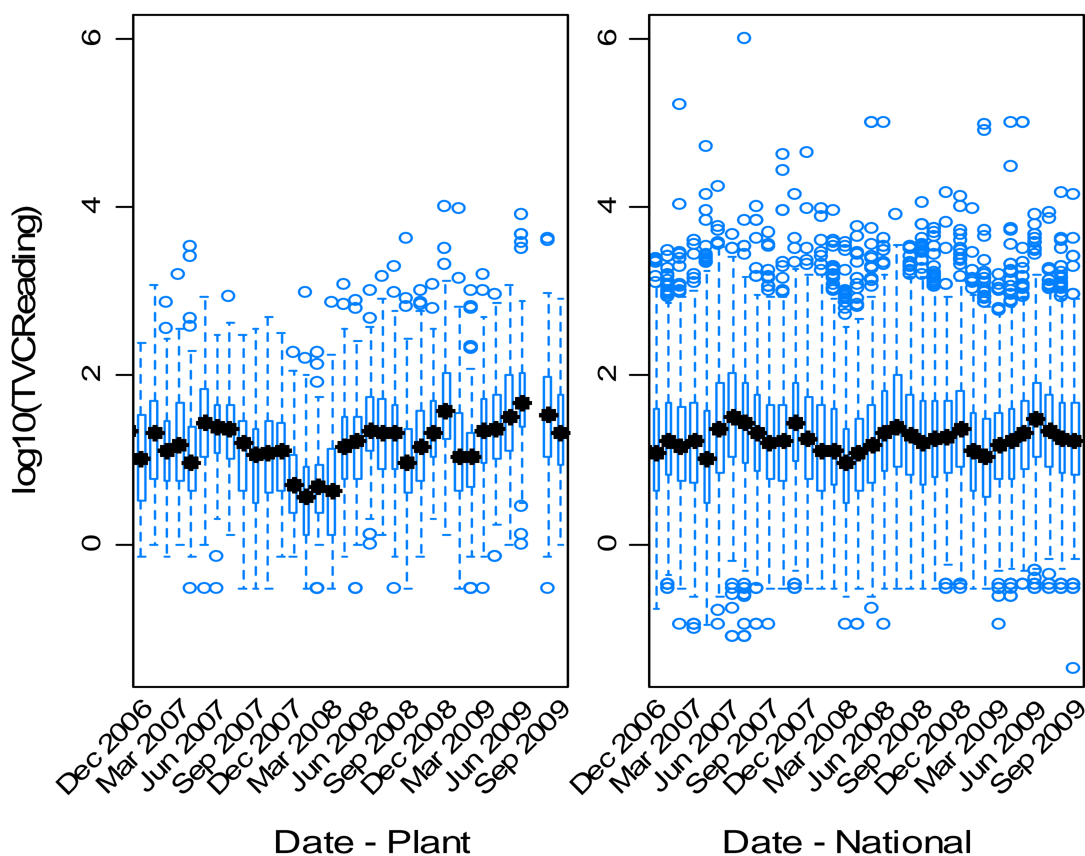


Of those observations which are drawn separately, only those that are along way from the rest of the data should be of concern and investigated further to determine if the value was a data entry error or if there was an identifiable reason for the extreme value.

Example 9

A box plot of monthly Total Viable Counts for a given establishment and all establishments is presented in Figure 3.

Figure 3: Box plot of monthly TVC for a particular plant and nationally



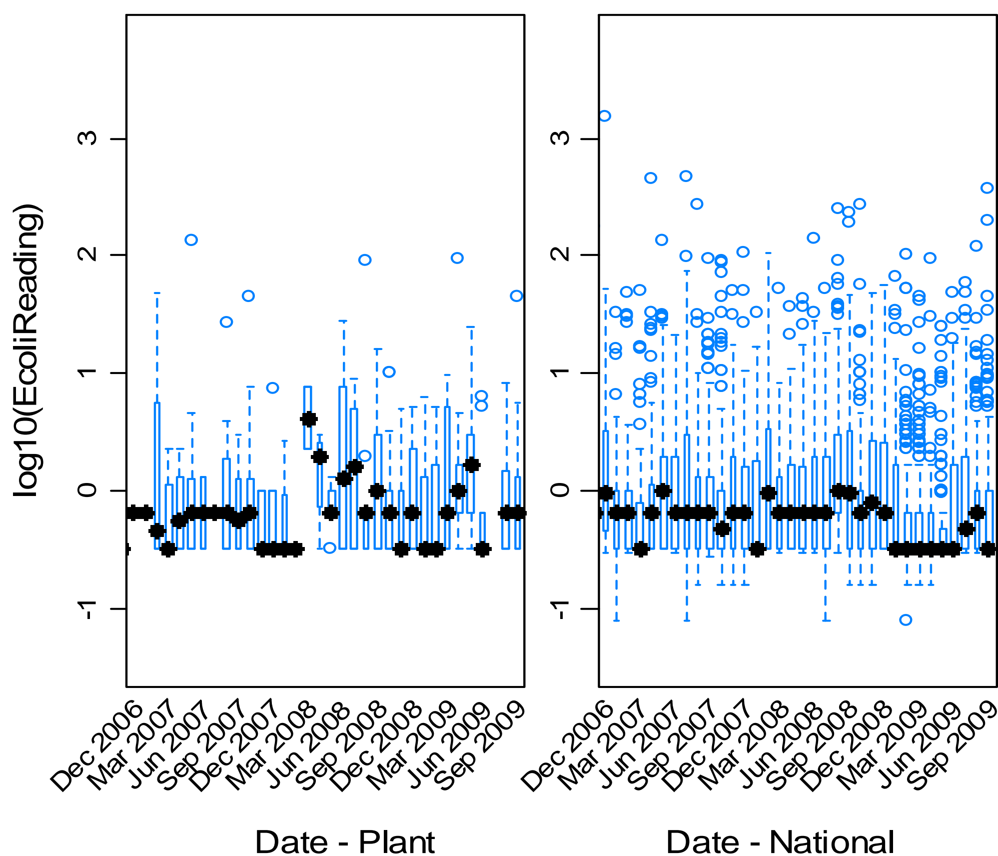
Interpretation

- The concentration of TVC at this plant is similar to that found nationally over the three year sampling period (around 1 – 1.5 log_{10} cfu/cm^2 on average).
- There does appear to be a similar pattern in concentrations over time, with lower TVCs being recorded in March 2008 at both the plant and nationally. However, this is more apparent at the plant.
- Note also that there were some consistent increases in the medians (black dots) in late 2008 and early 2009. This raises the question why these may have occurred.

Example 10

A box plot of monthly *E. coli* counts for a given establishment and all establishments is presented in Figure 4.

Figure 4: Box plot of monthly *E. coli* positive concentrations for a particular plant and nationally



Interpretation

- The concentration of *E. coli* at this plant is similar to that found nationally over the three year sampling period.
- Many of the positive samples are equal to the limit of detection. If not, the concentrations are generally low and at this plant with a maximum of 2 log or 100 cfu/cm^2.
- Note that the extreme points often fall a "long way" from the rest of that month's data. This makes them clear outliers and consequently their validity should be investigated (where possible – especially for the most recent month).

Time plot

In this plot all *E. coli* concentrations, including the negative samples (i.e. those below the limit of detection) are summarised graphically on the original (untransformed) scale. This plot can be used to compare an individual plant's level of *E. coli* compared to that found nationally over the same sampling period. It may also assist in summarising the number of 'alerts' a plant has had over the sampling period compared to those found nationally. These values should also be carefully checked against the plant's records to ensure that they were not the result of data entry errors.

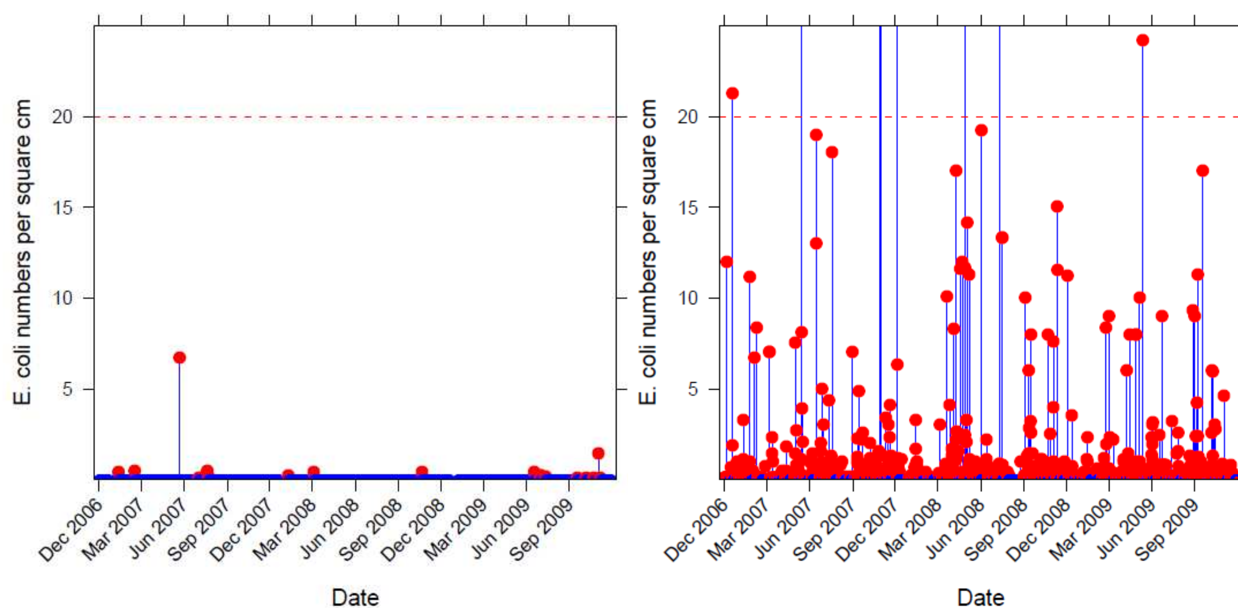
In the time plot:

- Positive tests are represented as red dots; negative tests as blue open circles.
- Red (dashed) horizontal lines show the 'm' and 'M' values for that species. The values of 'm' and 'M' for each species are defined in Appendix 1 of AQIS Meat Notice 2003/6.
  - Observations below the defined value 'm' are considered to have **Acceptable** levels of *E. coli*
  - Observations above the defined value 'M' are considered to have **Unacceptable** levels of *E. coli*
  - The observations between 'm' and 'M' are considered to have **Marginal** levels of *E. coli*.

Example 11

The time plot for the concentration of *E. coli* for an establishment and nationally is presented in Figure 5.

Figure 5: Time plot of *E. coli* concentrations at the plant and nationally.



Interpretation

- This plant had only a few positive *E. coli* samples over the three year period, with the majority of positives being observed between June and December 2009.
- The *E. coli* concentrations at this plant appear to be very low.

Appendix 2 – Example National Report

E. coli and *Salmonella* Monitoring Report

Species STEER/HEIFER

Reporting Period:

01 Dec 2006 to 30 Nov 2009

Generated

December 18, 2009 at 21:01

Prepared by
SARDI Food Safety Research
33 Flemington Street
GLENSIDE, SA 5065, Australia
Ph: +61 8 8207 7945
Fax: +61 8 8207 7854

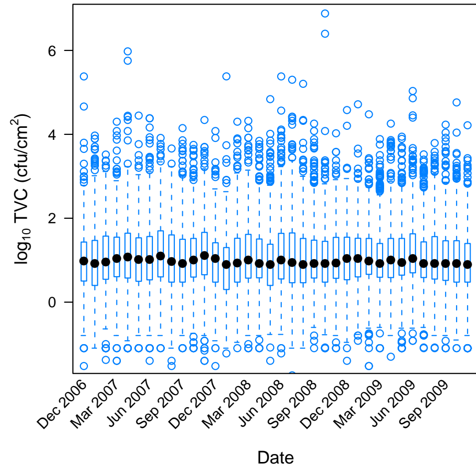
1 Total Viable Count Summary

Table 1: Total Viable Count prevalence summary for all establishments.

	National
Tests	37103
Positives	30539
Percent +ve	82.3
Lower Bound	81.9
Upper Bound	82.7

Table 2: Total Viable Count summary for all establishments.

	National
Positives	30539
Minimum	−1.745
Q1	0.518
Median	0.925
Mean (+ve)	1.055
Q3	1.519
90th Percentile	2.079
95th Percentile	2.447
99th Percentile	3.204
Maximum	6.881
SD	0.803



(a)

Figure 1: Box plot of monthly Total Viable Counts for all establishments.

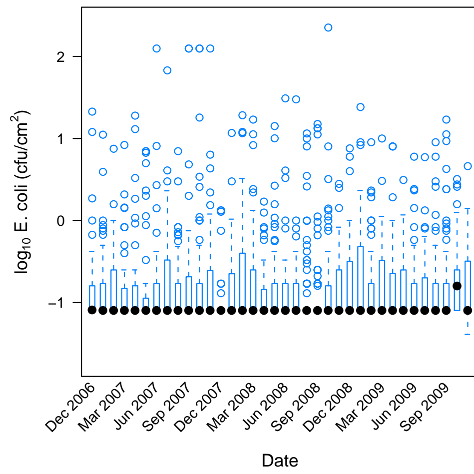
2 *E. coli* Summary

Table 3: *E. coli* prevalence summary for all establishments.

	National
Tests	50001
Positives	2076
Percent +ve	4.15
Lower Bound	3.98
Upper Bound	4.33

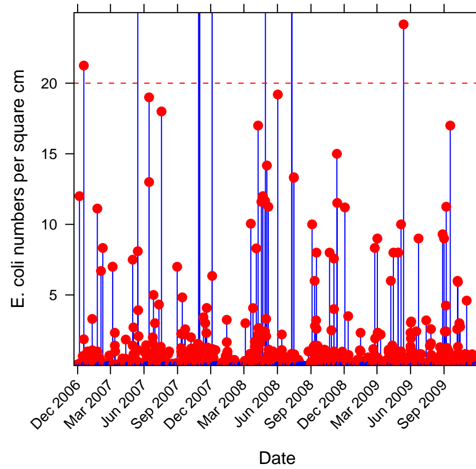
Table 4: *E. coli* summary for all establishments.

	National
Positives	2076
Minimum	−1.387
Q1	−1.097
Median	−1.097
Mean (+ve)	−0.816
Q3	−0.770
90th Percentile	−0.097
95th Percentile	0.324
99th Percentile	1.124
Maximum	2.684
SD	0.521



(a)

Figure 2: Box plot of monthly *E. coli* positive concentrations for all establishments.



(a)

Figure 3: Time plot of *E. coli* tests for all establishments — positive tests are presented as red points; negative tests are represented as blue circles.

3 *Salmonella* Summary

Table 5: *Salmonella* prevalence summary for all establishments.

	National
Tests	11872
Positives	18
Percent +ve	0.152
Lower Bound	0.090
Upper Bound	0.240

4 Glossary of Terms

4.1 Prevalence summary

Tests: The total number of samples (TVC, *E. coli* or *Salmonella*) in the ESAM database during the reporting period.

Positives: The number of samples with positive concentrations (ie. concentrations > 0).

Percent +ve: $100 \times$ Positives/Tests.

Lower Bound & Upper Bound: Lower and Upper 95% Confidence Bounds. The "true" prevalence is expected to be in this range.

4.2 TVC and *E. coli* concentration summary

All concentration data are converetd into logarithms with base 10, given by $\log_{10}$ cfu/cm^2.

Minimum: Minimum concentration.

Q1 or 1st Quartile: 25% of the data are less than this value, 75% are more.

Q3 or 3rd Quartile: 75% of the data are less than this value, 25% are more.

Median: 50% of the data are less than this concentration, 50% are more,

90th Percentile: 90% of the data are less than this value, 10% are more.

95th Percentile: 95% of the data are less than this value, 5% are more.

99th Percentile: 99% of the data are less than this value, 1% are more.

Maximum: Maximum concentration.

Mean: The average.

Standard Deviation (SD): A measure of spread (or variability) about the mean.

4.3 Box plot

A graphical tool to assess the data.

- The solid dot is the median.
- The box contains half the data.
- The lower and upper bounds of the box are the 1st and 3rd quartile.
- The inter-quartile range (IQR) = Q3 - Q1
- The length of the whiskers is calculated by $\pm 1.5\times$ IQR. The end of the whiskers corresponds to the observation in the dataset that is closest to this defined value.
- Observations falling outside the extent of the whiskers are indicated separately. Values falling far outside the whiskers indicate unusual or extreme values.

4.4 Time plot of *E. coli* concentrations over time

This plot useful to compare the level of *E. coli* at individual plants over time compared to that found nationally over the same sampling period.

- Positive tests are represented as red dots; negative tests as blue open circles.
- Red (dashed) horizontal lines show the 'm' and 'M' values for that species.
  - The value of 'm' and 'M' for each species is defined in Appendix 1 of AQIS Meat Notice 2003/6.
  - Observations **below** the defined value 'm' are considered to have Acceptable levels of *E. coli*.
  - Observations **above** the defined value 'M' are considered to have Unacceptable levels of *E. coli*.
  - The observations between 'm' and 'M' are considered to have Marginal levels of *E. coli*.