Journal of Official Statistics, Vol. 13, No. 4, 1997, pp. 367–383

Developing an Estimation Strategy for a Pesticide Data Program

Phillip S. Kott and D. Andrew Carr[1]

The Agricultural Marketing Service's Pesticide Data Program (PDP) is a cooperative effort of U.S. Department of Agriculture (USDA) and several state agencies. The ultimate purpose of the program is to make scientific statements about the distribution of certain pesticide residues in particular products (mostly fresh fruits and vegetables) consumed by the U.S. public. Developing a statistically defensible estimation strategy for the PDP required overcoming a number of thorny problems. Chief among them was the non-random nature of the "sample" of participating states. Also of concern was the level-of-detection/level-of-quantification issue: not all potential levels of pesticide residue can be detected by a given lab; moreover, certain detectable levels are not quantifiable. A graphical method was developed to display parameter estimates (means and percentiles) in light of the detection/quantification problem. Included on the graphs (as an option) are fairly robust, model-based estimates of confidence intervals.

Key words: Target population; inferential population; level of detection; level of quantification; percentile; distribution.

1. Introduction

The Agricultural Marketing Service's Pesticide Data Program (PDP) has been a cooperative effort of U.S. Department of Agriculture (USDA) and several state agencies. This article deals primarily with the PDP in 1993, although the 1992 survey is discussed briefly and plans for subsequent years are also addressed.

The ultimate purpose of the 1993 PDP was to make scientific statements about the distribution over the year of certain pesticide residues, in particular fresh fruits and vegetables, consumed by the U.S. public: broccoli, carrots, celery, green beans, lettuce, potatoes, oranges, grapefruit, apples, bananas, grapes, and peaches. These twelve commodities were chosen because of their high consumption levels in the U.S., especially among children.

For each type of fresh produce, say apples, two distinctions need to be drawn. The first is between the *sample* of apples analyzed by the PDP and the *target* population about which assumption-free statistical statements can be made. The target population for the 1993 PDP was all apples distributed to supermarkets, independent grocers, hotels, restaurants, institutions, and (on rare occasion) final consumers through listed wholesalers in participating states in 1993. The second distinction is between this target population and the *inferential* population of the apples actually consumed by the U.S. public in 1993.

[1] Research Division, National Agricultural Statistics Service, 3251 Old Lee Highway, Room 305, Fairfax, VA 22030, U.S.A.

We follow Groves (1989, p. 82) in our use of the modifiers ''target'' and ''inferential.'' Groves also draws a distinction between the *frame population* and the target population. Following up on the apple example, the frame population would be restricted to apples originating from the PDP's incomplete list of wholesalers, while the target population might include apples provided by a putative list of all U.S. wholesalers. Little is gained by focusing on this distinction *per se*. Consequently, we have simply defined the target population to coincide with the frame population here.

The sampling and estimation strategy discussed respectively in Sections 2 and 3 are directed at the PDP's target population. As a result of this strategy, rigorous statistical statements can be made concerning the distribution of pesticide residues within the target population using 1993 PDP survey data. It is considerably more problematic to make precise statistical statements about the distribution of pesticides within the inferential population *per se*. This is not to say that there exists significant differences between the target and inferential populations. On the contrary, the USDA suspects – but cannot prove – that such differences will be so minor as to be completely captured in the (estimated) standard errors of the estimates. We turn to the issue of estimates for the inferential population in Section 4.

Before leaving the question of the population of interest, it is important to understand that the population unit under discussion in this article is *not* a single discrete element of the product in question (e.g., a particular apple). Rather, it is a homogenized five pound ''sample'' of the product. This quantity of a commodity was required for the necessary laboratory work.

Another significant area of concern in estimating pesticide residues is the level-of-detection/level-of-quantification issue: not all potential levels of pesticide residue can be detected by a given lab; moreover, certain detectable levels are not quantifiable. A graphical method is discussed in Section 5 to display parameter estimates (means and percentiles) in light of the detection/quantification problem. Section 6 contains a brief conclusion.

2. Sampling

2.1. *States in the PDP*

The most obvious divergence between the PDP target and inferential populations is the number of states not covered by the PDP. Had the USDA chosen participating states at random, this divergence would disappear. Such an approach, while statistically valid, was not practical for operational reasons.

There were six states in the 1992 PDP – New York, Michigan, Florida, Texas, California, and Washington. For operational reasons, they had to be retained in the 1993 program. The USDA could have, *in principle*, increased its target population to all wholesale distributions in the U.S. by randomly adding *new* states to the PDP when it expanded from six to nine states for 1993. It would have been a mistake to do so, however, because the resulting U.S. level estimators would have a disappointingly low level of accuracy. Statistical modeling, despite its inherent dependence on underlying assumptions, was the most practical method available for linking the states in the PDP with the U.S. as a whole and producing reasonably accurate U.S. level estimates.

Since the three states added to the PDP were to be selected purposefully rather than randomly, it was reasonable to add them in areas under-represented by the 1992 program. Our choices were heavily influenced by the U.S. Census Bureau's separation of the country into four geographic regions and nine divisions.

The Mountain division (MT, ID, WY, CO, NM, AZ, UT, NV) had no representative in 1992. In that division, Colorado has by far the largest population of any state not bordering a PDP state. Thus, it was added to the program in 1993.

The second addition was Ohio, a large East North Central state bordering three other Census divisions. The state office in Michigan did not have the laboratory capacity to adequately represent this densely populated division (OH, MI, IN, IL, WI) by itself in 1993.

Unlike Michigan, New York could comfortably represent not only the Middle Atlantic division (NY, NJ, PA) but New England as well. The USDA does not believe there are meaningful differences between the wholesale produce in New York and in other states in the Northeast region.

In contrast to New York, Florida was not believed to be that good a representative of the large number of states in the South Atlantic division (DE, MD, DC, VA, WV, NC, SC, GA, FL). Accordingly, North Carolina was added to the program for 1993.

The PDP contains no West North Central state (MN, IA, MO, ND, SD, NE, KS). This is because three of these (MN, IA, MO) are typical Midwestern states, which will be effectively represented by Michigan and Ohio. Three (ND, SD, NE) are sparsely populated states, which are demographically similar to Mountain states. The last (Kansas) borders Colorado.

There is also no East South Central state because all four states in this division (KY, TN, AL, MS) have modest populations, while the South region as a whole, which also includes the South Atlantic and West South Central (AR, LA, OK, TX) divisions, have three participants in the 1993 PDP.

2.2. *Site selection*

The primary sampling units (PSU's) for the 1993 PDP were sites, either chain store distribution centers or terminal markets. Sites were selected each month independently across the states in the program. Probability proportional to size (pps) sampling was used for the most part in 1993. Ideally, pps site selection gives each pound of product, rather than each site, an equal chance of being in the PDP sample. This is an efficient sampling design – that is, it produces estimates with the low mean squared errors – when every pound of a commodity has the same potential distribution of pesticide residues.

It is not possible to predict in advance how much product is at every site on a state list. Nevertheless, past experience served as a guide in developing size measures for pps site selection that reflect present conditions reasonably well.

Each state in the 1993 PDP was responsible for developing the list of sites and size measures used for pps site selection. Some states used the same measures of size for all commodities. Others have access to and employ separate measures for different commodities or pairs of commodities (see Kott 1994a).

Sites were often chosen for a pair of commodities at a time. Since each pesticide/commodity pair was to be analyzed independently, this practice did not bias the results.

Most states used the same measures of size for every month in a quarter, but some modified the size measures from month to month. The use of less than perfect size measures does not systematically bias PDP estimates. Their use merely renders the estimates less accurate, but still much more accurate than had equal probability site selection been used.

It was decided that no site no matter how large would be selected twice in the same month for the same commodity. When a site had a measure of size that would allow this possibility under a strict pps sampling regime, it was selected with certainty and pps principles applied to the remaining sites on the state list.

2.3. *Other sampling issues*

Once a site was selected in 1993 for a particular commodity in a particular month, a week was selected at random, and the site was visited during that week. In order not to bias the selection process, a pallet of the commodity was selected using simple random sampling. This essentially gave each pound of product regardless of variety (e.g., a Fuji apple) or origin (e.g., New Zealand) an equal chance of selection.

Once a pallet was selected, a predetermined quantity of the commodity was chosen at the convenience of the collector from the same lot as the sampled pallet. A lot is a fairly homogenous grouping containing product of the same variety and generally from the same place of origin. The placement of product within pallets and pallets within lots by the wholesaler is an arbitrary process. There is no reason to suspect pesticide residues have anything but random variation within a lot. On the other hand, the randomized selection of lots described above assures us that the distributions of product varieties and origins within the sample match those of the population – plus or minus sampling error.

In addition to collecting a random sample of the commodity in question, sites were asked to provide the sales or distributions of the commodity to supermarkets, independent grocers, hotels, restaurants, institutions, and (on rare occasion) final consumers over the previous seven days. It was felt that this time interval was short enough to be reasonably represented by the sampled product. Since sales and distributions can vary considerably from day to day, a shorter time span for quantity information would have led to estimates with larger mean squared error. Although there is a theoretical potential for bias due to pesticide residues in a commodity being systematically different on the day of collection and the previous seven days, this potential appears small enough to ignore entirely.

Of more concern was a moratorium on data collection that occurred in August 1993 in most states and in July in California. This moratorium allowed labs to catch up with backlogs in their workloads. How the absence of data from a particular month was handled in estimation is discussed in the next sections.

3. Estimation

3.1. *Population parameters*

The 1993 PDP survey can be used to estimate the distribution of a particular pesticide, say Thiabendazole, in the population of a particular commodity, say apples, at the wholesale level in participating states. Each pesticide/commodity pair can be viewed independently for this purpose.

Since the goal is the estimation of the distribution of pesticide residues in a commodity population, every sample of that commodity taken in 1993 should have been tested for the pesticide in question. On occasion, the PDP finds a pesticide residue in a commodity that is not generally tested for that pesticide. Such findings, while of some interest to scientists, have no value when estimating population values.

The 1993 PDP employed a number of chemical laboratories to evaluate sampled commodities. These labs have varying levels of detection and quantification. A *level of detection* (LOD) for a pesticide/commodity pair is the lowest concentration (in parts per billion) of pesticide residue detectible in a commodity. A *level of quantification* (LOQ) is the lowest concentration of pesticide residue in a commodity that can be given a particular numerical value. For example, the LOD for a particular pesticide residue may be 30 parts per billion (ppb), while the LOQ is 100 ppb. As a result, when the residue concentration in a test sample is below 30 ppb, the lab will not be able to detect it. Furthermore, when the concentration is between 30 and 100 ppb, the lab will be able to detect the presence of the residue, but not determine the exact concentration. For all the labs in the PDP, the LOD for a PDP pesticide residue was approximately a fixed percentage of the LOQ.

The LOD and LOQ for a pesticide varied across the PDP labs. This rendered the simple estimation of the fraction of commodity population A with detectible residues of pesticide B dubious. One parameter that appears estimatable from 1993 PDP survey data is the proportion of the target population with pesticide concentrations at or above the highest lab LOD. Unfortunately, this assumes each lab's LOD for a particular pesticide was a constant in 1993, which was not always the case. Nevertheless, for estimation purposes in Section 3.2. we will treat each lab's average LOD for a pesticide as if it were fixed and constant.

In the 1994 PDP survey data base (and the 1995 and 1996 data bases as well), the LOD for each lab test has been recorded, and so the constant lab LOD assumption is unnecessary. An additional assumption we make here also applies to the analysis of 1994 PDP data. Namely, we assume that 1993 laboratory results are free of appreciable measurement errors. Such errors, even when free of systematic (i.e., directional) biases, can have a biasing effect on distributional estimates.

The pth *percentile* of the distribution of pesticide residues in a commodity population is said to be Q_p when at least p per cent of the commodity population has pesticide residues less than or equal to Q_p while at least 100-p per cent has pesticide residues greater than or equal to Q_p.

It is common for a sample to be used to estimate the mean of a distribution. The existence of nonzero LOD's made it difficult to estimate the mean value of pesticide residuals within a commodity's 1993 target population in a statistically reasonable manner. An alternative approach was adopted in which a range of values was estimated for the mean. The lower bound of this range occurs when every pesticide concentration observed below the analyzing lab's average LOD was set at 0; the upper bound occurs when every such observation is set at the lab's average LOD. We will discuss this further in Section 3.3.

3.2. Estimating PAHLOD's

One parameter that can be estimated with 1993 PDP data is the proportion of a commodity's target population at or above a particular concentration of pesticide residue – the highest lab (average) LOD. We will name this fraction the ‘‘PAHLOD’’ for ‘‘*p*roportion at or *a*bove the *h*ighest *l*evel *o*f *d*etection’’ and denote it by *P*.

There is a difficulty in determining the PAHLOD (in addition to the variable lab LOD problem) that can best be demonstrated with an example. The highest lab LOD for Thiabendazole is 180 parts per billion (ppb). This is for the New York and Texas labs. The LOD for another lab is 76 ppb, but its LOQ is 250 ppb. Thus, the presence of Thiabendazole can be detected in an apple sample from this second lab, but the USDA may be unable to determine whether the concentration is above or below the 180 ppb *PAHLOD threshold.* We discuss a method of handling this problem later in the section.

Consider the ideal situation in which all LOD’s are equal and no data is missing. Let S denote the 1993 PDP sample of a commodity and k denote a particular observation from the sample. A nearly unbiased estimate of P, the fraction of (say) apples in the target population with Thiabendazole residues at or above the LOD would be

$$p = \frac{\sum_{k \varepsilon S} \pi_k{}^{-1} V_k x_k}{\sum_{k \varepsilon S} \pi_k{}^{-1} V_k} \tag{1}$$

where Π_k is the probability of selecting the site from which k was selected, V_k is the weekly volume of the site, and x_k is 1 if sample k has a concentration at or above the PAHLOD threshold, 0 otherwise (we are ignoring the fact that the number of days in a month varies).

Let us call $A_k = \pi_k^{-1} V_k$ the *analysis weight* of observation k, and $a_k = A_k/(\Sigma_{j \varepsilon S} A_j)$ the *normalized* analysis weight. In effect, observation k represents $100{*}a_k$ per cent of the target population.

The nature of the a_k suggests the following procedure for handling an observation j measured at a lab with an LOD below and an LOQ above the PAHLOD threshold: Treat j as a set of m observations with concentrations uniformly distributed from the LOD to the LOQ and with a common normalized analysis weight a_j/m. This is not a new idea. For example, Wendelberger (1995, p. 39) discusses this approach in a different context (handling observations below the LOD when the LOD and LOQ coincide).

Unfortunately, the 1993 PDP did not record the LOD for detections below the LOQ. The LOQ, by contrast, was recorded. Since in all labs but one the LOD is 30 per cent of the LOQ, an observation j in this range was broken into four observations, each with normalized analysis weight $a_j/4$. One would be given a pesticide residue concentration at 35 per cent of the LOQ, one at 55 per cent, one at 75 per cent, and one at 95 per cent (an analysis of 1992 data suggests that using eight rather than four breaks has little effect on the results). The last lab has LOD’s that are consistently 50 per cent of the LOQ. Observations measured at this lab were broken up into three observations with concentrations at 55 per cent, 75 per cent and 95 per cent of the LOQ, respectively (and weights of $a_j/3$).

The approach to the varying LOD problem described above is based on the heroic assumption that the distribution of pesticide residue concentrations between the LOD and LOQ of a particular lab is uniform. Although this is a questionable assumption, the USDA’s methodology does have the virtue of being relatively simple. Moreover, it is clearly superior to the often-used practice of imputing a value for the concentration. In the 1992 PDP data base, for example, a value of 1/2 the LOQ has been recorded when an observation had a pesticide residue concentration between the lab's LOD and LOQ.

From now on, we will refer to the *augmented sample* S^* for a pesticide/commodity pair. S^* contains four (or three) observations in place of the original observation in S when the original has a residue concentration between its lab's LOD and LOQ.

One last adjustment was made to the weights in the augmented sample to compensate for the August (July in California) data collection moratorium. The analysis weights, the A_k, for observations in July and September (June and August in California) were multiplied by 1.5. This methodology allows for the possibility that pesticide residues may be seasonal. Nevertheless, if the residue distribution of a particular pesticide in a given commodity is different in August (July) from that in July and September (June and August), the estimated PAHLOD for the pesticide/commodity pair will likely be biased. Analysis of 1992 data suggested little to fear from this.

Table 1 contains estimated PAHLOD's for 17 commodity/pesticide pairs based on the 1993 PDP survey. All pairs had PAHLOD's of at least 15 per cent. Section 4 will describe how the confidence intervals for the PAHLOD's were derived. It is important to remember that the estimated PAHLOD's are for five-pound samples of the product, say apples, rather than for individual apples.

3.3. Estimating means

If all lab LOD's were zero, then a nearly unbiased estimate of M, the mean (average) pesticide residue concentration in a commodity's target population would be

$$m = \frac{\sum\limits_{k \varepsilon S^*} a_k y_k}{\sum\limits_{k \varepsilon S^*} a_k} \tag{2}$$

where y_k is the concentration of observation k (assuming, of course, that our assumption about the distribution of observations below the LOQ is correct).

As noted in Section 3.1., m cannot be calculated in practice because we do not have y_k values for observations with residues below the analyzing lab's LOD. If, however, we set all such y_k values to zero, then Equation (2) can be used to determine the value $m_{\min}$, which estimates a lower bound for M. Similarly, if we set all such y_k values to the analyzing lab's LOD, then Equation (2) can be used to determine $m_{\max}$, which estimates an upper bound for M. Table 1 contains $m_{\min}$ and $m_{\max}$ estimates for 17 commodity pairs. Section 4 will describe how 95 per cent confidence intervals were derived for this table.

The approach taken here makes no assumptions about the real values for observations less than the LOD. This frees the analyst to make his or her own assumption. Wendelberger (1995) provides a review of alternative censored-data techniques.

Unlike the situation with proportions, estimated means apply to both five-pound samples of, say, apples, and to individual apples.

3.4. Estimating percentiles

Let y_k be the concentration for observation k, where concentrations measured below a

Table 1. Estimated residue means and proportion by pesticide/commodity

Pesticide/ commodity	Highest lab level of detection HLOD (ppm)	Proportion at or above the HLOD[1]	Mean	
			Minimum estimate[2] (ppm)	Maximum estimate[3] (ppm)
APPLES				
Benomyl	0.05	19.0 (12.9–26.8)	0.040 (0.011–0.070)	0.081 (0.051–0.11)
Diphenylamine	0.13	46.3 (37.0–56.0)	0.54 (0.38–0.70)	0.55 (0.39–0.71)
Thiabendazole	0.18	52.9 (44.8–60.7)	0.44 (0.36–0.52)	0.48 (0.40–0.55)
CARROTS				
DDE	0.014	19.7 (15.9–24.6)	0.0092 (0.0068–0.0115)	0.014 (0.012–0.016)
CELERY				
Dicloran	0.035	16.6 (12.1–22.1)	0.051 (0.019–0.083)	0.062 (0.030–0.095)
GRAPEFRUIT				
Thiabendazole	0.18	15.1 (9.8–22.4)	0.084 (0.053–0.114)	0.12 (0.09–0.15)
GRAPES				
Iprodione	0.088	16.7 (13.3–20.9)	0.065 (0.031–0.099)	0.087 (0.053–0.120)
GREEN BEANS				
Acephate	0.07	15.1 (7.6–27.7)	0.065 (0.007–0.12)	0.082 (0.024–0.140)
Endosulfans	0.088	15.4 (11.0–21.3)	0.044 (0.024–0.064)	0.058 (0.038–0.078)
Methamidophos	0.028	15.3 (7.8–27.6)	0.018 (0.003–0.033)	0.026 (0.011–0.041)
ORANGES				
Thiabendazole	0.18	22.8 (14.6–33.9)	0.11 (0.07–0.16)	0.15 (0.10–0.19)
PEACHES				
Benomyl	0.05	18.3 (14.7–22.7)	0.038 (0.005–0.070)	0.078 (0.046–0.111)
Dicloran	0.035	44.9 (39.4–50.6)	0.66 (0.28–1.04)	0.67 (0.29–1.05)
Iprodione	0.088	57.2 (51.4–63.0)	0.51 (0.29–0.73)	0.52 (0.30–0.74)
Parathion-Methyl	0.028	28.3 (21.5–36.2)	0.026 (0.013–0.039)	0.031 (0.018–0.043)
POTATOES				
Chloropropham	0.12	45.2 (40.2–50.1)	0.74 (0.53–0.96)	0.76 (0.54–0.97)

Note: 95% confidence intervals for estimated means and proportions are shown in parentheses.

[1] The estimated percentage of the commodity population with residue concentrations at or above the HLOD listed in the previous column.

[2] The estimated mean residue when a concentration below the individual lab level of detection is valued at zero.

[3] The estimated mean residue when a concentration below the individual lab level of detection is valued at that level of detection.

Table 2. Estimated residue distributions by commodity/pesticide pair

Commodity/ pesticide	Population percentiles			
	75th (ppm)	80th (ppm)	85th (ppm)	90th (ppm)
APPLES				
Benomyl	*	*	0.075 (0–0.14)	0.13 (0.08–0.21)
Diphenylamine	0.83 (0.61–1.10)	0.98 (0.75–1.40)	1.2 (0.9–1.9)	1.8 (1.2–2.5)
Thiabendazole	0.62 (0.49–0.78)	0.76 (0.59–1.00)	0.91 (0.72–1.14)	1.1 (0.9–1.5)
CARROTS				
DDE	*	*	0.0195 (0.0143–0.0220)	0.0247 (0.0209–0.0370)
CELERY				
Dicloran	*	*	0.055 (0.010–0.130)	0.13 (0.06–0.23)
GRAPEFRUIT				
Thiabendazole	*	*	0.18 (0.14–0.24)	0.23 (0.17–0.31)
GRAPES				
Iprodione	*	*	0.10 (0.07–0.14)	0.18 (0.12–0.24)
GREEN BEANS				
Acephate	*	*	0.073 (0–0.440)	0.18 (0.02–0.94)
Endosulfans	*	*	0.097 (0.044–0.140)	0.14 (0.11–0.33)
Methamidophos	*	*	0.035 (0–0.110)	0.067 (0.010–0.230)
ORANGES				
Thiabendazole	*	0.19 (0.14–0.30)	0.24 (0.17–0.38)	0.31 (0.23–0.46)
PEACHES				
Benomyl	*	*	0.095 (0–0.140)	0.14 (0.10–0.22)
Dicloran	0.36 (0.19–0.52)	0.6 (0.4–1.1)	1.3 (0.6–2.1)	2.3 (1.4–3.6)
Iprodione	0.54 (0.46–0.66)	0.66 (0.56–0.78)	0.79 (0.68–0.95)	1.1 (0.9–1.4)
Parathion-Methyl	0.029 (0.018–0.037)	0.037 (0.028–0.05)	0.046 (0.036–0.090)	0.074 (0.046–0.14)
POTATOES				
Chloropropham	0.93 (0.66–1.30)	1.4 (1.0–1.7)	1.9 (1.5–2.2)	2.5 (2.2–2.9)

Note: 95% confidence intervals for estimated percentiles are shown in parentheses.
*Denotes that the value falls below the highest lab level of detection.

lab's LOD are set to 0, and concentration between an LOD and LOQ are handled as discussed in the previous subsection. The estimated pth percentile is

$$y_{(p)} = (y_{(pL)} + y_{(pU)})/2 \tag{3}$$

where $y_{(pL)}$ and $y_{(pU)}$ are determined by the conditions

$$\sum_{y_k \le y_{(pL)}} a_k \ge p/100 \text{ and } \sum_{y_k \ge y_{(pL)}} a_k > 1 - (p/100)$$

$$\sum_{y_k \le y_{(pU)}} a_k > p/100 \text{ and } \sum_{y_k \ge y_{(pU)}} a_k \ge 1 - (p/100)$$

and the summations in the above equations are over subsets of the augmented sample S^* (note: p in Equation (3) need not be the estimated PAHLOD from Equation (1)).

Percentiles can be computed with the statistical package SAS (SAS Institute, Inc. 1985) using PROC UNIVARIATE. The original analysis weights (the A_k) can be captured in the FREQ statement (the FREQ statement truncates input values to whole numbers, but analysis weights tend to be very large).

Table 2 displays selected percentile estimates for the 17 commodity/pesticide pairs in Table 1. It should be noted that when an estimated percentile falls below the PAHLOD threshold, it becomes biased downward because of the possibility that some samples have positive residues below the testing lab's LOD. Such estimates were excluded from the table. Observe that only two commodity/pesticide pairs had PAHLOD's above 0.5. As a result, we decided not to include medians in the table.

Section 5 will describe how the 95 per cent confidence intervals were derived for this table. It should be remembered that the percentiles in Table 1 have been estimated for five-pound samples of apples, not for particular apples.

3.5. Estimating variances

The augmented sample was partitioned into eleven "clusters" in the following manner. New York and California were broken into four clusters. New York City and Long Island formed one cluster, while the remainder of New York made up a second. The Anaheim District (which covers Southern California) was the third cluster and the rest of California the fourth. Each of the other respective states constituted a single cluster.

Let $S(i)^*$ denote the subset of the sample in cluster i. It is possible to rewrite the estimated PAHLOD in Equation (1) as

$$p = \frac{\sum_i a_{i+} p_{i+}}{\sum_i a_{i+}} \tag{4}$$

where $a_{i+} = \sum_{S(i)^*} a_k$, and $p_{i+} = \sum_{S(i)^*} a_k x_k / \sum_{S(i)^*} a_k$.

The difference between the estimator p and the target value P is

$$p - P = \sum_i a_{i+} d_{i+}$$

where $d_{i+} = p_{i+} - P$ (recall that $\sum_i a_{i+} = \sum_S a_k = 1$). Observe that the d_{i+} are all independent random variables. They have the same mean only when the state PAHLOD's are all equal. Thus, the following is likely to be, if anything, an upwardly biased estimator of the variance of p:

$$var(p) = (11/10)\left(\sum_i a_{i+}^2 e_{i+}^2\right) \tag{5}$$

where $e_{i+} = p_{i+} - p$. The factor 11/10 is an *ad hoc* adjustment for the downward bias in e_{i+}^2 as an estimator for d_{i+}^2. Equation (5) is very similar to the so-called ''linearization variance estimator'' in Fuller (1975, p. 123; Equation (7) with $L = 1$).

It is a simple matter to adapt the variance formula in Equation (5) to $m_{\min}$ and $m_{\max}$. One need only replace each x_k by y_k, where y_k is defined in Section 3.3.

3.6. Missing data

Missing data in the 1993 PDP data set can take three forms. A site chosen for particular commodity may have had none of it in stock on the day the site was visited. Alternatively, the site may have had the commodity in stock but chose not to allow the state agent to collect a sample of it. Finally, the site may not have been willing or able to provide weekly volume data on the commodity.

Let us consider these in reverse order. When a site failed to report weekly volume information, a value was imputed for it based on its measure of size and other available data (from the same month and, where reasonable, state). When a site had the commodity in question but did not allow any of it to be sampled, observations on the commodity from the same state and month was reweighted to compensate for the loss. Finally, when a site did not have any of the commodity of interest in stock, the ideal solution from a model-free statistical point of view would be to simply give the site an analysis weight of zero and leave it at that. This solution, however, does not make efficient use of expensive laboratory resources. For that reason, a protocol was adopted to select replacement sites in many states. Replacement sites were given the average analysis weight in the state/month combination. Weights were then scaled so that the total of the analysis weights in the state/month equaled what it would have been had no replacement protocol been in effect.

4. The Inferential Population

There are many potential sources of bias in using the sample data from the 1993 PDP to draw inferences about the distribution of pesticide residues in a target commodity population of product distributed to supermarkets, independent grocers, hotels, restaurants, institutions, and (on rare occasions) final consumers through listed wholesalers in PDP participating states in 1993. Many of these were discussed in the text. The most troublesome potential sources of bias was the August (July) moratorium on data collection and the assumption needed to handle the varying lab LOD/LOQ problem.

Assuming that the various sources of potential biases have no appreciable affect on its estimates, there remains the problem of linking estimates for a PDP target population to a real inferential population of interest, namely, product consumed by the U.S. public in 1993. There are two obvious disconnects between a 1993 PDP target population and a real inferential population of interest. First, a 1993 PDP target population is made up of product at the wholesale level. Even if the entire population of a commodity consumed by the U.S. public in 1993 passed through a wholesaler, the pesticide residue concentrations measured in the program may be biased. This is because many pesticides break down chemically over time. Thus, the concentration of Thiabendazole in a given apple sample may be higher when it is at the wholesaler than when it is ultimately eaten by a U.S. consumer.

There is little we can do about this potential source of bias except to note that it is likely to be very small – the bulk of pesticide degradation occurs before a commodity reaches the wholesaler. Moreover, if anything, pesticide residue estimates for a 1993 target population of wholesalers slightly overestimates the residues in the real inferential population at the consumer level.

The second obvious disconnect between a 1993 PDP target population and a real inferential population of interest is the states not in the PDP. To deal with this problem, we note that the data in the PDP are collected at the wholesale not the farm level. Thus, the differences in pesticide residue concentrations across states may be smaller than one would naively speculate. Nevertheless, there may be differences, and these potential differences should be statistically modeled.

To model the potential effects of state differences on the accuracy of our estimates, we assumed a simple random-effects model. Recall that clusters were very similar to states (except that NY and CA were each divided into two clusters). In principle, each cluster i has its own fraction of product at or above the PAHLOD threshold for a pesticide/commodity pair, P_{i+}. We model the differences across clusters by assuming that P_{i+} is itself a random variable with mean P_1 – the PAHLOD for the entire inferential population. If, 1, this random-effects model is correct, and, 2, each p_{i+} is an unbiased estimator for its respective P_{i+}, then p in Equation (1) is an unbiased estimator of P_1. Although theoretically superior model-based estimators of P_1 may exist, p has the practical advantage of simultaneously estimating the inferential parameter P_1 and the target parameter P.

It is easy to show that Equation (5) can now serve as an estimate of the variance of p if the P_{i+} are uncorrelated across the clusters in the PDP.

A slightly better variance estimator under the model in terms of bias may be

$$var'(p) = var(p)\frac{(10/11)\left[\rho\sum a_{i+}^2 + (1-\rho)\sum \alpha_i\right]}{\rho\left(\sum a_{i+}^2 - 2\sum a_{i+}^3 + \left[\sum a_{i+}^2\right]^2\right) + (1-\rho)\left(\sum \alpha_i - 2\sum a_i\alpha_i + \sum a_i^2 \sum \alpha_i\right)} \qquad (6)$$

where $\alpha_i = \sum_{k\varepsilon S(i)} a_k^2$ and ρ is the "state effect." See Kott (1994b) for a discussion of the rationale behind Equation (6). Based on preliminary analysis ρ was set equal to 0.015, but it turned out that $var'(p)$ in Equation (6) was not very sensitive to the choice of ρ.

It is common practice to say the $p \pm 2\sqrt{var(p)}$ is a 95 per cent confidence interval for the true PAHLOD in the target population, P. This practice assumes that $var(p)$ was estimated with at least 60 degrees of freedom. As it happens, $var'(p)$ is based on at most ten degrees of freedom (the eleven clusters minus one). With this in mind, the degrees of freedom for $var'(p)$ were estimated using an alternative approach discussed in Kott (1994b). For the 16 commodity/pesticide pairs covered by Tables 1 and 2, these degrees of freedom estimates ranged from roughly 3.7 to 9.0.

An asymmetric 95 per cent confidence interval for P can be determined by including all values P^* for which

$$(p - P^*)^2 \leq t_{df}^2 var(P^*)$$

where t_{df} is the 97.5 per cent value for a Student's t distribution with df degrees of freedom, and

$$var(P^*) = var(p)(1 - P^*)P^*/[(1 - p)p] \qquad (7)$$

Equation (7) is based on the implicit assumption that the design effect is the same for every estimated proportion (the design effect of an estimate, q, of a proportion is defined to be $nVar(q)/q(1 - q)$, where n is the sample size and $Var(q)$ is the variance of q; the design effect of q under simple random sampling from a very large population is 1. We will not pursue the plausibility of this assumption further here. 95 per cent confidence intervals for estimated PAHLOD's calculated as described above are displayed in Table 1.

Let us rename the PAHLOD, P_0, and its estimate, p_0. The value P can now represent the proportion of the commodity population with residue concentrations at or below an arbitrary level. Its estimate is p. 95 per cent confidence intervals for every conceivable P can be determined by including all P^* for which $(p - P^*)^2 \leq t_{df}^2 var(P^*)$, where $var(P^*)$ is defined by Equation (7).

If the same model and assumptions hold for every concentration level at or above the PAHLOD threshold, then estimated percentiles for the target population also apply to the inferential population. Moreover, the determination of confidence intervals for the percentiles is straight-forward (see Section 5). Finally, as in Section 3.5 it is a trivial matter to adapt Equation (6) for $m_{\min}$ and $m_{\max}$. 95 per cent confidence intervals for these estimates are displayed in Table 1.

5. Displaying Estimates

Figure 1 displays several things simultaneously. The estimated per cent of the target population, apple ''samples,'' with Thiabendazole residues at or below a specific concentration level is marked with a large dot. For example, 70 per cent of the commodity population have Thiabendazole concentrations at or below (approximately) 0.5 parts per million (ppm). Stated another way, the cumulative distribution at 0.5 ppm is 70 per cent.

The Thiabendazole concentration at a specific estimated percentile is marked in the figure with either a large dot or a solid line. For example, both the estimated 89th and 90th percentiles are 1.1 ppm. This happens because more than 1 per cent of the commodity population is estimated to have the same residue concentration. The percentile calculation method inherent in this figure is equivalent to definition 3 in SAS.

The highest lab level of detection (HLOD) is marked with a dashed line. Percentile and proportion estimates below the HLOD line are likely to be biased downward. As described in Section 3.3, the estimated mean is a range of values denoted by the grey shaded area.

Figure 1 was composed using the statistical package S-PLUS (Statistical Sciences, Inc. 1991). In order to calculate weighted statistics, the original analysis weight (the A_k) for each observation in the PDP data set was rounded to an integer and then the PDP sample point was repeated that many times. This resulted in a data set that was too large for

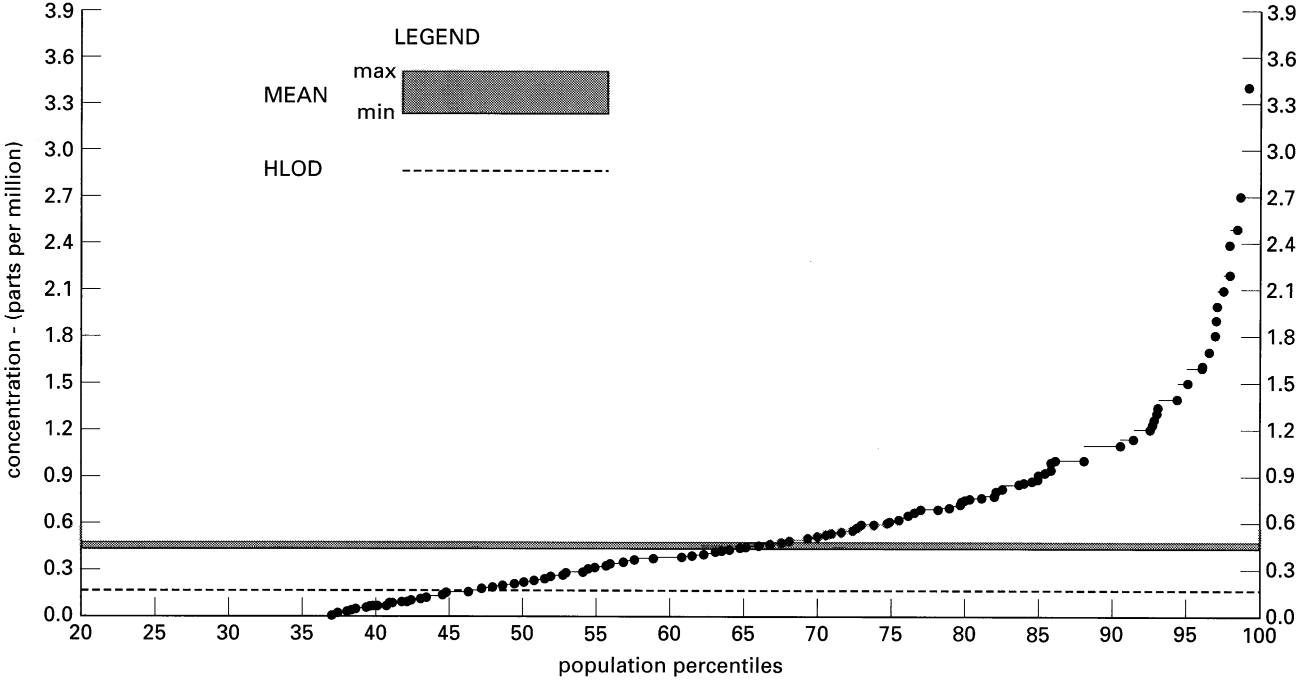

Fig. 1. Thiabendazole residue on apples

Note. Copies of the original figures are available from co-author Phillip S. Kott (pkott@nass.usda.gov).

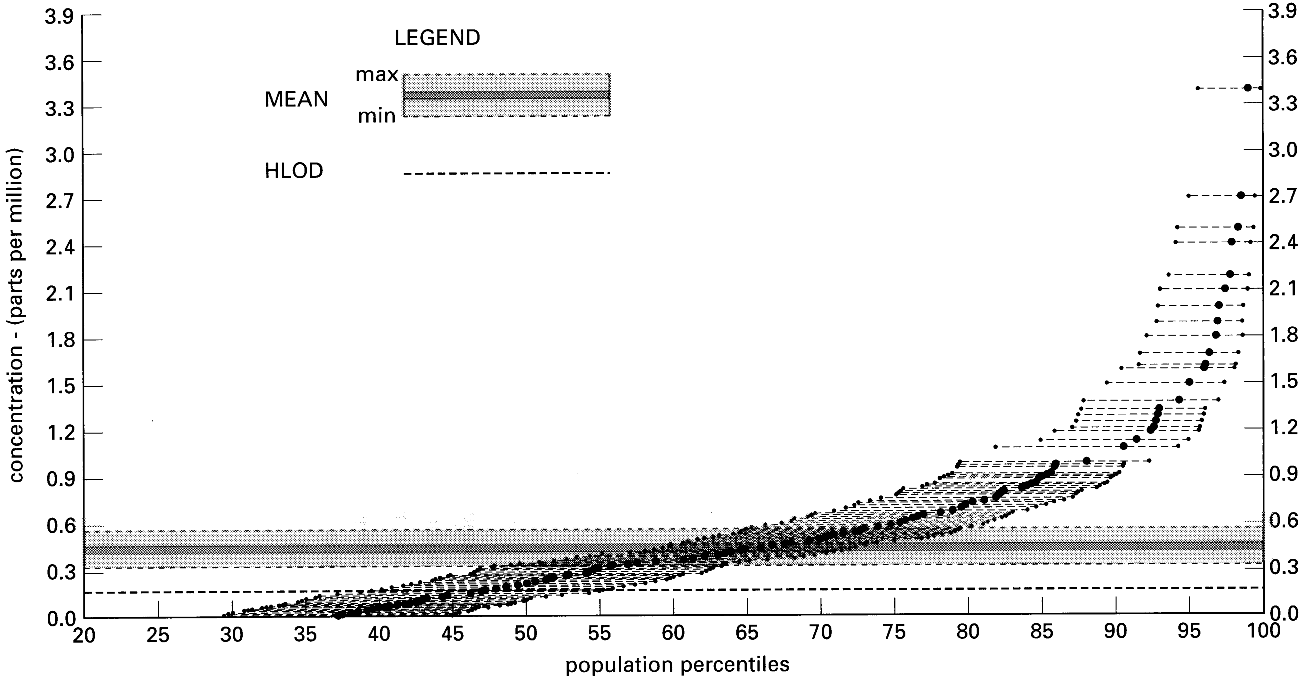

Fig. 2. Thiabendazole residue on apples with 95% confidence intervals

Note. Copies of the original figures are available from co-author Phillip S. Kott (pkott@nass.usda.gov).

S-PLUS to process. Consequently, the data set was sorted by concentration, and a systematic sample of 1,000 observations was drawn. The results were compared to those from SAS and proved to be identical to (at least) three significant digits.

Figure 2 adds 95 per cent confidence intervals to the information in Figure 1 calculated as described in Section 4. The union of the confidence intervals for the two mean estimates is displayed in the new figure. Dashed lines represent estimated 95 per cent confidence intervals for estimated proportions. These confidence intervals cover both the large dots and the points on the solid line.

Woodruff (1952) observed that the region marked out by the 95 per cent confidence intervals for the estimated proportions can also produce 95 per cent confidence intervals for the percentiles. For example, the 95 per cent confidence interval for the 90th percentile in Figure 2 can be seen to range from approximately 0.9 ppm to 1.5 ppm.

Figures 3 through 17 (not presented here, but available from the authors upon request), reveal estimates and confidence intervals for the other pesticide/commodity pairs covered in Tables 1 and 2. The 95 per cent confidence intervals for the selected percentiles displayed in Table 2 come from these figures.

6. Concluding Remarks

As stated in the introduction, ''[t]he ultimate purpose of the 1993 Pesticide Data Program was to make scientific statements about the distribution over the year of certain pesticide residues in particular fresh fruits and vegetables consumed by the U.S. public.'' The text describes the steps undertaken by the USDA to develop a sampling and estimation strategy that produced nearly unbiased estimates of the distribution of pesticide residues at the wholesale level for the nine states in the program (taken as a whole) in 1993 under a small number of supporting assumptions, few of which are subject to much serious debate.

With additional, stronger assumptions, these same calculations can also serve as estimates of the distribution of residues at the consumption level for the U.S. Although these additional assumptions can be questioned more seriously, they can also be defended as both reasonable and necessary given the resource limitations of the program. Moreover, even with more resources, we feel the public would be better served by expanding the PDP to cover additional commodities and pesticides or by increasing the number, quality, and consistency of the PDP laboratories rather than by focusing only on purely statistical issues like the choice of sample states.

7. References

Fuller, W.A. (1975). Regression Analysis for Sample Surveys. Sankhyā, 37, Series C, Part 3, 117–132.

Groves, R.M. (1989). Survey Errors and Survey Costs. New York: John Wiley and Sons.

Kott, P.S. (1994a). Site Selection Sample Design for the 1993 Pesticide Data Program. Internal USDA document. Available upon request.

Kott, P.S. (1994b). A Hypothesis Test of Linear Regression Coefficients with Survey Data. Survey Methodology, 20, 159–164.

SAS Institute Inc. (1985). SAS™ Procedures Guide for Personal Computers, Version 6 Edition. Cary, NC: SAS Institute Inc, 343–357.

Statistical Sciences, Inc. (1991). S-PLUS™ User's Manual. Seattle: Statistical Sciences, Inc.

Wendelberger, J.R. (1995). Methods for Handling Values Below the Detection Limits. Proceedings of the American Statistical Association, Section on Statistics and the Environment, 38–43.

Woodruff, R.S. (1952). Confidence Intervals for Median and Other Position Members. Journal of the American Statistical Association, 47, 635–646.

Received September 1995
Revised January 1997