

**7TH OECD INTERNATIONAL TRADE STATISTICS EXPERT MEETING ITS
and OECD-EUROSTAT MEETING OF EXPERTS IN TRADE-IN-SERVICES STATISTICS
(TIS)**

Tour Europe - Paris La Défense, Salle des Nations, 11 - 14 September 2006

UNSD Practice

This document has been prepared by Ronald Jansen, UNSD

item 8 c)



UNITED NATIONS
DEPARTMENT OF ECONOMIC AND SOCIAL AFFAIRS

STATISTICS DIVISION
TRADE STATISTICS BRANCH

7TH OECD INTERNATIONAL TRADE STATISTICS EXPERT (ITS) MEETING
Paris, 11-13 September 2006

item 8c of the provisional agenda

Trade Volumes and Unit Values

Note by UNSD

In the processing of detailed merchandise trade statistics, we need good estimates for unit values in two instances, namely (1) to estimate quantities in those cases where a reporting country provided commodity values, but no quantities, and (2) to check if received quantity data is valid. In fact, in the latter case we use an interval of acceptable unit values.

As basic data for the calculations discussed in this note, we used the value and quantity reported by a particular country or area for a given commodity, year, flow, unit and partner country. This means that exports of cars by Germany to USA in 2004 gives us one unit value, and exports of cars by Germany to Nigeria in 2004 gives us another unit value. Per commodity, year, flow and unit we would generally have around 1,000 unit values by weight and around 500 by another quantity unit (like heads, liters of square meters). These other quantity units are recommended by WCO for each HS subheading. Note that in for a large number of HS subheadings the recommended unit is in fact kilograms.

We did the calculations described below for all HS commodities (about 5,000), years 2000 to 2004, separately for imports and exports, and for weight and supplementary quantity (where WCO recommends a unit other than weight). In all, we did these calculations on more than 50,000 different samples using some where around 50 million records from UN Comtrade. The annex to this note describes in more detail the methodology and provides a few examples.

Standard Unit Value

For each combination of quantity unit, commodity, year and flow (imports/exports) a best estimate for unit value can be calculated, which we call Standard Unit Value (SUV). For instance, the SUV for exports of butter in 2004 is US\$ 2.38 per kilogram. This estimate will be used to generate trade volume in case we have no information other than the commodity value for a given year and flow.

Currently, we have only calculated SUVs on a World basis. It may be possible that SUVs at a regional or sub-regional level provide better estimates compared to a World estimate. Other refinements, for instance commodity by region, are thinkable. For now, we have made no attempts to do such calculations.

Criteria for accepting or rejecting SUVs as good estimates

In the process of calculating SUVs two types of decisions were taken, namely (i) which unit values should be regarded as outlier and be left out of the sample for SUV calculation and (ii) – once the SUV has been calculated – which SUV is good enough as an estimate. For the outlier test we used the well-known box plot technique (established by John Tukey¹), in which one obtains lower and upper limits of acceptable unit values by taking 1.5 times the inter-quartile range and subtract, respectively add these to the 25th and 75th quartile. These calculations are done on the log-transformed data, since those are more centrally distributed than the original unit value data, of which the distribution is usually skewed.

Once a SUV has been calculated, it is still not immediately obvious if this SUV is a good estimate or not. In the case of Butter, the 25th percentile was US\$ 1.82 per kilogram and the 75th was US\$ 3.37. So, 50% of the (more than one thousand) data points stayed within a unit value range of little more than one dollar difference. In addition, only 2 observations were found to be too low and 3 too high as a unit value with limits of US\$ 0.80 at the lower end and US\$ 7.78 at the high end. For such a commodity, the SUV can be very confidently used to estimate missing weight.

However, there are many cases where the distribution of unit values is less consistent. For instance, the sample of unit values of pure-bred breeding horses, mules or hinnies ranges from a few hundred dollars per animal up to a few hundred thousand dollars per animal. In addition, a quarter of the unit values is lower than US\$ 3,000 per animal, whereas another quarter is higher than US\$ 33,000. Finally, the arithmetic mean is higher than the 75th percentile, which shows large skewness. Another element to consider is that the total number of records of this sample was only 116.

All of these observations mentioned point to selection criteria: (1) number of observations, (2) spread, and (3) skewness. In addition, we started looking at (4) the number of reporting countries and (5) multi-modality. The last one is maybe the least intuitive one, but is easy to explain. A commodity does not necessarily consist of just one product. On the contrary, in general a commodity consists of a basket of products. Whereas – for example – beer, milk, butter and crude oil are relatively homogeneous commodities, automated data processing machines may range from calculators to mainframe computers. When calculating SUVs we assume that there is one highly preferred product with a limited price range, which the large majority of the countries export (or import). In such case,

¹ See Tukey, J.W. (1977), *Exploratory Data Analysis*, Reading, MA: Addison Wesley Publishing Company.

the SUV would be a decent estimate even if there is variability in the basket of commodities. However, if more than one mode exist (a group of inexpensive and a distinct group of more expensive products), then the SUV would be a bad estimate either way.

The annex gives some mathematical background to these mentioned five criteria. For the moment we have set the limits of acceptability at (1) at least two reporting countries, (2) a sample of at least 30 unit values, (3) a relative standard deviation of less than 1.75, (4) a relative standard deviation of less than 3 and multimodality less than 2, and (5) cumulative value of excluded observations (outliers) less than 10% of total commodity value.

Applying these criteria leaves us about 85% of SUVs by weight (more than 4,000 commodities per year and flow) and about 80% by WCO recommended unit. These SUVs and the lower and upper limits derived from the SUV calculations are currently used in our joint UN/OECD data processing system CoprA.

It should be noted that, whereas we may establish SUVs for a large majority of commodities, we only use SUVs in those cases where no quantities were reported and none of the other rules for deriving quantities could be used. For an overview of the set of rules in deriving quantities, please refer to the joint UN/OECD paper at the 2006 meeting describing the data processing system.

Further analysis and use of unit values

Over the next months we will do further testing to maybe exclude more SUVs based on some additional criteria. We will also analyze the correlation between export and import unit values for the same commodities, as well as the time series of unit values for the same commodity, year and flow over a period of at least five years. As mentioned, we will attempt to group SUVs by region and to possibly detect strong grouping of SUVs by commodity, flow and region.

Further ahead, we are thinking of establishing a database with unit values which eventually could be used to derive trade indexes by country and commodity.

Annex

Methodology for calculation of Standard Unit Values

1. Introduction

- A ‘**standard unit value**’ (SUV) is defined for each basic heading of a commodity classification (for instance the 6-digit level of HS), by year, trade flow (imports/exports), and different quantity unit.
- The SUV serves two main **objectives**:
 1. It is used to **estimate volume** of trade when only monetary values are available
 2. It also provides a **benchmark** against which the quality of new value/volume data pairs can be assessed.
- A **sample** of unit values for each combination of unit, commodity, flow, and year is available in **UN Comtrade** when dividing values by their respective quantities. Based on that sample, a Standard Unit Value (SUV) can be calculated for each unit/commodity/flow/year, using the median unit value of value/quantity ratios.
- The methodology to calculate Standard Unit Values can be applied for several commodity classifications. At the moment, work has been completed for HS88, HS96, and HS02 classifications, and there is work in progress for the different revisions of the SITC classification.
- Section 2 of this annex summarizes some **features of the unit value data** using descriptive statistics. It provides alternative **measures of location, dispersion, and skewness** for the sample distribution of unit values, and illustrates these findings with some specific examples.
- The **main conclusions** from the descriptive analysis of Section 2 are:
 1. Unit value data for most commodities exhibit **high degree of variability**
 2. The distribution of unit values is usually **asymmetric** around its mean (skewness is usually positive; right-tailed).
 3. The data is affected by the presence of **outliers**.
 4. A **log-transformation** of the unit value data significantly reduces asymmetry, and therefore is more appropriate **to construct confidence intervals and rejection thresholds for outliers**.
- Section 3 sets up the criteria used to determine whether the available sample of unit values of a specific commodity/flow/year can be relied upon to determine a Standard Unit Value. Such criteria impose maximum acceptable limits on the asymmetry, spread and/or multimodality of the sample distribution.

2. Descriptive Statistics: Main Features of Unit Value Data

2.1. Assessment of variability

- A first measure of variability in unit value data for each commodity/flow/year sample is their relative standard deviation, RSD, which is defined as the ratio of the standard deviation (s) divided by the arithmetic mean (\bar{x})
- An analogous non-parametric measure of variability is the relative inter-quartile range, which is defined as

$$RIQ = \frac{(Q_3 - Q_1)}{M},$$

where M represents the median and Q_1 and Q_3 the 25th and 75th percentiles of the unit value sample, respectively.

- For a majority of commodities, the unit values calculated on the basis of value and quantity data exhibit a high degree of variability, as measured by the relative inter-quartile range (see Figure 1).
- In particular, 65% of 51,945 commodity-specific unit value samples available for imports and exports in the period 2000-2004 (using only the recommended quantity units of measurement) have a relative inter-quartile range greater than one. This variability shall be taken into account when assessing the reliability of commodity-specific Standard Unit Values for volume estimation and quality-checks purposes.

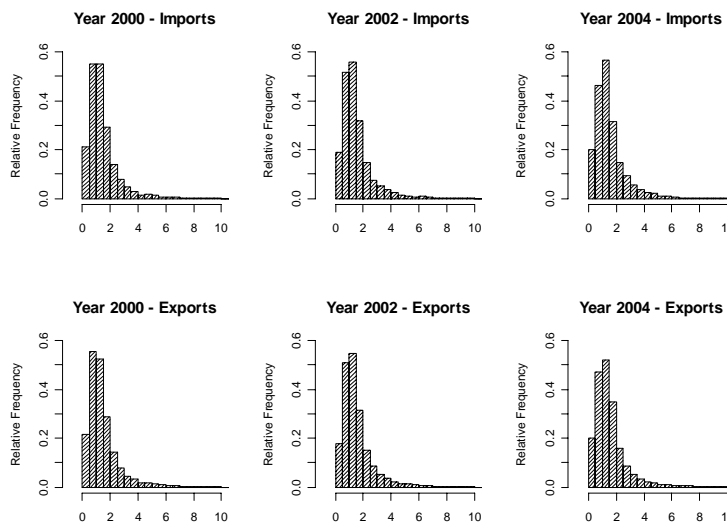


Figure 1. Distribution of the relative inter-quartile range of unit values among commodities

2.2. Assessment of asymmetry

- A non-parametric measure of skewness (or asymmetry) in the distribution of each unit value sample is provided by the Bowley skewness coefficient, which is defined as

$$B = \frac{(Q_3 - M) - (M - Q_1)}{Q_3 - Q_1} = \frac{(Q_3 - 2M + Q_1)}{Q_3 - Q_1},$$

Its value is bounded between -1 and +1, and it is equal to zero if the median is located exactly in the middle of the inter-quartile range.

- Examination of the samples (see Figure 2) reveals that the distribution of unit value data is typically skewed to the right (i.e., $B > 0$). More specifically, 92% of the commodity/flow/year samples of unit values have a positive Bowley coefficient, and in about 50% of the samples this coefficient is greater than 0.32.

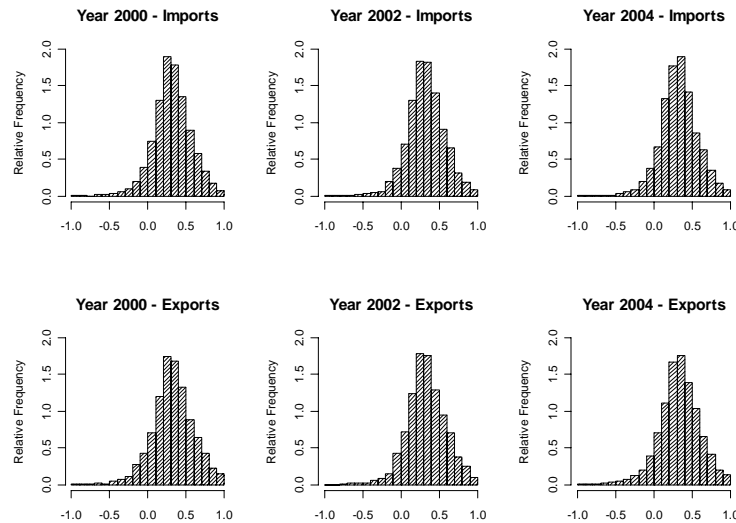


Figure 2. Distribution of the Bowles skewness of unit values among different commodities

- After applying a **logarithmic transformation** to the unit value data in each commodity/flow/year sample, their skew is typically near zero, as is shown in Figure 3. Moreover, approximately 50% of the transformed unit value samples have a Boewley coefficient of skewness that is bounded between -0.10 and 0.17, indicating that the logarithmic transformation is successful in restoring symmetry.

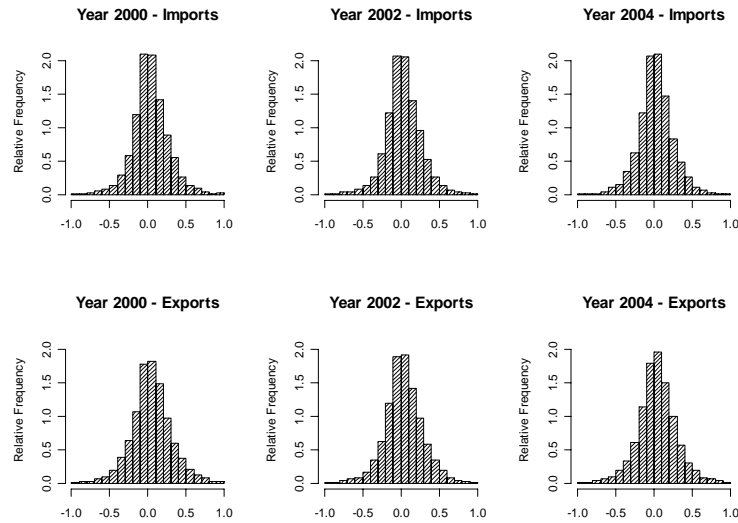


Figure 3. Distribution of the Bowles skewness of unit values among different commodities, after applying logarithmic transformation

2.3. Identification of outliers

- Data points that seem to be inconsistent with the general characteristics of the sample are called **outliers**. These are values “that lie far from the middle of the distribution in either direction.”
- Outliers may arise for several reasons:
 1. **Errors** in data entry or processing.
 2. **Atypical circumstances** in the data generating process
 3. **Intrinsic variability** of the data generating process.
- Methods of outlier detection are useful for both conducting **data quality checks** and understanding the **reliability** and **intrinsic characteristics** of the data generating process.
- The method for outlier detection adopted in this report is based on the idea that most values are expected in the **inter-quartile range**, which is the interval between Q_1 and Q_3 .
- On the log-transformed sample, the left and right **thresholds for anomalous values** are determined by adding to or subtracting from Q_1 or Q_3 , respectively, a symmetric step equal to one and a half times the inter-quartile range.
- Using this criterion, about 4.7% of the observations in the unit value samples were diagnosed as outliers and disregarded from further calculations to obtain Standard Unit Values.

2.4. Assessing multimodality

- Determining a single Standard Unit Value from for the all transactions classified under a single commodity/flow/year is problematic if the data sample comes from various **heterogeneous** subpopulations.
- This form of heterogeneity is frequently reflected in the **presence of multiple modes** in the sample of individual unit values used to calculate a Standard Unit Value. Ideally, Standard Unit Values should be calculated from uni-modal samples.
- To assess the degree of multimodality in the samples of unit values available for each commodity/flow/year, the following **multimodality index** based on the histogram of the log-transformed data is proposed²:

$$\text{Multimodality index} = \frac{(m_1 + \dots + m_k)^2}{m_1^2 + \dots + m_k^2},$$

where k is the number of modes in the sample histogram and m_j is the mass weight attached to its j th mode (i.e., the number of data points falling in j th mode's cell, divided by the total number of individual unit values used to construct the histogram). If there is only one mode (i.e., if $k = 1$), the multimodality index takes the value of one; if there are two equally relevant modes (i.e., if $k = 2$, with $m_1 = m_2$), the index is equal to two; etc.

2.5. Some specific examples

- Some specific examples are given after the next section. They refer to export unit values in 2004 of several commodities. They provide an overview of the main features typically encountered in unit value data.
- In each table, the outlier detection criteria discussed above is applied to the sample of unit values for the corresponding commodity. Measures of location, spread, skewness, and multimodality are also presented, both before and after removing outliers.
- The left plot under each table contains the **histograms of the unit value data before removing outliers** (in logarithmic scales), as well as a **box-plot** indicating:
 1. The location of the **inter-quartile range** (the length of the "box")
 2. The location of the **median** (the bold vertical line dividing the box in two parts)
 3. The location of the **acceptance thresholds** that are used to detect outliers (represented by the extremes of the "whiskers").
- The plot to the right shows the **histograms of the unit value data after removing outliers** (in logarithmic scale).

² In defining the multimodality index, the histogram of the log-transformed data is constructed by assigning each data point to one of **ten equally-spaced cells** on the log-transformed scale.

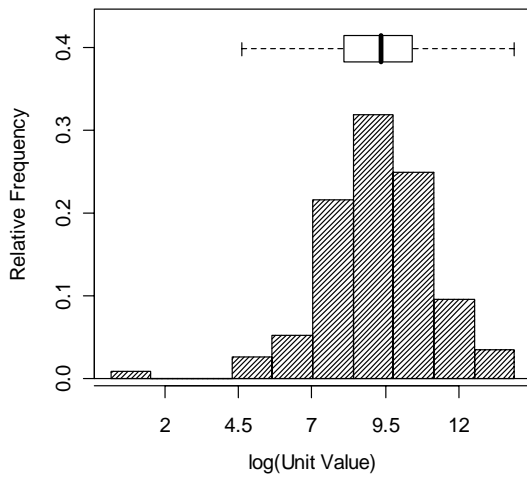
3. Standard Unit Values

- The Standard Unit Value (SUV) of a specific commodity/flow/year is defined as the **median unit value (after removing outliers)**.
- However, this is considered to be reliable for estimation purposes if and only if the sample of unit values on which it is based fulfills the following **reliability criteria**:
 1. The data must come from more than two reporting countries/regions.
 2. There must be at least 30 observations in the sample
 3. The relative standard deviation must be less than or equal to 1.75, or
 4. The relative standard deviation is between 1.75 and 3, provided that its multimodality index is less than 2
 5. The trade value corresponding to outliers must be less than 10% of the total trade value.
- The resulting Standard Unit Values for different classifications are available in the table views SuvH0, SuvH1, and SuvH2 of the StandardUnitValues data base of the UNSD.
- The SQLscripts used to generate Standard Unit Values are divided in several modules:
 - ❖ Module 0.1 - Original Classification used by each Country.sql
 - ❖ Module 0.2 - Correct Original Classification Used by Country.sql
 - ❖ Module1 - Commodity Catalog v2.sql
 - ❖ Module2 - Create Functions and Output Tables v1.sql
 - ❖ Module3 - SUV calculator v3.sql
 - ❖ Module4 – MultimodalityCalculator v1.sql
 - ❖ Module5 - SUV statistics overview.sql
 - ❖ Module6 - SUV statistics overview by HS Classification.sql
 - ❖ Module7 - DescriptiveStatisticsViews.sql

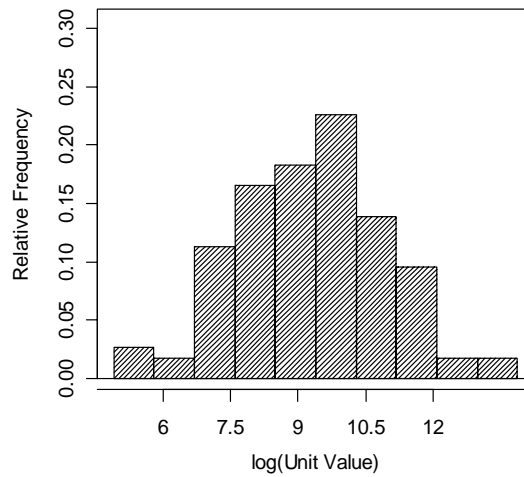
**HS2 010110 - Live horses/asses/mules/hinnies: pure-bred breeding animals
(Quantity unit: 5)**

Number of observations:	116	
Total quantity:	245,279	
Total value:	704,651,457	
Number of left outliers:	1	
Total quantity:	218,927	
Total value:	261,771	
Number of right outliers:	0	
Total quantity:		
Total value:		
Detection of outliers		
Left threshold:	100.23	
Right threshold:	1,070,785.29	
Descriptive statistics	Before removing outliers	After removing outliers
Min:	1.20	135.02
Q1:	3,249.17	3267.41
Median:	11,333.80	11418.23
Q3:	33,032.71	33138.91
Max:	1,057,588.00	1,057,588.00
Arithmetic mean:	43,607.90	43,987.09
Geometric mean:	10,086.37	10,911.26
Bowley measure of skewness		
<i>Original data:</i>	0.46	0.45
<i>Log-transformed data:</i>	-0.08	-0.08
Multimodality index:	1.05	1.23

Before removing outliers

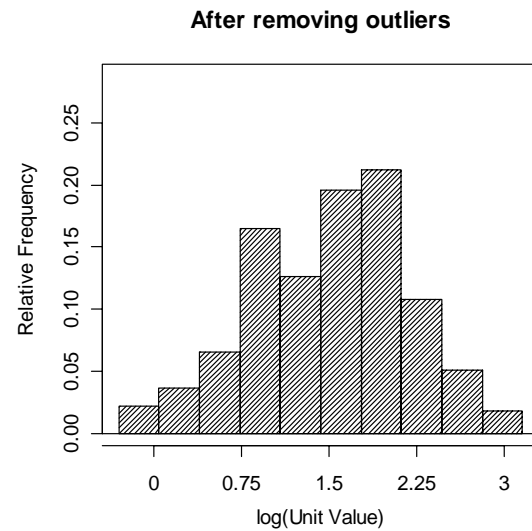
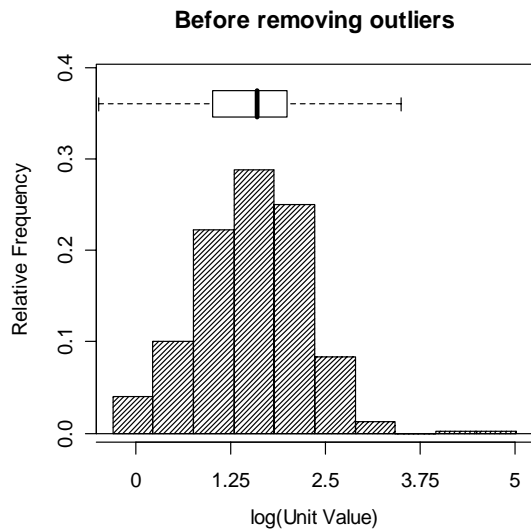


After removing outliers



**HS2 020130 - Meat of bovine animals, fresh/chilled, boneless
(Quantity unit: 8)**

Number of observations:	549	
Total quantity:	1,654,095,730	
Total value:	7,447,097,377	
Number of left outliers:	0	
Total quantity:		
Total value:		
Number of right outliers:	2	
Total quantity:	6,699	
Total value:	442,065	
Detection of outliers		
Left threshold:	0.62	
Right threshold:	32.99	
Descriptive statistics	Before removing outliers	After removing outliers
Min:	0.74	0.74
Q1:	2.75	2.75
Median:	4.98	4.93
Q3:	7.43	7.38
Max:	151.90	23.49
Arithmetic mean:	5.99	5.63
Geometric mean:	4.60	4.55
Bowley measure of skewness		
<i>Original data:</i>	0.04	0.06
<i>Log-transformed data:</i>	-0.19	-0.18
Multimodality index:	1.01	1.97

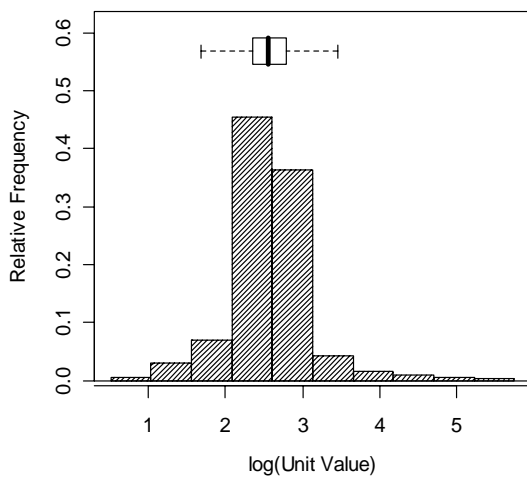


HS2 030541 - Pacific salmon /Atlantic salmon / Danube salmon [see list of conventions for s ...

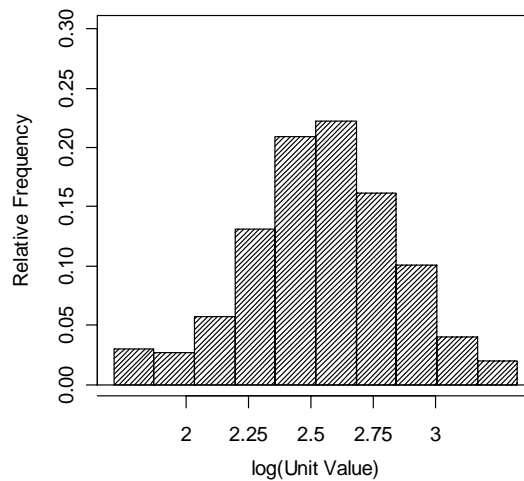
(Quantity unit: 8)

Number of observations:	328	
Total quantity:	40,093,899	
Total value:	485,711,880	
Number of left outliers:	14	
Total quantity:	920,284	
Total value:	3,846,846	
Number of right outliers:	17	
Total quantity:	61,846	
Total value:	3,020,481	
Detection of outliers		
Left threshold:	5.40	
Right threshold:	31.57	
Descriptive statistics	Before removing outliers	After removing outliers
Min:	1.67	5.53
Q1:	10.47	10.68
Median:	12.93	12.91
Q3:	16.28	15.76
Max:	315.53	27.92
Arithmetic mean:	16.24	13.46
Geometric mean:	13.12	12.85
Bowley measure of skewness		
<i>Original data:</i>	0.16	0.12
<i>Log-transformed data:</i>	0.05	0.03
Multimodality index:	1.00	1.27

Before removing outliers



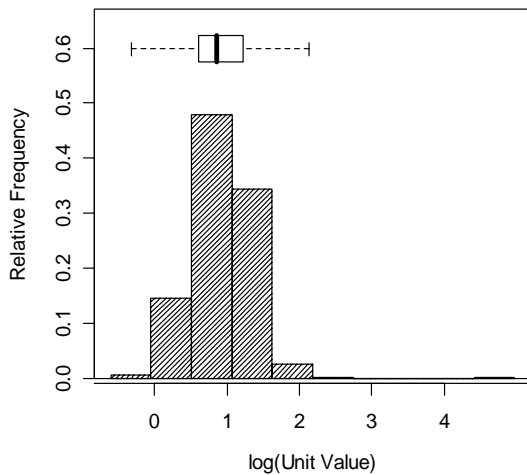
After removing outliers



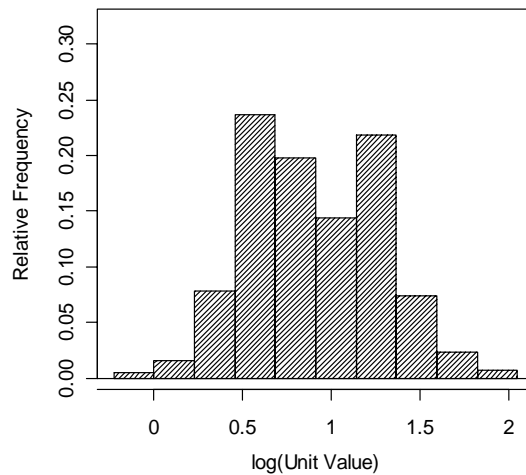
HS2 040510 - Butter
(Quantity unit: 8)

Number of observations:	1,028	
Total quantity:	1,093,578,404	
Total value:	3,013,485,587	
Number of left outliers:	3	
Total quantity:	239,628	
Total value:	143,057	
Number of right outliers:	2	
Total quantity:	3,175	
Total value:	53,801	
Detection of outliers		
Left threshold:	0.73	
Right threshold:	8.44	
Descriptive statistics	Before removing outliers	After removing outliers
Min:	0.55	0.80
Q1:	1.82	1.83
Median:	2.38	2.38
Q3:	3.37	3.37
Max:	143.92	7.77
Arithmetic mean:	2.80	2.66
Geometric mean:	2.48	2.48
Bowley measure of skewness		
<i>Original data:</i>	0.28	0.29
<i>Log-transformed data:</i>	0.14	0.14
Multimodality index:	1.00	2.00

Before removing outliers



After removing outliers



HS2 070410 - Cauliflowers & headed broccoli, fresh/chilled
(Quantity unit: 8)

Number of observations:	243	
Total quantity:	952,325,382	
Total value:	632,506,089	
Number of left outliers:	16	
Total quantity:	205,381,754	
Total value:	37,441,669	
Number of right outliers:	4	
Total quantity:	1,281,087	
Total value:	4,165,206	
Detection of outliers		
Left threshold:	0.24	
Right threshold:	2.87	
Descriptive statistics	Before removing outliers	After removing outliers
Min:	0.02	0.25
Q1:	0.61	0.65
Median:	0.80	0.82
Q3:	1.13	1.14
Max:	5.88	2.79
Arithmetic mean:	0.95	0.94
Geometric mean:	0.77	0.85
Bowley measure of skewness		
<i>Original data:</i>	0.26	0.28
<i>Log-transformed data:</i>	0.11	0.14
Multimodality index:	1.04	1.21

