Standard Error Calculation for HUD Property Physical Inspection Scores

Shawn Jacobson
Shawn Jacobson, 1280 Maryland Ave SW. Suite 800, Washington, DC 20024
Shawn_D._Jacobson@HUD.GOV

Any opinions expressed in this paper are those of the author and do not constitute policy of the Department of Housing and Urban Development.

Key words: Sample Design, Scoring, Truncated data, Jackknife, General variance function

1. Introduction

The Department of Housing and Urban Development is responsible for assessing the physical condition of the public and assisted housing portfolios. This includes:

- Approximately 14,000 properties owned by public housing authorities (public housing), and
- Approximately 28,000 properties that are FHA-insured and/or HUD-assisted (multifamily housing).

To meet this responsibility, HUD conducts physical inspections, observing the condition of the properties by recording deficiencies. Due to cost constraints, only a sample of dwelling units is inspected in most properties (those with at least 6 dwelling units). In addition, an inspection is limited to a sample of buildings in some properties. Thus, inspection scores are subject to sampling error.

After an overview of the sample design and the scoring methodology, this paper discusses the procedure used to estimate physical inspection score sampling errors. Some possible uses of these error estimates are also discussed. The sections of this paper are as follows:

2. Inspection sample design;
3. Scoring methodology;
4. Sample error estimation using the jackknife method;
5. Sample error estimation using a model;
6. Uses of sample error estimates; and
7. Conclusion and future study.

2. Inspection sampling design

The first step in sampling for inspections is to draw a sample of dwelling units. Any building that contains a selected dwelling unit is automatically selected for inspection. Buildings with no dwelling units (common buildings) are also automatically selected for inspection. If necessary, the sample of buildings is augmented by including some buildings that did not contain selected dwelling units.

For a given inspection, i, the size of the dwelling unit sample, n_i, is computed as:

$$n_i = \text{Ceiling}(M / (1 + M/N_i)) \quad (1)$$

Here:

- M, the infinite population sample size, is 26.471,
- N_i is the number of dwelling units on inspection i's target property, and
- Ceiling indicates rounding up to the next highest integer.

N_i is found by summing the number of dwelling units N_{ij} across all buildings j on the target property.

Once n_i is calculated, the sampling interval, R_i, is computed as:

$$R_i = N_i / n_i. \quad (2)$$

R_i is then used to divide the population of buildings into three classes:

- Large residential buildings ($N_{ij} >= R_i$),
- Small residential buildings ($N_{ij} < R_i$ & $N_{ij} > 0$), and
- Common buildings ($N_{ij} = 0$).

Buildings are then randomly ordered within building type.

Once this is done, each residential building ($N_{ij} > 0$) is given a number range from N_{ijl} to N_{iju} such that:

$$N_{ij} = N_{iju} - N_{ijl}. \quad (3)$$

Then, a systematic sample of numbers r_k (k = 1 to n_i) is selected so that $0 < r_i \leq R_i$, and $r_k = r_i + R_i*(k-1)$. The number of dwelling units selected from each building j, n_{ij}, is:

$$n_{ij} = \sum_{k=1}^{n_i} (N_{ijl} < r_k) * (r_k \leq N_{iju}). \quad (4)$$

This is the number of r_k values between N_{ijl} and N_{iju}.

All common buildings and all residential buildings where $n_{ij} > 0$ are selected for inspection. The sample of buildings is augmented if:

- m_{ir} (selected residential buildings) < n_i, and

- M_{ir} (residential buildings on the property> m_{ir}. If needed, the augmentation sample, m_{ira}, is:

$$M_{ira} = \min(n_i, M_{ir}) - m_{ir} \quad (5)$$

Thus, the residential building sample is augmented up to the number of selected dwelling units or the number of residential buildings on the property whichever is the smallest. If $N_{ij} > 0$ and $m_{ir} < M_{ir}$, the probability of selecting building j, P_{ij}, is:

$$P_{ij} = P1_{ij} + (1-P1_{ij})*P2_{ij}. \quad (6)$$

Here:
- $P1_{ij} = \min(N_{ij}/R_i, 1)$, and
- $P2_{ij} = m_{ir}/(M_{ir} - m_{ir})$. Otherwise, $P_{ij} = 1$.

3. Scoring methodology

Each inspection score is comprised of 5 area scores. These areas are
- Site (a = 1)
- Building Exterior (a = 2)
- Building Systems (a = 3)
- Common areas (a = 4) and
- Dwelling units (a = 5).

Each area (except site) is comprised of several sub-areas.
- For building exterior, building systems, and common areas, each inspected building is a sub-area.
- For dwelling units, each inspected dwelling unit is a sub-area.

Each sub-area score would be 100 if no deficiencies were found. Each observed deficiency reduces the score based on:
- Severity (a big crack in a wall is more severe than a small one);
- Criticality (clogged drains are more critical than damaged cabinets); and
- Item importance (the bathroom is more important than the laundry area).

If deficiencies drive the sub-area score below 0, the sub-area score is set to 0. Thus, sub-area scores are bounded by 0 and 100. Connell (1999) gives a technical discussion of the computation of sub-area scores.

Area scores are weighted averages of their component sub-area scores. For area a, in inspection i the area score, S_{ia}, is computed as:

$$S_{ia} = \sum_{k \text{ in } a} W_{iak} * B_{iak} S_{iak} \Big/ \sum_{k \text{ in } a} W_{iak} * B_{iak}. \quad (7)$$

Here:
- S_{iak} is the sub-area score.
- W_{iak} is the sub-area weight, and
- B_{iak} is the amenity weight (weighted items possessed by the sub-area),

For sub-area k in area a in inspection i. The formula used to compute sub-area weights, W_{iak} is:
- 1 (for a = 1 or 5),
- $M_{iak,j} / P_{iak,j}$ (for a = 2 or 3), and
- $1/P_{iak,j}$ (for a = 4).

Here, $M_{iak,j}$ is the size ($N_{ij,k}$ for residential buildings and $N_i/M_i * B_{i4k,j}$, M_i being the number of buildings on the property, for common buildings), and $P_{iak,j}$ is the selection probability of building j containing sub-area k. Thus, building level sub-areas are weighted to account for those buildings not selected for inspection; in addition, building exterior and system sub-areas are weighted to reflect building size.

It will be useful to define the relative weight of sub-area k within area a of inspection i, W'_{iak}, as:

$$W'_{iak} = W_{iak} * B_{iak} / \sum_{k \text{ in } a} W_{iak} * B_{iak} \quad (8)$$

The inspection score, S_i, is a weighted average of S_{ia} and is computed as follows:

$$S_i = \sum_{a=1}^{5} \alpha_{ia} * W_{ia} * S_{ia} \Big/ \sum_{a=1}^{5} \alpha_{ia} * W_{ia}. \quad (9)$$

Here:
- α_{ia} is the area nominal weight, and
- W_{ia} is an adjustment to account for the absence of items from some sub-areas.

The weight adjustment, W_{ia}, is computed as:

$$W_{ia} = \sum_{k \text{ in } a} W_{iak} * B_{iak} / \sum_{k \text{ in } a} W_{iak} \quad (10)$$

Table 1
Nominal weight of areas

Area		Scored Areas	
Name	#	5	2
Site	1	0.15	0.00
Building Exterior	2	0.15	0.00
Building systems	3	0.20	0.36
Commons	4	0.15	0.00
Dwellings	5	0.35	0.64

The nominal weight, α_{ia}, depends on the area and the number of areas used in the scoring method. The 5-area scoring method is used for multifamily housing and was used for the first several years of public housing inspections (see Federal Register 2000). The 2-area scoring method is used to score current public

housing inspections (see Federal Register 2001). Table 1 gives the nominal weights for each score method.

It will be useful to define the normalized area weight for area a in inspection i, W'_{ia}, as:

$$W_{ia} = \alpha_{ia} * W_{ia} / \sum_{a=1}^{5} \alpha_{ia} * W_{ia}. \quad (11)$$

4. Sample error estimation using the jackknife method

Originally, three methods were considered for computing standard errors of physical scores (Jacobson 1999). Of these, the jackknife method (Wolter 1985) was selected because:

- It did not require estimates of standard deviations for sub-areas, and
- It would implicitly account for any correlation between sub-area scores.

In this implementation of the jackknife method, n_i replicates are created by excluding a dwelling unit. Replicate r will contain all sub-areas except:

- The rth dwelling unit's sub-area (if dwelling unit r's building has $P_{ij'r} = 1$), and
- The rth dwelling unit's sub-area and the sub-areas for building j'r containing dwelling unit r otherwise.

Note that for purposes of discussing replicates, the author defines building j'r as the building containing dwelling unit r.

The area a score for replicate r, $S_{i(r)a}$, is:

$$S_{i(r)a} = \sum_{k \text{ in } a} W'_{iak} * \delta_{i(r)ak} * S_{iak} / \sum_{k \text{ in } a} W'_{iak} * \delta_{i(r)ak} \quad (12)$$

Here, $\delta_{i(r)ak} = 1$ if sub-area k is included in replicate r and 0 otherwise.

Once $S_{i(r)a}$ is computed, the replicate r inspection score, $S_{i(r)}$, is computed as:

$$S_{i(r)} = \sum_{a=1}^{5} \alpha_{ia} * W_{i(r)a} * S_{i(r)a} / \sum_{a=1}^{5} \alpha_{ia} * W_{i(r)a} \quad (13)$$

The replicate r weight adjustment for area a, $W_{i(r)a}$, is:

$$W_{i(r)a} = \sum_{k \text{ in } a} W_{iak} * \delta_{i(r)ak} * B_{iak} / \sum_{k \text{ in } a} W_{iak} * \delta_{i(r)ak} \quad (14)$$

Once the replicate inspection scores are calculated, the standard error of the inspection score, se_i is computed as:

$$se_i = F_i * ((n_i - 1)/n_i) * \sum_{r=1}^{n_i} (S_{i(r)} - S_i)^2 \quad (15)$$

Here, F_i (the finite population correction factor for inspection i) is $(N_i - n_i)/N_i$.

As of April 24, standard errors had been computed for 6,639 inspections:

- 2,887 were scored under the 5-area method, and
- 3,752 were scored under the 2-area method.

Table 2 gives summary statistics for these standard errors.

It is apparent that scores based on 2 areas have larger standard errors than do scores based on all 5 areas. This is because 2-area scores put more emphasis on the dwelling unit area, always subject to sample error, than do 5-area scores. The next section shows that a model can be fit for both 2-area and 5-area scores.

Table 2
Summary of Standard Error estimates

Statistic	Scoring Method	
	5-area	2-area
Inspections	2,887	3,752
Mean	1.37	2.62
Standard deviation	0.94	1.68
Percentiles		
5^{th}	0.01	0.02
10^{th}	0.15	0.39
25^{th}	0.52	1.21
50^{th}	1.35	2.52
75^{th}	2.09	3.92
90^{th}	2.68	4.88
95^{th}	2.96	5.31

5. Standard error estimation using a model

Because only part of an inspection is subject to sample error, the standard error estimate should be related to the weight of areas where sampling occurs. Also, because sub-area scores are bounded by 0 and 100, sampled areas with scores near the boundaries should have relatively low standard errors because the component sub-area scores must be homogeneous. This leads to the attempt to model standard error estimates documented in this section.

The model was based on 4,154 inspections that had

- Dwelling units that were not inspected;
- A sufficient sample of dwelling units to produce an acceptable score; and
- An inspection date before February 22, 2002.

The modeled standard error, se'_i, is computed as:

$$se'_i = \sqrt{\sum_{a=2}^{5} F_{ia} * W'^{2}_{ia} * D_a^2 * S_{ia} * (100 - S_{ia})}. \quad (16)$$

F_{ia} accounts for the sample size and finite correction factor. For a = 2, 3, or 4,

$$F_{ia} = \sum_{k \text{ in } a}((1/P_{iak,j}) - 1)/(1/P_{iak,j}))*W'^{2}_{iak}. \quad (17)$$

For a = 5,

$$F_{ia} = F_i * \sum_{k \text{ in } a} W'^{2}_{iak}, \text{ for a = 5} \quad (18)$$

D_a is the distribution factor that could range from 0 (all sub-areas have the same score) to 1 (all sub-areas either have the maximum score, 100, or the minimum score, 0). To compute distribution factors, areas from inspections were chosen if:

- The inspection was used to create the model;
- There were at least 20 inspected sub-areas in the area; and
- $S_{ia}*(100-S_{ia}) > 100$.

For each chosen area, the distribution ratio, D_{ia}, was computed as:

$$D_{ia} = \sqrt{\frac{\sum_{k \text{ in } a}(S_{iak} - S_{ia.})^2/n}{S_{ia.}*(100-S_{ia.})}}. \quad (19)$$

The model distribution factor, D_a, was found by taking the median, within each area, of the distribution ratios, D_{ia}. These factors are as follows:

- .4906 for building exteriors;
- .8957 for building systems;
- .8334 for common areas; and
- .5841 for dwelling units.

Thus, the sub-area scores for building exteriors and dwelling units tend to be relatively homogeneous while the sub-area scores for building systems and common areas tend to be near 100 or 0. The distribution factor and the square root term that follows constitute the model estimate of the standard deviation of sub-area scores.

For the inspections used in model creation, the average model standard error, se'_i, was 2.03 compared to the average system generated standard error, se_i, of 2.13. When the analysis was limited to the 3,963 model inspections with a model standard error of at least 0.25, the standard error ratio, se_i/se'_i, followed the distribution given in table 3.

It appears that the model used here is reasonably good. For more than 70% of cases the system generated standard error is between .75 and 1.5 times the model standard error.

Table 3
Distribution of Standard Error Ratios
for Inspections used in Model Creation

Ratio Range	Score Type	
	5-area	2-area
< 0.5	9	8
0.5 to < 0.75	16	18
0.75 to < 0.9	12	11
0.9 to 1.1	24	25
> 1.1 to 1.25	19	20
>1.25 to 1.5	18	16
>1.5	2	2
Total	2,268	1,695

Table 4 gives the standard error ratio distribution for the 1,975 inspections used to test the model. These inspections:

- Were not used in model creation, and
- Had a model standard error of at least 0.25.

Note that the distributions in Tables 3 and 4 are very similar.

Table 4
Distribution of Standard Error Ratios
for Other Inspections

Ratio Range	%Inspections	
	5-area	2-area
< 0.5	10	9
0.5 to < 0.75	18	20
0.75 to < 0.9	12	13
0.9 to 1.1	24	22
> 1.1 to 1.25	18	18
>1.25 to 1.5	16	15
>1.5	2	2
Total	1,154	821

The model is almost as good when applied to inspections not used to create the model as it was for inspections used for model creation. Again, for about 70% of inspections, system generated standard errors are between 0.75 and 1.5 times the model standard error. The next section gives evidence that inspections in the other 30% may have more data quality issues than do inspections where the system and model standard errors are similar.

6. Uses of standard error estimates

Currently, standard errors are used along with inspection scores to assist in the administration of HUD's multifamily housing portfolio. These standard errors are also being used to determine the optimal number of dwelling units to be selected for future inspections.

It may also be possible to use system generated and model generated standard errors to assure the quality of inspections. System generated standard errors are based on reported sub-area scores. By contrast, model generated standard errors are based on estimates of sub-area score standard deviation given a typical distribution of sub-area scores for an area. The model also assumes no correlation between sub-area scores in different areas.

A disparity between these estimates of standard error indicates either an atypical distribution of sub-area scores or correlation (positive or negative) between the sub-area scores in different areas.

Some atypical distributions may be legitimate. For instance, if a property provides two types of housing (elderly or families), then a relatively large standard error may reflect the fact that the first type of housing is easier to manage than the latter.

However, abnormally large or small standard errors may indicate data quality problems. A large standard error may indicate an inspector's varying diligence in finding deficiencies over the inspection (for example, the inspector may want to leave the property before dark and may therefore rush through the last part of the inspection). By contrast, a very low standard error may indicate that an inspector is very diligent in finding some deficiencies but is not at all diligent in finding others. The result is artificially homogeneous sub-area scores caused by the same deficiencies being reported in every sub-area.

For purposes of this study, each inspector was scored to see if there is a link between atypical standard error ratios and data quality. The standard error ratio of inspections was compared to the data quality score of the inspector who conducted the inspection. The data quality score reflects the inspector's inspection record with regard to:

- Time taken to conduct inspections;
- Average inspection score;
- Consistency in reporting the existence of items;
- Changes in scores based on technical reviews;
- Engineer rejection of inspection; and
- Comparison with on site inspection quality control.

For purposes of this study, the inspectors' data quality score was rated using color codes: "Green" for the fewest data quality concerns through "Yellow" and "Orange" to "Red" for the most data quality concerns. Table 5 shows the relationship between standard error ratios and inspector color codes for the 5,644 inspections conducted by inspectors with enough experience to be scored.

Table 5
Distribution of Inspector Ratings
Within Standard Error Ratio Class

Ratio Range	%Distribution			
	Green	Yellow Orange	Red	Total
< 0.5	59	25	16	484
0.5 to < 0.75	64	19	17	981
0.75 to < 0.9	75	17	8	655
0.9 to 1.1	75	15	11	1,377
> 1.1 to 1.25	76	16	8	1,086
>1.25 to 1.5	73	17	10	957
>1.5	57	28	15	104
Total	71	17	12	5,644

Although the relationship is far from deterministic, inspections with atypical standard error ratios (<0.75 or >1.5) tend to be conducted by inspectors with data quality concerns (color other than green). Of the 660 inspections conducted by "Red" inspectors, 252 (38%) have a standard error ratio, se_i/se'_i, less than 0.75 and 20 (3%) have $se_i/se'_i > 1.5$. By contrast, of the 4,037 inspections conducted by "Green" inspectors, only 23% have se_i/se'_i less than 0.75 and only 1.5% have se_i/se'_i greater than 1.5. Inspectors in the marginal "Yellow" and "Orange" categories also are more likely to produce inspection scores with atypical standard error ratios than are inspectors in the "Green" category.

7. Conclusion and future study

This paper gives a description of the sample design for HUD inspections and a description of how they are scored. The method for computing system generated standard errors is described in some detail.

A relatively complex generalized variance function used to derive model generated estimates of standard error for these inspections is described. Finally, a ratio of the system generated to the model generated standard error estimate was computed for each studied inspection that had a model generated standard error of at least 0.25 points. Table 5 shows that this ratio may be useful for finding problems with the quality of inspection data.

This merits some future analysis of inspections with atypical standard errors. The scores from these inspections could be compared with the scores for inspections with typical standard error ratios. If inspections with atypical standard error ratios are higher than other inspection scores, it could indicate that deficiencies are missed in these inspections. Likewise, the difference between the current and previous inspections for a property can be investigated. If atypical standard error ratio inspections have relatively high score differences, this could indicate that deficiencies are being missed.

Inspected items could be ordered by the time when they were recorded (timestamp analysis) to see if there is a relationship between inspection order and:

- Deficiency status (no deficiency, a deficiency, or item does not exist), and
- Time taken to record the item (elapsed time between the previously recorded and current item).

It would prove useful to find out if inspections with atypically low standard errors have a typical distribution of deficiencies. A typical distribution would come from inspections with typical standard errors.

Acknowledgments

I would here like to thank Terrence Connell, Ching Yu, Rod Harris, Fred Lau, Brian Fitzpatrick, and James Cruickshank for their hard work in developing the inspector rating system used in this paper.

In addition, I would like to thank Terrence Connell, Claudia Yarus, and Janice Kuhl for their invaluable help in proofreading this paper.

References

- Connell, Terrence, (1999), "Physical Inspection Scoring 2.1", June 4, 1999, Available from HUD on request.
- Federal Register (2000), "Public Housing Assessment System Physical Condition Scoring Process, Notice", June 28, 2000, Vol. 65 No. 12, Pages 39987—40005.
- Federal Register (2001), "Public Housing Assessment System Physical Condition Scoring Process, Interim Scoring, Correction and Republication, Notice", Vol. November, 26, 2001, 66 No. 22, Pages 59083—59124.
- Jacobson, Shawn, (1999), "Variance Calculation Methodology for Physical Scores", November 10, 1999, available from HUD on request.
- Wolter, Kurt, (1985), *Introduction to Variance Estimation*, Springer-Verlag, New York City, NY, Pages 153—200.