Explanatory Guide for the National Microbiological Database Reporting System

Government of South Australia

SARDI
SOUTH AUSTRALIAN RESEARCH AND DEVELOPMENT INSTITUTE

Explanatory Guide for the National Microbiological Database Reporting System

Background

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

The NMD provides useful information for benchmarking Australia's performance and can be used in market access negotiations. Meat and Livestock Australia provided funding for a reporting system for the NMD to be developed and to give feedback to establishments on their consolidated ESAM results. SARDI Food Safety now provides establishments with regular (monthly) reports so that they can compare their performance with that found nationally for the same reporting period.

This guide has been developed to assist QA staff interpret their establishment's monthly reports.

The Data

The NMD includes a combination of descriptive or **categorical variables** and **continuous measurement variables**.

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

- *Species* – Calf, Cow/Bull, Steer/Heifer, Sheep, Lambs, Goat Skin on, Goat Skin Off, Pig Skin on or Pig Skin Off
- *Dressing* – Bed, Conventional, Gravity Rail or Inverted
- *Shift* - First, Second or Third
- *Swabbed* - Hot or Cold

At present the only categorical variable used in the reports is *Species*.

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 species at each establishment. The species are: Calf, Cow/Bull, Steer/Heifer, Sheep, Lambs, Goat Skin On, Goat Skin Off, Pig Skin On and Pig Skin Off. A separate report on *E.coli* O157 is also generated for each establishment testing for *E.coli* O157.

Summary tables and graphs are presented for prevalence and positive concentrations for the individual establishment and the national baseline.

A 3-year moving window of data is used to generate the reports. Every month the oldest month is dropped from the dataset and the newest month added. The information presented in the tables is summarised over the given 3-year sampling period.

As of August 2010 monthly summaries for the most recent two months are reported to assist establishments verify the accuracy of the data entered into the NMD.

Definition of Terms

Prevalence

Prevalence refers to the percentage of samples 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, i.e. tested positive for *Salmonella*. This is then divided by the total number of tests performed during the sampling period and multiplied by 100 to give the percentage of positive tests. In the reports, prevalence is referred to as **Percent +ve**.

Percent +ve: Positives/Tests*100.

Where:

Tests: The total number of samples in the NMD during the reporting period.

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

95% Confidence Interval

A **Confidence Interval** gives a 'ballpark' of where the *true prevalence* may be. The true prevalence is unknown as not every carcase is tested and neither is the total surface area of each carcase.

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

If nothing in the process changes, then it is expected that the prevalence in the future will lie within those bounds 95% of the time.

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. As such, the widths of the confidence intervals for the national summaries are likely to be smaller than those found at an individual establishment.

Median and Mean

The median and the mean are both measures to determine values that are typical. They describe the 'middle' of the dataset and what could be called an 'expected' concentration in the data.

The **median** describes the midpoint concentration of the data. As such, 50% of values are less than the median and 50% of values are greater than the median.

The median is a 'resistant' measure, because it is not influenced by extreme observations.

An alternate measure is the **mean** or average. It is calculated by adding all values in the sampling period and dividing by the number of tests. Because the mean is calculated from the actual values it is easily influenced by extreme observations.

Standard Deviation

The **standard deviation** is another measure of variability which is commonly used in conjunction with the mean. The standard deviation is also susceptible to the effects of unusually large or small observations.

- For those who are not mathematically inclined, it is sufficient to know that the standard deviation is calculated by looking at how far each observation is from the mean of the observations.
- For the mathematically inclined, a description of the calculation can be found on Wikipedia at

  http://en.wikipedia.org/wiki/Standard_deviation#With_sample_standard_deviation.

It is important to note that the standard deviation will be small if the observations are very consistent, i.e. they lie close together. However, if the observations are inconsistent, i.e. sometimes high and sometimes low, then the standard deviation will be large.

A slaughter and dressing process with a small standard deviation produces more consistent and predictable end product.

Log_{10} Transformation

It is standard practice for microbiological concentration data to be transformed using a mathematical logarithm function (base 10), denoted by $\log_{10}$. After this transformation, the distribution of concentrations is made more symmetrical.

The effect of the $\log_{10}$ transformation on the concentration data is also shown in Table 1.

Table 1: 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

Note that:

- A 1 $\log_{10}$ reduction equates to a 90% reduction in the concentration on the original scale (i.e. a change from 1000 to 100 or 10,000 to 1000, etc).
- A 2 $\log_{10}$ reduction equates to a 99% reduction in the concentration on the original scale (i.e. a change from 10,000 to 100 or 100,000 to 1000 etc).

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

The original count can be obtained from the transformed value by performing the reverse transformation. This is done by calculating 10 to the power of the value of interest. For example, 2.176 $\log_{10}$ cfu/cm^2 (the $\log_{10}$ value) = $10^{2.176}$ cfu/cm^2 = 150 cfu/cm^2 (the original value).

Limit of Detection and 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 surface area that each ml of the undiluted sample represents.

For cows/bulls, steers/heifers and pigs 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 estimated to be **less than** the limit of detection or <0.08 cfu/cm^2.

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

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

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

Therefore, it is expected that the minimum $\log_{10}$ concentration (i.e. 1 colony) reported in the tables should be equal to the $\log_{10}$ of the sampling factor for that species. In other words, the 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, steers/heifers and pigs and
- -0.48 $\log_{10}$ cfu/cm^2 for sheep, lambs, calves and goats.

If minimum values less than these are reported in the summary tables, then QA staff should check the original records to ensure that the data were entered correctly and/or the dilution and method of analysis used by the laboratory (see also Example 7).

Percentiles and Quartiles

Percentiles are those points where only a pre-determined percentage of values exceed the percentile. They are useful for making statements about the process.

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.

A special percentile is the **median**, which equals the 50th percentile. Two other special percentiles are the 25th and the 75th percentiles – these are also known as **quartiles**.

The **first or lower quartile (Q1)** is the value which divides the data set in such a way that one quarter (25%) of the data fall below and three quarters (75%) of the data fall above Q1.

The **third or upper quartile (Q3)** is the value which divides the data set in such a way that three quarters (75%) of the data fall below and one quarter (25%) of the data fall above Q3.

Inter-Quartile Range (IQR)

The **inter-quartile range (IQR)** is the difference between **Q3** and **Q1** (IQR = Q3 – Q1). It 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.

Box Plot

A box plot is a convenient way to present groups of data. It utilises the descriptive statistics from above, namely the minimum, Q1, median, Q3 and maximum. Box plots help to identify differences between groups (in the reports between months) of concentrations, showing the spread of the data within each sampling month. They can also help to identify seasonal trends and extreme or unexpected concentrations. In the box plots, the $\log_{10}$ concentrations are summarised on a monthly basis over the 3-year period for only **positive samples**. See Appendix 1 for some additional details on box plots.

Time Plot

In the time plots, all observations are used on the original (untransformed) scale, including negative samples (or those less than the limit of detection).

This plot can be used to detect ‘hot spots’ or periods where contamination is more common. They may also assist in summarising the number of ‘alerts’ an establishment has had over the sampling period.

Examples: Main Concepts

Example 1: Median and Mean

The following log_{10} concentration data (11 values) are ordered from lowest to highest:

-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 6^{th} largest observation and equals **0.69** log_{10} cfu/cm^2. Five samples are below the median and five are above. If there were an even number of observations we would have two midpoints. The median is then defined as being halfway between these two midpoints.

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

Note that the mean and median concentrations are very similar. Therefore, it can be concluded that there are few (if any) highly unusual observations in the data. Of course, this is easily verified in such a small dataset, but more difficult in a larger dataset, such as those obtained from the NMD. In those cases, unusual observations are assessed better with the box plots.

Example 2: Effect of data entry error on median and mean

This is the same data, but the largest observation has been changed from 1.57 to ***4.57***, as might occur if a data entry error had been made.

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

The **median** again corresponds to the 6^{th} largest observation and equals **0.69** log_{10} cfu/cm^2. Five observations are below the median and five are above. It is unaffected by the change in the largest value.

The **mean** is again 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 concentration is influenced (shifted upwards in this case) by an unusual (or extreme) observation.

Example 3: Inter-Quartile Range (IQR)

The log_{10} concentration (n=11) data presented in **Example 1** is 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, as shown earlier, is **0.69** log_{10} cfu/cm^2.

To determine the **lower quartile** (**Q1**), the median of the five observations **less** than the median value of 0.69 is taken. This gives a value of ***0.23*** log_{10} cfu/cm^2.

To determine the **upper quartile** (**Q3**), the median of the five observations **above** the median value of 0.69 is taken. This gives a value of ***1.10*** log_{10} cfu/cm^2.

To determine the **inter-quartile range**, the difference between Q3 and Q1 is taken. That is, the **IQR = 1.10 – 0.23 = 0.87**.

Example 4: Standard Deviation (SD)

The following example contains two data sets of $\log_{10}$ concentrations.

Data 1	0.8	0.9	1.0	1.1	1.2
Data 2	0.4	0.7	1.0	1.3	1.6

Both data sets have the same sample mean of 1.0 $\log_{10}$ cfu/cm^2. However, the standard deviation for Data 1 equals 0.16 $\log_{10}$ cfu/cm^2, while that for Data 2 equals 0.47 $\log_{10}$ cfu/cm^2. Therefore, the process from which Data 1 originate is more likely to produce more consistent results.

Examples: Interpretation of the Reports

Example 1: Prevalence summary for TVC

A TVC prevalence summary is shown in Table 1.

Table 1: Total Viable Count prevalence summary for an establishment and nationally

	This Establishment			National
	2 Months Ago	Last Month	Last 3 Years	Last 3 Years
Tests	**36**	**26**	**1472**	**37265**
Positives	**34**	**26**	**1462**	**30535**
Percent +ve	**94.44**	**100.00**	**99.32**	**81.94**
Lower bound	81.34	86.77	*98.75*	*81.55*
Upper bound	99.32	100.00	*99.67*	*82.33*

Interpretation:

- Of the 30535 positive TVC samples found nationally, 1462 (1462/30535 or 4.79%) were from this establishment.
- The prevalence (Percent +ve) at this establishment was higher than the national prevalence during the same period (99.32% versus 81.94%).
- From the national summary we can be fairly confident that the national prevalence falls between *81.55* and *82.33*%.
- From the establishment summary (Last 3 Years) it can be concluded that the slaughter and dressing process results in between *98.75* and *99.67*% of carcases with TVC above the limit of detection.

Note that the prevalence summary considers positive samples only and does not take into account the actual TVC value for the sample. This means that even though 100% of the samples could be positive, they could all be at the limit of detection (which would be acceptable) or (the other extreme) they could all be unacceptable.

Example 2: Prevalence summary for *E. coli*

An *E.coli* prevalence summary is shown in Table 2.

Table 2: E. coli prevalence summary for an establishment and nationally.

	This Establishment			National
	2 Months Ago	Last Month	Last 3 Years	Last 3 Years
Tests	**31**	**24**	**1015**	**49299**
Positives	**2**	**1**	**42**	**1817**
Percent +ve	**6.45**	**4.17**	**4.14**	**3.69**
Lower bound	0.79	0.10	*3.00*	*3.52*
Upper bound	21.42	21.12	*5.55*	*3.86*

Interpretation:

- Of the 1817 *E. coli* detections nationally, 42 (42/1817 or 2.31%) were from this establishment.
- The prevalence (Percent +ve) at this establishment was slightly higher than the national prevalence during the same period (4.14% versus 3.69%).
- From the national summary we can be fairly confident that the national prevalence falls between *3.52* and *3.86%*.
- From the summary for this establishment it can be concluded that the slaughter and dressing process results in between *3.00* and *5.55*% of carcases with *E. coli* above the limit of detection.

Note that the prevalence summary is only part of the story – it doesn't tell you how 'bad' the contamination was.

Example 3: Prevalence summary for *Salmonella*

A *Salmonella* prevalence summary is shown in Table 3.

Table 3: *Salmonella* prevalence summary for an establishment and nationally

	This Establishment			National
	2 Months Ago	Last Month	Last 3 Years	Last 3 Years
Tests	**11**	**14**	**136**	**5745**
Positives	**0**	**1**	**2**	**41**
Percent +ve	**0.000**	**7.143**	**1.471**	**0.714**
Lower bound	0.000	0.181	*0.179*	*0.513*
Upper bound	28.491	33.868	*5.211*	*0.967*

Interpretation:

- This establishment had 2 positive *salmonella* tests over the last three years, one of which occurred in the last month.
- The national prevalence of *Salmonella* is low (about 0.7%).
- The prevalence is higher at this establishment than that found nationally (1.471% compared to 0.714%).
- Compared to the national level, this establishment appears to perform worse.
- Based on the bounds of the confidence intervals, it may be expected that at this establishment as few as 17-18 carcases in 10,000 (*0.179*%) to as many as 52 carcases in 1000 (*5.211*%) contain *Salmonella*.

Example 4: Log_{10} TVC concentration summary

A log_{10} TVC concentration summary for an export establishment is compared against national levels in Table 4.

Table 4: Total Viable Count summary for an establishment and nationally (log_{10} cfu/cm^2) for samples where TVC was greater than the limit of detection.

	This Establishment			National
	2 Months Ago	Last Month	Last 3 Years	Last 3 Years
Minimum	-1.097	-0.284	-1.097	-1.745
Q1	**-0.103**	**0.735**	**0.813**	**0.545**
Median	**0.530**	**1.779**	**1.447**	**0.928**
Mean (+ve)	0.648	1.671	1.375	1.063
Q3	**1.439**	**2.249**	**2.000**	**1.519**
90th Percentile	1.968	2.895	2.415	2.079
95th Percentile	2.581	3.610	2.716	2.462
99th Percentile	3.240	3.904	3.266	3.318
Maximum	3.380	3.940	4.380	6.881
SD	1.123	1.113	0.883	0.804

Interpretation:

- TVC concentrations for the last three years for this establishment were generally higher/worse than those found nationally when comparing **Q1**, **Median**, Mean, **Q3**, 90^{th} percentile and 95^{th} percentile.
- Microbiologists often consider a difference of 0.5 to 1 log_{10} to be of practical significance. So based on the medians (for this establishment for the last three years and nationally), this establishment could be considered to be worse than the national baseline.

Example 5: *E. coli* concentration summary

An *E. coli* concentration summary for an export establishment is compared against national levels in Table 5. Note that it is important to summarise the concentrations in context of the prevalence Table for *E. coli*, discussed in Example 2.

Table 5: *E. coli* summary for an establishment and nationally ($\log_{10}$ cfu/cm^2) for samples where *E. coli* was detected.

	This Establishment			National
	2 Months Ago	Last Month	Last 3 Years	Last 3 Years
Minimum	-1.097	1.114	-1.097	-1.097
Q1	**-0.532**	**1.114**	**-1.044**	**-1.097**
Median	**0.032**	**1.114**	**-0.770**	**-1.097**
Mean (+ve)	0.032	1.114	-0.557	-0.784
Q3	**0.597**	**1.114**	**-0.496**	**-0.620**
90th Percentile	0.936	1.114	0.515	-0.081
95th Percentile	1.048	1.114	1.159	0.384
99th Percentile	1.139	1.114	1.298	1.147
Maximum	1.161	1.114	1.383	3.079
SD	1.597	NA	0.688	0.545

Interpretation:

- Example 2 demonstrated that 4.14% of samples from this establishment were found to have *E. coli* above the limit of detection, compared to only 3.69% nationally (Percent +ve).
- It can be concluded from the above summary for the national data that although *E.coli* was detected in 1817 samples (from Table 2), 50% of these had concentrations that were at the limit of detection (using the **Median**). *Consequently, QA Managers may wish to check that data are entered correctly, namely that values "< 0.08" are entered as "0" rather than as "0.08".*
- In addition, for this establishment over the last three years, 75% of samples where *E. coli* was detected had concentrations less than $10^{-0.496}$ = 0.32 cfu/cm^2 (using Q3) and 90% were less than $10^{0.515}$ = 3.27 cfu/cm^2 (using the 90th percentile).
- There is no standard deviation for the data for the last month because there was only one *E. coli* detection. Calculating a standard deviation requires two or more observations.

Example 6: Box plot of monthly TVC

Box plots of monthly Total Viable Counts for one establishment and all establishments are shown in Figure 1.

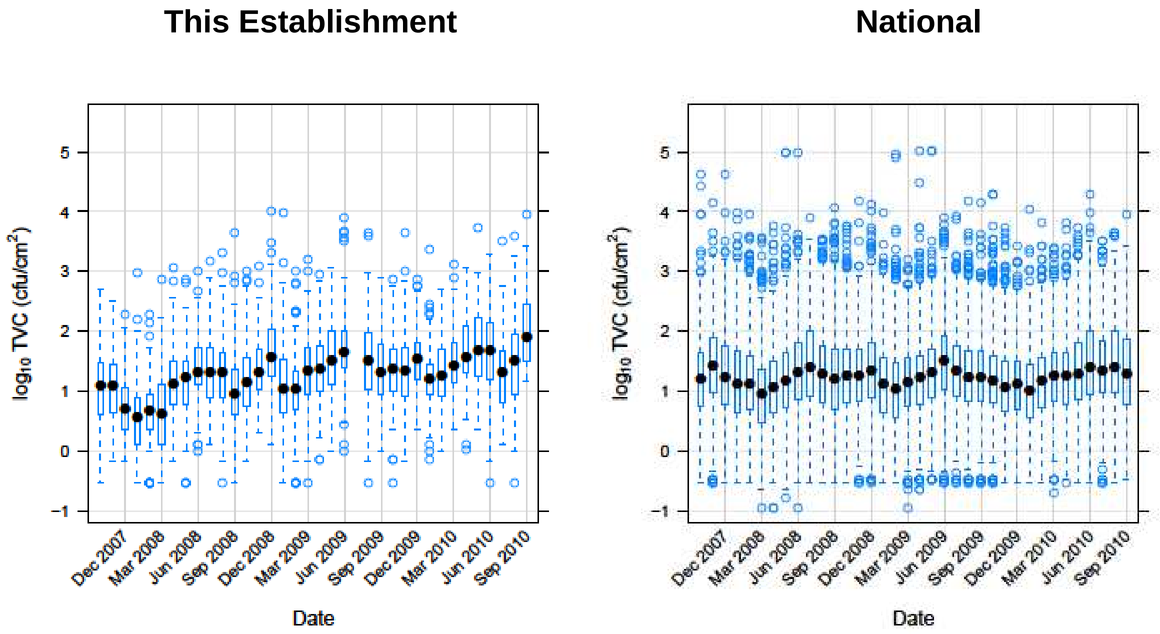

Figure 1: Box plot of monthly Total Viable Counts for an establishment and all establishments

Interpretation:

- The concentrations of TVC at this establishment are reasonably similar to those found nationally over the 3 year sampling period (the monthly median is around 1-1.5 $\log_{10}$ cfu/cm^2).
- Over the three years, the median $\log_{10}$ TVC concentration appears to have increased. In particular there appears to be an increase over the most recent three months.
- There are short term trends, for example, the consistent increase from September 2008 to December 2008. *Could there be a reason for these?*

Example 7: Box plot of monthly *E. coli* Counts

Box plots of monthly *E. coli* counts for one establishment and all establishments are shown in Figure 2.

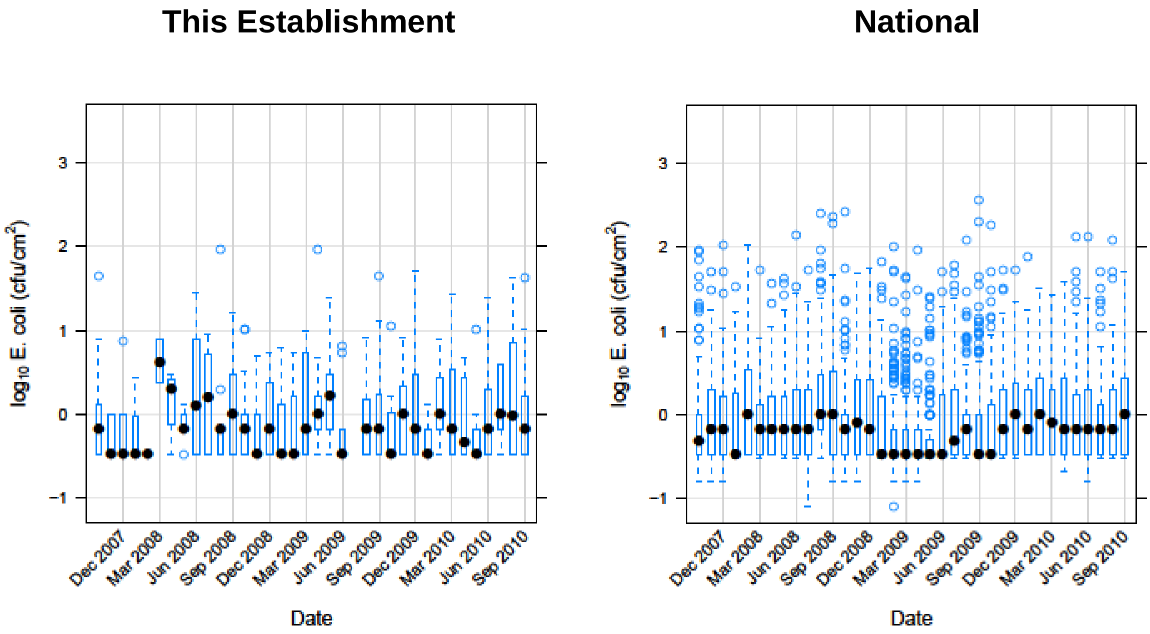

Figure 2: Box plot of monthly *E. coli* positive concentrations for an establishment and all establishments

Interpretation:

- The concentrations for samples where *E. coli* was detected at this establishment are similar to those found nationally over the 3 year sampling period, although some months show some higher values (e.g. March 2008).
- The lowest *E. coli* values should be -0.48 $\log_{10}$ cfu/cm^2 for this species, unless different sized areas were sampled or different dilutions were undertaken. Consequently, concentrations below the limit of detection should be checked to ensure they are 'real'. Note that this is not an issue for this establishment.
- There are several months for which the extreme observations (indicated by small circles) for this establishment are considerably higher than the remainder of that month's *E. coli* concentrations, e.g. April 2009 and May 2010. These extreme observations should, as far as possible, be investigated.

Example 8: Time plot for the concentration of *E. coli*

The time plot for the concentration of *E. coli* for one establishment and all establishments are shown in Figure 3. In the time plot:

- Tests where *E. coli* was detected are represented as red dots.
- Tests where *E. coli* was not detected are represented as blue open circles.
- Red (dashed) horizontal lines show the marginal 'm' and unacceptable 'M' limits for that species, as defined in Appendix 1 of AQIS Meat Notice 2003/6. For the species in Figure 3, the values are m=5 and M=100.
  - Observations below or equal to 'm' are considered to have **Acceptable** levels of *E. coli*
  - Observations above the '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*.

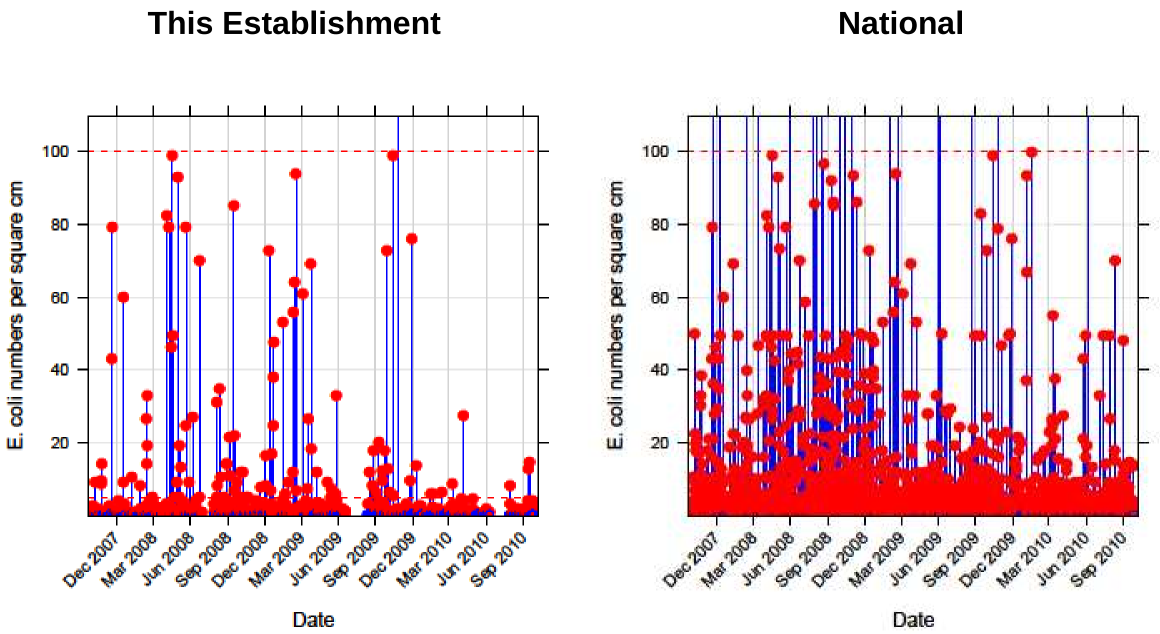

Figure 3: Time plot of *E. coli* tests for an establishment and all establishments – positive tests are presented as red points; negative tests are represented as blue circles

Interpretation:

- There seem to be some "clusters" of high values around May 2008, February/March 2009 and October/November 2009. *What could be causing these?*
- This establishment had one *E.coli* concentration that is considered unacceptable, and two that border on unacceptable.
- There have been much lower concentrations in the last few months. *Has the process been improved?*

Example 9: *E. coli* O157 concentration summary

An *E.coli* O157 prevalence summary is shown in Table 6.

Table 6: *E. coli* O157 prevalence summary for an establishment and all establishments (on-plant and AQIS verification tests)

	This Establishment			National
	2 Months Ago	Last Month	Last 3 Years	Last 3 Years
Tests	**127**	**68**	**2744**	**92351**
Positives	**1**	**0**	**12**	**106**
Percent +ve	**0.79**	**0.00**	**0.44**	**0.11**
Lower bound	0.02	0.00	*0.23*	*0.09*
Upper bound	4.31	5.28	*0.76*	*0.14*

Interpretation:

- Of the 106 *E. coli* O157 detections nationally, 12 (12/106 or 11.32%) were from this establishment. The time plot for this establishment is shown in Example 10 and highlights the contribution this establishment makes to the national counts.
- The prevalence (Percent +ve) at this establishment was four times higher than the national prevalence during the same period (0.44% versus 0.11%).
- From the national summary we can be fairly confident that the national prevalence falls between *0.09* and *0.14*%.
- From the summary for this establishment it can be concluded that the slaughter and dressing process results in between *0.23* and *0.76*% of carcases with *E. coli* O157 above the limit of detection.

Example 10: Time plot for the prevalence of *E. coli* O157

Time plots of monthly *E. coli* O157 counts for one establishment and all establishments are shown in Figure 4. In the time plot:

- If *E. coli* O157 was detected in one or more tests for a particular month, then a red dot is used to indicate the number of detections in that month.
- If tests were conducted but *E. coli* O157 was not detected in a particular month, then a blue dot is used to indicate that there were zero detections in that month.
- If no tests were conducted in a particular month then there is no dot for that month.

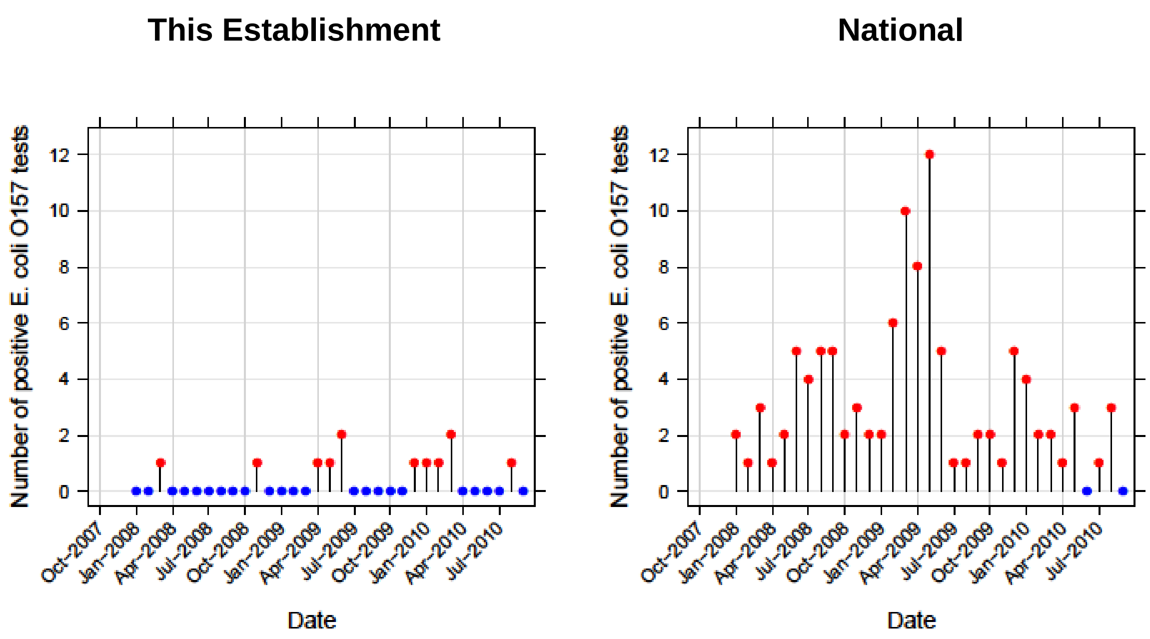

Figure 4: Time plot of confirmed detections of *E. coli* O157 for an establishment and all establishments (on-plant and AQIS verification tests) - no dot indicates no tests, blue dots indicate no detections and red dots indicate detections

Interpretation:

- As mentioned in Example 9, this establishment was responsible for 11.32% of all tests where *E.coli* O157 was detected.
- There are two clusters of detections (April 2009 to June 2009 and December 2009 to March 2010) and several apparently unrelated detections. *Is there a reason for the clusters?*

Appendix 1: The Box plot

To construct a box plot:

- Draw a box from Q1 to Q3. Half of the data will fall within this box.
- Draw the median (a dot inside the box).
- Draw the "whiskers", the length of which is at most 1.5*Inter-quartile Range. The lower whisker starts at the first observation *greater than or equal to* 1.5*Inter-quartile Range and ends at the box. The upper whisker starts at the box and ends at the last observation *less than or equal to* 1.5*Inter-quartile Range. The whiskers indicate the maximum and minimum after excluding "extreme" counts, and can be thought of as the limit of expected values.
- Observations falling outside the whiskers are indicated separately. Values falling far outside the whiskers indicate potentially unusual or extreme observations. They should be investigated (if possible) to determine the reason for such an unusual observation.